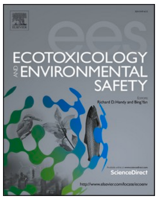
Contents lists available at ScienceDirect

Ecotoxicology and Environmental Safety

journal homepage: www.elsevier.com/locate/ecoenv



The application of machine learning to air pollution research: A bibliometric analysis

Yunzhe Li [a], Zhipeng Sha [a], Aohan Tang [a,*], Keith Goulding [b], Xuejun Liu [a]

[a] *Beijing Key Laboratory of Farmland Soil Pollution Prevention and Remediation, College of Resources and Environmental Science, China Agricultural University, Beijing 100193, China*

[b] *Sustainable Soils and Crops, Rothamsted Research, Harpenden AL5 2JQ, UK*

ARTICLE INFO

Edited by HAO ZHU

Keywords:
Air pollution
Machine learning
Bibliometric analysis
Research focuses
Development

ABSTRACT

Machine learning (ML) is an advanced computer algorithm that simulates the human learning process to solve problems. With an explosion of monitoring data and the increasing demand for fast and accurate prediction, ML models have been rapidly developed and applied in air pollution research. In order to explore the status of ML applications in air pollution research, a bibliometric analysis was made based on 2962 articles published from 1990 to 2021. The number of publications increased sharply after 2017, comprising approximately 75% of the total. Institutions in China and United States contributed half of all publications with most research being conducted by individual groups rather than global collaborations. Cluster analysis revealed four main research topics for the application of ML: chemical characterization of pollutants, short-term forecasting, detection improvement and optimizing emission control. The rapid development of ML algorithms has increased the capability to explore the chemical characteristics of multiple pollutants, analyze chemical reactions and their driving factors, and simulate scenarios. Combined with multi-field data, ML models are a powerful tool for analyzing atmospheric chemical processes and evaluating the management of air quality and deserve greater attention in future.

1. Introduction

Air pollution is harmful to human health and ecosystem stability, causing 7 million premature deaths and a $2.9 trillion global economic loss every year (IQAir, 2020). The global premature mortality burden resulting from ambient $PM_{2.5}$ and ozone exposure continues to increase (Chowdhury et al., 2020). In response to mounting risks, ground monitoring and remote sensing of pollutant emissions, transformation and deposition have played an important role in revealing the dynamic changes in air pollution (Cetin et al., 2018; Elsunousi et al., 2021; Hu et al., 2017; Kuerban et al., 2020; Li et al., 2023; Sevik et al., 2019; Wen et al., 2022; Xu et al., 2022a), and there has been a resultant explosion in monitoring data. Making full use of these data could provide a sound scientific basis for effective control of air pollution and policymaking.

The atmosphere is a changeable and open environmental system containing a range of pollutants and complex chemical processes, in which the relationships between components and controlling factors are not simply linear. Traditional statistical regressions models, such as parametric regression models, have limitations for fitting nonlinear relationships when dealing with the large and growing quantity of data in atmospheric science, including poor prediction accuracy for nonlinear problems analyzed using 'Big Data' (Feng et al., 2011). Some regression models are also complicated with too many explanatory variables, and their requirements for variable distribution are strict. These disadvantages are obstacles when trying to dig deeper into complex data and obtaining more valuable information. More convenient and accurate methods are urgently required for effective data analysis.

As artificial intelligence rapidly develops, machine learning (ML) is playing an increasingly important role in dealing with very large data sets. To date, commonly used ML algorithms are supervised learning algorithms such as artificial neural networks (ANN), random forest (RF) and support vector machine (SVM). These are mainly based on sample data for computer modeling for classification and regression (Zhong et al., 2021). ML models can establish direct relationships between data and weaken the impact of outliers (Ucun Ozel et al., 2020), delivering a higher prediction accuracy and robustness when dealing with large data sets (Chen et al., 2022; Yuchi et al., 2019). They produce a better fit to the data with a smaller root mean square error (RMSE), especially in

[*] Correspondence to: No. 2 Yuanmingyuan West Road, Beijing 100193, China.
E-mail address: aohantang@cau.edu.cn (A. Tang).

https://doi.org/10.1016/j.ecoenv.2023.114911
Received 24 November 2022; Received in revised form 27 March 2023; Accepted 10 April 2023
Available online 15 April 2023

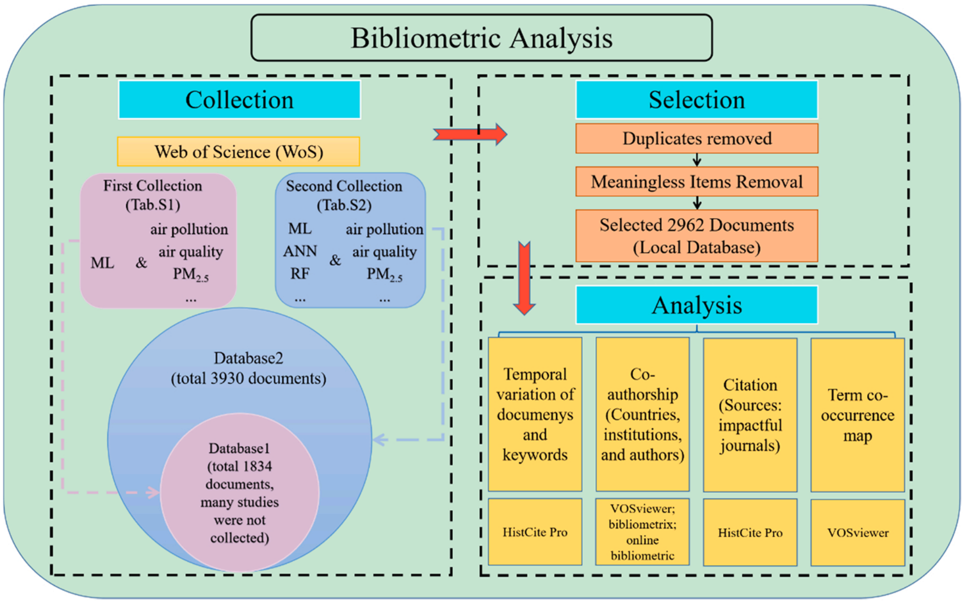

Bibliometric Analysis
Collection
Web of Science (WoS)
First Collection (Tab.S1)
ML & air pollution air quality $PM_{2.5}$ ...
Second Collection (Tab.S2)
ML ANN RF ... & air pollution air quality $PM_{2.5}$ ...
Database2 (total 3930 documents)
Database1 (total 1834 documents, many studies were not collected)
Selection
Duplicates removed
Meaningless Items Removal
Selected 2962 Documents (Local Database)
Analysis
Temporal variation of documenys and keywords
Co-authorship (Countries, institutions, and authors)
Citation (Sources: impactful journals)
Term co-occurrence map
HistCite Pro
VOSviewer; bibliometrix; online bibliometric
HistCite Pro
VOSviewer

Fig. 1. Flow chart of the bibliometric analysis.

nonlinear situations (Chen et al., 2022; Feng et al., 2011; Zimmerman et al., 2018). Also, multiple types data like integers and strings can be introduced when model construction. ML models are relatively simple, less time-consuming and low-cost compared to numerical models. These advantages make ML models increasingly popular in atmospheric science research (Liao et al., 2021; Zheng et al., 2021). Combined with multiple fields of data, ML models are key components for many tasks, such as short-term forecasting (Yan et al., 2021), analysis of the chemical behavior of pollutants (Huang et al., 2021) and impact assessment (Lv et al., 2023).

Bibliometric analysis is a valuable tool for analyzing the research status of a specific field (Qin et al., 2022a; Zhang and Chen, 2020). It can not only provide an analysis of the development of themes in research (Zhang et al., 2020b), but also show interrelations between countries, institutions and authors by revealing social networks (van Eck and Waltman, 2010). With the aim of advancing the understanding of ML, we made a bibliometric analysis of its applications and the development in global air pollution research. This provides further insight into the most appropriate application of ML in atmospheric science research in the future.

2. Materials and methods

2.1. Data collection

Literature recorded in the Web of Science (WoS) core database from 1990 to 2021 was collected. The search and analysis processes are shown in Fig. 1. Keywords used for this initial selection are shown in Tab. S1, and a total of 1834 documents were collected in database1. However, some documents were missed because the names of specific ML algorithms rather than “machine learning” were used in the titles or keywords of these papers. Therefore, keywords were expanded with specific ML algorithm names such as ANN, RF, and SVM. The keywords used in the second collection together with the retrieved paper numbers are shown in Tab. S2. A total of 3930 documents were collected in database2. The percentages of research articles, conference papers and reviews were 75.4%, 23.9% and 1.3%, respectively. Peer-reviewed publications identified as “article” and “review” were selected for further analysis (Bao et al., 2021; Qin et al., 2022a; Zhang et al., 2020b). After removing duplicate documents and irrelevant items, a total of 2962 publications were selected and formed into a new database named “Local Database” for subsequent analysis. The required information from each document, such as keywords, abstract, authors, source, reference, etc., was downloaded in “txt” format for analysis.

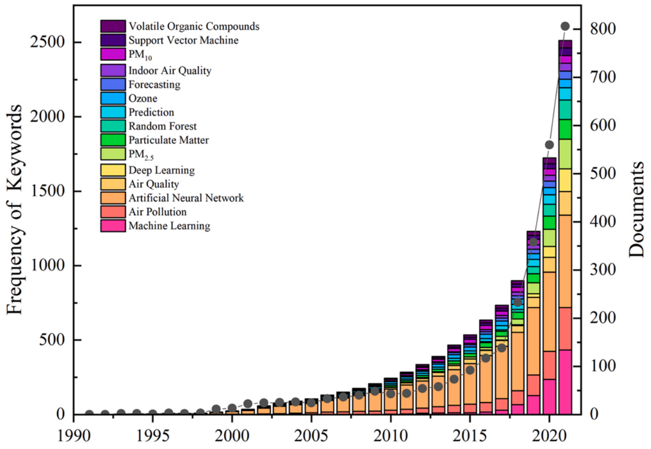


Fig. 2. Number of publications and frequency of occurrence of keywords from 1990 to 2021

2.2. Bibliometric analysis methods

Based on the Local Database, four types of analytical and visualization software were used for bibliometric analysis. Firstly, HistCite Pro (v2.1) was used to output informetric indicators, including the annual variation of publications and keyword occurrence, as well as citations of literature and journals. Then, three mature types of visualization software were introduced for social-relation analysis to reveal active groups, researchers, and their collaboration networks. The “bibliometrix” package in R, and an online bibliometric analysis software (https://bibliometric.com/) described the country distribution and collaboration networks. VOSviewer (v1.6) was used to construct bibliometric maps of institutions and authors. Finally, a term co-occurrence map was created from titles and abstracts for cluster analysis in VOSviewer to show the main directions and focuses of current research.

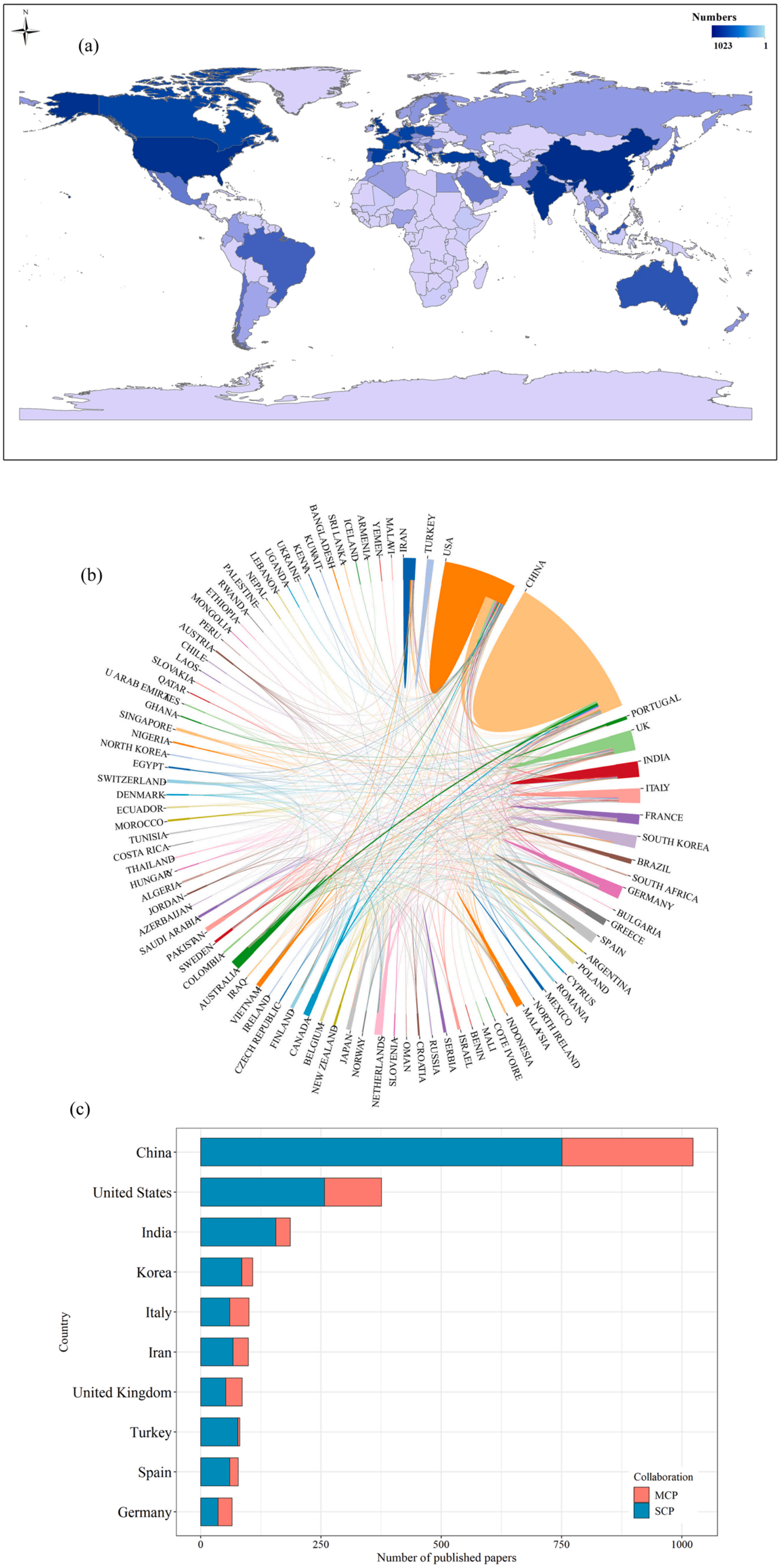


Fig. 3. (a) Distribution of publications by country; (b) Collaboration network of countries; (c) Analysis of single country publications (SCP) and multiple country publications (MCP).

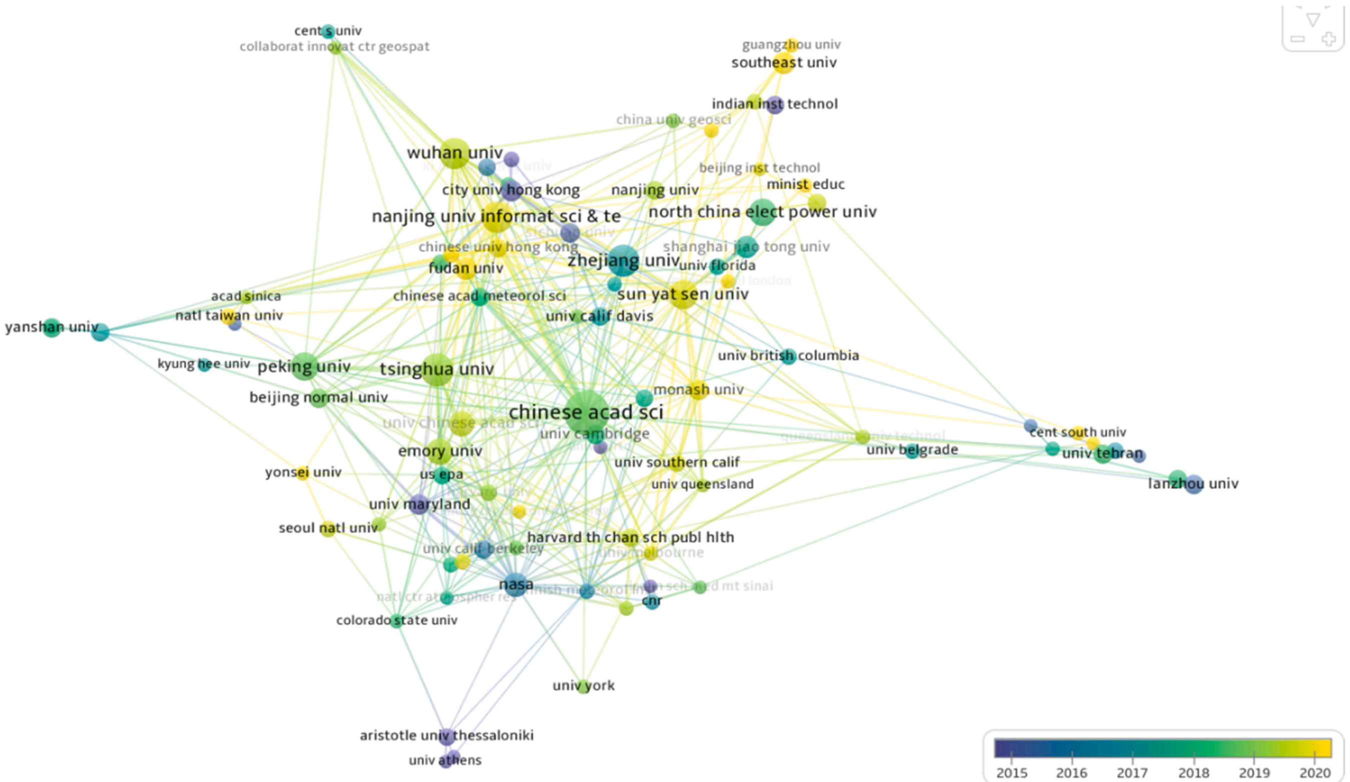


Fig. 4. Collaboration networks of the 96 institutions that each published > 10 papers. Note: The larger the marker, the more articles published; the lighter the color the more recent the publication.

3. Results and discussion

3.1. Temporal changes in the number of publications and keywords

The number of publications in the Local Database and the frequency of occurrence of the top 15 keywords are shown in Fig. 2. The publications increased from 1990 to 2021, with the first published research in 1993, applying ML to predict ambient SO_2 concentrations (Boznar et al., 1993). Research publications totaled less than 100 each year and increased slowly before 2015, with most research focusing on pollutant prediction by means of ANN. An explosive growth of publications occurred after 2017, with the total each year approaching 800 in 2021, illustrating the rapidly increasing interest in ML applications to air pollution research.

Changes in keywords are mainly reflected in two topics: ML algorithms and research targets. With the rapid increase of publications after 2015, the types of ML algorithms also increased. The phrase "machine learning" is now found in many papers, with ML algorithms such as "RF", "SVM" and "deep learning (DL)" appearing more and more frequently in recent years. The increasing use of multiple ML algorithms probably results from the requirement for better prediction accuracy. ANN has been used to predict atmospheric pollutant concentrations at a certain point in the past (Gardner and Dorling, 1999; Yi and Prybutok, 1996). With algorithm innovation, some emerging ML models can obtain more accurate prediction with less bias. For instance, DL has the capacity to achieve hourly forecasting and grid mapping to display high spatial-temporal resolution prediction of pollutants (Brokamp et al., 2017; Najafi et al., 2016; Qin et al., 2019; Yan et al., 2021). RF supports the analysis of the relationships between target pollutants and variables and so pollution formation mechanisms (Li et al., 2022b; Ye et al., 2022).

Fine particulate matter ($PM_{2.5}$), ozone, inhalable particulate matter (PM_{10}) and volatile organic compounds (VOCs) have been the main target pollutants, among which $PM_{2.5}$ has been of the most concern, with a greatly increased frequency of ML model applications after 2017. The application of ML models for indoor air quality control also gradually increased after 2015 (Ren and Cao, 2019; Yuchi et al., 2019). Generally, the main applications of ML were for predicting chemical characteristics of pollutants, including concentrations and their spatial and temporal variation.

3.2. Influential countries and groups

A total of 101 countries have conducted air pollution research using ML. The distribution and collaborations of active groups are shown in Fig. 3. China leads with the most publications (1023), followed by the USA (376) and India (186). Fig. 3b shows the collaborations of each country, in which cooperation between China and the USA have been the most frequent (180), followed by China and the UK (59) and the USA and the UK (39). Multiple country publications (MCP) are an indicator of international collaboration. However, the proportion of MCP for the top three countries was less than 30% (Fig. 3c), which shows that most research was conducted independently.

A total of 2836 institutions contributed to the application of ML to air quality research, of which 96 published more than 10 articles. The collaborations in these 96 papers are shown in Fig. 4: the bigger the size of the label, the more publications; the lighter the color, the more recent the publication. The Chinese Academy of Sciences published the most papers with 147 documents, followed by Tsinghua University (China) (56), and Wuhan University (China) (53). Most of the affiliations with the higher numbers of publications were in China, and Chinese institutions accounted for 70% of the top 20 institutions.

Local and global citation scores are the citation number in the Local Database and global WoS core databases, respectively. The ratio between local and global citation scores ($R_{(L/G)}$) was calculated to indicate the impact of each institution. A high $R_{(L/G)}$ means that an institution has had a significant impact on air pollution research by applying ML. The results in Tab. S3 show that Emory University has the highest $R_{(L/G)}$ value of the top 20 institutions, followed by Peking University and

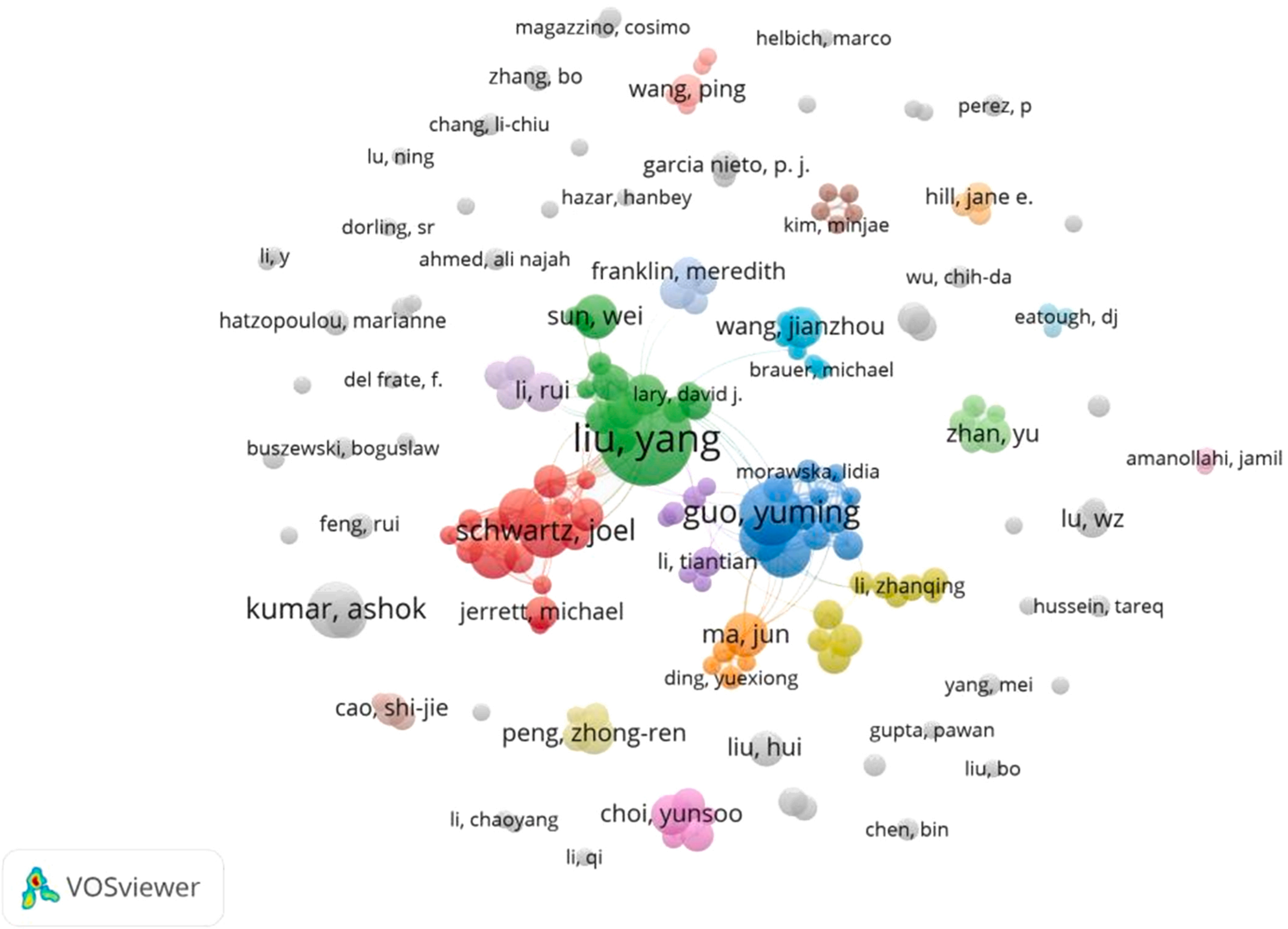


Fig. 5. Collaboration network of authors.

Sichuan University. Most Chinese institutions in the top 20 have both more publications and a higher $R_{(L/G)}$, suggesting China has had the greatest influence in this field. In Fig. 4, the lighter colors of Sun Yat-sen University (China), Fudan University (China) and Southeast University (China) show that the number of publications from these organizations has increased rapidly in recent years, suggesting that they could be important centers for ML applications in atmospheric research in the future.

Regarding authors, 11030 were recorded, but only 169 (1.5%) could be described as ‘active’, publishing more than 5 papers. The co-author relationships and details of active authors are shown in Fig. 5 and Tab. S4. Liu Yang ranked first with 27 publications, followed by Guo Yuming (18) and Kumar Ashok (16). Authors concentrated in the middle of Fig. 5 have multiple collaborations, but most authors have few if any collaborations. This indicates that associations between authors are generally weak and large research groupings have not yet formed. Greater collaboration should be advocated. The average publication year of each author is shown in Tab. S4, with 76% of authors was after 2019, suggesting that the application of ML to air pollution research is still a novel field.

3.3. Journals with most impact

The influence of journals can be assessed by analyzing the number of publications and citations in each, which can guide researchers as to where best to submit manuscripts. A total of 686 journals have published papers applying ML to air pollution, but only 54 journals published more than 10 papers in this field. The details of the 20 journals that published the most papers are shown in Tab.S5. European and American journals contributed the most publications. Although papers from China dominate, Chinese journals published relatively few of these, highlighting the need for relevant and high impact journals based in China. Most journals are in the environmental science category. However, some such as *IEEE Access* and *Sensors* are not classified as environmental by the JCR, indicating that some of the research is interdisciplinary. These papers involved topics such as ML algorithm modification and efficiency improvement for chemical detection (e.g. Al-Janabi et al., 2019; Spinelle et al., 2017).

The journal *Atmospheric Environment* published by far the most papers (154), followed by *Science of the Total Environment* (97) and *Environmental Pollution* (72). $R_{(L/G)}$ and $R_{(year)}$ (the ratio of ML-related publications after 2017 to total publications in that journal) of the top journals are shown in Fig. 6. A high $R_{(year)}$ value indicates that a journal has a strong interest in air pollution research that applies ML and has received an increasing number of articles in the 5 years to 2021. The $R_{(L/G)}$ and $R_{(year)}$ values of *Environmental Pollution* and *IEEE Access* exceeded the mean values and ranked third and fourth for the number of publications. These two journals have a strong interest in air pollution research that applies ML and are regularly selected by those researching ML-related to air pollution, suggesting that they are very suitable for future research submissions. Some journals with high $R_{(L/G)}$ values were distributed in the lower right quadrant of Fig. 6. These had high citations in the Local Database, and most were classic scientific journals publishing in atmospheric science, such as *Atmospheric Environment, Atmospheric Pollution Research,* and *Atmospheric Chemistry and Physics*, or were influential journals in the environmental field such as *Science of the Total*

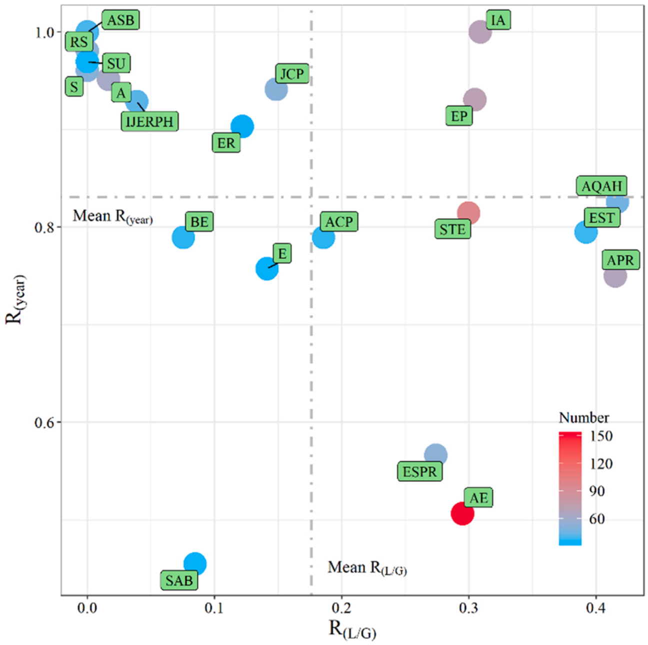


Fig. 6. Comparison of R(L/G) and R(year) for the 20 journals publishing the most ML-related papers. (AE: Atmospheric Environment; STE: Science of the Total Environment; EP: Environmental Pollution; IA: IEEE Access; APR: Atmospheric Pollution Research; A: Atmosphere; ESPR: Environmental Science and Pollution Research; JCP: Journal of Cleaner Production; RS: Remote Sensing; S: Sensors; AQAH: Air Quality Atmosphere and Health; IJERPH: International Journal of Environmental Research and Public Health; ASB: Applied Sciences-Basel; EST: Environmental Science & Technology; ACP: Atmospheric Chemistry and Physics; BE: Building and Environment; E: Energy; SAB: Sensors and Actuators B-Chemical; SU: Sustainability; ER: Environmental Research).

Environment and *Environmental Science & Technology*. Most papers published in these journals were concerned with applying ML to chemical processes of air pollutants formation.

3.4. Research focuses

Cluster analysis was conducted using VOSviewer to better understand the foci of ML-related articles. As shown in Fig. 7, publications were divided into four categories. Cluster 1 contains papers reporting ‘sensors’, ‘classification’, ‘detection’, ‘VOCs,’ and similar topics. Research in Cluster 1 was mainly focused on applying ML models to identifying specific signals in software for improving the efficiency of chemical detection. ML models have been applied not only to VOC detection but also to $PM_{2.5}$, carbon monoxide (CO) and nitrogen dioxide (NO_2) monitoring (Spinelle et al., 2017; Srivastava, 2003; Wang et al., 2014; Zhang et al., 2021a; Zimmerman et al., 2018), in which ANN is the most widely used machine learning algorithm. The different sensitivities of each sensor resulted from the range of detection materials used that respond differently to the target gas (Penza and Cassano, 2003). It is difficult to meet the detection requirements with a single sensor, but the cost will increase when using multiple sensors (Lee et al., 2002). Therefore, ML algorithms are introduced in a data processing module for calibration and improving detection efficiency. Initially, unique signal values for the pure gas are recorded for each sensor and input to the data processing module with a high frequency. Then ML algorithms use the input data as “training data” for model development (Zhang et al., 2021a). The constructed models are packaged in the instrument operating software. When a gas mixture is detected, ML models can analyze the data and rapidly identify each gas with a similar signal value (Barash et al., 2012; Zhang et al., 2012).

Words such as ‘emission’, ‘NO_x’, ‘optimization’ and ‘engine’ are distributed in Cluster 2. Research in this cluster focused on vehicle exhaust mitigation, where ML algorithms, especially ANN, are used to optimize parameters for reducing engine emissions (Hosamani et al., 2021; Lv et al., 2013). Typical exhaust emissions contained particulate matter (PM), nitrogen oxides (NO_x), hydrocarbons (HC) and CO (Gugulothu et al., 2021; Norouzi et al., 2020; Roy et al., 2014), which are influenced by engine type and fuel blending (Deh Kiani et al., 2010; Gugulothu et al., 2021). An effective design is critical for improving fuel efficiency and reducing emissions. To select the optimal design, ML models have been used to avoid consideration of the instantaneous combustion process and predict exhaust emissions quickly at low cost (Arcaklioğlu and Çelıkten, 2005). Usually, data collected from preliminary experiments can provides a rough range of parameters for model optimization. Parameters are then adjusted based on ML models for reducing emissions without having to repeat experiments.

Words such as ‘forecasting’ and ‘root mean square error’ (RMSE) are prominent in Cluster 3, while ‘particulate matter’, ‘China’, ‘observation’ and ‘estimate’ dominate in Cluster 4. These two fields of research both focused on the application of ML algorithms for fast and accurate prediction of pollutants, combined with historical meteorological, geographic and atmospheric pollutant data (Benhaddi and Ouarzazi, 2021; Hu et al., 2017). Commonly used ML methods were conventional neural network algorithms, especially DL, including convolutional neural networks (CNN) and long short-term memory (LSTM), with mean square error (MSE) and RMSE as evaluation indicators (Sun et al., 2020; Zhong et al., 2021). However, the purposes of these models are not the same. Alone or associated with atmospheric numerical models, most ML models in Cluster 3 were used to improve the accuracy of short-term forecasting for atmospheric contamination, especially $PM_{2.5}$ (Huang et al., 2021; Yan et al., 2021). In Cluster 4, most ML models were used to reveal the chemical characteristics of pollutants. Early research mainly focused on the spatiotemporal distribution of $PM_{2.5}$ (Chen et al., 2018; Gupta and Christopher, 2009; Hu et al., 2017), but gradually expanded to the chemical characterization of specific components and precursors in $PM_{2.5}$ with the development of ML models (Brokamp et al., 2017; Xu et al., 2017). Cluster 4 contained most articles (31.4%) and these had a more recent publication date (Fig. S1), indicating that applying ML for analyzing the chemical characteristics of atmospheric pollutants has become a recent focus of research.

The five most highly cited articles in the past 30 years were analyzed to provide more insight into recent research priorities (Tab. S6). Among these, Feng et al. (2015) was one of the earliest publications, in which ANN was combined with geographic models and wavelet transformation to improve the accuracy of forecasting $PM_{2.5}$ concentrations in North China. Following this, Di et al. (2016) and Hu et al. (2017) used RF and CNN, respectively, to estimate $PM_{2.5}$ concentrations with high spatial resolution across the United States, based on meteorological parameters, satellite aerosol optical depth and land-use variables. Li et al. (2017) used LSTM layers to explore the spatiotemporal correlation of pollutants for more accurate predictions in Beijing, China. Chen et al. (2018) used meteorological parameters, remote sensing data and land use information to create a $PM_{2.5}$ prediction model with RF and estimated the historical trend of $PM_{2.5}$ pollution across China. These five papers all focused on applying ML for analyzing the chemical characteristics of regional pollution, similar to the research papers that group into Cluster 4. This shows again that using ML models to analyze the chemical characteristics of atmospheric pollutants has been an important research focus.

3.5. Developments

In recent years, ML models have been applied ever more widely owing to the continuous innovation of algorithms. Analyzing articles in the impactful journals (Tab. S5) and highly cited papers (contained but not limited to Tab. S6), provides new insights for developments of ML applications in air pollution research.

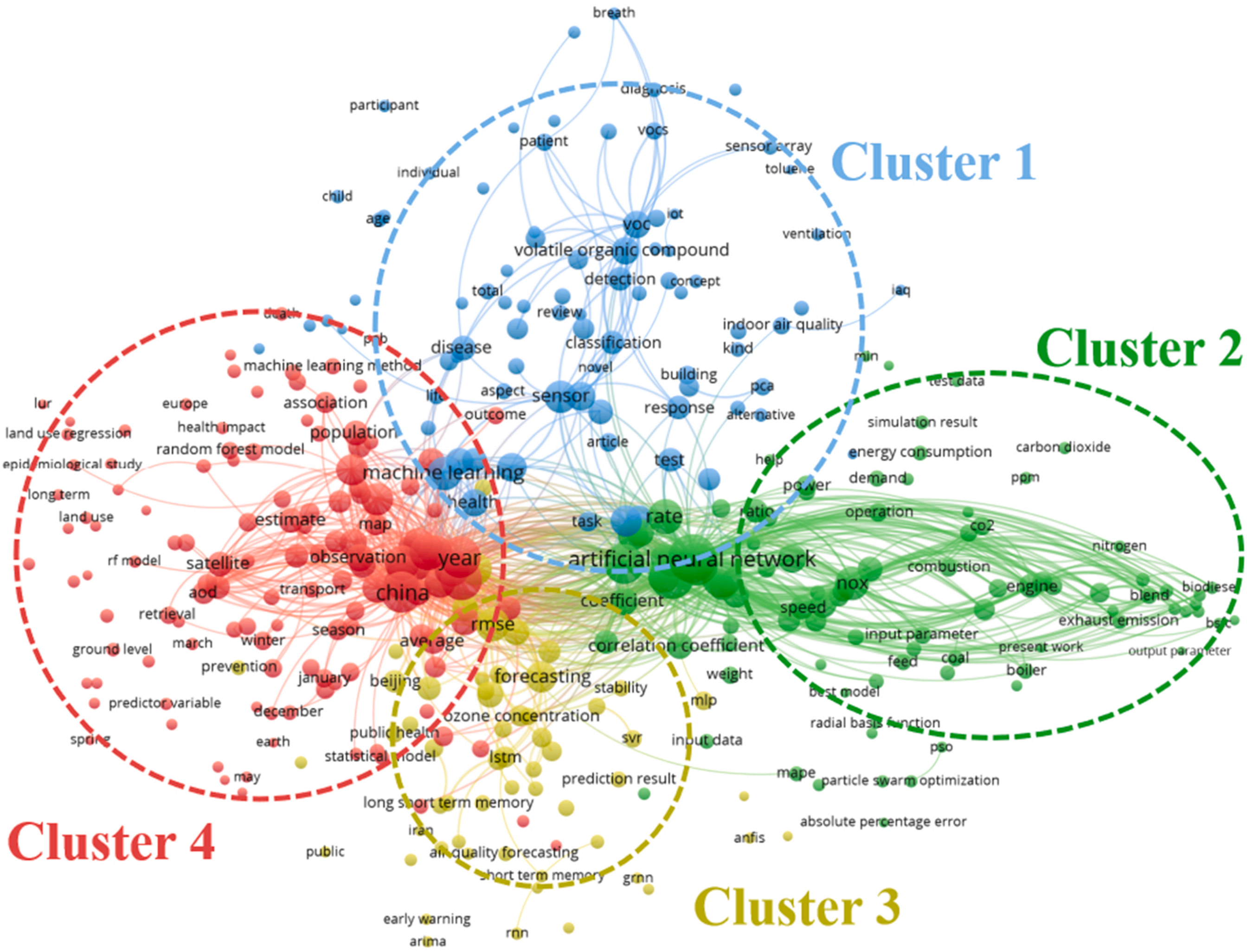


Fig. 7. Network visualization of a term co-occurrence map. The bigger the label, the more publications.

One important development is a focus on more diversified targets across larger domains. As noted in Section 3.4, the main application of ML models has been to predict and analyze the chemical characteristics of pollutants. Initially, ML models were used because they were an emerging and exciting way to obtain accurate and rapid predictions of atmospheric pollutant concentrations such as SO_2, NO_x, particulate matter and ozone at a specific place (Hooyberghs et al., 2005; Moseholm et al., 1993). Subsequently, rapid development has enabled the models to fill gaps resulting from incomplete monitoring. Not limited to $PM_{2.5,}$ concentrations and the spatial-temporal variation of multiple pollutants can be predicted at large scales, including carbonaceous materials, inorganic water-soluble ions and metal elements in $PM_{2.5}$ (Li et al., 2020a; Zhang et al., 2020c; Zhu et al., 2021), and their gaseous precursors such as ammonia, ozone and nitrous acid (Cui and Wang, 2021; He et al., 2021; Ren et al., 2022). ML models have also been used to quantify cloud condensation nuclei, which are important for aerosol formation but difficult to monitor directly (Nair and Yu, 2020), and to estimate emission and deposition fluxes of pollutants to explore their source and sink. Quantitative predictions of NH_3, NO_x, VOCs and even greenhouse gas emissions have been achieved (Bakay and Ağbulut, 2021; Li et al., 2021; Xu et al., 2021; Zhang et al., 2021b). Current research has also revealed the regional pattern of reactive nitrogen and/or sulfate in bulk deposition using ML models (Li et al., 2020b; Lu et al., 2020). ML models have become effective approaches to explore the transformation and transport of pollutants in atmosphere. With the development of ML models, it has been possible to predict multiple pollutants simultaneously, and we can expect their use to provide even more detail such as the vertical distribution of individual pollutants.

Simulating chemical reactions and diagnosing driving factors are important capabilities of ML models for dissecting chemical processes. Combined with quantum chemical methods, ML models can be used to analyze atmospheric chemical processes such as new particle formation, and heterogeneous and photochemical transformations of pollutants (Kubečka et al., 2022; Xia et al., 2022). For example, well-constructed ML models have been used to construct potential energy surfaces in order to obtain thermodynamic information about atmospheric radical reactions and cluster configurations in atmospheric nucleation (Liu et al., 2022; Stocker et al., 2020; Zhang et al., 2020a), providing more detailed parameters for thermodynamic and numerical models (Anderson et al., 2022; Xia et al., 2022). The importance of variables can be identified via the contribution of quantifying factors that can be used to explore the key factors affecting pollutant formation.Tree-based models were commonly applied, such as regression tree, RF and gradient boost regression tree (Carslaw and Taylor, 2009; Xu et al., 2021). When tree-based models are created, the importance of variables should be tested to reduce Gini importance or variance (Strobl et al., 2008). This can assist in the evaluation of factor weight, revealing the sensitivity of pollutants to these factors. For example, Zhang et al. (2020c) studied the chemical components of PM_1 (particulate matter less than 1 μm) by a similar approach with RF models and found that components were sensitive to relative humidity. Based on importance diagnose of variables, Ye et al. (2022) showed that a prediction bias for ozone concentrations in chemical transport models resulted from the underestimation of dry deposition and cloud optical depth. Further, coupled with a SHapley Additive ExPlanation (SHAP) approach, any ML model can provide a clear explanation of variable importance (evaluating contributions in physical units such as μg/m^3), which can support the dimension reduction of variables for the source apportionment of pollutants (Hou et al., 2022; Qin et al., 2022b; Zhang et al., 2022).

Scenario simulation with ML models enables researchers to evaluate control strategies. As with traditional numerical models, researchers can evaluate the effectiveness of management measures for air pollution mitigation at regional or even global scales. For example, Li et al. (2022a) predicted a decreasing trend of global aerosol pollution from 2020 to 2100, based on a number of projections in the Coupled Model Intercomparison Project Phase 6. Xu et al. (2022b) forecasted ammonia emissions and their resultant health impacts in 2030 under several scenarios in which nitrogen fertilizer was replaced by agricultural wastes. ML models are also an alternative for clarifying the factors driving air quality change over the short term. Research has simulated atmospheric pollutant variation under various emission scenarios to assess the impact of reduced anthropogenic activity during COVID-19 lockdowns periods (He et al., 2021; Shi et al., 2021). Compared to traditional numerical models, ML models are simpler and easier for new users. Moreover, data from economic, energy-related and other fields can be incorporated into ML models to further identity key driving factors, providing more comprehensive evaluations of regional or national air quality management. More scenario simulations using ML models are expected in the future.

Although ML has shown great potential in air pollution research, there are still some issues worth noting when considering its future application. Firstly, outputs of ML models are highly dependent on the quality of training data. Representativeness rather than number of variables is the basis for better model performance; pre-process variables must be carefully selected, and model users must have enough experience and patience to debug the models effectively. Secondly, outputs are merely the results of computer calculations and sometimes may contradict current understanding due to the “black box” effect if there is no good understanding and what the model does. Therefore, after model construction, the simulations should be fully verified and interpreted based on atmospheric theory.

4. Conclusion

Publications that apply ML to air pollution research have increased rapidly in the past 30 years and grew exponentially after 2017. Dominant research groups were based mainly in China and the United States, which together accounted for more than 50% of all articles. Research institutions and groups were relatively scattered, and large research collaborations have not yet formed. However, the application of ML to air pollution research has reached maturity, with a strong focus on the chemical characteristic analysis of pollutants, short-term pollutant forecasts, improving pollutant detection efficiency and optimizing design for emission reduction. Based on representative training data and effective verification, ML models supported by advanced algorithms have considerable potential to comprehensively interpret air pollution formation and control, including exploring the chemical characteristics of multiple pollutants, quantifying atmospheric chemical process and their driving factors, and simulating scenarios.

CRediT authorship contribution statement

Yunzhe Li: Data curation, Writing – original draft preparation, Visualization. **Zhipeng Sha:** Data curation, Visualization. **Aohan Tang:** Conceptualization, Methodology,Supervision. **Keith Goulding:** Writing – review & editing. **Xuejun Liu:** Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

Data will be made available on request.

Acknowledgments

This work was supported by the Key Program of the National Natural Science Foundation of China (No. 41730855), the major project of Inner Mongolian Natural Science Foundation (No. 2019ZD02), and the High-level Innovative Talent Project of China Agricultural University.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.ecoenv.2023.114911.

References

Al-Janabi, S., Mohammad, M., Al-Sultan, A., 2019. A new method for prediction of air pollution based on intelligent computation. Soft Comput. 24, 661–680.

Anderson, D.C., Follette-Cook, M.B., Strode, S.A., Nicely, J.M., Liu, J., Ivatt, P.D., Duncan, B.N., 2022. A machine learning methodology for the generation of a parameterization of the hydroxyl radical. Geosci. Model Dev. 15, 6341–6358.

Arcaklioğlu, E., Çelıkten, İ., 2005. A diesel engine’s performance and exhaust emissions. Appl. Energ. 80, 11–22.

Bakay, M.S., Ağbulut, Ü., 2021. Electricity production based forecasting of greenhouse gas emissions in Turkey with deep learning, support vector machine and artificial neural network algorithms. J. Clean. Prod. 285, 125324.

Bao, Y., Mehmood, K., Saifullah, Yaseen, M., Dahlawi, S., Abrar, M.M., Khan, M.A., Saud, S., Dawar, K., Fahad, S., Faraj, T.K., 2021. Global research on the air quality status in response to the electrification of vehicles. Sci. Total Environ. 795, 148861.

Barash, O., Peled, N., Tisch, U., Bunn Jr., P.A., Hirsch, F.R., Haick, H., 2012. Classification of lung cancer histology by gold nanoparticle sensors. Nanomedicine 8, 580–589.

Benhaddi, M., Ouarzazi, J., 2021. Multivariate time series forecasting with dilated residual convolutional neural networks for urban air quality prediction. Arab. J. Sci. Eng. 46, 3423–3442.

Boznar, M., Lesjak, M., Mlakar, P., 1993. A neural network-based method for short-term predictions of ambient SO_2 concentrations in highly polluted industrial areas of complex terrain. Atmos. Environ. Part B. Urban Atmos. 27, 221–230.

Brokamp, C., Jandarov, R., Rao, M.B., LeMasters, G., Ryan, P., 2017. Exposure assessment models for elemental components of particulate matter in an urban environment: A comparison of regression and random forest approaches. Atmos. Environ. 151, 1–11.

Carslaw, D.C., Taylor, P.J., 2009. Analysis of air pollution data at a mixed source location using boosted regression trees. Atmos. Environ. 43, 3563–3570.

Cetin, M., Onac, A.K., Sevik, H., Sen, B., 2018. Temporal and regional change of some air pollution parameters in Bursa. Air Qual. Atmos. Health 12, 311–316.

Chen, G., Li, S., Knibbs, L.D., Hamm, N.A.S., Cao, W., Li, T., Guo, J., Ren, H., Abramson, M.J., Guo, Y., 2018. A machine learning method to estimate $PM_{2.5}$ concentrations across China with remote sensing, meteorological and land use information. Sci. Total Environ. 636, 52–60.

Chen, X., Zheng, H., Wang, H., Yan, T., 2022. Can machine learning algorithms perform better than multiple linear regression in predicting nitrogen excretion from lactating dairy cows. Sci. Rep. 12, 12478.

Chowdhury, S., Pozzer, A., Dey, S., Klingineullei, K., Lelieveld, J., 2020. Changing risk factors that contribute to premature mortality from ambient air pollution between 2000 and 2015. Environ. Res. Lett. 15, 074010.

Cui, L., Wang, S., 2021. Mapping the daily nitrous acid (HONO) concentrations across China during 2006-2017 through ensemble machine-learning algorithm. Sci. Total. Environ. 785, 147325.

Deh Kiani, M.K., Ghobadian, B., Tavakoli, T., Nikbakht, A.M., Najafi, G., 2010. Application of artificial neural networks for the prediction of performance and exhaust emissions in SI engine using ethanol- gasoline blends. Energy 35, 65–69.

Di, Q., Kloog, I., Koutrakis, P., Lyapustin, A., Wang, Y., Schwartz, J., 2016. Assessing $PM_{2.5}$ exposures with high spatiotemporal resolution across the continental United States. Environ. Sci. Technol. 50, 4712–4721.

van Eck, N.J., Waltman, L., 2010. Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics 84, 523–538.

Elsunousi, A.A.M., Sevik, H., Cetin, M., Ozel, H.B., Ozel, H.U., 2021. Periodical and regional change of particulate matter and CO_2 concentration in Misurata. Environ. Monit. Assess. 193, 707.

Feng, X., Li, Q., Zhu, Y., Hou, J., Jin, L., Wang, J., 2015. Artificial neural networks forecasting of $PM_{2.5}$ pollution using air mass trajectory based geographic model and wavelet transformation. Atmos. Environ. 107, 118–128.

Feng, Y., Zhang, W., Sun, D., Zhang, L., 2011. Ozone concentration forecast method based on genetic algorithm optimized back propagation neural networks and support vector machine data classification. Atmos. Environ. 45, 1979–1985.

Gardner, M.W., Dorling, S.R., 1999. Neural network modelling and prediction of hourly NOx and NO_2 concentrations in urban air in London. Atmos. Environ. 33, 709–719.

Gugulothu, S.K., Ramachander, J., Kumar, A.K., 2021. Predicting the engine trade-off study and performance characteristics using different blends of methyl Ester fish oil and higher alcohol with aid of artificial neural network based multi objective optimization. Heat. Mass Transf. 57, 1121–1138.

Gupta, P., Christopher, S.A., 2009. Particulate matter air quality assessment using integrated surface, satellite, and meteorological products: 2. A neural network approach. J. Geophys. Res-Atmos. 114, D20205.

He, Y., Pan, Y., Gu, M., Sun, Q., Zhang, Q., Zhang, R., Wang, Y., 2021. Changes of ammonia concentrations in wintertime on the North China Plain from 2018 to 2020. Atmos. Res. 253, 105490.

Hooyberghs, J., Mensink, C., Dumont, G., Fierens, F., Brasseur, O., 2005. A neural network forecast for daily average PM concentrations in Belgium. Atmos. Environ. 39, 3279–3289.

Hosamani, B.R., Abbas Ali, S., Katti, V., 2021. Assessment of performance and exhaust emission quality of different compression ratio engine using two biodiesel mixture: Artificial neural network approach. Alex. Eng. J. 60, 837–844.

Hou, L., Dai, Q., Song, C., Liu, B., Guo, F., Dai, T., Li, L., Liu, B., Bi, X., Zhang, Y., Feng, Y., 2022. Revealing drivers of haze pollution by explainable machine learning. Environ. Sci. Technol. Lett. 9, 112–119.

Hu, X., Belle, J.H., Meng, X., Wildani, A., Waller, L.A., Strickland, M.J., Liu, Y., 2017. Estimating $PM_{2.5}$ concentrations in the conterminous United States using the random forest approach. Environ. Sci. Technol. 51, 6936–6944.

Huang, G., Li, X., Zhang, B., Ren, J., 2021. $PM_{2.5}$ concentration forecasting at surface monitoring sites using GRU neural network based on empirical mode decomposition. Sci. Total. Environ. 768, 144516.

IQAir, 2020. World air quality report. Switzerland.

Kubečka, J., Christensen, A.S., Rasmussen, F.R., Elm, J., 2022. Quantum machine learning approach for studying atmospheric cluster formation. Environ. Sci. Technol. Lett. 9, 239–244.

Kuerban, M., Waili, Y., Fan, F., Liu, Y., Qin, W., Dore, A.J., Peng, J., Xu, W., Zhang, F., 2020. Spatio-temporal patterns of air pollution in China from 2015 to 2018 and implications for health risks. Environ. Pollut. 258, 113659.

Lee, D.S., Kim, Y.T., Huh, J.S., Lee, D.D., 2002. Fabrication and characteristics of SnO_2 gas sensor array for volatile organic compounds recognition. Thin Solid Films 416, 271–278.

Li, H., Dai, Q., Yang, M., Li, F., Liu, X., Zhou, M., Qian, X., 2020a. Heavy metals in submicronic particulate matter (PM_1) from a Chinese metropolitan city predicted by machine learning models. Chemosphere 261, 127571.

Li, H., Yang, Y., Wang, H., Wang, P., Yue, X., Liao, H., 2022a. Projected aerosol changes driven by emissions and climate change using a machine learning method. Environ. Sci. Technol. 56, 3884–3893.

Li, R., Cui, L., Fu, H., Zhao, Y., Zhou, W., Chen, J., 2020b. Satellite-based estimates of wet ammonium (NH_4-N) deposition fluxes across China during 2011-2016 using a space-time ensemble model. Environ. Sci. Technol. 54, 13419–13428.

Li, S., Chen, C., Yang, G.L., Fang, J., Sun, Y., Tang, L., Wang, H., Xiang, W., Zhang, H., Croteau, P.L., Jayne, J.T., Liao, H., Ge, X., Favez, O., Zhang, Y., 2022b. Sources and processes of organic aerosol in non-refractory PM_1 and $PM_{2.5}$ during foggy and haze episodes in an urban environment of the Yangtze River Delta, China. Environ. Res. 212, 113557.

Li, X., Peng, L., Yao, X., Cui, S., Hu, Y., You, C., Chi, T., 2017. Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation. Environ. Pollut. 231, 997–1004.

Li, Y., Jia, M., Han, X., Bai, X.-S., 2021. Towards a comprehensive optimization of engine efficiency and emissions by coupling artificial neural network (ANN) with genetic algorithm (GA). Energy 225.

Li, Y., Hong, T., Gu, Y., Li, Z., Huang, T., Lee, H.F., Heo, Y., Yim, S.H.L., 2023. Assessing the spatiotemporal characteristics, factor importance, and health impacts of air pollution in seoul by integrating machine learning into Land-use Regression modeling at high spatiotemporal resolutions. Environ. Sci. Technol. 57 (3), 1225–1236.

Liao, K., Huang, X., Dang, H., Ren, Y., Zuo, S., Duan, C., 2021. Statistical approaches for forecasting primary air pollutants: a review. Atmosphere 12 (6), 686.

Liu, Y., Xie, H.B., Ma, F., Chen, J., Elm, J., 2022. Amine-enhanced methanesulfonic acid-driven nucleation: Predictive model and cluster formation mechanism. Environ. Sci. Technol. 56, 7751–7760.

Lu, X., Yuan, D., Chen, Y., Fung, J.C.H., Li, W., Lau, A.K.H., 2020. Estimations of long-term nss-SO_4^{2-} and NO_3^- wet depositions over East Asia by use of ensemble machine-learning method. Environ. Sci. Technol. 54, 11118–11126.

Lv, Y., Liu, J., Yang, T., Zeng, D., 2013. A novel least squares support vector machine ensemble model for NOx emission prediction of a coal-fired boiler. Energy 55, 319–329.

Lv, Y., Tian, H., Luo, L., Liu, S., Bai, X., Zhao, H., Zhang, K., Lin, S., Zhao, S., Guo, Z., Xiao, Y., Yang, J., 2023. Understanding and revealing the intrinsic impacts of the COVID-19 lockdown on air quality and public health in North China using machine learning. Sci. Total. Environ. 857, 159339.

Moseholm, L., Taudorf, E., Frosig, A., 1993. Pulmonary function changes in asthmatics associated with low-level SO_2 and NO_2 air pollution, weather, and medicine intake. An 8-month prospective study analyzed by neural networks. Allergy 48, 334–344.

Nair, A.A., Yu, F., 2020. Using machine learning to derive cloud condensation nuclei number concentrations from commonly available measurements. Atmos. Chem. Phys. 20, 12853–12869.

Najafi, G., Ghobadian, B., Moosavian, A., Yusaf, T., Mamat, R., Kettner, M., Azmi, W.H., 2016. SVM and ANFIS for prediction of performance and exhaust emissions of a SI engine with gasoline–ethanol blended fuels. Appl. Therm. Eng. 95, 186–203.

Norouzi, A., Aliramezani, M., Koch, C.R., 2020. A correlation-based model order reduction approach for a diesel engine NOx and brake mean effective pressure dynamic model using machine learning. Int. J. Engine Res. 22, 2654–2672.

Penza, M., Cassano, G., 2003. Application of principal component analysis and artificial neural networks to recognize the individual VOCs of methanol/2-propanol in a binary mixture by SAW multi-sensor array. Sens. Actuators B-Chem. 89, 269–284.

Qin, D., Yu, J., Zou, G., Yong, R., Zhao, Q., Zhang, B., 2019. A novel combined prediction scheme based on CNN and LSTM for urban $PM_{2.5}$ concentration. IEEE Access 7, 20050–20059.

Qin, F., Li, J., Zhang, C., Zeng, G., Huang, D., Tan, X., Qin, D., Tan, H., 2022a. Biochar in the 21st century: a data-driven visualization of collaboration, frontier identification, and future trend. Sci. Total. Environ. 818, 151774.

Qin, X., Zhou, S., Li, H., Wang, G., Chen, C., Liu, C., Wang, X., Huo, J., Lin, Y., Chen, J., Fu, Q., Duan, Y., Huang, K., Deng, C., 2022b. Enhanced natural releases of mercury in response to the reduction in anthropogenic emissions during the COVID-19 lockdown by explainable machine learning. Atmos. Chem. Phys. 22, 15851–15865.

Ren, C., Cao, S.-J., 2019. Development and application of linear ventilation and temperature models for indoor environmental prediction and HVAC systems control. Sustain. Cities Soc. 51, 101673.

Ren, X., Mi, Z., Cai, T., Nolte, C.G., Georgopoulos, P.G., 2022. Flexible bayesian ensemble machine learning framework for predicting local ozone concentrations. Environ. Sci. Technol. 56, 3871–3883.

Roy, S., Banerjee, R., Bose, P.K., 2014. Performance and exhaust emissions prediction of a CRDI assisted single cylinder diesel engine coupled with EGR using artificial neural network. Appl. Energy 119, 330–340.

Sevik, H., Cetin, M., Ozel, H.B., Akarsu, H., Zeren Cetin, I., 2019. Analyzing of usability of tree-rings as biomonitors for monitoring heavy metal accumulation in the atmosphere in urban area: a case study of cedar tree (Cedrus sp. Environ. Monit. Assess. 192, 23.

Shi, Z., Song, C., Liu, B., Lu, G., Xu, J., Van Vu, T., Elliott, R.J.R., Li, W., Bloss, W.J., Harrison, R.M., 2021. Abrupt but smaller than expected changes in surface air quality attributable to COVID-19 lockdowns. Sci. Adv. 7, eabd6696.

Spinelle, L., Gerboles, M., Villani, M.G., Aleixandre, M., Bonavitacola, F., 2017. Field calibration of a cluster of low-cost commercially available sensors for air quality monitoring. Part B: NO, CO and CO_2. Sens. Actuators B-Chem. 238, 706–715.

Srivastava, A.K., 2003. Detection of volatile organic compounds (VOCs) using SnO_2 gas-sensor array and artificial neural network. Sens. Actuators B-Chem. 96, 24–37.

Stocker, S., Csanyi, G., Reuter, K., Margraf, J.T., 2020. Machine learning in chemical reaction space. Nat. Commun. 11, 5505.

Strobl, C., Boulesteix, A.L., Kneib, T., Augustin, T., Zeileis, A., 2008. Conditional variable importance for random forests. BMC Bioinforma. 9, 307.

Sun, Z., Wang, C., Ye, Z., Bi, H., 2020. Long short-term memory network-based emission models for conventional and new energy buses. Int. J. Sustain. Transp. 15, 229–238.

Ucun Ozel, H., Gemici, B.T., Gemici, E., Ozel, H.B., Cetin, M., Sevik, H., 2020. Application of artificial neural networks to predict the heavy metal contamination in the Bartin River. Environ. Sci. Pollut. Res. Int. 27, 42495–42512.

Wang, B., Cancilla, J.C., Torrecilla, J.S., Haick, H., 2014. Artificial sensing intelligence with silicon nanowires for ultraselective detection in the gas phase. Nano. Lett. 14, 933–938.

Wen, Z., Wang, R., Li, Q., Liu, J., Ma, X., Xu, W., Tang, A., Collett Jr., J.L., Li, H., Liu, X., 2022. Spatiotemporal variations of nitrogen and phosphorus deposition across China. Sci. Total. Environ. 830, 154740.

Xia, D., Chen, J., Fu, Z., Xu, T., Wang, Z., Liu, W., Xie, H.B., Peijnenburg, W., 2022. Potential application of machine-learning-based quantum chemical methods in environmental chemistry. Environ. Sci. Technol. 56, 2115–2123.

Xu, P., Li, G., Houlton, B.Z., Ma, L., Ai, D., Zhu, L., Luan, B., Zhai, S., Hu, S., Chen, A., Zheng, Y., 2022b. Role of Organic and Conservation Agriculture in Ammonia Emissions and Crop Productivity in China. Environ Sci Technol 56, 2977–2989.

Xu, X., Ouyang, X., Gu, Y., Cheng, K., Smith, P., Sun, J., Li, Y., Pan, G., 2021. Climate change may interact with nitrogen fertilizer management leading to different ammonia loss in China's croplands. Glob. Chang. Biol. 27, 6525–6535.

Xu, Y., Yang, W., Wang, J., 2017. Air quality early-warning system for cities in China. Atmos. Environ. 148, 239–257.

Xu, W., Zhao, Y., Wen, Z., Chang, Y., Pan, Y., Sun, Y., Ma, X., Sha, Z., Li, Z., Kang, J., Liu, L., Tang, A., Wang, K., Zhang, Y., Guo, Y., Zhang, L., Sheng, L., Zhang, X., Gu, B., Song, Y., Van Damme, M., Clarisse, L., Coheur, P.F., Collett Jr., J.L., Goulding, K., Zhang, F., He, K., Liu, X., 2022a. Increasing importance of ammonia emission abatement in $PM_{2.5}$ pollution control. Sci. Bull. 67, 1745–1749.

Yan, R., Liao, J., Yang, J., Sun, W., Nong, M., Li, F., 2021. Multi-hour and multi-site air quality index forecasting in Beijing using CNN, LSTM, CNN-LSTM, and spatiotemporal clustering. Expert Syst. Appl. 169, 114513.

Ye, X., Wang, X., Zhang, L., 2022. Diagnosing the model bias in simulating daily surface ozone variability using a machine learning method: the effects of dry deposition and cloud optical depth. Environ. Sci. Technol. 56, 16665–16675.

Yi, J., Prybutok, V.R., 1996. A neural network model forecasting for prediction of daily maximum ozone concentration in an industrialized urban area. Environ. Pollut. 92, 349–357.

Yuchi, W., Gombojav, E., Boldbaatar, B., Galsuren, J., Enkhmaa, S., Beejin, B., Naidan, G., Ochir, C., Legtseg, B., Byambaa, T., Barn, P., Henderson, S.B., Janes, C. R., Lanphear, B.P., McCandless, L.C., Takaro, T.K., Venners, S.A., Webster, G.M., Allen, R.W., 2019. Evaluation of random forest regression and multiple linear regression for predicting indoor fine particulate matter concentrations in a highly polluted city. Environ. Pollut. 245, 746–753.

Zhang, J., Glezakou, V.A., Rousseau, R., Nguyen, M.T., 2020a. NWPEsSe: an adaptive-learning global optimization algorithm for nanosized cluster systems. J. Chem. Theory Comput. 16, 3947–3958.

Zhang, J., Xue, Y., Sun, Q., Zhang, T., Chen, Y., Yu, W., Xiong, Y., Wei, X., Yu, G., Wan, H., Wang, P., 2021a. A miniaturized electronic nose with artificial neural network for anti-interference detection of mixed indoor hazardous gases. Sens. Actuators B-Chem. 326, 128822.

Zhang, L., Tian, F., Nie, H., Dang, L., Li, G., Ye, Q., Kadri, C., 2012. Classification of multiple indoor air contaminants by an electronic nose and a hybrid support vector machine. Sens. Actuators B-Chem. 174, 114–125.

Zhang, R., Wang, H., Tan, Y., Zhang, M., Zhang, X., Wang, K., Ji, W., Sun, L., Yu, X., Zhao, J., Xu, B., Xiong, J., 2021b. Using a machine learning approach to predict the emission characteristics of VOCs from furniture. Build. Environ. 196, 107786.

Zhang, Y., Chen, Y., 2020. Research trends and areas of focus on the Chinese Loess Plateau: A bibliometric analysis during 1991–2018. Catena 194, 104798.

Zhang, Y., Pu, S., Lv, X., Gao, Y., Ge, L., 2020b. Global trends and prospects in microplastics research: a bibliometric analysis. J. Hazard. Mater. 400, 123110.

Zhang, Y., Vu, T.V., Sun, J., He, J., Shen, X., Lin, W., Zhang, X., Zhong, J., Gao, W., Wang, Y., Fu, T.M., Ma, Y., Li, W., Shi, Z., 2020c. Significant changes in chemistry of fine particles in wintertime Beijing from 2007 to 2017: Impact of clean air actions. Environ. Sci. Technol. 54, 1344–1352.

Zhang, Z., Xu, B., Xu, W., Wang, F., Gao, J., Li, Y., Li, M., Feng, Y., Shi, G., 2022. Machine learning combined with the PMF model reveal the synergistic effects of sources and meteorological factors on $PM_{2.5}$ pollution. Environ. Res. 212, 113322.

Zheng, L., Lin, R., Wang, X., Chen, W., 2021. The development and application of machine learning in atmospheric environment studies. Remote Sens-Basel 13, 4839.

Zhong, S., Zhang, K., Bagheri, M., Burken, J.G., Gu, A., Li, B., Ma, X., Marrone, B.L., Ren, Z.J., Schrier, J., Shi, W., Tan, H., Wang, T., Wang, X., Wong, B.M., Xiao, X., Yu, X., Zhu, J.J., Zhang, H., 2021. Machine learning: New ideas and tools in environmental science and engineering. Environ. Sci. Technol. 55, 12741–12754.

Zhu, J.J., Chen, Y.C., Shie, R.H., Liu, Z.S., Hsu, C.Y., 2021. Predicting carbonaceous aerosols and identifying their source contribution with advanced approaches. Chemosphere 266, 128966.

Zimmerman, N., Presto, A.A., Kumar, S.P.N., Gu, J., Hauryliuk, A., Robinson, E.S., Robinson, A.L., 2018. A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring. Atmos. Meas. Tech. 11, 291–313.