

This is a peer-reviewed prepublished version of the paper to be published in **Vol. 26, No. 1, March 2027**.

A data-driven review of air pollution: Causes, impacts and prediction models

Evans Osei-Opoku, Oyeniya Akeem Alimi, Robyn Thompson and Kuburat Oyeranti Adefemi Alimi

RECEIVED

16 June 2026

CITE THIS ARTICLE

Osei-Opoku, E., Alimi, O.A., Thompson, R. and Alimi, K.O.A., 2027. A data-driven review of air pollution: Causes, impacts and prediction models. *Nature Environment and Pollution Technology*, 26(1), D1954.

<https://doi.org/10.46488/NEPT.2027.v26i01.D1954>

EARLY-ACCESS STATEMENT

This manuscript, after peer review, is being released online in its original, unedited form to provide early access to its findings. Prior to final publication, it will undergo additional editing and formatting. Readers may note that, as an unedited version, the manuscript may have linguistic and other minor errors. The DOI will be active after the final publication of the paper.

PUBLICATION STATUS

PEER REVIEW

Completed

ONLINE VERSION

Prepublished

FINAL ISSUE

March 2027

© The Author(s) 2027. This article is Open Access under the Creative Commons Attribution 4.0 International License, allowing use, sharing, adaptation and reproduction with proper credit. Images or third-party material are covered by this licence unless otherwise noted. For uses beyond the licence, permission must be obtained from the copyright holder. Licence details: creativecommons.org/licenses/by/4.0/.

Review

A data-driven review of air pollution: Causes, impacts and prediction models

Evans Osei-Opoku^{1,2†}, Oyeniyi Akeem Alimi², Robyn Thompson¹ and Kuburat Oyeranti Adefemi Alimi³

¹ Department of Information Technology, Durban University of Technology, Durban 4000, South Africa; 22176496@dut4life.ac.za; robynt@dut.ac.za

² Innovation for Sustainability Group, Department of Information Systems, Durban University of Technology, Durban 4000, South Africa; OyeniyiA@dut.ac.za

³ Mangosuthu University of Technology, Umlazi, South Africa; adefemi.kuburat@mut.ac.za

†Corresponding author: Evans Osei-Opoku; 22176496@dut4life.ac.za

Abstract

Air pollution remains a major global environmental challenge driven by rapid urbanisation, industrialisation, fossil fuel combustion, and large-scale deforestation, resulting in elevated levels of key pollutants including nitrogen dioxide, carbon monoxide, sulphur dioxide, and particulate matter. These pollutants have significant impacts on human health, climate systems and ecosystems, directly affecting multiple Sustainable Development Goals (SDGs), including health, sustainable cities, and climate action. This study presents a comprehensive review of recent literature on air pollution, with particular emphasis on data-driven approaches for air quality prediction and forecasting. A narrative review was conducted using recent studies from major scientific databases, with emphasis on modelling techniques, datasets, and methodological frameworks. The findings reveal a clear progression from classical statistical models to more sophisticated machine learning and deep learning methods, which demonstrate improved ability in modelling nonlinear, high-dimensional air quality data. The study also highlights the growing importance of multi-source data integration and preprocessing techniques in enhancing temporal robustness. Despite these advances, several challenges remain, including data quality, model interpretability, and scalability for real-time applications. The review concludes by identifying critical research gaps and providing strategic insights for researchers, policymakers, and urban planners to support the development of effective, data-driven air quality management systems aligned with sustainable development objectives.

Keywords

Air pollution; Air quality monitoring; Machine learning; Deep learning; Sustainable development goals

1. INTRODUCTION

The degradation of air quality has become a major global issue of concern with wide-ranging implications for human health, ecosystems, and socio-economic development. According to the World Health Organisation (WHO), unsafe ambient air quality affects roughly 99% of the global population, a crisis linked to close to seven million premature deaths each year (Ji et al. 2025). Air quality degradation occurs when biological, chemical, or physical agents alter the natural makeup of the air, thereby threatening ecological systems and human well-being. Rapid urbanisation and industrialisation have significantly intensified air pollution through increased fossil fuel combustion, deforestation, and inefficient waste management practices. In particular, rural-to-urban migration driven by economic opportunities has led to population concentration in cities, resulting in increased emissions, environmental degradation, and ecological imbalance (Agyemang 2025, World Bank 2024, European Parliament 2023, National Aeronautics and Space Administration (NASA) 2023, Liang and Gong 2020). Consequently, air pollution has become not only an environmental issue but also a significant challenge to sustainable urban development. In response, governments and international organisations have implemented various regulatory frameworks and policy instruments aimed at mitigating pollution and promoting sustainable development. Mechanisms such as Environmental Impact Assessment (EIA) and Strategic Environmental Assessment (SEA) have been widely adopted to integrate environmental considerations into decision-making processes (Dias et al. 2025, Legislative Affairs Office of the State Council of the People's Republic of China 2019, Evans 2013). Additionally, governing bodies such as the United States Environmental Protection Agency (EPA) have set air quality standards for key pollutants, while several countries have introduced economic instruments such as carbon taxes to curb emissions (Elvan 2018). In addition, China introduced the practice in 1979 and later reinforced it as part of the EIA Law (Huang 2012). Similarly, agencies such as the US EPA set the national ambient air quality standards for six critical air pollutants: nitrogen dioxide (NO₂), carbon monoxide (CO), sulphur dioxide (SO₂), ozone (O₃), lead and particulate matter (PM₁₀ & PM_{2.5}) (U.S. Environmental Protection Agency 2025). These standards are intended to protect public health and welfare by regulating ambient concentrations of criteria air pollutants. Recognising the threats imposed by poor air quality, some countries have introduced levies and taxes on companies for pollutant greenhouse gas emissions. For example, the 2019 carbon tax act by the South African government imposes an environmental tax on establishments that burn fossil fuels and similar carbon-related pollutants (South African Revenue Service 2024). Ghana also introduced a similar emission levy in 2024 (Ernst & Young Global 2024) while other African nations are following similar policy measures. Despite these initiatives, air pollution remains the leading environmental health risk globally.

WHO considers poor air quality to be the single greatest environmental threat to human health, driving an estimated seven million premature deaths on a yearly basis (Gu et al. 2024). Extensive research has shown that air pollutants, such as NO₂, CO, SO₂, and PM_{2.5} and PM₁₀, are strongly associated with respiratory and cardiovascular illnesses, including asthma and chronic obstructive pulmonary disease (COPD) (European Environment Agency 2025, Tran et al. 2023, Manisalidis et al. 2020, National Park Service 2018). He et al. (2022) estimated that two popular air pollutants: NO₂ and PM_{2.5} contributed to approximately 55,000 and 417,000 premature deaths across approximately 40 European countries in 2018 respectively. The 2019 Global Burden of Disease (GBD) study reported that air pollution was the leading cause of premature deaths in the world, accounting for about 6.67 million deaths in 2019 (Guo et al. 2024). In addition to human health, air pollution has

multidimensional impacts on sustainability, including on climate change, ecosystem degradation, and economic losses (Zhang et al. 2022b, Coelho et al. 2021, Kaur and Pandey 2021, Sivarethinamohan et al. 2021). Furthermore, Zhang et al. (2022a) reported that, only a small fraction (12%) of the global population lives in areas that comply with the WHO air quality guidelines, highlighting the need for actionable monitoring and intervention measures. These challenges have led to the need for reliable monitoring and forecasting of air pollution to enable informed decision-making and proactive environmental management. Traditional statistical models, such as Auto-regressive Integrated Moving Average (ARIMA) and Seasonal ARIMA (SARIMA), have been widely applied for forecasting air quality (He et al. 2024, Li et al. 2019). A significant drawback of these models is their inability to capture the dynamic, non-linear characteristics of atmospheric pollution. Such data are highly complex, driven by the interplay of human behaviour, topographical features, and shifting weather patterns (Izah et al. 2024). The interactions between these factors create non-linear dynamics that affect atmospheric properties. Furthermore, these factors contribute to the poor forecasting performance of traditional models. Consequently, there has been a growing shift towards data-driven approaches, particularly machine learning and deep learning models, which offer enhanced capabilities in modelling high-dimensional and nonlinear time-series data (Leong et al. 2020, Zhang et al. 2022a). Unlike conventional statistical models, data-driven models have the capacity to learn the dynamic characteristics of time series air pollutant features (Yu et al. 2020, Pak et al. 2020, Du et al. 2019). With air pollution data that has millions of time-series observations with many weather and pollutant features, researchers have proposed forecasting and prediction models based on data driven automated models (Zhu and Xie 2023, Jiang et al. 2021, Jung et al. 2020).

With the increasing concerns about air pollution and its effects, numerous studies have examined this problem from different standpoints. Kaur and Pandey (2021) analysed the relationship between air pollution, climate change and public health in Indian cities. Similarly, Coelho et al. (2021) evaluated climate extreme scenarios that impact air pollution. While the study extensively evaluated the impact of climate change on future air quality, it does not adequately address the aspect on data-driven intelligent forecasting models nor model interpretability. By examining China as a focal point, Zhang et al. (2022b) analysed how the spatial distribution of atmospheric contaminants intersects with socioeconomic variables to impact population health. Similarly, the health and socio-economic effects and future directions were reviewed by Sivarethinamohan et al. (2021).

Although scholarly interest in atmospheric degradation is expanding, current reviews focus primarily on environmental and health impacts or on climate-related aspects, with limited integration of data-driven forecasting techniques, dataset characteristics, and model interpretability. In particular, there remains a lack of a comprehensive synthesis that combines modelling approaches with practical considerations such as data integration, preprocessing, and real-world implementation. Unlike previous reviews, which primarily compare predictive accuracy, this review critically evaluates the methodological rigour of forecasting studies by examining validation design, temporal robustness, data leakage risk, benchmark selection, and model interpretability. Building on this critical synthesis, this review proposes a methodological evaluation framework that assesses forecasting studies not only according to predictive accuracy but also according to validation design, temporal robustness, data leakage risk, interpretability, and deployment readiness. To address these gaps, this study provides a comprehensive review of recent advances in air pollution research with emphasis on the air pollution data-driven intelligent models for forecasting. The study reviews the approaches to modelling, data and methodological evaluation frameworks, focusing on interpretability, multi-source data integration and applicability. It aims to

provide strategic insights for researchers, policy makers and urban planners, supporting decision-making for data-driven air quality management.

This review evaluates the strengths and limitations of the current forecasting models and datasets from the last five years of published research in the most prominent scientific databases, such as Scopus and Web of Science. The main contributions of this study are as follows: (1) an updated overview of air pollution causes and impacts; (2) a detailed analysis of state-of-the-art forecasting models; (3) a methodological evaluation framework for comparing forecasting studies based on temporal robustness, validation strategy, interpretability, and deployment readiness; and (4) recommendations for the development of transparent, scalable, and effective predictive systems. The remainder of this paper is organised as follows: Section 2 presents the methods adopted in the study. Section 3 examines the economic, ecological, and health consequences of atmospheric degradation while evaluating deep learning, machine learning, and statistical approaches to forecasting. Section 4 proposes the methodological evaluation framework, identifies key research gaps and outlines future research opportunities, and Section 5 concludes the study.

2. MATERIALS AND METHODS

In this study, a narrative review methodology was adopted to examine recent developments in air pollution research, with particular emphasis on data-driven forecasting approaches. To ensure the reliability and credibility of the analysis, the literature later analysed was retrieved from prominent academic databases including Springer Nature, ScienceDirect, IEEE Xplore, Web of Science, and Scopus. The review primarily considered works published in the last five years, starting from January 2020 to December 2025. To cover recent advances in modelling techniques, literature searches were conducted using the following keywords: “air pollution prediction”, “air quality index (AQI)”, “machine learning”, “deep learning”, as well as by the following general terms: “air pollution”, “air quality monitoring”, “pollution forecasting”, and “atmospheric modelling”. The combined searches resulted in approximately 1,000 records being identified. The results which were subsequently screened to identify the most relevant studies. The retrieved literature was subsequently examined for relevance to the objectives of this review. Studies unrelated to air pollution forecasting, non-English publications, non-peer-reviewed articles, conference abstracts without accompanying full papers, duplicate records, and studies lacking a statistical, machine learning, or deep learning modelling component were excluded during the screening process. Particular emphasis was placed on peer-reviewed forecasting studies published between 2020 and 2025, which form the basis of the comparative synthesis presented in Section 3.2 and Tables 4–6. Institutional, regulatory, and policy sources cited elsewhere in this review (e.g., WHO guidelines, national environmental protection agency websites, carbon tax and emissions-levy notices) fall outside this eligibility window and are used solely to establish regulatory and contextual background, and are not treated as part of the modelling evidence base.

Following the screening process, a curated set of studies was chosen for comprehensive synthesis and detailed analysis. The selected studies were further assessed based on their relevance to air pollution monitoring, forecasting methodologies, and data utilisation. Following the removal of duplicates and the rigorous application, a final set of approximately 160 studies were retained for critical synthesis. Fig. 1 shows the flow diagram for the screening process. Particular attention was given to studies that addressed modelling frameworks, dataset characteristics, and methodological procedures used in air quality prediction. Traditional statistical models

and data-driven approaches such as machine learning and deep learning algorithms were both considered. The review systematically examines literature to identify common modelling workflows such as data collection, data pre-processing, model development, validation, and performance evaluation. Furthermore, the review critically assesses the methodological limitations and strengths of the datasets, validation strategies, and forecasting models used in the literature examined. This approach provides a comparative view of the methods developed so far and identifies some trends, challenges, and opportunities for future research on air pollution forecasting. The findings of this review are presented and synthesised in the next section, which examines air pollution and the related forecasting models found in the literature.

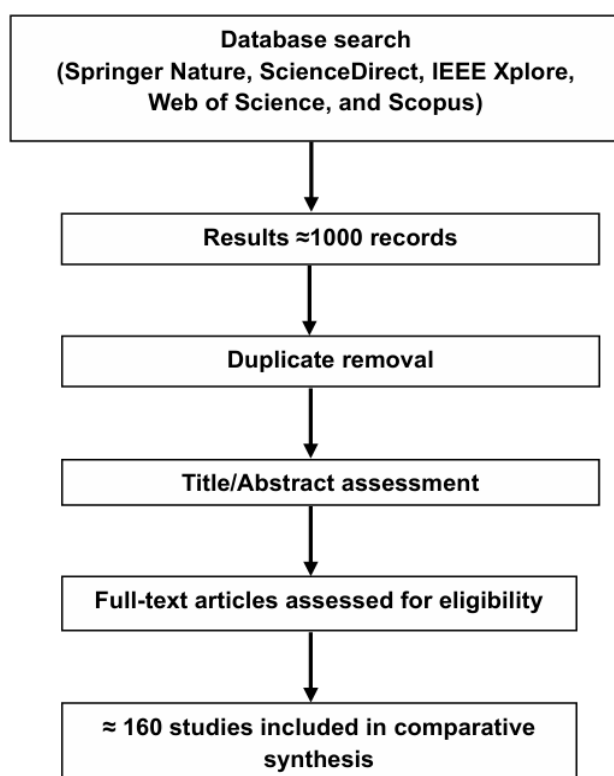


Fig. 1: Screening Process for the records

3. Literature Synthesis and Discussion

This section summarises the main findings of the literature reviewed in order to provide both contextual and technical insights on air pollution. Section 3.1 gives an overview of air pollution, particularly its main sources, and its broader implications for sustainability, including impacts on human health, the environment and socio-economic systems. Section 3.2 explores the major modelling approaches for air pollution forecasting, including recent studies on statistical, machine learning, and deep learning techniques. These discussions collectively offer a comprehensive picture of the drivers of air pollution and how air pollution behaviour is monitored and predicted over time, using data-driven approaches.

3.1. Overview of Air Pollution

The global surge in industrial expansion, urban migration, and population growth has significantly increased fossil fuels combustion and the generation of municipal and industrial waste. Consequently, the atmosphere has experienced a substantial increase in toxic emissions, endangering climatic equilibrium, built infrastructure, biological diversity, and human health. As a result, air pollution is recognised as a severe environmental and health problem that requires global solutions to mitigate its impacts. In response, many countries have implemented policies, regulatory frameworks, and strategic initiatives aimed at reducing emissions and promoting environmental sustainability. For example, China has introduced several top-down policies designed to curb emissions and accelerate the transition to sustainable energy systems. At the 75th United Nations General Assembly in 2020, China committed to peaking CO₂ emissions before 2030 and achieving carbon neutrality by 2060 (Yang et al. 2021, McGrath 2020). This commitment is supported by increased investment in green energy technologies and targeted interventions in major metropolitan regions such as Beijing, Hebei, and Tianjin, where air pollution reduction has been prioritised (Ao et al. 2023, Chen et al. 2023). Similarly, the Canadian Environmental Protection Act aims to address air pollution by minimising greenhouse gas emissions in an effort to have net-zero emissions by the year 2050 (Environment and Climate Change Canada 2024).

As illustrated in Fig. 2, the sources of air pollution are divided into anthropogenic (human-induced) and natural sources (Sompornrattanaphan et al. 2020). Anthropogenic sources are linked to emissions from industrial processes, transportation systems, and indoor activities; natural sources are related to activities like dust storms, volcanic eruptions, and wildfires. The pollutants produced are usually classified as gaseous pollutants, including carbon monoxide (CO), CO₂, SO₂, O₃ and NO₂ and particulate matter which are fine solid or liquid particles in the atmosphere. Particulate matter is also divided by aerodynamic diameter into PM₁₀ (particles < 10µm) and PM_{2.5} (particles < 2.5µm) particles, which are more harmful to health because they penetrate deeper into the lungs.

The primary atmospheric contaminants, alongside their origins and environmental and health impacts, are outlined in Table 1. CO₂ is included in Table 1 as a co-emitted combustion by-product and climate-relevant greenhouse gas. Unlike CO, NO₂, SO₂, O₃, Pb, and particulate matter, it is not one of the criteria pollutants regulated under ambient air-quality standards. To guide global efforts in reducing exposure to harmful pollutants, the WHO has published air quality guidelines (World Health Organization 2021) since 1987. These guidelines serve as a benchmark for the development of national air quality standards and environmental policy development. The updated WHO Air Quality Guidelines (AQGs) released in 2021 introduced significantly stricter limits compared to the 2005 guidelines, reflecting growing evidence of the adverse health effects of even low concentrations of pollutant exposure. For example, the acceptable annual mean concentration for PM_{2.5} was halved from 10 to 5 µg/m³, while the NO₂ limit was tightened from 40 to 10 µg/m³. These revisions highlight the increasing urgency for more effective air quality monitoring, forecasting, and mitigation strategies.

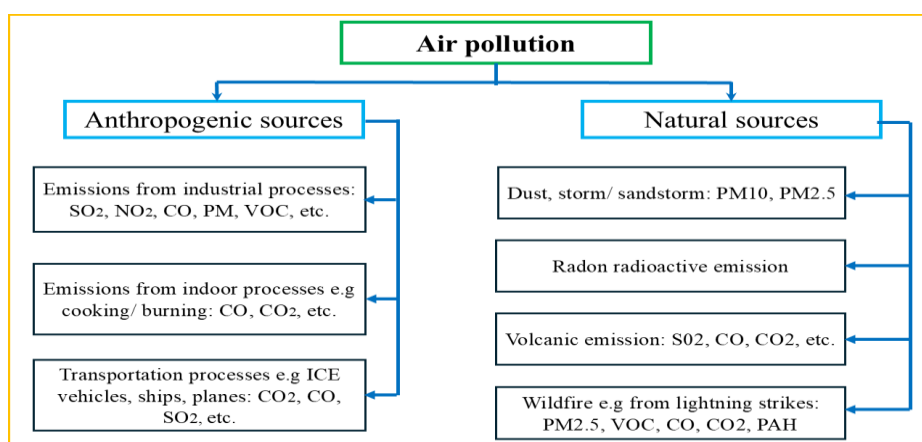


Fig. 2: Classification of Air Pollutants

Table 1: Common Air Pollutants (Gu et al. 2024, Manisalidis et al. 2020, Kaur and Pandey 2021, Sompornrattanaphan et al. 2020, Garbagna et al. 2025).

Pollutant	Primary Sources	Potential Health & Environmental Impact
CO	Emissions from vehicles, industrial machinery, manufacturing	Interferes with oxygen transport. Can lead to respiratory failure, neurological harm, even death
CO ₂	Burning fossil fuels, decomposition of organic matter	High levels can cause headaches, dizziness, and severe oxygen deprivation (asphyxia)
NO ₂	Fossil fuel combustion	It causes respiratory illness, worsens asthma, contributes to acid rain, ecosystem damage, and haze
SO ₂	Industrial emissions, coal-fired power plants, volcanic activity	Triggers respiratory issues, increases heart disease risk, causes acid rain, reduced visibility, haze
O ₃	Secondary pollutant formed by photochemical reactions between NO _x and volatile organic compounds in sunlight; precursor emissions from vehicles, industry, and chemical solvents	Causes breathing problems, damages vegetation, and disrupts photosynthesis
PM _{2.5}	Emitted from human activities and chemical transformation	Linked to cardiovascular disease and complications
PM ₁₀	Emitted from human activities and chemical transformation	Serious risks to cardiovascular health

3.1.1. Causes of Air Pollution

Air pollution originates from both anthropogenic (human-induced) and natural processes (Singh et al. 2025). Anthropogenic emission sources are commonly classified as mobile and stationary (fixed). Mobile sources, such as vehicles, ships, and aircraft powered by internal combustion engine, represent a significant share of urban pollution, including NO₂, CO, CO₂, and particulate matter (Filonchyk and Peterson 2024). In contrast, stationary sources comprise power plants, manufacturing facilities, refineries, and other industrial installations that release pollutants from identifiable emission points such as stacks and vents (Karri et al. 2024). These sources are often responsible for sustained emissions in industrialised and densely populated urban areas. To monitor and quantify emissions from these sources, many countries have developed regulatory tools and emission modelling frameworks. For instance, the United States EPA utilises the Motor Vehicle Emissions Simulator (MOVES) model to estimate both micro- and macro-scale emission factors for vehicle emission factors (U.S. Environmental Protection Agency 2024). Tables 2 and 3 summarise common anthropogenic and natural sources of air pollution respectively.

Table 2: Anthropogenic Sources of Air Pollution (Manisalidis et al. 2020, Kaur and Pandey 2021, Singh et al. 2025, Filonchyk and Peterson 2024).

Source	Description
Stationary	The emissions from Power plants, factories, waste incinerators, and home heating systems are among the sources of these emissions. In most developing countries, the widespread use of traditional biomass fuels, which include crop waste, animal dung, and wood, is a major contributor of air pollution.
Mobile	Also, transportation systems, such as cars, ships, and airplanes contribute to air pollution. These sources release particulate matter, NO _x , CO, and volatile organic compounds (VOCs).
Agricultural and Forestry Practices	Controlled or prescribed burning is commonly used in agriculture and forest management. Although fire plays a role in ecosystem renewal, such practices release smoke, CO, and particulate matter into the atmosphere.
Chemical and Industrial Products	Solvents, paints, aerosol sprays, varnishes, and other industrial chemicals release VOCs that contribute to the development of smog and secondary pollutants.
Waste Management	Organic waste decomposes in landfills to produce methane (CH ₄), a potent greenhouse gas and flammable material that can blow up in small areas
Military Activities	Hazardous pollutants can be released into the atmosphere during military operations as nuclear weapons, explosives, toxic gases, and rocket propellants are used.

Table 3: Natural Sources of Air Pollution (Singh et al. 2025).

Pollutant	Primary Sources
Dust Storms	In arid or semi-arid regions with little vegetation, wind erosion can create large quantities of airborne dust particles.
Radon Gas	Radon is a naturally occurring radioactive gas generated from the decay of uranium in soil and rock. It may build up indoors, especially in poorly ventilated spaces and is recognised as the second leading cause of lung cancer after tobacco smoke.
Biogenic Emissions from Vegetation	Vegetation emits VOCs that react with anthropogenic pollutants (e.g., NO _x and SO ₂) to form ozone and smog.
Methane Emissions from Livestock	Enteric fermentation in ruminant animals, particularly cattle, leads to the release of methane, a significant contributor to greenhouse gas accumulation.
Volcanic Activity	Eruptions emit SO ₂ , ash, and other compounds that degrade air quality locally and globally.
Wildfires	Natural vegetation fires emit smoke, CO, and fine PM, with regional and sometimes global effects.

3.1.2. Effect on Sustainability

Air pollution is widely recognised as a multidimensional sustainability challenge that directly affects several United Nations Sustainable Development Goals (SDGs) (Zhou et al. 2024). In particular, it poses a substantial threat to SDG 3 (Good Health and Well-being), as it is a primary driver of millions of premature deaths annually and is strongly associated with cardiovascular and respiratory diseases (Ogwu and Izah 2024). Furthermore, outdoor air pollution and particulate matter have been classified as carcinogenic to humans by the International Agency for Research on Cancer (IARC) (Maji et al. 2023). Air pollution also undermines SDG 11 (Sustainable Cities and Communities). Rapid urbanisation and increased rural-to-urban migration have led to growing urban populations, intensifying pollution levels and reducing overall urban liveability. Elevated concentrations of air pollution contribute to unhealthy living conditions, environmental degradation, and increased risks to both human health and ecosystems (Lei et al. 2024). In addition, air pollution presents a significant challenge to SDG 13 (Climate Action). Certain pollutants, such as black carbon and tropo-spheric ozone, act as short-lived climate forcers that contribute to global warming and exacerbate climate change impacts (Lei et al. 2024).

Framing air pollution within the context of the SDGs highlights the need for integrated and multidisciplinary solutions that at the same time address human health, environmental sustainability, and socio-economic development.

3.1.3. Effect on Health

A substantial body of literature has identified PM_{2.5}, PM₁₀, SO₂, NO₂, as the principal air pollutants and ground-level ozone, as major contributors to adverse health outcomes, including lung cancer, respiratory illnesses, cardiovascular diseases, and increased mortality (World Health Organization (WHO) 2023). Short-term exposure to these pollutants typically results in conditions such as coughing, wheezing, and shortness of breath, whereas long-term exposure significantly increases the risk of serious health effects such as asthma, chronic obstructive pulmonary disease (COPD) and ischemic heart disease (Manisalidis et al. 2020, World Health Organization (WHO) 2023). The health risks are particularly severe in low- and middle-income countries (LMICs), where air pollution levels frequently exceed WHO guidelines and healthcare systems are often less equipped to manage the resulting disease burden (Syuhada et al. 2023). Empirical studies further highlight the scale and geographical variability of these impacts. For example, the 2017 India State-Level Disease Burden Initiative reported substantial increases in mortality, disease burden, and reduced life expectancy attributable to air pollution exposure (Balakrishnan et al. 2019). High annual concentrations of PM_{2.5} have been associated with substantial premature mortality and years of lost life in Jakarta (Syuhada et al. 2023, Institute for Health Metrics and Evaluation (IHME) 2020). Regional analyses reinforce the uneven distribution of risk: South America recorded approximately 70,600 cardiovascular deaths and 1.74 million disability-adjusted life years (DALYs) associated with PM_{2.5} exposure in 2019 (de Oliveira Salerno et al. 2023), while South Asia reported around 1.02 million deaths, largely caused by residential fuel combustion, industrial emissions, and power generation (Chatterjee et al. 2023). Similarly, global burden estimates indicate approximately 1.67 million premature deaths linked to air pollution in India alone (Jaganathan et al. 2024). In Africa, 1.1 million people died in 2019 as a result of contaminated air, with almost 700,000 caused by household pollution and an estimated loss of 1.96 billion IQ points among children (Jaganathan et al. 2024). Such statistics reinforce that while the medical repercussions of poor air quality are universal, their root causes and specific impacts vary drastically by geography.

The effects of air pollution are not experienced uniformly across demographic groups. While young children are particularly vulnerable because of their developing respiratory systems, older adults (aged 50 years and above) are also at risk from developing non-communicable diseases like COPD, stroke, diabetes, and cardiovascular disease (Health Effects Institute 2024). Comparative studies indicate that populations in LMICs are disproportionately affected compared to those in high-income countries, largely due to rapid urbanisation, weaker environmental regulations, and continued reliance on biomass fuels (Rentschler and Leonova 2023). In several major cities, including Delhi and Dhaka, annual PM_{2.5} levels often surpass the WHO recommended limit (5 µg/m³), with concentrations typically ranging from 90 to 100 µg/m³, which is approximately 20 times the recommended limit (Chetna et al. 2023). Similarly, some cities in sub-Saharan Africa, such as Lagos, have consistently high levels of Air Quality Index (AQI) (>150) in the “Unhealthy” and “Very Unhealthy” ranges (Suriano et al., 2024, Odubango et al., 2024). These differences reflect regional variations in exposure and population vulnerability, driven by differences in industrialisation, urbanisation, energy use, and environmental governance. Fig. 3 illustrates the mortality rates attributed to ambient PM_{2.5} and O₃ exposure in 2021 by demographic group (Health Effects Institute 2024).

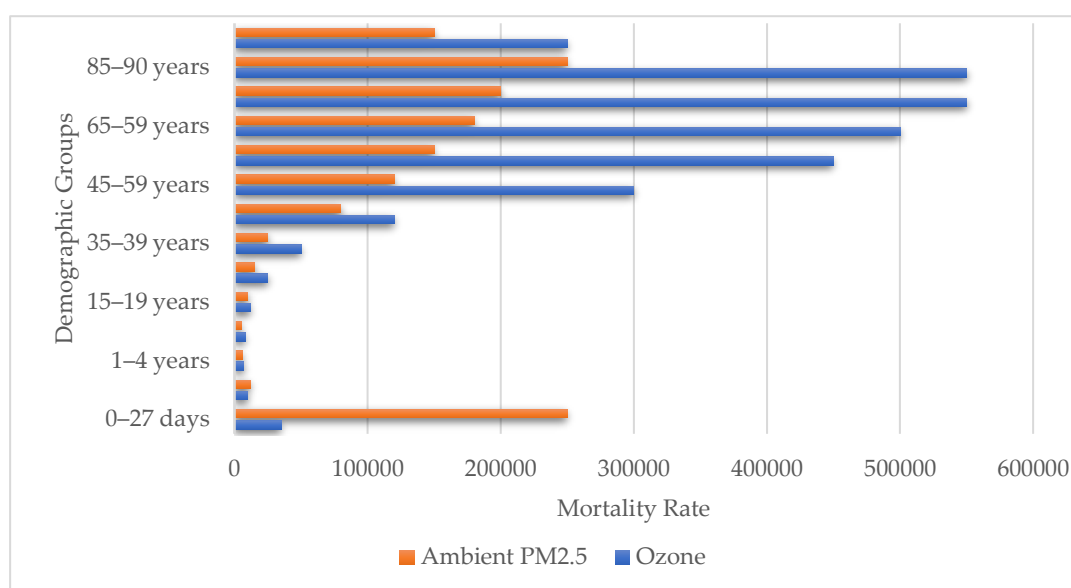


Fig. 3: Mortality Rate in 2021 by Demographic Group

3.1.4. Effect on the Environment

Beyond its effects on human health, air pollution poses significant threats to both biotic and abiotic components of ecosystems. Air pollutants arising from anthropogenic and natural sources contribute to environmental degradation at multiple scales. Elevated concentrations of pollutants such as SO_2 contribute to the formation of haze and photochemical smog, which reduce air quality and visibility while disrupting ecological processes (Wang et al. 2020). In addition, atmospheric reactions involving SO_2 and NO_2 lead to the formation of acid rain. Acid deposition has far-reaching environmental consequences, including soil acidification, freshwater acidification, forest degradation, and reduced agricultural productivity (Environmental Protection Agency (EPA) 2022). Increased concentrations of particulate matter (PM_{10} and $\text{PM}_{2.5}$) further contribute to environmental degradation by reducing visibility and impairing transportation systems, often resulting in disruptions such as traffic congestion (Zhang et al. 2021).

Air pollution also adversely affects vegetation and plant physiology. The Air Pollution Tolerance Index (APTI) has been widely used to assess plant responses to pollutant exposure. High levels of air pollutants can alter physiological and biochemical characteristics in vegetation including chlorophyll content, pH levels, ascorbic acid, and relative water content, ultimately reducing plant health, growth, and productivity (Singh et al. 2023, Bandara and Dissanayake 2021, Achakzai et al. 2017). Furthermore, tropospheric ozone stunts vegetative development and interferes with photosynthetic processes, which ultimately reduces agricultural yields and jeopardises global food security (Nowroz et al. 2024, Ramya et al. 2023). Secondary pollutants, such as smog formed through reactions between anthropogenic emissions and naturally occurring volatile organic compounds (VOCs), further degrade both air and water quality (Zhou et al. 2023). In regions with intensive mining and industrial activities, such as South Africa, pollutants including mercury are released into the environment through processes such as gold mining and coal combustion. These contaminants accumulate in aquatic systems and can have severe ecological consequences through bioaccumulation and biomagnification, ultimately threatening aquatic biodiversity and ecosystem stability.

3.1.5. Effect on the Economy

Air pollution is closely associated with a variety of economic activities, such as transportation powered by ICE vehicles, industrialisation, fossil fuel combustion, urban and agricultural waste burning, and deforestation (Arshad et al. 2024, Herath Bandara and Thilakarathne 2025, Pandya et al. 2022, Han et al. 2023). These activities, often compounded by inadequate environmental regulations and inefficient transport infrastructure, contribute significantly to increased emissions and environmental degradation (Al Alim et al. 2025, Lopez-Aparicio et al. 2025). The resulting economic impacts are substantial and vary across regions, often manifesting as reduced productivity, increased healthcare costs, and measurable losses in gross domestic product (GDP). The global financial burden of atmospheric degradation is staggering. For instance, World Bank assessments revealed that contaminated air costs Indonesia more than USD 220 billion annually, representing approximately 7% of the country's GDP (Awe et al. 2021). In India, the economic costs associated with premature mortality and morbidity due to air pollution have been estimated at USD 28.8 billion and USD 8.0 billion, respectively, accounting for about 1.4% of the national GDP (Pandey et al. 2021). These figures highlight the significant financial burden imposed by pollution-related health outcomes.

At a more localised level, studies have demonstrated the direct relationship between air pollution exposure and economic losses. For example, Syuhada et al. (2023) examined the health and economic impacts of PM_{2.5} and ozone (O₃) pollution across five districts in Jakarta, reporting increased mortality rates, hospital admissions, and associated economic costs, particularly among vulnerable populations such as children. Similarly, in Africa, air pollution imposes substantial economic losses. Fisher et al. (2021) estimated that in 2019, air pollution contributed to economic losses of approximately USD 3.02 billion in Ethiopia (1.16% of GDP), USD 1.63 billion in Ghana (0.95% of GDP), and USD 349 million in Rwanda (1.19% of GDP). These findings underscore the significant and geographically uneven economic burden of air pollution across both developed and developing regions. The impacts extend beyond direct financial losses to include reduced labour productivity, increased healthcare expenditure, and long-term constraints on economic growth. Addressing air contamination is consequently essential not just for safeguarding ecological and human health, but also for ensuring long-term, sustainable economic progress.

The substantial health, environmental, and economic consequences of air pollution underscore the need for accurate monitoring and reliable forecasting systems that enable timely interventions, optimise mitigation strategies, and support evidence-based environmental policy and sustainable urban planning. The following section therefore reviews contemporary statistical, machine learning, and deep learning approaches for air quality forecasting. Given the substantial health, environmental, and economic effects of poor air quality, the demand for precise predictive modelling and atmospheric tracking has surged, as explored in the subsequent section.

3.2. Air Pollution Prediction Models

Air quality and pollutant concentrations have become critical indicators of both environmental integrity and human well-being (Chen et al. 2024, Alkhanani 2025). Consequently, the monitoring and forecasting of air quality have received increasing attention from governments, research institutions, and environmental stakeholders (Relvas et al. 2025, Nugroho et al. 2025, Zhu et al. 2023). Effective air pollution forecasting relies heavily on time-series data, where historical observations are used to estimate current conditions and anticipate future

pollutant concentrations (Rosca et al. 2025, Lawrence and Bhathmanabhan 2026, Chan et al. 2025, Rodríguez-García et al. 2025). In response to this growing need, the number of air quality monitoring and data collection stations has expanded significantly worldwide in recent years (Sicard et al. 2023). Alongside this expansion, a wide range of modelling approaches have been developed to support air quality monitoring and forecasting. These methods range from traditional statistical models to recent data-driven techniques, such as machine learning and deep learning, that are increasingly employed to capture complex temporal patterns and improve air quality forecasting. Thus, the following subsections provide a detailed overview of these approaches, starting with statistical models, then moving to machine learning and more specifically to supervised learning and deep learning methods.

3.2.1. Statistical Models

Classical statistical frameworks formed the foundation to the early evolution of atmospheric forecasting and are still relevant in air quality forecasting (Lei et al. 2019, Liao et al. 2021). These models are typically used to examine past time-series data to identify patterns, such as seasonal, cyclical and long-term trends in pollutant concentrations (Devasekhar and Natarajan 2023). Among the most widely used techniques are the Autoregressive Moving Average (ARMA) model (Bartholomew 1971), the Autoregressive Integrated Moving Average (ARIMA) model, and its seasonal extension, the Seasonal Autoregressive Integrated Moving Average (SARIMA) model (Samal et al. 2019).

Several studies have applied these models to air pollution forecasting, though with varying degrees of methodological transparency. Pant et al. (2023) applied a seasonal ARMA model to forecast the AQI in Dehradun, India, capturing periodic variations in pollutant levels. However, the model's performance was reported solely via the Akaike Information Criterion ($AIC = 630.19$), with no RMSE or MAE provided thereby limiting to an independent assessment of forecasting accuracy. Compared to ARMA, ARIMA models incorporate an integration component that enables the modelling of non-stationary data, making them more suitable for real-world air quality datasets. Borhani et al. (2023) utilised the Box–Jenkins ARIMA framework to analyse SO_2 , $PM_{2.5}$, and PM_{10} concentrations in Tehran between 2011 and 2020, identifying seasonal peaks during winter and a gradual long-term decline in SO_2 levels. The study reported $R^2 = 0.74$ and $RMSE = 2.18$ ppb, but failed to explicitly specify a train–test partition, and used only 21 monitoring stations, which the authors acknowledged did not provide comprehensive spatial coverage of the city.. Likewise, Meenakshi Sundaram et al. (2022) developed a multi-pollutant air quality index prediction model with ARIMA with an RMSE below 8 and a 90% accuracy. However, the study did not report a standard regression performance metric and does not have a clear train–test split which has a high risk of temporal data leakage if a random partitioning method was employed. To address the seasonality in a more explicit way, SARIMA models have been widely adopted because they explicitly model seasonal patterns of the cyclical variations that are present in the time-series data. Bunnag (2024) used SARIMA and SARIMA-GARCH models to predict PM_{10} levels and analyse its relationship with NO_2 and O_3 concentrations in Bangkok, obtaining MAE values of 15.30 (SARIMA) and 17.28 (SARIMA-GARCH). The models, however, were fitted to the entire 2008–2023 dataset without a separate hold-out period, so reported errors are in-sample errors and not true out-of-sample forecasting errors. Unlike other reviewed statistical studies, Bhatti et al. (2021) applied a SARIMA model to $PM_{2.5}$ data from Lahore, Pakistan, using a

clearly defined chronological train–test split (2014–2018 training, 2019 testing) and producing a mean percentage error of -0.3%, which may hide larger individual prediction errors not apparent at the mean level.

Overall, these studies indicate that statistical models such as ARMA, ARIMA, and SARIMA perform reasonably well for short-term and seasonal forecasting, particularly when dealing with univariate datasets or moderately complex systems. However, a closer methodological evaluation reveals limitations beyond their well-documented difficulty in capturing nonlinear relationships and multi-factor interactions. Several studies rely exclusively on in-sample fit statistics such as AIC rather than held-out error metrics (Pant et al. 2023; Bunnag 2024), train–test partitioning is frequently undisclosed or ambiguous, creating a risk of temporal leakage (Borhani et al. 2023; Meenakshi Sundaram et al. 2022), and none of the reviewed studies benchmark their models against even a naive persistence baseline. This pattern of inadequate validation is consistent with concerns raised in adjacent machine-learning review literature that caution that, the apparent effectiveness of a modelling technique cannot be meaningfully assessed without first scrutinizing the dataset generation process, feature/validation strategy, and test conditions underlying the reported result (Alimi et al. 2020).

Collectively, these findings suggest that the credibility of statistical forecasting models depends as much on rigorous validation design as on the underlying modelling approach itself. Studies employing transparent chronological validation and clearly defined testing procedures provide more convincing evidence of predictive capability than those relying solely on in-sample statistics or unspecified data partitions. Consequently, future statistical forecasting studies should prioritise transparent reporting of evaluation protocols, leakage-aware validation, external testing, and comparisons against naïve persistence baselines to ensure that reported performance reflects genuine forecasting skill rather than favourable evaluation protocols. Table 4 summarizes selected recent studies that have applied statistical models for air pollution prediction across different regions, including a critical appraisal of their reported validation strategies.

Table 4: Overview of Recently Proposed Statistical Models for Air Pollution Forecasting/Prediction.

Ref.	Feature	Model/AI- gorithm	Pollutant	Dataset	Summary of Contribution
Pant et al. (2023)	Historical AQI Data; Seasonal Trend Analysis	Seasonal ARMA	PM ₁₀	Dehradun, India (2017–2021); CPCB (Uttarakhand Govt.)	Created AQI seasonality modelling on ARMA and underscored regional AQI prediction requirements in local health-planning. It did not break it down into components of pollutants but only used AQI.
Borhani et al. (2023)	Pollution Data and Meteoro- logical Data	ARIMA	SO ₂	Tehran Air Quality Control Company (AQCC), Depart- ment of Environment (DOE) and Iran Me- teorological Organi- zation (IRIMO)	Examined 10 years of multi-station data; mapped pollution hotspots with ArcGIS/IDW; best-fit ARIMA (3,1,4) achieved good accuracy for SO ₂ ($R^2=0.74$, RMSE=2.18). Limited to univariate ARIMA (no traf- fic/emission data); struggled with PM ₁₀ predictions due to meteorological complexity.
Meenakshi Sundaram et al. (2022)	Pollutant Data	ARIMA	AQI	India Pollution Con- trol Board (IPCB) for the years 2015– 2020	Added various pollutants to predict AQI, focused on minimising RMSE and health concern to the popula- tion. No seasonal modelling or neural comparison is provided; information about the data is not detailed.

Bunnag (2024)	Pollutant Data (NO ₂ , O ₃); His- torical PM ₁₀	SARIMA, SARIMA- GARCH	PM ₁₀	Ban- somdejchaopraya Rajabhat University Station (O2T), span- ning from January 2008 to July 2023	The results demonstrated that SARIMA was accu- rate and provided moderate correlation with NO ₂ and O ₃ in comparison to the SARIMA-GARCH. This study focused on only a single pollutant, and the me- teorological data was not observed.
	Federal and Provin- cial Environmental Protection Agencies collected air quality monitoring data for 6 years (2014–2019), La- hore, Pakistan			Developed optimal SARIMA (1,0,2) (1,0,0) model with very low mean error (≈0%); produced 3-year forecasts with confidence bands; identified Afghani- stan as a significant pollution source using HYSPLIT. Relied only on historical PM _{2.5} data (no meteorological or socioeconomic predictors); miss- ing station data reduced temporal robustness and continuity.	
Bhatti et al. (2021)	Pollutant Data	SARIMA	PM _{2.5}		

3.2.2. Machine Learning Models

Machine learning models have emerged as a highly favoured paradigm for forecasting atmospheric conditions driven by the rapid advancements in big data analytics and artificial intelligence (Tello-Leal et al. 2024, Chalavada et al. 2025, Chen et al. 2025). Traditional statistical models often struggle to properly model the non-linear, high dimensional, dynamic nature of air pollution data, particularly when modelling large-scale time-series datasets with meteorological and pollutant variables (Lawrence and Bhathmanabhan 2026). As a result, the scientific community have increasingly adopted machine learning paradigms, because of their capacity to mine large, heterogenous datasets and extract highly intricate, hidden patterns (Mubeen et al. 2025). Among the various artificial intelligence approaches, supervised learning techniques are the most widely applied in air pollution prediction because they utilise historically labelled datasets to learn relationships between historical predictor variables and pollutant concentrations or air quality indices (AQI) (Oumoulyte et al. 2025, Kumar et al. 2025). In the recent literature, machine learning approaches for air pollution forecasting are generally classified into two broad categories: traditional supervised learning models and deep learning models.

The development of machine learning-based air pollution prediction systems typically involves three major stages: data collection, data preprocessing, and predictive modelling. The quality, completeness, and representativeness of the dataset play a fundamental role in determining the reliability and generalisability of predictive models (Hua et al. 2024). Maleki et al. (2022) similarly highlight that the strength of the machine learning models depends heavily on the quality of the datasets used to train and validate the models. Consequently, the methodological quality of data preparation and validation is as important as the selection of the predictive algorithm itself. The following subsections therefore review traditional supervised learning and deep learning approaches while critically evaluating their methodological strengths and limitations.

3.2.2.1. Data Collection

Recent studies increasingly integrate multiple data sources to improve the spatial and temporal representation of air pollution patterns (Chen et al. 2022, Cao et al. 2020, Chatterjee et al. 2025). One of the most widely used sources is air quality monitoring stations, which provide high-resolution, time-stamped measurements of pollutants such as PM_{2.5}, PM₁₀, NO₂, SO₂, and CO (Simo et al. 2020). Operated by governmental agencies and research institutions, these stations deliver accurate ground-truth data critical for local pollution monitoring,

model training and validation (Petrić et al. 2025). However, model accuracy and coverage are greatly affected by the density, spatial distribution, and maintenance of the monitoring stations.

Many studies also use satellite remote sensing data, such as Moderate Resolution Imaging Spectroradiometer (MODIS), Sentinel, and Aerosol Optical Depth (AOD) imagery (Miller et al 2026, Kathirolu et al 2023, Ayek et al 2025). Satellite data provides a large spatial coverage, more comprehensive data sources, and can provide information about atmospheric transport processes and the distribution of pollutants in the region, which is beneficial for monitoring areas with limited monitoring infrastructure (Ma et al. 2022). Meteorological variables also play a fundamental role in air quality modelling. Variables such as relative humidity, temperature, wind speed, wind direction, atmospheric pressure, directly influence the dispersion and accumulation of pollutants. Integrating these variables enables predictive models to capture atmospheric dynamics more effectively and improves the representation of temporal variations in pollutant concentrations. Additional data sources frequently used in the literature include traffic records, industrial emission records, socioeconomic indicators, land use and land cover (LULC) data, and environmental incident records, including wildfires and dust storms. These complementary datasets can be used to improve the ability of models to link pollution trends to environmental events and human activity (Hait and Sahu 2024, Du et al. 2023, Gaur and Singh 2023).

When primary data are unavailable or insufficient, open-access repositories provide valuable alternatives for air quality research. OpenAQ and other platforms aggregate air quality data from governments and research institutes from over 100 countries, facilitating analyses between countries and improving the generalisability of models (OpenAQ 2024). Similarly, the UCI Machine Learning Repository offers benchmark datasets such as Beijing Air Quality dataset that are widely used for algorithm development and testing (Chen 2017). Other widely used resources include the World Air Quality Index (WAQI) platform and AirVisual, API based providing real-time AQI data for global cities (AQICN 2025, IQAIR 2025). Furthermore, the Copernicus Atmospheric Monitoring Service (CAMS) provides satellite-based atmospheric composition and air quality products suitable for large-scale forecasting and environmental monitoring (Peuch et al. 2022). Although these repositories greatly enhance data accessibility and reproducibility, differences in spatial resolution, measurement protocols, temporal coverage, and data quality should be carefully considered when integrating data from multiple sources.

3.2.2.2. Data Preprocessing

Following data collection, a data pre-processing stage is typically required to ensure that the data is of high quality and to enhance temporal robustness. Data preprocessing techniques include feature selection, feature extraction, normalisation and dimensionality reduction are essential to enabling machine learning models to learn effectively from the dataset (Alimi et al. 2020). The primary objective of pre-processing is to transform raw datasets into structured datasets suitable for training, testing, and validation.

One of the most recurrent issues with air pollution data is the missing values, which is frequently due to the malfunction of the measuring equipment, regular maintenance, or communication failures during data transmission. To address the issue imputation, interpolation and forward/backward filling methods are common approaches employed by researchers (Zhang and Thorburn 2022). Outlier detection is also crucial since the model could be affected by abnormal sensor data or environmental variations (Diallo et al. 2024, AlSalehy and Bailey 2025). Frequently used statistical techniques include z-score analysis and inter-quartile range (IQR) methods together with optimisation strategies like Genetic Algorithms (GA) and Ant Colony Optimisation (Dongre et al.

2025). Another crucial preprocessing step is temporal and spatial alignment of heterogeneous datasets, integrating satellite observations with ground-based monitoring data (Cabello-Solorzano et al. 2023). In addition, predictor variables are often normalised using techniques such as min-max scaling or z-score standardisation to ensure that they are within similar numerical ranges, thereby improving model stability and convergence in the training process (De Amorim et al. 2023, Rahmad Ramadhan and Anne Mudya 2024). To address limited or imbalanced datasets, several studies employ data augmentation techniques such as the Synthetic Minority Oversampling Technique (SMOTE) and Generative Adversarial Networks (GANs) to create synthetic samples and enhance class balance (Barid et al. 2024, Rajakumari and Lanjewar 2023).

Although preprocessing generally improves temporal robustness, the methodological choices made during this stage can substantially influence model reliability and reproducibility. For example, performing feature selection, normalisation, or imputation before data partitioning can introduce temporal data leakage, resulting in overly optimistic performance estimates. Consequently, preprocessing operations should be applied within the training data and subsequently transferred to the validation and test datasets to ensure fair and reproducible model evaluation.

3.2.2.3. Supervised Learning Models

Traditional supervised learning models are still among the most widely used approaches for predicting pollutant concentration largely due to their ability to learn the relationships among the pollutants and the labelled input variables (Zhang et al. 2022, Essamlali et al. 2024). Commonly applied models are Multiple Linear Regression (MLR), Random Forest (RF), Support Vector Regression (SVR), Artificial Neural Networks (ANN), Gradient Boosting (GB), and eXtreme Gradient Boosting (XGBoost). Devasekhar and Natarajan (2023) employed linear regression models to predict annual PM_{2.5} concentrations in Hyderabad, India. However, the reported models yielded negative R² values across several predictor combinations (e.g., R² = -0.07 for humidity, -0.17 for dew point using AdaBoost), indicating performance worse than a simple mean-based estimator, and the train–test protocol was not clearly specified. These findings should therefore be interpreted with considerable caution rather than as a representative example of supervised learning's applicability to PM_{2.5} prediction. Lei et al. (2022) compared MLR, SVR, GB, and RF models for air quality prediction in Macao using a strict chronological split across three independent test years (2019, 2020, 2021), where RF proved the most robust to the COVID-19-driven regime shift in 2020 (R² for PM_{2.5} declining from 0.88 to 0.65) while SVR degraded more severely (R² = 0.57). This represents one of the more rigorous validation designs among the reviewed studies. Maltare and Vahora (2023) reported an SVR-RBF model with R² = 0.999 and RMSE = 4.94 for AQI prediction outperforming both LSTM and SARIMA approaches. However, an R² this close to unity on a variable ranging from 0 to 500, together with an LSTM RMSE of 4.08, is more plausibly indicative of overfitting or a data-leakage artefact than of genuine predictive superiority, particularly as the margin between the competing models is marginal. Similarly, Zhao et al. (2022) enhanced SVR performance through Q-learning and Bees Swarm Optimisation for automated feature selection, achieving improved RMSE, MAPE, and NSE metrics compared to Ridge Regression and MLR. Nevertheless, R² varied considerably across the three cities tested (0.880–0.981), suggesting the presence of site-specific factors not captured by the feature-selection procedure, and no comparison against temporal models such as LSTM was reported.

Ensemble tree-based methods have also become increasingly popular as a result of their temporal robustness and capability to capture nonlinear relationships. Agibayeva et al. (2022) demonstrated the efficiency of RF models in forecasting PM_{2.5} concentrations during the COVID-19 period. However, the study evaluated same-day prediction using contemporaneous rather than lagged predictors. Consequently, the reported R² (up to 0.99) largely reflect contemporaneous correlation between PM_{2.5} and PM₁₀ and not to the model's forecasting accuracy. This distinction is important when interpreting temporal robustness. Similarly, Serrao (2023) applied Regression Trees and SVR to model the spatiotemporal variation in Delhi, which led to a competitive R² of 0.874. However, the time horizon of the study was substantially longer (365 days), which is a major drawback when compared with other similar studies, which generally used time horizons of 1-7 days, and hence long-term forecasts are inherently more uncertain. In other studies, hyperparameter tuning and advanced feature engineering methods were used to enhance the performance of RF, with the models often achieving better results than XGBoost or Support Vector Machines (SVM) (Alolayan et al. 2023, Isaev et al. 2022). However, Isaev et al. (2022) did not report an exact train–test split and did not provide a baseline comparison to evaluate the added value of the proposed approach, as the RMSE values spanned a wide range for each pollutant (e.g., 2.0 for CO versus 45 for NO₂). Alolayan et al. (2023) found that R² values increased significantly for the proxy emission features (33–61% relative increase), but remained relatively low (absolute R² 0.49–0.75). Ren et al. (2020) developed a sound methodological framework by testing 13 algorithms with 6 validation strategies: 10-fold spatial, leave-month-out temporal, and peak-specific validation. Compared with linear models, RMSE decreased by 37% for both RF and XGBoost, and sample-weight tuning maintained predictive accuracy during periods of peak ozone levels (>70 ppb). The study also showed the interpretability of ML models by using Partial Dependence Plots (PDP) and Individual Conditional Expectation (ICE), showing physically meaningful nonlinear relationships, such as NO_x–ozone chemistry.

Paul et al. (2022) created the Canadian Optimised Statistical Smoke Exposure Model (CanOSSEM) using Random Forest to estimate daily PM_{2.5} from biomass smoke, with an overall RMSE of 2.96 µg/m³, across a large geographical area. However, leave-one-region-out and leave-one-year-out validation found the model performance deteriorated, as RMSE rose to 15.78 µg/m³ during the extreme 2018 wildfire season, and systematically underestimated the number of severe interface wildfires, including the 2016 Fort McMurray wildfire. To solve the data imbalance problem, Asgari et al. (2022) used a Distributed Random Forest with spatiotemporal partitioning for AQI classification in Tehran, which achieved a relative improvement in the precision of minority-class compared to Naive Bayes of 44–82%. However, this limited the global accuracy of the model to 68%, and the lack of a persistence baseline prevented full assessment of the framework's practical benefit. Hybrid methods of neural networks and traditional statistical methods have also been investigated. Rahman et al. (2023) combined ANN and Triple Exponential Smoothing (TES) to forecast PM_{2.5} levels in Malaysia and obtained significantly smaller prediction errors when compared to individual models.

In general, the reviewed studies showed that supervised machine learning models consistently achieve high predictive accuracy and outperform the conventional statistical models in dealing with high-dimensional and nonlinear air pollution datasets. This conclusion should be interpreted in the light of the methodological inconsistencies noted above, as near-unity R² values (Maltare and Vahora 2023) should be treated with suspicion of overfitting or leakage, and studies with concurrent predictors (Agibayeva et al. 2022) should not be

classified as forecasting models, and at least one study (Devasekhar and Natarajan 2023) reported negative R^2 values, which should be interpreted with caution when assessing algorithm performance.

The collective findings from the reviewed works suggest that the dataset quality, feature engineering, and validation strategy are more important determinants of temporal robustness in supervised learning, as opposed to the choice of algorithm alone. Ensemble methods such as Random Forest and Gradient Boosting consistently demonstrate reliable performance when evaluated using rigorous temporal or spatial validation schemes, whereas exceptionally high reported accuracies obtained without transparent validation procedures should be interpreted cautiously. These findings reinforce the importance of methodological transparency and reproducibility as prerequisites for meaningful comparison between competing machine learning models. Table 5 summarizes recent studies that have applied machine learning techniques for atmospheric pollutant forecasting across various geographic contexts. Collectively, these findings demonstrate that robust validation design, transparent reporting, and appropriate benchmarking are more reliable indicators of forecasting quality than exceptionally high accuracy metrics reported in isolation.

Table 5: Overview of Conventional Machine Learning Based Models for Air Pollution Forecasting/Prediction.

Ref.	Feature	Model/Algorithm	Pollution	Dataset	Contribution in Summary
Devasekhar and Natarajan (2023)	Meteorological Data	LR	PM _{2.5}	Air quality Data in India	The findings demonstrated a regional variance in the prediction, with an advisable need to combine diverse approaches to enhance accuracy. The model involved only annual forecasting of PM _{2.5} with no real-time or hourly prediction analysis. It did not consider more than one pollutant and limited long-term forecasting.
Lei et al. (2022)	Meteorological Data Surface Observation	MLR	PM _{2.5} and PM ₁₀	Macao Meteorological and Geophysical Bureau (SMG), and the upper-sounding data were obtained from Hong Kong King's Park Station (Number 45004)	MLR outperformed complex models during normal periods (2019–2021). The model underperformed during COVID-19 anomalies as it lacked temporal robustness to sudden emissions changes. In addition, it paid more attention to method comparison instead of the feature analysis or system integration.
Maltare and Vahora (2023)	Meteorological Data	SVR	AQI (CO, PM _{2.5} and O ₃)	SAFAR January 2015 to January 2021. It is observed from 7 different air quality monitoring stations of Ahmedabad	Compared SARIMA, SVM, and LSTM; SVM with RBF kernel achieved the best performance. Highlighted the importance of preprocessing for improving accuracy. Did not address feature impact, real-time deployment, or multivariate dependencies.
Zhao et al. (2022)	Meteorological Data	SAC-QBSO-SVR	AQI	AQIStudy January 1st, 2019, to December 31st, 2020	Proposed spatio-temporal framework with feature selection, reinforcement learning, and trigonometric tuning in order to enhance performance. Model demonstrated strong AQI predictions across 34 cities in the Yangtze River Delta. High complexity of the algorithms and focused only on AQI, not individual pollutants.
Agibayeva et al. (2022)	Meteorological Data	RF	PM _{2.5}	Two air pollution monitoring stations in Astana, Kazakhstan into two periods (i.e., heating (October–April) and non-heating (May–September))	Random Forest outperformed MLR (R^2 up to 0.98) for PM _{2.5} forecasting and the authors used a DALY based health risk assessment in an attempt to measure the burden to public health. The model examined only a single year data at two sites with regards to PM _{2.5} , reducing generality and temporal stability.
Serrao (2023)	Meteorological and Pollutant Data	RT-SVR	SO ₂ , NO ₂ , PM _{2.5} , O ₃ , PM ₁₀ ,	Central Pollution Control Board (CPCB) of India from January 2015 to June 2023	Designed a hybrid RT-SVR model that will be more resistant to fluctuations in pollution data and will have more temporal robustness as compared to the individual models. Not widely

			Benzene, CO, and Temperature		validated; not extensively empirically tested and not broken-down pollution-wise
Isaev et al. (2022)	Meteorological Data	RF	NO ₂ , PM _{2.5} , NO, PM ₁₀ , CO,	Kyrgyzhydromet	Used weather observations to predict pollution and highlighted PM _{2.5} reduction during COVID-19 lockdown (54% decrease). Nevertheless, limited model optimisation and results may not generalise to non-lockdown scenarios.
Alolayan et al. (2023)	Meteorological Data Power Plant Fuel Consumption	RF	SO ₂ , NO ₂	six power plants and five air quality monitoring stations in the State of Kuwait	Compared weather-only, fuel-only, and combined models where the combined model improved NO ₂ prediction by 66% and SO ₂ by 27%. It also identified feature equivalence between fuel and meteorology. However, it was limited to only two pollutants and only the daily average concentrations were used, thereby limiting its use in real-time forecasting.
Paul et al. (2022)	Satellite, Emissions, Meteorology	RF	PM _{2.5}	Government of Canada Open Data	Developed a nationwide random forest model for PM _{2.5} exposure from biomass smoke. They generated fine resolution (5 km ²) exposure estimates that can be used in health studies. The study focused solely on PM _{2.5} making it vulnerable to large wildfire events, as well as the complexity limits real-time use.
				US EPA Air Quality System	
Ren et al. (2020)	Meteorological Data LULC	RF	O ₃	USGS National Elevation Dataset	Compared 13 ML and LUR models for nationwide ozone forecasting where RF and XGBoost outperformed conventional LUR. The study applied advanced validation and variable weighting. Focused only on ozone and the ML models are computationally costly and complex.
Rahman et al. (2023)	Meteorological and Pollutant Data	ANN-TES	PM _{2.5}	The stationary source emission data were retrieved from EPA's 2011 National Emission Inventory Malaysia's Department of Environment (DOE) and 65 air quality monitoring stations from January 2018 to December 2019	Formed a hybrid model in forecasting region specific PM _{2.5} in Malaysia due to outperforming other models in both high and medium pollution areas. Data variability results in a huge decline in performance in areas of low pollution.
Asgari et al. (2022)	Meteorological Data	RF, Hadoop and Spark	AQI	Tehran Air Quality Control	Proposed a distributed forecasting framework (Spark + Hadoop) for large, imbalanced spatio-temporal datasets where two-level partitioning improved the load balance. Distributed RF enhanced accuracy under data imbalance. High system complexity may limit scalability outside big-data infrastructure.

While traditional supervised learning models have demonstrated strong predictive capabilities, recent studies increasingly favour deep learning approaches because of their ability to learn complex temporal and nonlinear patterns from large-scale air pollution datasets.

3.2.2.4. Deep Learning Models

Deep learning models represent a specialised subset of machine learning techniques that employ multi-layered ANNs to learn complex and nonlinear patterns from large-scale datasets (Ekundayo and Ezugwu 2025, Taye 2023). Owing to the high-dimension nature, temporal dependencies, and volume of air pollution and meteorological datasets, deep learning approaches have become increasingly prominent in air quality prediction and forecasting tasks (Tello-Leal et al. 2024, Booth et al. 2025). In particular, recurrent neural networks (RNNs) and their variants have shown strong capability in modelling sequential time-series data, making them well-suited for forecasting pollutant concentrations and Air Quality Index (AQI) levels (Andrade et al. 2025, Chan et al. 2025, Waseem et al. 2022). Several studies report that deep learning models yield better temporal robustness than traditional machine

learning, though the magnitude of the reported gains warrants scrutiny in a number of cases. Aljanabi et al. (2021) compared a Multi-Layer Perceptron (MLP) model with conventional models such as SVR and XGBoost for forecasting NO_2 , SO_2 , and PM_{10} concentrations. After applying preprocessing techniques including noise reduction and Pearson correlation-based feature selection, the MLP model outperformed the competing approaches across multiple evaluation metrics. Wai and Yu (2023) evaluated ANN models against Computational Fluid Dynamics (CFD) methods for modelling urban PM_{10} dispersion, achieving comparable accuracy ($R = 0.82$ versus 0.81) at a fraction of the computational cost providing an important practical advantage for operational deployment, though the ANN's black-box nature was explicitly acknowledged by the authors.

Hybrid and ensemble deep learning approaches have also gained significant attention. Scagliotti et al. (2022) combined an ANN with a Self-Organizing Map (SOM) to predict $\text{PM}_{2.5}$ concentrations in Buenos Aires, reducing RMSE from 11.4 to $1.8 \mu\text{g}/\text{m}^3$. This dramatic improvement should, however, be understood in light of the fact that the task was using concurrent predictors rather than genuine forecasting, and the resulting model may be overfitted to the specific station and limited 2015–2017 data period. Aguilera et al. (2023) developed a new spatiotemporal ensemble approach (DRF, GBM, DL) combining wildfire data, LULC, meteorological data, and AOD measurements, obtaining an R^2 of 0.83 using a strict spatial cross-validation. Among deep learning approaches, LSTM networks have emerged as the most widely adopted models for air pollution prediction owing to their capacity to capture long-term temporal dependencies in sequential data. Gu et al. (2022) proposed HIPML, a hybrid Deep Neural Network (DNN)–NARMAX framework incorporating Orthogonal Forward Regression (OFR) for feature selection and tested using Beijing air quality data. The proposed model consistently outperformed benchmark approaches across multiple forecasting horizons. Notably, the study is among the few reviewed studies to provide interpretable, feature-level insight into model behaviour, representing a notable advance over black-box alternatives. However, Alghieth et al. (2021) and Arsov et al. (2020) reported that LSTM models outperformed ARIMA, Gated Recurrent Unit (GRU), and conventional RNN models in PM_{10} forecasting tasks. Alghieth et al. (2021), however, the reported $R^2 = 0.999$ and $\text{MAE} = 0.98$ for hourly PM_{10} prediction in a city where concentrations range from near-zero to over $1000 \mu\text{g}/\text{m}^3$ is implausibly precise, and in the absence of any baseline comparison it is not possible to determine whether this reflects genuine model ability, data leakage, overfitting to an undersized test set, or an unreported scaling artefact. Gladkova and Saychenko (2022) compared ARIMA, Prophet, and LSTM models across Russian cities and reported superior temporal robustness from the LSTM approach, though the underlying error values were reported in normalised rather than raw $\mu\text{g}/\text{m}^3$ units, complicating interpretation of absolute accuracy.

Several recent studies have enhanced LSTM architecture by incorporating additional contextual variables and optimisation techniques. Espinosa et al. (2021) and Cican et al. (2023) integrated meteorological, pollutant, and population data into LSTM and GRU frameworks for NO_2 forecasting. Cican et al. (2023) is particularly noteworthy for its methodological candour, reporting $R^2 < 0.8$ for both architectures across an exhaustive 5,832-configuration hyperparameter search, and explicitly acknowledging the limitations of the available predictor set rather than overstating model performance. Kachhara and Jain (2022) proposed a hybrid LSTM-ARIMA framework for AQI prediction across Indian cities, combining the strengths of statistical and deep learning approaches. Tao et al. (2023) integrated GEOS-Chem simulations and satellite-derived AOD data within an LSTM framework for air quality prediction in Iraq, reporting $R^2 = 0.92$ with rigorous spatial validation using MapCurve and Cramer's V statistics. Drewil and Al-Bahadili (2022) utilised Genetic Algorithms (GA) to optimise LSTM hyperparameters for forecasting $\text{PM}_{2.5}$,

PM₁₀, CO, and NO₂ concentrations, reporting RMSE improvements over Bi-LSTM and CLSTM benchmarks. However, the results were drawn from separate studies using different datasets, time periods, and evaluation protocols rather than from a like-for-like re-implementation. Additionally, the GA search itself used a population size and generation count of only four, which is generally considered too small for meaningful hyperparameter optimisation both of which factors weaken the claimed superiority. Advanced architectures combining multiple deep learning techniques have also demonstrated promising results. Yasmin et al. (2023) proposed AQIPred, a hybrid MLP-LSTM framework applied to the Beijing Multi-Site Air Quality dataset, achieving an R² of 0.95 and reportedly outperforming multiple state-of-the-art models. The accompanying MAE (0.0075) and MAPE (0.42%), however, are implausibly small for AQI values ranging from 0 to 500 and are more likely artefacts of data scaling or normalisation than genuine indicators of prediction accuracy. Also, the improvement over standalone LSTM and MLP baselines is, on closer review, marginal. Jesupriya et al. (2025) employed a Bidirectional LSTM (BiLSTM) combined with a multi-head attention mechanism to classify AQI severity levels in Indian cities, achieving a validation accuracy of 99.2%. However, the model's binary hazard-trigger component recorded a false-negative rate of 36.12%, which is a materially significant limitation for a safety-critical, real-time early-warning application, since missed hazard detections carry direct public-health consequences that the headline accuracy figure does not convey.

Overall, the reviewed studies demonstrate that deep learning models provide advantages in modelling complex, nonlinear, and high-dimensional air pollution datasets, and their ability to automatically extract temporal and spatial features has contributed to improved forecasting accuracy relative to conventional statistical and traditional machine learning models in several well-validated cases (e.g., Lei et al. 2022, Aguilera et al. 2023, Gu et al. 2022, Tao et al. 2023). This advantage should not, however, be generalised uncritically across the full body of reviewed work: a substantial subset of the headline LSTM and hybrid results including near-unity R² values, sub-1% MAPE figures, and cross-study "outperformance" claims computed on non-identical datasets. These results may reflect unaddressed concerns about data leakage, overfitting, artefacts related to scaling, or lack of comparability, and one high accuracy classification model hides a safety-relevant failure mode within the aggregated metric (Jesupriya et al. 2025). The current paradigm of offline training and online application is demonstrated to yield performance that is not necessarily adaptable to the current and dynamic operating conditions. The effective use of these techniques requires not only the availability of large amounts of data, significant computing resources, and careful hyperparameter tuning, but also reporting the results of validity checks in a transparent way that prevents leakage.

Overall, the reviewed evidence suggests that the advantages of deep learning models are conditional rather than universal. They are most useful when modelling large-scale nonlinear spatiotemporal datasets obtained under strict validation procedures and adequate training data. On the other hand, findings that are accompanied by unusually high prediction performance with unclear validation processes or across non-comparable datasets should be treated with caution. Future research should therefore place equal emphasis on model interpretability, uncertainty quantification, reproducibility, and real-world deployment alongside further improvements in predictive accuracy. An overview of recent deep learning implementations for air quality forecasting across diverse geographical areas, together with a critical appraisal of their reported validation strategies, is presented in Table 6.

Table 6: Overview of deep learning models for air pollution forecasting/prediction.

Ref.	Feature	Model/Algorithm	Pollution	Dataset	Contribution in Summary
Aljana bi et al. (2021)	Meteorological and Pollution Data	MLP	NO ₂ , SO ₂ and PM ₁₀	Daily Pollution Concentrations in King Al-Hussein Public Parks Station Dataset, 2019. ASU Weather Station Dataset, 2019.	MLP outperformed SVR, XGBoost, and DTR with R ² values above 90%. The study also examined sensor multiplicity and feature reduction to optimise performance. Limited exploration of deep learning methods beyond MLP (e.g., LSTM, hybrid approaches), restricting broader applicability.
Wai and Yu (2023)	Meteorological Parameters and PM ₁₀ Concentrations	ANN	PM ₁₀	Hong Kong Environmental Protection Department (EPD)	ANN showed high accuracy and faster execution compared to CFD in modelling PM ₁₀ dispersion in urban canyons. Accuracy of CFD model, however, makes it difficult to apply in the real-life scenario. Also, the accuracy of the ANN model deteriorated in the unseen environments.
Scagliotti et al. (2022)	Meteorological Data	ANN-SOM	PM _{2.5}	Github Repository	Use of self-organizing maps and mixed ML methods improved predictor selection and enhanced MLP forecasting performance in Buenos Aires. Nonetheless, it only focused on PM forecasting, limiting its transferability to other pollutants or cities with different environmental conditions.
Aguilera et al. (2023)	AOD, Meteorological Data, LULC and Wildfire Data	ANN-ensemble	PM _{2.5}	Wildfire PM _{2.5} California	Developed an ensemble model tailored to wildfire-specific PM _{2.5} imputation, offering a novel way to isolate smoke-related pollution. Lack of validation against direct smoke-PM _{2.5} measurements and absence of links to health outcome analysis limit reliability and impact.
Gu et al. (2022)	Air pollution, Season and Weather Measurements	HIP-ML	PM _{2.5}	UC Irvine Machine Learning Repository, (n.d.). UCI Machine Learning Repository 2012 to 2015 in Beijing, China	Proposed a Hybrid interpretable model (HIP-ML) combining DNN and NARMAX achieved strong short-term predictive accuracy (1-hr correlation = 0.987), especially effective for peak PM _{2.5} events. High complexity, restricted to PM _{2.5} , dependent on rich historical data; requires advanced feature engineering and expert tuning, reducing usability in real-world or resource-limited settings.
Alghith et al. (2021)	Meteorological Data	LSTM	PM ₁₀	General Authority of Meteorology and Environmental Protection in Saudi Arabia	Created LSTM based PM ₁₀ prediction model of Riyadh; obtained a good model of PM ₁₀ with high accuracy (R ² = 0.999); used data normalisation and data interpolation as a preprocessing step. Time-consuming data cleaning was the only aspect that was performed at PM ₁₀ .
Arsov et al. (2020)	Meteorological and Pollution Data	RNN-LSTM	PM ₁₀	Skopje Air Pollution (GitLab)	RNN and LSTM outperformed ARIMA for O ₃ and PM ₁₀ forecasts, particularly when meteorological inputs were included. Limited to short-term and regional predictions, lacking interpretability and generalisability.
Gladkova and Saychenko (2022)	Meteorological and Pollution Data	LSTM	PM _{2.5}	World Air Quality Index project 2021 (Air Quality Open Data Platform), Russia	The study compared ARIMA, Prophet, and LSTM for PM _{2.5} trends and highlighted the role of data visualisation and trend analysis in air quality management. It was restricted to long-term predictions; heavily dependent on data quality, which was often inadequate.
Espinosa et al. (2021)	Meteorological and Pollution Data	LSTM	NOx, NO ₂	2015 to 2017 Wrocław Contamination data are recorded by the Provincial Environment Protection Inspectorate traffic data are collected by the Traffic Public Transport Management Department of the Roads and City Maintenance Board	Combined deep learning (LSTM, GRU) with statistical models (RF), emphasising temporal robustness via statistical validation. High computational cost and heavy feature engineering requirements reduce scalability.

meteorological data belong to the Institute of Meteorology					
Cican et al. (2023)	Meteorological and Pollutant Data	LSTM	NO ₂	Romanian National Air Quality Monitoring Network 1 August 2021 to 1 July 2022 Visual Crossing Weather	Compared LSTM and GRU for NO ₂ in Bucharest, showing similar accuracy with GRU slightly outperforming; handled missing data effectively. $R^2 < 0.8$; focused only on NO ₂ without multi-pollutant consideration.
Kachhara and Jain (2022)	Pollutant Data	LSTM-ARIMA	AQI	Central Pollution Control Board (CPCB) of India	Analysed six years of data across 23 Indian cities, finding XGBoost superior to LSTM, RNN, and others for PM _{2.5} prediction in terms of the linearity of predicted and measured values. Focused mainly on model comparison without exploring interpretability or causal insights.
Tao et al. (2023)	Meteorological Data, AOD and Soil Data	LSTM	PM _{2.5}	AOD Atmospheric Composition Analysis Group ERA5 reanalysis, Copernicus Climate Data Store	Developed high-resolution PM _{2.5} model in Iraq using Bayesian-optimised LSTM with simulated annealing for predictor selection; achieved high accuracy (KGE = 0.89). However, the calibration restricted to early summer months, limiting temporal generalisability.
Drewil and Al-Bahadili (2022)	Pollution Data	GA-LSTM	PM _{2.5} , PM ₁₀ , CO, and NO _x	Historical dataset from the Kaggle website India from 2017 to 2020	Hybrid GA-LSTM efficiently forecasted multiple pollutants (PM ₁₀ , PM _{2.5} , CO, NO _x) with improved accuracy and computation efficiency. But due to demands of computation power and reliance on GA it may not apply on real-time applications.
Yasmin et al. (2023)	Meteorological and Pollutant Data	MLP-LSTM	AQI	Beijing Multi-Site Air-Quality Dataset (2013–2017), UCI Machine Learning Repository.	Effective hybrid neural network system that can effectively capture spatial and temporal interactions in the data of AQI and have greater accuracy as compared to the current models. It also had an objective analysis of preprocessing methods and feature-wise performance. Still, this complexity can represent an obstacle in the model implementation in low-resource environments.
Jesu-priya et al. (2025)	Meteorological and Pollutant Data	BiLSTM + Multi-Head Attention Hybrid DL	AQI	Time-Series Air Quality Data of India (2010–2023) from Kaggle	The model is an advanced interpretable deep learning model with real-time applicability that showed multi-output classification and the prediction of AQI severity. Experienced minor issues with high AQI level discrimination and does not have up to date integration of spatiotemporal or meteorological augmentations

Overall, the reviewed literature demonstrates substantial progress in the application of statistical, machine learning, and deep learning models for air pollution prediction and forecasting. Advances in data collection, preprocessing, feature engineering, and hybrid modelling techniques have significantly enhanced prediction accuracy and the ability to capture complex temporal and spatial patterns of pollution. Despite these advances, important challenges remain regarding the quality of data, model interpretability, computational complexity, uncertainty analysis, and the broader applicability of models to other environmental settings. These restrictions indicate promising future research directions, which are further discussed in the following section.

In addition, the reviewed literature reveals a notable methodological gap in the evaluation of model performance during extreme pollution events.. Validation is typically conducted under normal operating conditions using conventional train–test or cross-validation procedures, with few studies explicitly assessing model temporal robustness during rare events such as severe wildfire episodes or prolonged pollution spikes. Evidence from national-scale wildfire forecasting demonstrates that models performing well under conventional validation may degrade substantially under unprecedented conditions (e.g., Paul et al., 2022). This limitation represents a distinct methodological blind spot that extends beyond concerns of temporal data leakage. Furthermore, the

prevailing paradigm of offline model training followed by online deployment often limits the ability of forecasting systems to adapt to evolving environmental conditions and changing data distributions (Alimi et al. 2020). Addressing these challenges will require greater emphasis on rigorous validation protocols, uncertainty quantification, explainability, external validation, and adaptive learning strategies to improve the reliability and practical deployment of future air quality forecasting systems. These issues provide important directions for future research, which are discussed in the following section.

4. RESEARCH GAPS AND FUTURE WORKS

Despite significant advances in the use of statistical, machine learning, and deep learning approaches for air quality forecasting, several key challenges remain. Although recent research have reported high prediction performance, many existing methods still suffer from data quality issues, model interpretability limitations, scalability problems, and real-world deployment limitations. Addressing these challenges is important for developing robust, reliable, and transferable air quality prediction platforms. To provide a structured synthesis of the identified research gaps the shortcomings revealed in this review are organised into two linked themes rather than presented as an unordered catalogue. The first focusses on data-related challenges including quality, resolution, and representativeness of the datasets; and the second examines modelling- and deployment-level challenges including how models are built, validated, and operationalised once suitable data exists.

4.1. Dataset Characteristics

The performance of air pollution prediction models is highly dependent on the quality, completeness, and representativeness of the underlying datasets (Narkhede et al. 2025, Betancourt et al. 2021). However, several limitations associated with existing datasets continue to affect model reliability and forecasting accuracy:

- **Temporal Resolution:** Current air pollution datasets vary considerably in temporal granularity, ranging from high-frequency real-time measurements to daily or monthly averages. Many existing models are optimised for specific temporal resolutions and therefore struggle to effectively handle both short-term forecasting and long-term trend analysis simultaneously. Future studies should focus on developing adaptive, multi-resolution architectures that can more seamlessly assimilate disparate time-series data (Zhalehdoost and Taleai 2025).;
- **Spatial Heterogeneity:** Ground-based monitoring stations offer highly accurate local measurements, while satellite-derived data offer broader spatial coverage with a lower spatial resolution. It is difficult to integrate these heterogeneous data sources because of aggregation and downscaling processes can conceal significant spatial information and decrease the accuracy of the predictions (Zhalehdoost and Taleai 2025). Future research should therefore prioritise advanced data fusion methods to allow for the effective fusion of multi-source data while maintaining spatial fidelity.
- **Data Completeness:** Air pollution datasets frequently contain missing data due to sensor failure, maintenance disruptions, transmission problems and cloud cover affecting satellite imagery. While there are a number of imputation techniques, many of them are inefficient in reproducing extreme pollution events or sharp temporal shifts (Kalajdjieski et al. 2023). Future studies should explore more robust methods for preprocessing data and recovering missing data capable of handling extensive missing data without loss of temporal/spatial consistency.

- **Bias and Representativeness:** Existing datasets are predominantly collected in urban settings resulting in rural and resource-poor areas being underrepresented (El-Hajj 2026, Muneri et al. 2026). Furthermore, the severity of pollution events is relatively low when compared with normal air quality days, creating class imbalance issues in the training of the model (Bachtar et al. 2026, Muhammad et al. 2025). These limitations reduce the transferability and generalisability of predictive models across different geographical and environmental contexts. To enhance the model temporal robustness and representativeness, future research should therefore focus on balanced sampling, transfer learning, and larger, more geographically representative datasets (Shahbazi et al. 2023).

4.2. Modelling Challenges

Although recent machine learning and deep learning models have demonstrated good prediction accuracy, the literature reviewed reveals several methodological and practical constraints.

- **Model Interpretability:** Deep learning models are often criticised as “black-box” models, because their internal decision-making processes are difficult to interpret. While these models frequently achieve high predictive accuracy, their lack of transparency can limit stakeholder trust and limit their adoption in policy and regulatory contexts (Tasioulis and Karatzas 2023, Panda and Sinha 2024). Future research should incorporate Explainable Artificial Intelligence (XAI) techniques to improve the interpretability of the model and make the decision-making process more transparent.
- **Data Integration:** Although recent studies increasingly combine meteorological data, satellite images, traffic data and socioeconomic indicators the methods used to integrate these heterogeneous data sources remain inconsistent and fragmented. The absence of standardised data integration frameworks limits reproducibility and makes meaningful comparisons between studies difficult (Malings et al. 2024, Lilhore et al. 2025, Rowland and Karakuş 2023). Future research should therefore focus on unified multi-source data fusion frameworks that improve temporal robustness while supporting reproducibility and cross-study comparability.
- **Spatial and Temporal Generalisation:** Most existing models are developed and validated using local data or historical observations and may fail to accurately predict results in other geographical regions or under different climatic conditions. Future studies should therefore prioritise transferable, scalable modelling frameworks that support changing geographical and environmental landscapes.
- **Data Limitations:** Several reviewed studies relied on simulated datasets, open-source databases, and short-term data from short monitoring periods, with varying reliability of the equipment (Taştan 2025, Karmoude et al. 2025, Kamigauti et al. 2024). Such restrictions can result in inaccuracies in the predictions and reduce the model's temporal robustness. Future efforts should emphasise large-scale, high-quality, and more geographically diverse datasets by expanding monitoring networks and promoting data-sharing initiatives.
- **Model Reliability and Overfitting:** The review identified substantial variation in validation practices across studies. Many papers did not clearly report chronological train–test splits, relied solely on cross-validation, omitted benchmark comparisons, or reported implausibly high prediction accuracies that may indicate overfitting or data leakage. (Karmoude et al, 2025, Sawah et al, 2025). These shortcomings hinders reproducibility and reduces confidence in reported performance. Future studies should adopt leakage-aware validation protocols, benchmark models, uncertainty quantification, and standardised evaluation frameworks to improve the temporal robustness and consistency of the models.

- **Real-Time Deployment and Edge Computing:** Although many studies demonstrate excellent predictive capabilities in experimental settings, few address deployment in operational environments. Models are rarely evaluated under changing environmental conditions, extreme pollution episodes, or real-time streaming data, and few are optimised for the Internet of Things (IoT) and smart cities (Chan et al. 2024, Gill et al. 2025). The next generation of research should prioritise computationally efficient, adaptive, and low-latency forecasting systems capable of supporting continuous real-time tracking of air quality and prompt policy decisions.

Addressing these research methodological challenges are essential for the future of air pollution forecasting. In the future, efforts should focus on building interpretable, scalable and computationally efficient models that can integrate different datasets and be used under various environmental conditions. Such advancements will improve the reliability, transparency, and transferability of forecasting models while supporting evidence-based policymaking, and sustainable air quality management.

4.3. Practical Research Framework

The critical synthesis presented in Sections 3.2.1 to 3.2.3 revealed that many published forecasting studies report high temporal robustness but employ inconsistent validation procedures, limited external testing, insufficient uncertainty analysis, poor model interpretability, and inadequate benchmarking. To address these recurring methodological shortcomings, this review proposes the framework shown in Table 7. Each component of the framework is directly aligned with the limitations identified throughout the reviewed studies and provides practical guidance for the development of more reliable, transparent, and deployable air pollution forecasting systems.

Table 7: Proposed Research Framework.

Framework Component	Shortcoming Identified in Review	Recommendation
Data quality	Missing values, inconsistent preprocessing	Standardise preprocessing, feature engineering and data integration
Temporal validation	Random train/test splits causing leakage	Use chronological splits, rolling-origin evaluation and time-aware cross-validation
Spatial validation	Models tested on one city only	Validate across multiple regions and monitoring stations
Benchmarking	No comparison with naïve models	Compare against persistence, climatology and statistical baselines
Uncertainty estimation	Only RMSE/R ² reported	Report confidence intervals, prediction intervals and uncertainty estimates
Explainability	Black-box models dominate	Use SHAP, LIME, PDP or attention visualisation
External validation	Tested on one dataset only	Evaluate on independent datasets from different regions
Reproducibility	Missing implementation details	Publish datasets, preprocessing steps, code and hyperparameters

- **Data Quality:** Several reviewed studies reported missing data due to sensor failure, gaps in data transmission, and cloud cover, and imputation was found to be inadequate at detecting extreme pollution events (Kalajdjieski et al. 2023). The proposed framework suggests standardising the preprocessing, feature engineering, and data integration processes to enhance the reliability of the dataset.
- **Temporal Validation:** Several reviewed studies did not clearly separate training and test data. Bunnag (2024) estimated the parameters from the entire dataset without any held-out period, and Meenakshi Sundaram et al. (2022) was identified as high risk for temporal leakage from an unknown split. In view of this, the proposed

framework recommends chronological data partitioning along with rolling-origin validation as shown by Pant et al. (2023) and Bhatti et al. (2021), for realistic forecasting evaluation.

- **Spatial Validation:** Most reviewed models were developed and validated using data from a single city or a limited number of monitoring stations. Bhatti et al. (2021) focused on one city, and Asgari et al. (2022) was restricted to 21 stations judged insufficient for full coverage. The proposed framework therefore recommends validation across multiple regions and monitoring stations, following the leave-one-region-out approach of Ren et al. (2020), to establish spatial generalisability.
- **Benchmarking:** Many reviewed studies reported model performance without comparison to naïve or statistical baselines. Pant et al. (2023) specified none, and Asgari et al. (2022) compared only to Naive Bayes, judged a weaker proxy than a persistence baseline. The proposed framework therefore recommends benchmarking against persistence, climatology, and statistical baseline models.
- **Uncertainty Estimation:** Most of the reviewed studies reported point estimates like RMSE and R^2 (Pant et al. 2023, Borhani et al. 2023), while Bhatti et al. (2021) reported confidence intervals. The proposed framework thus suggests reporting of prediction intervals, confidence intervals, and other measures of uncertainty in addition to commonly reported accuracy measures.
- **Explainability:** In contrast to Ren et al. (2020) who highlighted a true nonlinearity between NO_x and ozone with explainability methods such as PDP and ICE, most of the reviewed models remained unexplainable, with only a few reporting some limited interpretation capability. The proposed framework suggests employing explainability tools like SHAP, LIME, partial dependence plot, or attention visualisation.
- **External Validation:** The reviewed models were mostly trained on a single dataset, with the exception of Paul et al. (2022), who used leave-one-region-out and leave-one-year-out validation for national-scale claims. The proposed framework therefore recommends external validation on independent datasets to test the generalisability of the findings for other situations than the one studied.
- **Reproducibility** Several reviewed studies lacked critical implementation information, such as hyperparameters, preprocessing, and code availability. The work of Ren et al. (2020) is an example of a complex hyperparameter tuning process which was not fully disclosed. The proposed framework thus suggests publishing with datasets, preprocessing pipelines, code and hyperparameters alongside published findings.

Collectively, these framework components address the principal methodological limitations identified throughout this review, namely inconsistent validation strategies, insufficient benchmarking, limited model interpretability, weak reproducibility, and inadequate evaluation under real-world conditions. Reproducibility underpins all five, since none of these properties can be verified without access to the data, code, and settings behind a reported result. Rather than focusing solely on predictive accuracy, the proposed framework encourages the development of forecasting systems that are robust, transparent, reproducible, and suitable for operational deployment.

Translating the themes above into a sequence of practical next steps, rather than a flat list of limitations, we propose the following priority order for the field:

- First, adopt consistent evaluation practice: chronological train–test splitting, rolling-origin validation, and comparison against naive baselines should become minimum reporting requirements, since without them, gains attributed to model architecture cannot be distinguished from gains attributable to leakage-prone evaluation.

- Second, close the data-representativeness gap identified in Section 4.1, particularly for rural and resource-constrained regions, since improved evaluation practice cannot compensate for training data that under-represents the conditions a model will be deployed in.
- Third, prioritize interpretability and uncertainty estimation for models intended to inform regulatory or public-health decisions, so that predictions carry a defensible confidence signal rather than a point estimate alone.
- Fourth, once the above are in place, invest in real-time and edge-deployable architectures, since deployment efficiency is of limited practical value if the underlying model's generalisation performance has not first been established under rigorous, leakage-free evaluation.

We present this ordering as a starting point for prioritisation rather than a strict dependency chain, since individual research groups will reasonably differ on sequencing given their available data and infrastructure

5. CONCLUSIONS

Atmospheric contamination continues to represent one of the most significant global threats to environmental sustainability and public health, with far-reaching consequences for climate systems, economic development, biodiversity, and human well-being. The rising concentration of pollutants such as NO₂, SO₂, CO₂, and PM_{2.5} and PM₁₀ highlight the urgent need for effective monitoring, prediction, and mitigation measures. This study presented a comprehensive review of the recent studies on air pollution focussing on its causes, sustainability impacts, and data-driven approaches for air quality prediction and forecasting. The review critically evaluated the methodologies, datasets, data pre-processing techniques, validation strategies and modelling frameworks employed in recent studies.

The findings indicate that traditional statistical models such as ARIMA and SARIMA remain useful for short-term and seasonal forecasting, while machine learning and deep learning models generally demonstrate superior temporal robustness for nonlinear, highly dimension, and large-scale time-series datasets. However, this apparent advantage should be treated with caution because many of the studies reviewed relied on random rather than chronological train–test, partitioning, lacked external validation or benchmark comparisons, and fail to evaluate performance during extreme pollution events. Consequently, reported prediction accuracies may overestimate true forecasting capability, and claims of superiority cannot yet be generalised across studies employing heterogeneous and often inadequately reported validation procedures.. Deep learning approaches, particularly LSTM-based methods and hybrid models, demonstrate promising temporal robustness for various pollutants and geographic areas.

The review also highlighted several key research gaps relating to dataset quality, model interpretability, validation practices, uncertainty quantification, data integration, scalability, and real-time deployment. These challenges underscore the need for more robust, transparent and transferable forecasting systems capable of integrating heterogeneous multi-source datasets while supporting reliable environmental decision-making. Unlike previous reviews that primarily compare predictive accuracy, this review introduces a structured methodological evaluation framework that assesses forecasting studies by considering validation design, temporal robustness, data leakage risk, interpretability, and practical deployment readiness. This broader perspective offers a more reliable basis for assessing forecasting models and identifying priorities for future air quality prediction research.

In summary, this review provides an integrated synthesis of recent advances in statistical, machine learning, and deep learning approaches for air quality prediction and forecasting while highlighting the methodological considerations necessary for reliable model evaluation.. The findings provide valuable guidance for researchers, policymakers, urban planners, and environmental practitioners seeking to develop robust, transparent, and operational forecasting systems that support evidence-based air quality management and contribute to the achievement of the. Future research should focus on explainable AI, uncertainty quantification, external and temporal validation, adaptive real-time forecasting, edge computing, and the development of large, standardised data repositories to improve the reliability, transparency, and practical deployment of next-generation air quality forecasting systems.

6. REFERENCES

- Achakzai, K., Khalid, S., Adrees, M., Bibi, A., Ali, S., Nawaz, R. and Rizwan, M., 2017. Air pollution tolerance index of plants around brick kilns in Rawalpindi, Pakistan. *Journal of Environmental Management*, 190, pp.252-258. DOI: <https://doi.org/10.1016/j.jenvman.2016.12.072>
- Agibayeva, A., Khalikhan, R., Guney, M., Karaca, F., Torezhan, A. and Avcu, E., 2022. An air quality modeling and disability-adjusted life years (DALY) risk assessment case study: Comparing statistical and machine learning approaches for PM2.5 forecasting. *Sustainability*, 14(24), p.16641. DOI: <https://doi.org/10.3390/su142416641>
- Aguilera, R., Luo, N., Basu, R., Wu, J., Clemesha, R., Gershunov, A. and Benmarhnia, T., 2023. A novel ensemble-based statistical approach to estimate daily wildfire-specific PM2.5 in California (2006–2020). *Environment International*, 171, p.107719. DOI: <https://doi.org/10.1016/j.envint.2022.107719>
- Agyemang, F., 2025. The right to public participation in advancing environmental sustainability in South African cities. *Journal of Environmental & Earth Sciences*, 7(6), pp.244-266. DOI: <https://doi.org/10.30564/jees.v7i6.9228>
- Al Alim, M., Hoque, M., Hore, S. and Hore, R., 2025. Environmental impacts of transportation in rapidly urbanizing South and Southeast Asian cities. *Smart Construction and Sustainable Cities*, 3(1), p.16. DOI: <https://doi.org/10.1007/s44268-025-00066-6>
- Alghieth, M., Alawaji, R., Saleh, S.H. and Alh, S., 2021. Air pollution forecasting using deep learning. *International Journal of Online & Biomedical Engineering*, 17(14). DOI: <https://doi.org/10.3991/ijoe.v17i14.27369>
- Alimi, O.A., Ouahada, K. and Abu-Mahfouz, A.M., 2020. A review of machine learning approaches to power system security and stability. *IEEE Access*, 8, pp.113512-113531. DOI: <https://doi.org/10.1109/ACCESS.2020.3003568>
- Aljanabi, M., Shkoukani, M. and Hijawi, M., 2021. Comparison of multiple machine learning algorithms for urban air quality forecasting. *Periodicals of Engineering and Natural Sciences*, 9(4), pp.1013-1028. DOI: <https://doi.org/10.21533/pen.v9.i4.1004>
- Alkhanani, M.F., 2025. Assessing the impact of air quality and socioeconomic conditions on respiratory disease incidence. *Tropical Medicine and Infectious Disease*, 10(2), p.56. DOI: <https://doi.org/10.3390/tropicalmed10020056>
- Alolayan, M.A., Almutairi, A., Aladwani, S.M. and Alkhamees, S., 2023. Investigating major sources of air pollution and improving spatiotemporal forecast accuracy using supervised machine learning and a proxy. *Journal of Engineering Research*, 11(3), pp.87-93. DOI: <https://doi.org/10.1016/j.jer.2023.100126>

- AlSalehy, A.S. and Bailey, M., 2025. Improving time series data quality: Identifying outliers and handling missing values in a multilocation gas and weather dataset. *Smart Cities*, 8(3), p.82. DOI: <https://doi.org/10.3390/smartcities8030082>
- Andrade, P.A.B., Zamora, P.E.C., Lara, A.E.C. and Piedra, J.P.P., 2025. A comprehensive evaluation of AI techniques for air quality index prediction: RNNs and transformers. *Ingenius*, (33), pp.60-75. DOI: <https://doi.org/10.17163/ings.n33.2025.06>
- Ao, Z., Fei, R., Jiang, H., Cui, L. and Zhu, Y., 2023. How can China achieve its goal of peaking carbon emissions at minimal cost? A research perspective from shadow price and optimal allocation of carbon emissions. *Journal of Environmental Management*, 325, p.116458. DOI: <https://doi.org/10.1016/j.jenvman.2022.116458>
- AQICN. 2025. Air quality programmatic APIs [Online]. Available: <https://aqicn.org/api/> (Accessed on 10 May 2025).
- Arshad, K., Hussain, N., Ashraf, M.H. and Saleem, M.Z., 2024. Air pollution and climate change as grand challenges to sustainability. *Science of the Total Environment*, 928, p.172370. DOI: <https://doi.org/10.1016/j.scitotenv.2024.172370>
- Arsov, M., Zdravevski, E., Lameski, P., Corizzo, R., Koteli, N., Mitreski, K. and Trajkovik, V., 2020, September. Short-term air pollution forecasting based on environmental factors and deep learning models. In *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*, pp.15-22. IEEE. DOI: <https://doi.org/10.15439/2020F211>
- Asgari, M., Yang, W. and Farnaghi, M., 2022. Spatiotemporal data partitioning for distributed random forest algorithm: Air quality prediction using imbalanced big spatiotemporal data on spark distributed framework. *Environmental Technology & Innovation*, 27, p.102776. DOI: <https://doi.org/10.1016/j.eti.2022.102776>
- Awe, Y.A., Larsen, B.K. and Sanchez-Triana, E., 2021. *The Global Health Cost of PM_{2.5} Air Pollution: A Case for Action Beyond 2021*. International Development in Focus. Washington, D.C.: World Bank Group. DOI: <https://doi.org/10.1596/978-1-4648-1816-5>
- Ayek, A.A.E., Loho, M.A. and Karmoka, S.F., 2025. Satellite air quality monitoring: An interactive and accessible tool for spatiotemporal analysis using Google Earth Engine. *Geo-spatial Information Science*, pp.1-16. DOI: <https://doi.org/10.1080/10095020.2025.2537352>
- Bachtiar, V.S., Purnawan, P., Raudina, A. and Habibah, H.R.I., 2026. Performance and application of air quality models in Indonesia: A systematic review of progress, challenges, and future directions. *Atmospheric Environment: X*, p.100457. DOI: <https://doi.org/10.1016/j.aeaoa.2026.100457>
- Balakrishnan, K., Dey, S., Gupta, T., Dhaliwal, R.S., Brauer, M., Cohen, A.J., Stanaway, J.D., Beig, G., Joshi, T.K., Aggarwal, A.N. and Sabde, Y., 2019. The impact of air pollution on deaths, disease burden, and life expectancy across the states of India: The Global Burden of Disease Study 2017. *The Lancet Planetary Health*, 3(1), pp.e26-e39. DOI: [https://doi.org/10.1016/S2542-5196\(18\)30261-4](https://doi.org/10.1016/S2542-5196(18)30261-4)
- Bandara, W.A.R.T.W. and Dissanayake, C.T.M., 2021. Most tolerant roadside tree species for urban settings in humid tropics based on air pollution tolerance index. *Urban Climate*, 37, p.100848. DOI: <https://doi.org/10.1016/j.uclim.2021.100848>
- Barid, A.J. and Wibowo, A., 2024. Optimization of the algorithms use ensemble and synthetic minority over-sampling technique for air quality classification. *Indonesian Journal of Electrical Engineering and Computer Science*, 33(3), pp.1632-1640. DOI: <https://doi.org/10.11591/ijeecs.v33.i3.pp1632-1640>
- Bartholomew, D.J., 1971. *Time Series Analysis Forecasting and Control*. DOI: <https://doi.org/10.2307/3008255>

- Betancourt, C., Stomberg, T., Roscher, R., Schultz, M.G. and Stadtler, S., 2021. AQ-Bench: A benchmark dataset for machine learning on global air quality metrics. *Earth System Science Data*, 13(6), pp.3013-3033. DOI: <https://doi.org/10.5194/essd-2020-380>
- Bhatti, U.A., Yan, Y., Zhou, M., Ali, S., Hussain, A., Qingsong, H., Yu, Z. and Yuan, L., 2021. Time series analysis and forecasting of air pollution particulate matter (PM_{2.5}): An SARIMA and factor analysis approach. *IEEE Access*, 9, pp.41019-41031. DOI: <https://doi.org/10.1109/ACCESS.2021.3060744>
- Booth, A., James, P., McGough, S. and Solaiman, E., 2025. Cross-regional deep learning for air quality forecasting: A comparative study of CO, NO₂, O₃, PM_{2.5}, and PM₁₀. *Forecasting*, 7(4), p.66. DOI: <https://doi.org/10.3390/su16167062>
- Borhani, F., Shafiepour Motlagh, M., Ehsani, A.H., Rashidi, Y., Ghahremanloo, M., Amani, M. and Moghimi, A., 2023. Current status and future forecast of short-lived climate-forced ozone in Tehran, Iran, derived from ground-based and satellite observations. *Water, Air, & Soil Pollution*, 234(2), p.134. DOI: <https://doi.org/10.1007/s11270-023-06138-6>
- Bunnag, T., 2024. Forecasting PM₁₀ caused by Bangkok's leading greenhouse gas emission using the SARIMA and SARIMA-GARCH model. *International Journal of Energy Economics and Policy*, 14(1), pp.418-426. DOI: <https://doi.org/10.32479/ijeeep.15275>
- Buvana, R., Sathya, A., Chathana, R., Jaysri, C.M. and Jashna, D.K., 2022, November. Air pollution prediction with machine learning. In *2022 1st International Conference on Computational Science and Technology (ICCST)*, pp.1-4. DOI: <https://doi.org/10.1109/ICCST55948.2022.10040297>
- Cabello-Solorzano, K., Ortigosa de Araujo, I., Peña, M., Correia, L. and Tallón-Ballesteros, A.J., 2023, August. The impact of data normalization on the accuracy of machine learning algorithms: A comparative analysis. In *International Conference on Soft Computing Models in Industrial and Environmental Applications*, pp.344-353. Cham: Springer Nature Switzerland. DOI: https://doi.org/10.1007/978-3-031-42536-3_33
- Cao, R., Li, B., Wang, Z., Peng, Z.R., Tao, S. and Lou, S., 2020. Using a distributed air sensor network to investigate the spatiotemporal patterns of PM_{2.5} concentrations. *Environmental Pollution*, 264, p.114549. DOI: <https://doi.org/10.1016/j.envpol.2020.114549>
- Chadalavada, S., Faust, O., Salvi, M., Seoni, S., Raj, N., Raghavendra, U., Gudigar, A., Barua, P.D., Molinari, F. and Acharya, R., 2025. Application of artificial intelligence in air pollution monitoring and forecasting: A systematic review. *Environmental Modelling & Software*, 185, p.106312. DOI: <https://doi.org/10.1016/j.envsoft.2024.106312>
- Chan, K., Matthews, P. and Munir, K., 2025. Time series forecasting for air quality with structured and unstructured data using artificial neural networks. *Atmosphere*, 16(3), p.320. DOI: <https://doi.org/10.3390/atmos16030320>
- Chan, Y.W., Kristiani, E., Fathoni, H., Chen, C.Y. and Yang, C.T., 2024. A smart edge computing infrastructure for air quality monitoring using LPWAN and MQTT technologies. *The Journal of Supercomputing*, 80(7), pp.9961-9985. DOI: <https://doi.org/10.1007/s11227-023-05837-5>
- Chatterjee, D., McDuffie, E.E., Smith, S.J., Bindle, L., Van Donkelaar, A., Hammer, M.S., Venkataraman, C., Brauer, M. and Martin, R.V., 2023. Source contributions to fine particulate matter and attributable mortality in India and the surrounding region. *Environmental Science & Technology*, 57(28), pp.10263-10275. DOI: <https://doi.org/10.1021/acs.est.2c07641>
- Chatterjee, K., Bala, B., Khan, M., Chitla, A., Reddy, K.N., Menshawi, A., Prasad, M., Kadurka, R.S., Altaf, M. and Jyothi, K.A., 2025. AiM: Urban air quality forecasting with grid-embedded recurrent MLP model. *Scientific Reports*, 15(1), p.42812. DOI: <https://doi.org/10.1038/s41598-025-27073-y>

- Chen, F., Zhang, W., Mfarrej, M.F.B., Saleem, M.H., Khan, K.A., Ma, J., Raposo, A. and Han, H., 2024. Breathing in danger: Understanding the multifaceted impact of air pollution on health impacts. *Ecotoxicology and Environmental Safety*, 280, p.116532. DOI: <https://doi.org/10.1016/j.ecoenv.2024.116532>
- Chen, G., Chen, S., Li, D. and Chen, C., 2025. A hybrid deep learning air pollution prediction approach based on neighborhood selection and spatio-temporal attention. *Scientific Reports*, 15(1), p.3685. DOI: <https://doi.org/10.1038/s41598-025-88086-1>
- Chen, L., Han, B., Wang, X., Zhao, J., Yang, W. and Yang, Z., 2023. Machine learning methods in weather and climate applications: A survey. *Applied Sciences*, 13(21), p.12019. DOI: <https://doi.org/10.3390/app132112019>
- Chen, L., Wang, J., Wang, H. and Jin, T., 2022. Urban air quality assessment by fusing spatial and temporal data from multiple study sources using refined estimation methods. *ISPRS International Journal of Geo-Information*, 11(6), p.330. DOI: <https://doi.org/10.3390/ijgi11060330>
- Chen, S., 2017. *Beijing Multi-Site Air Quality* [Dataset]. UCI Machine Learning Repository. DOI: <https://doi.org/10.24432/C5RK5G>
- Chetna, Dhaka, S.K., Longiany, G., Panwar, V., Kumar, V., Malik, S., Rao, A.S., Singh, N., Dimri, A.P., Matsumi, Y. and Nakayama, T., 2023. Trends and variability of PM_{2.5} at different time scales over Delhi: Long-term analysis 2007–2021. *Aerosol and Air Quality Research*, 23(5), p.220191. DOI: <https://doi.org/10.4209/aaqr.220191>
- Cican, G., Buturache, A.N. and Mirea, R., 2023. Applying machine learning techniques in air quality prediction—A Bucharest city case study. *Sustainability*, 15(11), p.8445. DOI: <https://doi.org/10.3390/su15118445>
- Coelho, S., Rafael, S., Lopes, D., Miranda, A.I. and Ferreira, J., 2021. How changing climate may influence air pollution control strategies for 2030? *Science of the Total Environment*, 758, p.143911. DOI: <https://doi.org/10.1016/j.scitotenv.2020.143911>
- De Amorim, L.B., Cavalcanti, G.D. and Cruz, R.M., 2023. The choice of scaling technique matters for classification performance. *Applied Soft Computing*, 133, p.109924. DOI: <https://doi.org/10.1016/j.asoc.2022.109924>
- de Oliveira Salerno, P.V., Briones-Valdivieso, C., Motairek, I., Dallan, L.P., Rajagopalan, S., Deo, S.V., Petermann-Rocha, F. and Al-Kindi, S., 2023. The cardiovascular disease burden attributable to particulate matter pollution in South America: Analysis of the 1990–2019 Global Burden of Disease. *Public Health*, 224, pp.169–177. DOI: <https://doi.org/10.1016/j.puhe.2023.07.035>
- Devasekhar, V. and Natarajan, P., 2023. Prediction of air quality and pollution using statistical methods and machine learning techniques. *International Journal of Advanced Computer Science and Applications*, 14(4). DOI: <https://doi.org/10.14569/IJACSA.2023.01404103>
- Diallo, A.A., Nderu, L., Malenje, B.M. and Kikuvi, G.M., 2024. Enhancing outlier detection in air quality index data using a stacked machine learning model. *Engineering Reports*, 6(11), p.e12936. DOI: <https://doi.org/10.1002/eng2.12936>
- Dias, A.M., Cook, C., Paglia, A.P. and Massara, R.L., 2025. Imperfect detection of terrestrial mammals in environmental impact assessment (EIA) baseline surveys may compromise decision-making and mitigation measures. *Environmental Impact Assessment Review*, 112, p.107850. DOI: <https://doi.org/10.1016/j.eiar.2025.107850>
- Dongre, P.K., Patel, V., Bhoi, U. and Maltare, N.N., 2025. An outlier detection framework for air quality index prediction using linear and ensemble models. *Decision Analytics Journal*, 14, p.100546. DOI: <https://doi.org/10.1016/j.dajour.2025.100546>

- Drewil, G.I. and Al-Bahadili, R.J., 2022. Air pollution prediction using LSTM deep learning and metaheuristics algorithms. *Measurement: Sensors*, 24, p.100546. DOI: <https://doi.org/10.1016/j.measen.2022.100546>
- Du, S., Li, T., Yang, Y. and Horng, S.J., 2019. Deep air quality forecasting using hybrid deep learning framework. *IEEE Transactions on Knowledge and Data Engineering*, 33(6), pp.2412-2424. DOI: <https://doi.org/10.1109/TKDE.2019.2954510>
- Du, W., Chen, L., Wang, H., Shan, Z., Zhou, Z., Li, W. and Wang, Y., 2023. Deciphering urban traffic impacts on air quality by deep learning and emission inventory. *Journal of Environmental Sciences*, 124, pp.745-757. DOI: <https://doi.org/10.1016/j.jes.2021.12.035>
- Ekundayo, O.S. and Ezugwu, A.E., 2025. Deep learning: Historical overview from inception to actualization, models, applications and future trends. *Applied Soft Computing*, 181, p.113378. DOI: <https://doi.org/10.1016/j.asoc.2025.113378>
- El-Hajj, M., 2026. Optimizing resource-constrained urban management: A LightGBM approach for high-fidelity prediction. *Frontiers in Sustainable Cities*, 8, p.1719036. DOI: <https://doi.org/10.3389/frsc.2026.1719036>
- Elvan, O.D., 2018. Analysis of environmental impact assessment practices and legislation in Turkey. *Environmental Science & Policy*, 84, pp.1-6. DOI: <https://doi.org/10.1016/j.envsci.2018.02.008>
- Environment and Climate Change Canada. 2024. Government of Canada actions to reduce air pollution. Retrieved September 22, 2025, from <https://www.canada.ca/en/environment-climate-change/campaigns/canadian-environment-week/clean-air-day/action-air-pollution.html>
- Environmental Protection Agency (EPA), 2022. *Acid Rain*. United States Environmental Protection Agency.
- Ernst & Young Global. 2024. Ghana imposes emissions levy. Available online: https://www.ey.com/en_gl/technical/tax-alerts/ghana-imposes-emissions-levy (Accessed on 22 September 2025).
- Espinosa, R., Palma, J., Jiménez, F., Kamińska, J., Sciavicco, G. and Lucena-Sánchez, E., 2021. A time series forecasting based multi-criteria methodology for air quality prediction. *Applied Soft Computing*, 113, p.107850. DOI: <https://doi.org/10.1016/j.asoc.2021.107850>
- Essamlali, I., Nhaila, H. and El Khaili, M., 2024. Supervised machine learning approaches for predicting key pollutants and for the sustainable enhancement of urban air quality: A systematic review. *Sustainability*, 16(3), p.976. DOI: <https://doi.org/10.3390/su16030976>
- European Environment Agency. 2025. Air pollution. Retrieved February 3, 2025, from <https://www.eea.europa.eu/en/topics/in-depth/air-pollution>
- European Parliament. 2023. Climate change: The greenhouse gases causing global warming. Retrieved March 23, 2025, from <https://www.europarl.europa.eu/topics/en/article/20230316STO77629/climate-change-the-greenhouse-gases-causing-global-warming>
- Evans, R., 2013. *The Environmental Impact Statement: A Rhetorical Analysis*. Ph.D. Thesis, Texas Tech University. <http://hdl.handle.net/2346/58625>
- Filonchuk, M. and Peterson, M.P., 2024. Analysis of air pollution from vehicle emissions for the contiguous United States. *Journal of Geovisualization and Spatial Analysis*, 8(1), p.16. DOI: <https://doi.org/10.1007/s41651-024-00180-6>
- Fisher, S., Bellinger, D.C., Cropper, M.L., Kumar, P., Binagwaho, A., Koudenoukpo, J.B., Park, Y., Taghian, G. and Landrigan, P.J., 2021. Air pollution and development in Africa: Impacts on health, the economy, and human capital. *The Lancet Planetary Health*, 5(10), pp.e681-e688. DOI: [https://doi.org/10.1016/S2542-5196\(21\)00201-1](https://doi.org/10.1016/S2542-5196(21)00201-1)

- Garbagna, L., Saheer, L.B. and Oghaz, M.M.D., 2025. AI-driven approaches for air pollution modelling: A comprehensive systematic review. *Environmental Pollution*, 373, p.125937. DOI: <https://doi.org/10.1016/j.envpol.2025.125937>
- Gaur, S. and Singh, R., 2023. A comprehensive review on land use/land cover (LULC) change modeling for urban development: Current status and future prospects. *Sustainability*, 15(2), p.903. DOI: <https://doi.org/10.3390/su15020903>
- Gill, S.S., Golec, M., Hu, J., Xu, M., Du, J., Wu, H., Walia, G.K., Murugesan, S.S., Ali, B., Kumar, M. and Ye, K., 2025. Edge AI: A taxonomy, systematic review and future directions. *Cluster Computing*, 28(1), p.18. DOI: <https://doi.org/10.1007/s10586-024-04686-y>
- Gladkova, E. and Saychenko, L., 2022. Applying machine learning techniques in air quality prediction. *Transportation Research Procedia*, 63, pp.1999-2006. DOI: <https://doi.org/10.1016/j.trpro.2022.06.222>
- Gu, X., Li, Z. and Su, J., 2024. Air pollution and skin diseases: A comprehensive evaluation of the associated mechanism. *Ecotoxicology and Environmental Safety*, 278, p.116429. DOI: <https://doi.org/10.1016/j.ecoenv.2024.116429>
- Gu, Y., Li, B. and Meng, Q., 2022. Hybrid interpretable predictive machine learning model for air pollution prediction. *Neurocomputing*, 468, pp.123-136. DOI: <https://doi.org/10.1016/j.neucom.2021.09.051>
- Guo, Q., He, Z. and Wang, Z., 2024. Monthly climate prediction using deep convolutional neural network and long short-term memory. *Scientific Reports*, 14(1), p.17748. DOI: <https://doi.org/10.1038/s41598-024-68906-6>
- Hait, M. and Sahu, P., 2024. Natural hazards as a source of environmental pollution in the global south. In *Biomonitoring of Pollutants in the Global South*, pp.153-190. Singapore: Springer Nature Singapore. DOI: https://doi.org/10.1007/978-981-97-1658-6_5
- Han, C., Hua, D. and Li, J., 2023. A view of industrial agglomeration, air pollution and economic sustainability from spatial econometric analysis of 273 cities in China. *Sustainability*, 15(9), p.7091. DOI: <https://doi.org/10.3390/su15097091>
- He, C., Kumar, R., Tang, W., Pfister, G., Xu, Y., Qian, Y. and Brasseur, G., 2024. Air pollution interactions with weather and climate extremes: Current knowledge, gaps, and future directions. *Current Pollution Reports*, 10(3), pp.430-442. DOI: <https://doi.org/10.1007/s40726-024-00296-9>
- He, H., Schäfer, B. and Beck, C., 2022. Spatial heterogeneity of air pollution statistics in Europe. *Scientific Reports*, 12(1), p.12215. DOI: <https://doi.org/10.1038/s41598-022-16109-2>
- Health Effects Institute, 2024. *State of Global Air 2024*. Boston, MA: Health Effects Institute.
- Herath Bandara, S.J. and Thilakarathne, N., 2025. Economic and public health impacts of transportation-driven air pollution in South Asia. *Sustainability*, 17(5), p.2306. DOI: <https://doi.org/10.3390/su17052306>
- Hua, V., Nguyen, T., Dao, M.S., Nguyen, H.D. and Nguyen, B.T., 2024. The impact of data imputation on air quality prediction problem. *PLOS One*, 19(9), p.e0306303. DOI: <https://doi.org/10.1371/journal.pone.0306303>
- Huang, Z., 2012. Health impact assessment in China: Emergence, progress and challenges. *Environmental Impact Assessment Review*, 32(1), pp.45-49. DOI: <https://doi.org/10.1016/j.eiar.2011.03.001>
- Institute for Health Metrics and Evaluation (IHME), 2020. *Global Burden of Disease Collaborative Network. Global Burden of Disease Study 2019*.
- IQAIR. 2025. AirVisual Air Quality Data: API. Retrieved May 22, 2025, from <https://www.iqair.com/us/air-quality-monitors/api>

- Isaev, E., Ajikeev, B., Shamyrganov, U., Kalnur, K.U., Maisalbek, K. and Sidle, R.C., 2022. Impact of climate change and air pollution forecasting using machine learning techniques in Bishkek. *Aerosol and Air Quality Research*, 22(3), p.210336. DOI: <https://doi.org/10.4209/aaqr.210336>
- Izah, S.C., Sylva, L., Ogwu, M.C., Shahsavani, A., Bazzazpour, S. and Rahmatinia, M., 2024. Modeling and statistical approaches for air pollution analysis. In *Air Pollutants in the Context of One Health: Fundamentals, Sources, and Impacts*, pp.425-456. Cham: Springer Nature Switzerland. DOI: https://doi.org/10.1007/698_2024_1138
- Jaganathan, S., Stafoggia, M., Rajiva, A., Mandal, S., Dixit, S., De Bont, J., Wellenius, G.A., Lane, K.J., Nori-Sarma, A., Kloog, I. and Prabhakaran, D., 2024. Estimating the effect of annual PM_{2.5} exposure on mortality in India: A difference-in-differences approach. *The Lancet Planetary Health*, 8(12), pp.e987-e996. DOI: [https://doi.org/10.1016/S2542-5196\(24\)00248-1](https://doi.org/10.1016/S2542-5196(24)00248-1)
- Jesupriya, J., Mahalakshmi, R. and Komalavalli, C., 2025. An intelligent trigger-enabled deep learning model for IoT-based environmental monitoring and real-time air quality evaluation. *International Journal of Applied Mathematics*, 38(10s), pp.492-510. DOI: <https://doi.org/10.12732/ijam.v38i10s.973>
- Ji, J.S., Dominici, F., Gouveia, N., Kelly, F.J. and Neira, M., 2025. Air pollution interventions for health. *Nature Medicine*, 31(9), pp.2888-2900. DOI: <https://doi.org/10.1038/s41591-025-03929-8>
- Jiang, X., Wei, P., Luo, Y. and Li, Y., 2021. Air pollutant concentration prediction based on a CEEMDAN-FE-BiLSTM model. *Atmosphere*, 12(11), p.1452. DOI: <https://doi.org/10.3390/atmos12111452>
- Jung, C.C., Lin, W.Y., Hsu, N.Y., Wu, C.D., Chang, H.T. and Su, H.J., 2020. Development of hourly indoor PM_{2.5} concentration prediction model: The role of outdoor air, ventilation, building characteristic, and human activity. *International Journal of Environmental Research and Public Health*, 17(16), p.5906. DOI: <https://doi.org/10.3390/ijerph17165906>
- Kachhara, C. and Jain, K., 2022. Air pollution prediction with machine learning. *NeuroQuantology*, 20(16), pp.2596-2605. DOI: <https://doi.org/10.48047/NQ.2022.20.16.NQ880262>
- Kalajdjieski, J., Trivodaliev, K., Mirceva, G., Kalajdziski, S. and Gievska, S., 2023. A complete air pollution monitoring and prediction framework. *IEEE Access*, 11, pp.88730-88744. DOI: <https://doi.org/10.1109/ACCESS.2023.3251346>
- Kamigauti, L.Y., Perez, G.M., Martin, T.C., de Fatima Andrade, M. and Kumar, P., 2024. Enhancing spatial inference of air pollution using machine learning techniques with low-cost monitors in data-limited scenarios. *Environmental Science: Atmospheres*, 4(3), pp.342-350. DOI: <https://doi.org/10.1039/d3ea00126a>
- Karmoude, M., Munhungewarwa, B., Chiraira, I., Mckenzie, R., Kong, J., Smith, B., Ayana, G., Njara, N., Mathaha, T., Kumar, M. and Mellado, B., 2025. Machine learning for air quality prediction and data analysis: Review on recent advancements, challenges, and outlooks. *Science of the Total Environment*, 1002, p.180593. DOI: <https://doi.org/10.1016/j.scitotenv.2025.180593>
- Karri, R.R., Vera, T., Hassan, S.K.M., Khoder, M.I., Dehghani, M.H., Mubarak, N.M. and Ravindran, G., 2024. Classification, sources, and occurrence of outdoor air pollutants: A comprehensive overview. *Health and Environmental Effects of Ambient Air Pollution*, pp.1-34. DOI: <https://doi.org/10.1016/B978-0-443-16088-2.00002-8>
- Kathiroli, P., Vijayalakshmi, V., Gopal, P. and Panda, S., 2023, December. Investigation of satellite data for monitoring air quality over remote sensing technology. In *International Conference on Artificial Intelligence and Speech Technology*, pp.351-366. Cham: Springer Nature Switzerland. DOI: https://doi.org/10.1007/978-3-031-75167-7_28

- Kaur, R. and Pandey, P., 2021. Air pollution, climate change, and human health in Indian cities: A brief review. *Frontiers in Sustainable Cities*, 3, p.705131. DOI: <https://doi.org/10.3389/frsc.2021.705131>
- Kumar, P., Choudhary, A., Joshi, P.K., Kumar, R.P. and Bhatla, R., 2025. Machine learning models for estimating criteria pollutants and health risk-based air quality indices over Eastern Coast coal mine complex belts. *Frontiers in Environmental Science*, 13, p.1589991. DOI: <https://doi.org/10.3389/fenvs.2025.1589991>
- Lawrence, S. and Bhathmanabhan, S., 2026. Harnessing deep learning for air pollution forecasting: Trends, techniques, and future prospects. *Artificial Intelligence Review*. DOI: <https://doi.org/10.1007/s10462-026-11496-8>
- Legislative Affairs Office of the State Council of the People's Republic of China, 2019. *Law of the People's Republic of China on Environmental Impact Assessment*. Beijing: China Labor and Social Security Bureau Press.
- Lei, M.T., Monjardino, J., Mendes, L., Gonçalves, D. and Ferreira, F., 2019. Macao air quality forecast using statistical methods. *Air Quality, Atmosphere & Health*, 12(9), pp.1049-1057. DOI: <https://doi.org/10.1007/s11869-019-00721-9>
- Lei, T.M., Cai, J., Molla, A.H., Kurniawan, T.A. and Kong, S.S.K., 2024. Evaluation of machine learning models in air pollution prediction for a case study of Macau as an effort to comply with UN Sustainable Development Goals. *Sustainability*, 16(17), p.7477. DOI: <https://doi.org/10.3390/su16177477>
- Lei, T.M., Siu, S.W., Monjardino, J., Mendes, L. and Ferreira, F., 2022. Using machine learning methods to forecast air quality: A case study in Macao. *Atmosphere*, 13(9), p.1412. DOI: <https://doi.org/10.3390/atmos13091412>
- Leong, W.C., Kelani, R.O. and Ahmad, Z.J.J.O.E.C.E., 2020. Prediction of air pollution index (API) using support vector machine (SVM). *Journal of Environmental Chemical Engineering*, 8(3), p.103208. DOI: <https://doi.org/10.1016/j.jece.2019.103208>
- Li, X., Jin, L. and Kan, H., 2019. Air pollution: A global problem needs local fixes. *Nature*, 570(7762), pp.437-439. DOI: <https://doi.org/10.1038/d41586-019-01960-7>
- Liang, L. and Gong, P., 2020. Urban and air pollution: A multi-city study of long-term effects of urban landscape patterns on air quality trends. *Scientific Reports*, 10(1), p.18618. DOI: <https://doi.org/10.1038/s41598-020-74524-9>
- Liao, K., Huang, X., Dang, H., Ren, Y., Zuo, S. and Duan, C., 2021. Statistical approaches for forecasting primary air pollutants: A review. *Atmosphere*, 12(6), p.686. DOI: <https://doi.org/10.3390/atmos12060686>
- Lilhore, U.K., Simaiya, S., Singh, R.K., Baqasah, A.M., Alroobaea, R., Alsafyani, M., Alhazmi, A. and Khan, M.M., 2025. Advanced air quality prediction using multimodal data and dynamic modeling techniques. *Scientific Reports*, 15(1), p.27867. DOI: <https://doi.org/10.1038/s41598-025-11039-1>
- Lopez-Aparicio, S., Grythe, H., Drabicki, A., Chwastek, K., Toboła, K., Górski-Niemas, L., Kierpiec, U., Markelj, M., Strużewska, J., Kud, B. and Santos, G.S., 2025. Environmental sustainability of urban expansion: Implications for transport emissions, air pollution, and city growth. *Environment International*, 196, p.109310. DOI: <https://doi.org/10.1016/j.envint.2025.109310>
- Ma, Y., Tian, Y., Ren, Y., Wang, Z., Wu, L., Pan, X., Ma, Y. and Xin, J., 2022. Long-term (2017–2020) aerosol optical depth observations in Hohhot City in Mongolian Plateau and the impacts from different types of aerosols. *Atmosphere*, 13(5), p.737. DOI: <https://doi.org/10.3390/atmos13050737>
- Maji, S., Ahmed, S., Kaur-Sidhu, M., Mor, S. and Ravindra, K., 2023. Health risks of major air pollutants, their drivers and mitigation strategies: A review. *Air, Soil and Water Research*, 16, p.11786221231154659. DOI: <https://doi.org/10.1177/11786221231154659>

- Maleki, F., Ovens, K., Gupta, R., Reinhold, C., Spatz, A. and Forghani, R., 2022. Generalizability of machine learning models: Quantitative evaluation of three methodological pitfalls. *Radiology: Artificial Intelligence*, 5(1), p.e220028. DOI: <https://doi.org/10.1148/ryai.220028>
- Malings, C., Knowland, K.E., Pavlovic, N., Coughlin, J.G., King, D., Keller, C., Cohn, S. and Martin, R.V., 2024. Air quality estimation and forecasting via data fusion with uncertainty quantification: Theoretical framework and preliminary results. *Journal of Geophysical Research: Machine Learning and Computation*, 1(4), p.e2024JH000183. DOI: <https://doi.org/10.1029/2024JH000183>
- Maltare, N.N. and Vahora, S., 2023. Air quality index prediction using machine learning for Ahmedabad city. *Digital Chemical Engineering*, 7, p.100093. DOI: <https://doi.org/10.1016/j.dche.2023.100093>
- Manisalidis, I., Stavropoulou, E., Stavropoulos, A. and Bezirtzoglou, E., 2020. Environmental and health impacts of air pollution: A review. *Frontiers in Public Health*, 8, p.14. DOI: <https://doi.org/10.3389/fpubh.2020.00014>
- McGrath, M., 2020. Climate change: China aims for 'carbon neutrality by 2060'. Retrieved March 25, 2025, from <https://www.bbc.com/news/science-environment-54256826>
- Meenakshi Sundaram, S., Sivakumar, P., Parthiban, M. and Revathy, U., 2022. Predicting air quality index values using ARIMA model. *International Journal of Early Childhood*, 14(7). DOI: <https://doi.org/10.48047/intjecse/V14I7.368>
- Miller, R., Olbrich, J., Manuel, W., Chen, J., Wurm, M. and Alvarez, C.I., 2026. Satellite-driven modelling of NO₂ and PM_{2.5} across Germany (2019–2024): A multi-sensor machine-learning approach. *Environmental Pollution*, p.127898. DOI: <https://doi.org/10.1016/j.envpol.2026.127898>
- Mubeen, M., He, S., Rahman, M.S., Wang, L., Zhang, X., Ahmed, B., He, Z. and Han, Y., 2025. Smart prediction and optimization of air quality index with artificial intelligence. *Journal of Environmental Sciences*, 158, pp.761-775. DOI: <https://doi.org/10.1016/j.jes.2025.02.041>
- Muhammad, M., Ul-Saufie, A.Z., Radi, N.F.A., Noor, N.M. and Gusnanto, A., 2025. Modified resampling strategy for extreme values in imbalanced air pollution data using moving block bootstrapping approach with relevance weighting (MBB-RW). *Scientific Reports*. DOI: <https://doi.org/10.1038/s41598-025-28416-5>
- Muneri, A.I., Madonsela, B.S. and Maphanga, T., 2026. Integration of GIS in air quality monitoring: A review of low-income neighbourhoods in South Africa. *Discover Public Health*, 23(1), p.513. DOI: <https://doi.org/10.1186/s12982-026-01715-6>
- Narkhede, G., Poonawala, M., Sonawane, A., Hiwale, A. and Singh, A.R., 2025. Air pollution prediction with advanced preprocessing and deep ensemble learning. *Atmospheric Pollution Research*, 16(10), p.102610. DOI: <https://doi.org/10.1016/j.apr.2025.102610>
- National Aeronautics and Space Administration (NASA). 2023. What is the greenhouse effect? Retrieved March 23, 2025, from <https://science.nasa.gov/climate-change/faq/what-is-the-greenhouse-effect/>
- National Park Service. 2018. Where does air pollution come from? Retrieved August 18, 2025, from <https://www.nps.gov/subjects/air/sources.htm#:~:text=Types%20of%20Sources,cities%2C%20and%20wood%20burning%20fireplaces>
- Njaime, M., Abdallah, F., Snoussi, H., Akl, J., Chaaban, K. and Omrani, H., 2025. Transfer learning based solution for air quality prediction in smart cities using multimodal data. *International Journal of Environmental Science and Technology*, 22(3), pp.1297-1312. DOI: <https://doi.org/10.1007/s13762-024-05722-5>
- Nowroz, F., Hasanuzzaman, M., Siddika, A., Parvin, K., Caparros, P.G., Nahar, K. and Prasad, P.V., 2024. Elevated tropospheric ozone and crop production: Potential negative effects and plant defense mechanisms. *Frontiers in Plant Science*, 14, p.1244515. DOI: <https://doi.org/10.3389/fpls.2023.1244515>

- Nugroho, H.Y.S.H., Basuki, T.M., Savitri, E., Supangat, A.B., Putra, P.B., Wahyuningrum, N., Adi, R.N., Setiawan, O., Nandini, R., Cahyono, S.A. and Auliyani, D., 2025. Advancing air quality monitoring systems towards sustainable green development: Insight for metropolitan cities in Indonesia. *Environmental and Sustainability Indicators*, 26, p.100649. DOI: <https://doi.org/10.1016/j.indic.2025.100649>
- Odubanjo, O.F., Falaiye, O.A., Orosun, M.M. and Sanni, M., 2024. Investigation of particulate matter air quality index (AQI) and risk assessment in some locations in Nigeria. *Journal of the Nigerian Society of Physical Sciences*, pp.2120-2120. DOI: <https://doi.org/10.46481/jnsps.2024.2120>
- Ogwu, M.C. and Izah, S.C. (eds.), 2024. *Sustainable Strategies for Air Pollution Mitigation: Development, Economics, and Technologies*. Springer. DOI: <https://doi.org/10.1007/978-3-031-74157-9>
- OPENAQ, 2024. *Open Air Quality Data: The Global State of Play*.
- Oumoulyte, M., El Allaoui, A., Farhaoui, Y. and Boughrous, A.A., 2025. Efficient air quality prediction models based on supervised machine learning techniques. In *E3S Web of Conferences*, vol. 632, p.02012. EDP Sciences. DOI: <https://doi.org/10.1051/e3sconf/202563202012>
- Pak, U., Ma, J., Ryu, U., Ryom, K., Juhyok, U., Pak, K. and Pak, C., 2020. Deep learning-based PM_{2.5} prediction considering the spatiotemporal correlations: A case study of Beijing, China. *Science of the Total Environment*, 699, p.133561. DOI: <https://doi.org/10.1016/j.scitotenv.2019.07.367>
- Panda, S. and Sinha, A., 2024. Advanced AI-driven approaches for predicting air quality: A comprehensive review. *Journal of Computational Analysis & Applications*, 33(6).
- Pandey, A., Brauer, M., Cropper, M.L., Balakrishnan, K., Mathur, P., Dey, S., Turkogulu, B., Kumar, G.A., Khare, M., Beig, G. and Gupta, T., 2021. Health and economic impact of air pollution in the states of India: The Global Burden of Disease Study 2019. *The Lancet Planetary Health*, 5(1), pp.e25-e38. DOI: [https://doi.org/10.1016/S2542-5196\(20\)30298-9](https://doi.org/10.1016/S2542-5196(20)30298-9)
- Pandya, S., Gadekallu, T.R., Maddikunta, P.K.R. and Sharma, R., 2022. A study of the impacts of air pollution on the agricultural community and yield crops (Indian context). *Sustainability*, 14(20), p.13098. DOI: <https://doi.org/10.3390/su142013098>
- Pant, A., Joshi, R.C., Sharma, S. and Pant, K., 2023. Predictive modeling for forecasting air quality index (AQI) using time series analysis. *Avicenna Journal of Environmental Health Engineering*, 10(1), pp.38-43. DOI: <https://doi.org/10.34172/ajehe.2023.5376>
- Paul, N., Yao, J., McLean, K.E., Stieb, D.M. and Henderson, S.B., 2022. The Canadian Optimized Statistical Smoke Exposure Model (CanOSSEM): A machine learning approach to estimate national daily fine particulate matter (PM_{2.5}) exposure. *Science of the Total Environment*, 850, p.157956. DOI: <https://doi.org/10.1016/j.scitotenv.2022.157956>
- Petrić, V., Račić, N., Hrga, I., Grgec, D., Marić, M., Krivohlavek, A., Anić, Z., Lovrić, M. and Jergović, M., 2025. Assessment of sensor data from an air quality monitoring network—The need for machine learning-based recalibration and its relevance in health impact analysis of local pollution events. *Atmosphere*, 16(12), p.1358. DOI: <https://doi.org/10.3390/atmos16121358>
- Peuch, V.H., Engelen, R., Rixen, M., Dee, D., Flemming, J., Suttie, M., Ades, M., Agustí-Panareda, A., Ananasso, C., Andersson, E. and Armstrong, D., 2022. The Copernicus Atmosphere Monitoring Service: From research to operations. *Bulletin of the American Meteorological Society*, 103(12), pp.E2650-E2668. DOI: <https://doi.org/10.5194/gmd-10-1-2017>

- Rahmad Ramadhan, L. and Anne Mudya, Y., 2024. A comparative study of Z-score and min-max normalization for rainfall classification in Pekanbaru. *Journal of Data Science*, 2024(04), pp.1-8. DOI: <https://doi.org/10.61453/jods.v2024no04>
- Rahman, E.A., Hamzah, F.M., Latif, M.T. and Azid, A., 2023. Forecasting PM_{2.5} in Malaysia using a hybrid model. *Aerosol and Air Quality Research*, 23(9), p.230006. DOI: <https://doi.org/10.4209/aaqr.230006>
- Rajakumari R Peter, R. and Ambadas Lanjewar, U., 2023. Improving air quality prediction with a hybrid Bi-LSTM and GAN model. *ECTI Transactions on Computer and Information Technology (ECTI-CIT)*, 17(3), pp.376-386. DOI: <https://doi.org/10.37936/ecti-cit.2023173.251670>
- Ramya, A., Dhevagi, P., Poornima, R., Avudainayagam, S., Watanabe, M. and Agathokleous, E., 2023. Effect of ozone stress on crop productivity: A threat to food security. *Environmental Research*, 236, p.116816. DOI: <https://doi.org/10.1016/j.envres.2023.116816>
- Relvas, H., Lopes, D. and Armengol, J.M., 2025. Empowering communities: Advancements in air quality monitoring and citizen engagement. *Urban Climate*, 60, p.102344. DOI: <https://doi.org/10.1016/j.uclim.2025.102344>
- Ren, X., Mi, Z. and Georgopoulos, P.G., 2020. Comparison of machine learning and land use regression for fine scale spatiotemporal estimation of ambient air pollution: Modeling ozone concentrations across the contiguous United States. *Environment International*, 142, p.105827. DOI: <https://doi.org/10.1016/j.envint.2020.105827>
- Rentschler, J. and Leonova, N., 2023. Global air pollution exposure and poverty. *Nature Communications*, 14(1), p.4432. DOI: <https://doi.org/10.1038/s41467-023-39797-4>
- Rodríguez-García, M.I., Carrasco-García, M.G., Cubillas Fernández, P.R., Rodrigues Ribeiro, M.D.C., Cardoso, P.J. and Turias, I.J., 2025. Air pollution forecasting using autoencoders: A classification-based prediction of NO₂, PM₁₀, and SO₂ concentrations. *Nitrogen*, 6(4), p.101. DOI: <https://doi.org/10.3390/nitrogen6040101>
- Rosca, C.M., Carbureanu, M. and Stancu, A., 2025. Data-driven approaches for predicting and forecasting air quality in urban areas. *Applied Sciences*, 15(8), p.4390. DOI: <https://doi.org/10.3390/app15084390>
- Rowley, A. and Karakuş, O., 2023. Predicting air quality via multimodal AI and satellite imagery. *Remote Sensing of Environment*, 293, p.113609. DOI: <https://doi.org/10.1016/j.rse.2023.113609>
- Samal, K.K.R., Babu, K.S., Das, S.K. and Acharaya, A., 2019, August. Time series based air pollution forecasting using SARIMA and Prophet model. In *Proceedings of the 2019 International Conference on Information Technology and Computer Communications*, pp.80-85. DOI: <https://doi.org/10.1145/3355402.3355417>
- Sawah, M.S., Elmannai, H., El-Bary, A.A., Lotfy, K. and Sheta, O.E., 2025. Improving air quality prediction using hybrid BPSO with BWAQ for feature selection and hyperparameters optimization. *Scientific Reports*, 15(1), p.13176. DOI: <https://doi.org/10.1038/s41598-025-95983-y>
- Scagliotti, A.F., Margarit, D.H., Reale, M.V. and Jorge, G.A., 2022. Influence of settings and predictors in neural network model performance: A Buenos Aires air quality case. *Procedia Computer Science*, 212, pp.348-357. DOI: <https://doi.org/10.1016/j.procs.2022.11.019>
- Serrao, V., 2023. Prediction of environmental pollution using hybrid model. *Indian Journal of Natural Sciences*, 14(8).
- Shahbazi, N., Lin, Y., Asudeh, A. and Jagadish, H.V., 2023. Representation bias in data: A survey on identification and resolution techniques. *ACM Computing Surveys*, 55(13s), pp.1-39. DOI: <https://doi.org/10.1145/3588433>
- Sicard, P., Agathokleous, E., Anenberg, S.C., De Marco, A., Paoletti, E. and Calatayud, V., 2023. Trends in urban air pollution over the last two decades: A global perspective. *Science of the Total Environment*, 858, p.160064. DOI: <https://doi.org/10.1016/j.scitotenv.2022.160064>

- Simo, A., Dzitac, S., Frigura-Iliasa, F.M., Musuroi, S., Andea, P. and Meianu, D., 2020. Technical solution for a real-time air quality monitoring system. *International Journal of Computers Communications & Control*, 15(4). DOI: <https://doi.org/10.15837/ijccc.2020.4.3891>
- Singh, A.K., Kumar, M., Bauddh, K., Singh, A., Singh, P., Madhav, S. and Shukla, S.K., 2023. Environmental impacts of air pollution and its abatement by plant species: A comprehensive review. *Environmental Science and Pollution Research*, 30(33), pp.79587-79616. DOI: <https://doi.org/10.1007/s11356-023-28164-x>
- Singh, V., Srivastava, R.K. and Bhatt, A.K., 2025. Sources of air pollution. In *Battling Air and Water Pollution: Protecting Our Planet's Vital Resources*, pp.17-29. Singapore: Springer Nature Singapore. DOI: https://doi.org/10.1007/978-981-96-4375-2_2
- Sivarethinamohan, R., Sujatha, S., Priya, S., Gafoor, A. and Rahman, Z., 2021. Impact of air pollution in health and socio-economic aspects: Review on future approach. *Materials Today: Proceedings*, 37, pp.2725-2729. DOI: <https://doi.org/10.1016/j.matpr.2020.08.540>
- Sompornrattanaphan, M., Thongngarm, T., Ratanawatkul, P., Wongsas, C. and Swigris, J.J., 2020. The contribution of particulate matter to respiratory allergy. *Asian Pacific Journal of Allergy and Immunology*, 38(1), pp.19-28. DOI: <https://doi.org/10.12932/AP-100619-0579>
- South African Revenue Service. 2024. Carbon tax. Retrieved March 26, 2025, from <https://www.sars.gov.za/customs-and-excise/excise/environmental-levy-products/carbon-tax/>
- Suriano, D., Akinnusotu, A., Abulude, F.O. and Oluwagbayide, S.D., 2024. Assessment of particulate matter (PM_{2.5}) and air quality index (AQI) in eight locations of Lagos State, Nigeria. DOI: <https://doi.org/10.20944/preprints202409.1991.v1>
- Syuhada, G., Akbar, A., Hardiawan, D., Pun, V., Darmawan, A., Heryati, S.H.A., Siregar, A.Y.M., Kusuma, R.R., Driejana, R., Ingole, V. and Kass, D., 2023. Impacts of air pollution on health and cost of illness in Jakarta, Indonesia. *International Journal of Environmental Research and Public Health*, 20(4), p.2916. DOI: <https://doi.org/10.3390/ijerph20042916>
- Tao, H., Jawad, A.H., Shather, A.H., Al-Khafaji, Z., Rashid, T.A., Ali, M., Al-Ansari, N., Marhoon, H.A., Shahid, S. and Yaseen, Z.M., 2023. Machine learning algorithms for high-resolution prediction of spatiotemporal distribution of air pollution from meteorological and soil parameters. *Environment International*, 175, p.107931. DOI: <https://doi.org/10.1016/j.envint.2023.107931>
- Tasioulis, T. and Karatzas, K., 2023. Reviewing explainable artificial intelligence towards better air quality modelling. In *Environmental Informatics*, pp.3-19. Cham: Springer Nature Switzerland. DOI: https://doi.org/10.1007/978-3-031-46902-2_1
- Taştan, M., 2025. Machine learning-based calibration and performance evaluation of low-cost Internet of Things air quality sensors. *Sensors*, 25(10), p.3183. DOI: <https://doi.org/10.3390/s25103183>
- Taye, M.M., 2023. Understanding of machine learning with deep learning: Architectures, workflow, applications and future directions. *Computers*, 12(5), p.91. DOI: <https://doi.org/10.3390/computers12050091>
- Tello-Leal, E., Ramirez-Alcocer, U.M., Macías-Hernández, B.A. and Hernandez-Resendiz, J.D., 2024. Evaluation of deep learning models for predicting the concentration of air pollutants in urban environments. *Sustainability*, 16(16), p.7062. DOI: <https://doi.org/10.3390/su16167062>
- Tran, H.M., Tsai, F.J., Lee, Y.L., Chang, J.H., Chang, L.T., Chang, T.Y., Chung, K.F., Kuo, H.P., Lee, K.Y., Chuang, K.J. and Chuang, H.C., 2023. The impact of air pollution on respiratory diseases in an era of climate change: A review of the current evidence. *Science of the Total Environment*, 898, p.166340. DOI: <https://doi.org/10.1016/j.scitotenv.2023.166340>

- U.S. Environmental Protection Agency, 2024. *Motor Vehicle Emission Simulator: MOVES5*. Office of Transportation and Air Quality. Ann Arbor, MI.
- U.S. Environmental Protection Agency. 2025. Reviewing National Ambient Air Quality Standards (NAAQS): Scientific and technical information. Retrieved September 22, 2025, from <https://www.epa.gov/naaqs>
- Wai, K.M. and Yu, P.K., 2023. Application of a machine learning method for prediction of urban neighborhood-scale air pollution. *International Journal of Environmental Research and Public Health*, 20(3), p.2412. DOI: <https://doi.org/10.3390/ijerph20032412>
- Wang, J., Li, J., Ye, J., Zhao, J., Wu, Y., Hu, J., Liu, D., Nie, D., Shen, F., Huang, X. and Huang, D.D., 2020. Fast sulfate formation from oxidation of SO₂ by NO₂ and HONO observed in Beijing haze. *Nature Communications*, 11(1), p.2844. DOI: <https://doi.org/10.1038/s41467-020-16683-x>
- Waseem, K.H., Mushtaq, H., Abid, F., Abu-Mahfouz, A.M., Shaikh, A., Turan, M. and Rasheed, J., 2022. Forecasting of air quality using an optimized recurrent neural network. *Processes*, 10(10), p.2117. DOI: <https://doi.org/10.3390/pr10102117>
- World Bank. 2024. Urban population. Retrieved September 14, 2025, from <https://data.worldbank.org/indicator/SP.URB.TOTL.IN.ZS>
- World Health Organization (WHO). 2023. Air pollution and health. Retrieved February 6, 2025, from [https://www.who.int/news-room/fact-sheets/detail/ambient-\(outdoor\)-air-quality-and-health](https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health)
- World Health Organization, 2021. *WHO Global Air Quality Guidelines: Particulate Matter (PM_{2.5} and PM₁₀), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide*. Geneva: World Health Organization.
- Yang, T., Dong, Q., Du, Q., Du, M., Dong, R. and Chen, M., 2021. Carbon dioxide emissions and Chinese OFDI: From the perspective of carbon neutrality targets and environmental management of home country. *Journal of Environmental Management*, 295, p.113120. DOI: <https://doi.org/10.1016/j.jenvman.2021.113120>
- Yasmin, F., Hassan, M.M., Hasan, M., Zaman, S., Angon, J.H., Bairagi, A.K. and Changchun, Y., 2023. AQIPred: A hybrid model for high precision time specific forecasting of air quality index with cluster analysis. *Human-Centric Intelligent Systems*, 3(3), pp.275-295. DOI: <https://doi.org/10.1007/s44230-023-00039-x>
- Yu, Z., Yang, K., Luo, Y. and Shang, C., 2020. Spatial-temporal process simulation and prediction of chlorophyll-a concentration in Dianchi Lake based on wavelet analysis and long-short term memory network. *Journal of Hydrology*, 582, p.124488. DOI: <https://doi.org/10.1016/j.jhydrol.2019.124488>
- Zhalehdoost, A. and Taleai, M., 2025. Unravelling the importance of spatial and temporal resolutions in modeling urban air pollution using a machine learning approach. *Scientific Reports*, 15(1), p.27708. DOI: <https://doi.org/10.1038/s41598-025-13639-3>
- Zhang, B., Rong, Y., Yong, R., Qin, D., Li, M., Zou, G. and Pan, J., 2022. Deep learning for air pollutant concentration prediction: A review. *Atmospheric Environment*, 290, p.119347. DOI: <https://doi.org/10.1016/j.atmosenv.2022.119347>
- Zhang, M., Shi, L., Ma, X., Zhao, Y. and Gao, L., 2021. Study on comprehensive assessment of environmental impact of air pollution. *Sustainability*, 13(2), p.476. DOI: <https://doi.org/10.3390/su13020476>
- Zhang, Y. and Thorburn, P.J., 2022. Handling missing data in near real-time environmental monitoring: A system and a review of selected methods. *Future Generation Computer Systems*, 128, pp.63-72. DOI: <https://doi.org/10.1016/j.future.2021.09.033>
- Zhang, Y., Zhai, S., Huang, J., Li, X., Wang, W., Zhang, T., Yin, F. and Ma, Y., 2022a. Estimating high-resolution PM_{2.5} concentration in the Sichuan Basin using a random forest model with data-driven spatial autocorrelation terms. *Journal of Cleaner Production*, 380, p.134890. DOI: <https://doi.org/10.1016/j.jclepro.2022.134890>

- Zhang, Z., Zhang, G. and Su, B., 2022b. The spatial impacts of air pollution and socio-economic status on public health: Empirical evidence from China. *Socio-Economic Planning Sciences*, 83, p.101167. DOI: <https://doi.org/10.1016/j.seps.2021.101167>
- Zhao, Z., Wu, J., Cai, F., Zhang, S. and Wang, Y.G., 2022. A statistical learning framework for spatial-temporal feature selection and application to air quality index forecasting. *Ecological Indicators*, 144, p.109416. DOI: <https://doi.org/10.1016/j.ecolind.2022.109416>
- Zhou, X., Zhou, X., Wang, C. and Zhou, H., 2023. Environmental and human health impacts of volatile organic compounds: A perspective review. *Chemosphere*, 313, p.137489. DOI: <https://doi.org/10.1016/j.chemosphere.2022.137489>
- Zhou, Y., Zhang, X., Zhang, C., Chen, B. and Gu, B., 2024. Mitigating air pollution benefits multiple sustainable development goals in China. *Environmental Pollution*, 349, p.123992. DOI: <https://doi.org/10.1016/j.envpol.2024.123992>
- Zhu, J., Lu, C. and Wei, Z., 2023. Perception of air pollution and the evaluation of local governments' environmental governance: An empirical study on China. *Atmosphere*, 14(2), p.212. DOI: <https://doi.org/10.3390/atmos14020212>
- Zhu, M. and Xie, J., 2023. Investigation of nearby monitoring station for hourly PM_{2.5} forecasting using parallel multi-input 1D-CNN-BiLSTM. *Expert Systems with Applications*, 211, p.118707. DOI: <https://doi.org/10.1016/j.eswa.2022.118707>

7. BACKGROUND INFORMATION FOR RECORD (NOT TO BE PRINTED)

Author Contributions: Conceptualisation, E.O.-O., O.A.A., R.T., and K.O.A.A.; methodology, E.O.-O., O.A.A., R.T., and K.O.A.A.; formal analysis, E.O.-O., O.A.A., R.T., and K.O.A.A.; investigation, E.O.-O., O.A.A., R.T., and K.O.A.A.; resources, E.O.-O., O.A.A., R.T., and K.O.A.A.; writing—original draft preparation, E.O.-O., O.A.A., R.T., and K.O.A.A.; writing—review and editing, E.O.-O., O.A.A., R.T., and K.O.A.A.; supervision, O.A.A., R.T., and K.O.A.A.; project administration, O.A.A., R.T., and K.O.A.A.; funding acquisition, O.A.A., R.T., and K.O.A.A.. All authors have read and agreed to the published version of the manuscript.” Authorship should be restricted to individuals who have made significant contributions to the research.

Funding: This study received no external funding.

Conflicts of Interest: The authors declare no conflicts of interest.