

Article

From Data to discovery: The indispensable roles of statistics and machine learning in environmental chemistry

Abdulwasiu Olawale Salaudeen^{1,*}, Yemisi Ajoke Olawore², Hajara Yakubu³, Habib Abdulkadiri⁴ and Akoji Godwin Mathias⁵

¹ Chemistry Unit of Mathematics Programme, National Mathematical Centre Abuja, Nigeria

² Biology Unit of Mathematics Programme, National Mathematical Centre Abuja, Nigeria

³ Chemistry Department, University of Abuja, Nigeria

⁴ Industrial Chemistry Department, Edo State University, Iyamho, Nigeria

⁵ Chemistry Department, Federal University Lokoja, Nigeria

* Correspondence: greenplanet5050@gmail.com

Received: 11 July 2026; Accepted: 09 September 2026; Published: 28 September 2026.

Abstract: Environmental chemistry has entered a data-rich era in which analytical instruments, sensor networks, and remote-sensing platforms generate enormous volumes of complex information. Statistical methods have long been the foundation of environmental analysis, offering hypothesis testing, regression, variance analysis, and multivariate approaches for exploring patterns in pollutant distribution and chemical interactions. While these classical approaches remain essential for experimental design, calibration, and uncertainty assessment, the emergence of machine learning has provided new opportunities to address the high dimensionality, nonlinearity, and heterogeneity of modern environmental datasets. From random forests and support vector machines to gradient boosting and deep neural networks, these tools have been applied to problems such as pollutant source apportionment, non-target screening in high-resolution mass spectrometry, spectral deconvolution, sensor fusion, and water-quality forecasting. Their capacity to recognize complex patterns can deliver predictive accuracy that surpasses traditional methods in many study-specific comparisons, yet their adoption has introduced new concerns related to overfitting, data leakage, lack of interpretability, and poor reproducibility. In order to fully harness the benefits of machine learning in environmental chemistry, it is critical to embed domain knowledge, validate models using robust statistical frameworks, quantify uncertainty, and establish standardized practices for transparency and reproducibility. This review examines the evolution of statistical and machine learning tools in environmental chemistry, analyzes their applications and comparative strengths, highlights challenges and pitfalls, and proposes future directions that emphasize hybrid physics–ML models, interpretability, reproducibility, and ethical use. By integrating statistical rigor with computational innovation, environmental chemistry can better support discovery, decision-making, and sustainable management of pollutants.

Keywords: machine learning, statistics, environment, data, pollution, multivariate

1. Introduction

Environmental chemistry has long depended on quantitative tools to describe the distribution, transformation, and impacts of contaminants, and classical statistical methods such as regression, analysis of variance, principal component analysis, and cluster analysis provided a reliable framework for interpreting small, controlled datasets and informing regulatory decisions. However, the growth of high-resolution instrumentation and dense monitoring networks has produced vastly larger and more complex datasets—high-resolution mass spectrometry outputs with thousands of features, multispectral and hyperspectral imagery, continuous sensor streams, and extensive remotely sensed time series—that violate many assumptions of traditional statistics (linearity, independence, normality) and demand more flexible, computationally intensive approaches [1,2]. These shifts in data scale and complexity have driven

a methodological evolution in environmental chemistry, as shown in Figure 1, where the progression from traditional statistics to contemporary machine learning techniques is conceptually illustrated.

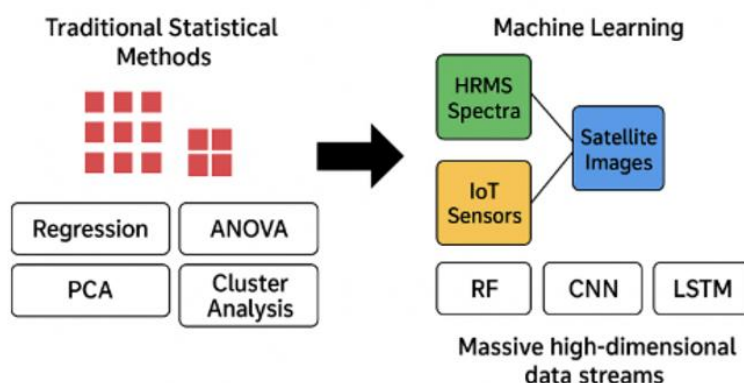


Figure 1: Evolution of data and methods in environmental chemistry

In response, machine learning (ML) has become a powerful complement to statistics in environmental chemistry: supervised methods (random forests, gradient boosting, support vector machines) and deep learning architectures can handle heterogeneous inputs, capture nonlinear interactions, and scale to high dimensionality, enabling tasks from pollutant classification and concentration prediction to automated prioritization of non-target mass spectral features and toxicity forecasting [1–4]. For example, ML has been applied to prioritize hazard-relevant features in non-target high-resolution mass spectrometry (HRMS) datasets, accelerating candidate ranking and reducing the manual burden of spectral annotation [3,5]. Community guidance and interlaboratory initiatives now emphasize quality assurance in non-target screening workflows, reflecting both the promise and the methodological challenges of these tools [5]. Table 1 provides a comparison of classical statistical methods and machine learning approaches.

Table 1. Comparison of classical statistical methods vs. machine learning approaches

Aspect	Classical Statistics	Machine Learning
Typical Dataset Size	Often small-to-moderate	Often moderate-to-large
Key Techniques	Regression, PCA, ANOVA	RF, XGBoost, CNN, GNN
Assumptions	Model-dependent; often linearity, independence, normality	Model-dependent; often fewer parametric distributional assumptions
Outputs	Parameter estimates, p-values	Predictions, feature importance
Main Limitation	Potential limits with highly nonlinear/high-dimensional data	Interpretability, data hunger

The dataset-size descriptions in Table 1 are indicative rather than prescriptive; method suitability depends on model complexity, sampling design, data structure, and validation strategy [1,2,29]. Despite ML's predictive strengths, classical statistics remain essential: experimental design, hypothesis testing, uncertainty quantification, and inferential frameworks are required to draw scientifically valid conclusions and to support regulatory decisions. Best practice therefore couples statistical rigor with ML's flexibility—using cross-validation rooted in statistical theory to avoid inflated performance estimates, and deploying interpretability techniques (feature importance, Shapley values) to translate model outputs into chemically meaningful drivers [1,6,7]. Hybrid approaches that embed mechanistic constraints into ML models (physics-informed networks, Bayesian ML) are gaining traction because they help retain physical plausibility and improve generalizability—for instance, in contaminant transport and groundwater flow modeling where physics-aware ML improves prediction while honoring conservation laws [8,9]. Uncertainty quantification methods are likewise being integrated into ML pipelines (prediction intervals, ensemble UQ) to reduce overconfidence and highlight out-of-distribution risk when transferring models across environmental regimes [10]. Machine learning is reshaping workflows across multiple subfields of environmental chemistry. In non-target screening, ML accelerates feature prioritization, retention time and spectral prediction, and candidate ranking, transforming laborious annotation into tractable pipelines [3,5].

Spectroscopic and chemometric analyses benefit from ML-assisted deconvolution and classification: infrared, Raman, and fluorescence spectra are efficiently mapped to pollutant signatures and concentration estimates, with ML improving robustness to instrumental drift and environmental variability [1,11,12]. Sensor networks exploit transfer learning, drift compensation, and recurrent models to correct for sensor drift, impute missing data, and detect anomalies, thereby improving the quality of continuous air and water monitoring streams and enabling citizen science and real-time management [8,13,14]. Remote sensing applications increasingly rely on deep learning—convolutional neural networks and related architectures—to extract water-quality proxies (turbidity, chlorophyll-a, bloom detection) from multispectral and hyperspectral data, enabling wide-area, near-real-time monitoring that complements in-situ sampling [7,12,15]. Practical examples include ML for harmful algal bloom forecasting using combined remote sensing and in-situ features [12,16].

Environmental metabolomics, exposomics, and toxicology are additional frontiers where ML excels. High-dimensional metabolomic signatures from LC-HRMS are analyzed with ML to classify exposure states, predict toxicological outcomes, and discover biomarkers, while ML-enhanced QSAR and graph-based neural networks improve ecotoxicity prediction and chemical prioritization, helping to reduce animal testing by highlighting high-risk compounds for targeted assays [8,17–20]. Environmental forensics and source attribution also benefit from ML's pattern-recognition capacity, where ensemble classifiers distinguish subtle compositional fingerprints of hydrocarbon mixtures or chemical discharge signatures more effectively than classical PCA alone [3,21]. In spatial mapping tasks—soil heavy-metal distribution or sediment contaminant patterns—ML models (random forests, gradient boosting) frequently outperform geostatistical stationarity-based methods by integrating heterogeneous predictors such as land use, topography, and remote sensing covariates [22,23].

These advantages are balanced by methodological and ethical challenges. Many environmental ML studies still display poor validation practices (insufficient external testing, reliance on internal cross-validation only), incomplete uncertainty reporting, and insufficient documentation of preprocessing and hyperparameter choices, all of which undermine reproducibility and model transferability [1,4,10]. Reporting checklists and community standards—such as the REFORMS recommendations and domain-specific checklists—have been proposed to improve transparency and reproducibility in ML-based science [24].

Explainability is vital for regulatory uptake; tools like SHAP and LIME enable attribution of model predictions to interpretable environmental drivers (e.g., soil pH, organic matter, precipitation), helping to bridge the gap between predictive performance and mechanistic insight [6,7,13,25]. Ethical considerations include geographic bias and equity (models trained on limited regions may underperform elsewhere), and environmental impacts of computation (the carbon footprint of large ML models), which argue for energy-efficient modeling and inclusive stakeholder engagement in model development [21,26].

Practical deployments underscore the need for hybrid solutions: physics-informed ML and Bayesian integration combine predictive power with uncertainty quantification and mechanistic constraints; transfer learning adapts models across sensors and regions with reduced retraining needs; and interlaboratory comparisons and guidance documents aim to harmonize non-target quantification and reporting protocols [5, 9,10,18]. Where time-series forecasting is needed, ML architectures such as LSTMs and temporal convolutional networks capture nonlinear dynamics and long-range dependencies more effectively than linear ARIMA models, improving forecasts of dissolved oxygen, algal bloom onset, and air pollutant episodes [10,24]. Ultimately, the most promising path forward in environmental chemistry is not a wholesale replacement of statistics by ML, but an integrative approach that combines rigorous experimental design, interpretable ML, community standards for reproducibility, and domain knowledge encoded through hybrid models—thereby delivering robust, transparent, and actionable insights for monitoring, management, and policy in a data-rich era.

The genuine contribution of this review is an integrative cross-domain synthesis rather than a new statistical estimator or machine-learning algorithm; it explicitly links statistical inference and uncertainty quantification with ML prediction, interpretability, and mechanistic constraints across environmental-chemistry workflows.

2. Methodology

2.1. Research design and scope

This paper adopts a structured narrative review design to synthesize and critically analyze the application of statistical and machine-learning methodologies in environmental chemistry and analysis. The review focuses on literature published between 2020 and 2025, reflecting the rapid methodological evolution and data-driven transformations in the field. The review is intended as a structured cross-domain synthesis rather than an exhaustive systematic-review protocol.

2.2. Literature search and data sources

Relevant publications were retrieved from major scientific databases, including Scopus, Web of Science, PubMed, DOAJ, CrossRef, Google Scholar, Ei Compendex, INSPEC, EMBASE and CNKI using a combination of targeted keywords: “machine learning in environmental chemistry,” “statistical modeling in pollution analysis,” “AI in water quality,” “multivariate analysis environmental data,” “physics-informed machine learning,” and “hybrid statistical-ML models.”

Inclusion criteria were:

1. Peer-reviewed journal articles, conference papers, or review papers published between 2020–2025;
2. Studies applying statistical and/or ML techniques in environmental chemistry contexts (soil, water, air, toxicology, sensors, or remote sensing);
3. Papers providing explicit methodological descriptions or performance comparisons between traditional and ML approaches.

Exclusion criteria included:

1. Studies focusing purely on computational theory without environmental application;
2. Articles without sufficient methodological transparency;
3. Duplicates and non-English papers.

A total of over 200 papers were initially screened, from which 126 key publications were shortlisted for full review following relevance and methodological depth assessment; 90 of these are cited directly in the final manuscript.

2.3. Data extraction and thematic categorization

Data from selected studies were systematically extracted into a comparative matrix summarizing:

- Type of environmental medium (soil, water, air, biota, or remote sensing);
- Data characteristics (temporal, spatial, high-dimensional, spectral);
- Statistical methods applied (e.g., PCA, ANOVA, regression, PMF);
- Machine-learning techniques employed (e.g., RF, SVM, CNN, LSTM, GNN);
- Validation methods (cross-validation, external test sets, uncertainty quantification);
- Key findings and performance metrics (R^2 , RMSE, accuracy).

Performance metrics were retained as reported by the cited studies; no pooled meta-analysis or cross-study normalization of performance metrics was performed. The extracted information was then organized thematically into major analytical domains:

1. Soil and Sediment Analysis
2. Water Quality and Hydrological Forecasting
3. Air Pollution and Source Apportionment
4. Toxicology and Chemoinformatics (QSAR/ML Models)
5. Remote Sensing and Environmental Monitoring
6. Sensor Calibration and Real-Time Data Analytics

This thematic approach enabled cross-comparison of the suitability, accuracy, and interpretability of statistical versus machine-learning approaches.

2.4. Synthesis and visualization

Findings were synthesized through comparative tables (e.g., Table 1 and Table 2) and conceptual figures (Figures 1, 2 and 3) summarizing the evolution, application domains, and interlinkages between statistical and ML methods. Trends were visualized qualitatively to illustrate the increasing use of ML in predictive analytics and the continuing indispensability of statistics for hypothesis testing and interpretability. Figure 2 is therefore an illustrative trend graphic rather than an exhaustive bibliometric count, and its numerical values are not used for inferential comparisons. Statements that one method outperforms another refer to within-study comparisons reported by the cited authors under their own datasets and validation designs and should not be interpreted as universal superiority across studies.

Table 2. Types of Data and Statistical/ML Methods Commonly Used in Environmental Chemistry

Data Type	Statistical Tool	ML/AI Tool	Example Application
Soil heavy metals	Kriging, regression	Random Forest, XGBoost	Mapping Pb and Cd hotspots
Water quality	ANOVA, PCA	LSTM, CNN	Nutrient dynamics, algal blooms
Air pollutants	ARIMA	RNN, hybrid DL	PM2.5, Ozone forecasting
Toxicity data	QSAR regression	Graph Neural Networks	Endocrine disruptors, PFAS
Remote sensing	Multivariate regression	CNN	Microplastics, algal bloom detection

3. Types of statistics and machine learning used in environmental

CHEMISTRY The landscape of statistical methods and machine learning in environmental chemistry has grown significantly in scope over the last two decades, with the period from 2021 to 2025 witnessing a rapid increase in methodological innovation and practical application [27]. Figure 2 illustrates increasing applications of machine learning across environmental compartments such as soil, water, air, toxicology.

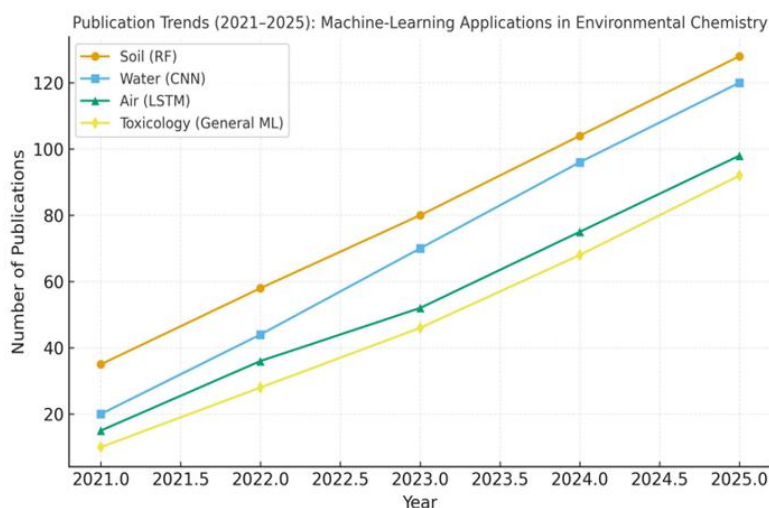


Figure 2. Increasing applications of ML (RF, CNN, LSTM) across soil, water, air, toxicology. (Illustrative trend; not an exhaustive bibliometric count)

Traditional statistics remains foundational, particularly because environmental data often require careful treatment of variability, uncertainty, and experimental design before advanced modeling can be undertaken. Classical approaches include regression models, correlation analyses, variance partitioning, and hypothesis testing, which are essential for identifying significant differences in pollutant concentrations, source contributions, or treatment effects [28]. Regression analysis, both simple and multiple, continues to provide insights into linear relationships between environmental parameters, such as the dependence of groundwater nitrate levels on agricultural practices. Logistic regression, meanwhile, has been widely applied in ecological risk assessments where outcomes are binary, such as the presence or absence of contamination above regulatory thresholds [29].

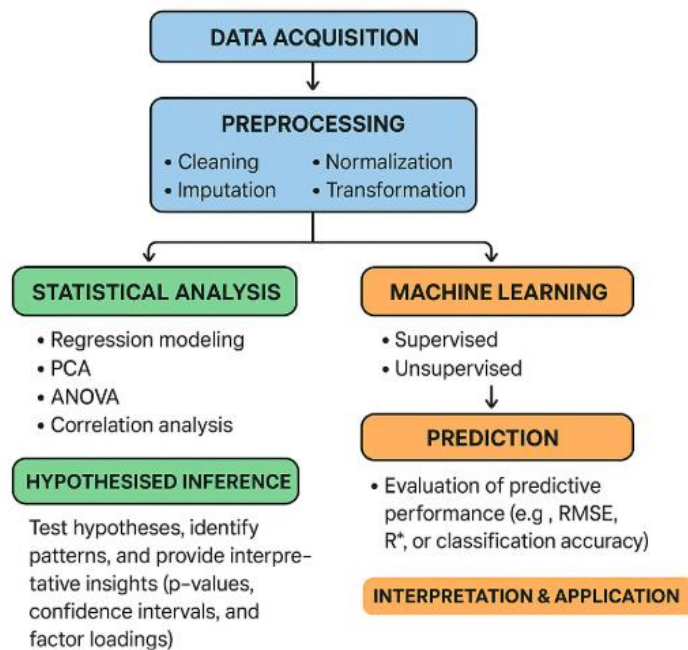


Figure 3. Simplified workflow of statistics vs machine learning in environmental chemistry

Multivariate statistics has played an equally central role, with principal component analysis (PCA), factor analysis, and cluster analysis being indispensable in reducing data dimensionality and highlighting patterns in complex chemical datasets [28]. PCA, for instance, remains a popular method in air quality studies where multiple pollutants interact and covary, enabling researchers to identify latent factors that represent emission sources such as traffic, industrial discharges, or biomass burning. Similarly, discriminant analysis has been widely used to classify contaminated versus uncontaminated sites or to differentiate between pollution sources in river catchments [30]. These statistical approaches, while sometimes limited in predictive power, are highly valued for their interpretability and ability to guide regulatory decision-making by providing clear, evidence-based explanations of data patterns [5].

The rise of machine learning has added a new dimension to environmental chemistry. Supervised learning methods such as random forests (RF), support vector machines (SVM), k-nearest neighbors (kNN), and gradient boosting machines (GBMs) are now commonplace in predictive modeling of pollutant behavior [3]. Random forest in particular has been extensively applied in soil and groundwater studies to predict heavy-metal concentrations, bioavailability, and ecological-risk indices, often outperforming traditional regression models in terms of accuracy [31,32]. SVM has gained traction in air-quality modeling where pollutant dispersion is influenced by complex meteorological and urban factors, offering high classification accuracy in distinguishing high- versus low-pollution episodes [33].

Deep learning approaches have gained momentum more recently, particularly convolutional neural networks (CNNs) and long short-term memory networks (LSTMs). CNNs have been applied to remote-sensing data for detecting environmental contamination and monitoring land-use changes [34,35], while LSTMs have demonstrated high potential in forecasting water-quality indicators such as dissolved oxygen, turbidity, and nutrient concentrations based on temporal sensor data [36,37]. These models excel in capturing nonlinear dependencies and temporal dynamics that cannot be readily modeled by linear statistics. Hybrid deep-learning frameworks, often integrating CNNs with LSTMs, are emerging as powerful tools for handling both spatial and temporal complexities in environmental datasets [9,37].

Unsupervised learning methods also play a critical role, particularly clustering algorithms such as k-means and hierarchical clustering, which have been used to group monitoring sites based on contamination profiles [28]. Self-organizing maps (SOMs) provide visual insights into high-dimensional datasets, supporting exploratory analyses that complement statistical tools such as PCA [5]. Reinforcement learning and Bayesian machine learning, although still relatively underexplored, are beginning to be applied in adaptive pollution-management systems where decisions must be continuously updated based on incoming data streams [38].

Explainability and interpretability are increasingly emphasized as essential components of environmental ML, ensuring that models remain transparent for policy and regulatory use [29,39]. Techniques such as SHAP and LIME allow scientists to attribute model predictions to interpretable environmental drivers—such as soil pH, organic matter, or precipitation—bridging the gap between predictive performance and mechanistic insight [3]. Concurrently, initiatives such as the NORMAN network are pushing for open data, standardized workflows, and reproducible machine-learning pipelines [4,40], reinforcing the reliability of AI-driven environmental assessments.

The diversity of statistical and machine-learning approaches reflects the complexity of environmental problems, with each method offering unique strengths. Statistics remains indispensable for hypothesis-driven analysis and interpretability, while machine learning increasingly dominates in predictive tasks and in dealing with large, noisy, and nonlinear datasets. The emerging trend is toward hybrid and physics-informed frameworks that combine mechanistic modeling with data-driven learning to maximize accuracy and generalizability [9,30]. Ultimately, the integration of these paradigms marks a pivotal evolution in environmental chemistry—transforming the field from descriptive analysis toward predictive and adaptive environmental intelligence [24,27]. Table 2 provides a glimpse into the type of data, statistical and machine learning techniques commonly employed in environmental chemistry.

4. Workflow of statistics and machine learning in environmental chemistry and analysis

The workflow of applying statistical and machine learning methods in environmental chemistry follows a structured sequence that ensures both data quality and analytical rigor. Figure 3 shows a simplified workflow of statistics vs machine learning techniques as applied in environmental chemistry. It typically begins with data acquisition, where raw inputs are obtained from laboratory instrumentation, field sensors, or remote sensing platforms. Analytical techniques such as liquid chromatography–mass spectrometry (LC-MS), atomic absorption spectroscopy (AAS), X-ray fluorescence (XRF), and continuous monitoring sensors generate large volumes of data that are inherently multivariate, noisy, and often incomplete [1,18,41,42]. Preprocessing is therefore the first crucial step, encompassing data cleaning, normalization, imputation of missing values, and transformation to meet model assumptions or to improve performance in machine learning pipelines [42–44].

Once preprocessing is complete, the workflow diverges into two principal branches: statistical analysis and machine learning. In the statistical branch, methods such as regression modeling, PCA, ANOVA, and correlation analysis are applied to test hypotheses, identify significant patterns, and provide interpretative insights. For example, ANOVA might be used to assess whether differences in pollutant concentrations across sites are statistically significant, while PCA could identify source contributions in air pollution datasets. This branch emphasizes hypothesis-driven inquiry and interpretability, with results often communicated in terms of p-values, confidence intervals, and factor loadings [45].

The machine learning branch, by contrast, emphasizes prediction and generalization. Here, data are split into training, validation, and test sets, and algorithms are trained to identify patterns and relationships that maximize predictive performance. Supervised models are evaluated based on metrics such as root mean square error (RMSE), coefficient of determination (R^2), or classification accuracy, while unsupervised models are assessed through clustering validity indices or silhouette scores [46–48]. Cross-validation is a standard practice to assess robustness and generalization, particularly in environmental applications where datasets are often small, spatially or temporally structured, and heterogeneous — and where the choice of validation strategy (random k-fold vs. spatial/probability test splits vs. specialized CV) can materially change perceived model performance and model selection. Practitioners are advised to use spatially-aware validation or tailored test-sample strategies when spatial autocorrelation or extrapolation is relevant [47,48].

Importantly, the workflow is not a dichotomy but rather a continuum where statistics and machine learning can interact and complement one another. For instance, PCA may be used as a preprocessing step to reduce dimensionality before applying a machine learning algorithm such as random forest. Likewise, statistical models may be employed to validate and interpret the outputs of machine learning models, supporting assessment of whether results are accurate and explainable. Recent work demonstrates practical methods for transforming high-performance predictive models (e.g., random forests) into more interpretable analyses that can reveal mechanisms and drivers behind predictions [48].

The final stage of the workflow is interpretation and application. In the statistical branch, outputs typically inform environmental policies or guide management interventions by identifying significant drivers of pollution. In the machine learning branch, outputs often serve predictive purposes, such as forecasting future pollution episodes, predicting contaminant transport, or assessing ecological risks under changing environmental conditions [18,44,49,50]. When combined, the workflow provides a conceptual framework that balances the interpretability of statistics with the predictive power of machine learning, offering a combined toolkit for modern environmental analysis. Recent advances in end-to-end LC-MS/exposomics pipelines illustrate how rigorous preprocessing plus integrated statistical/ML downstream analysis improves reproducibility and enables both mechanistic inference and prediction in environmental chemistry [18,44].

Figure 3 illustrates the parallel yet complementary paths of statistical analysis and machine learning pipelines. This dual-pathway workflow is increasingly being adopted in cutting-edge environmental studies, particularly those addressing complex, multidimensional challenges such as climate change impacts on water quality, combined pollutant exposures, and emerging contaminants like PFAS and microplastics [18,47].

5. Integrating statistics and machine learning in environmental chemistry

The integration of statistics and machine learning in environmental chemistry has emerged as one of the most promising strategies for overcoming the limitations of each approach when applied in isolation [51]. Traditional statistical methods are valued for their transparency, interpretability, and theoretical grounding in probability distributions and inference, making them highly suitable for regulatory purposes and hypothesis-driven research. However, many commonly used parametric models rely on assumptions such as linearity, independence, and homoscedasticity, which may fail in the face of complex environmental systems characterized by non-linear interactions, multicollinearity, and high levels of uncertainty [52]. Machine learning, by contrast, can handle non-linear, high-dimensional datasets and may deliver superior predictive accuracy in appropriately validated applications [53]. Its limitation, however, lies in the so-called “black-box” nature of many algorithms, which makes interpretability difficult and hinders acceptance in regulatory frameworks where transparency is critical [54].

Hybrid approaches that combine the interpretability of statistics with the predictive power of machine learning are increasingly being adopted in environmental chemistry. One common integration strategy involves using statistical methods as a preprocessing or dimensionality-reduction step prior to machine learning [51,53,55,56]. For example, principal component analysis (PCA) can be applied to reduce the dimensionality of spectroscopic data, which is then used as input features for machine-learning models such as random forests or neural networks. This not only improves computational efficiency but also helps to highlight the most important variables driving the predictive outcomes [51,53].

Another integration approach involves embedding machine learning into statistical frameworks, creating hybrid models intended to retain interpretability while improving predictive performance. Generalized additive models (GAMs) and mixed-effects models, for instance, can be enhanced with machine-learning components that capture non-linearities or spatiotemporal dependencies [57,58]. Bayesian hierarchical models, long used in environmental risk assessment, are also being expanded to incorporate machine-learning priors or data-driven components, enabling them to better adapt to large and heterogeneous datasets [58].

The integration also extends into the validation and interpretation stage. Machine-learning models can be evaluated through statistical tests that assess whether their predictions align with theoretical expectations and observed patterns. Conversely, statistical analyses can help to explain the features identified by machine learning as most predictive, thus bridging the gap between performance and understanding [59–61]. This integration has proven particularly valuable in fields such as water-quality modeling, where machine-learning models may achieve high accuracy in forecasting parameters such as dissolved oxygen, but require statistical post-analysis to interpret which environmental variables—temperature, nutrient load, or hydrological conditions—are most influential [61].

This synergy is not merely methodological but conceptual, reflecting a broader epistemological shift in environmental chemistry. By integrating statistics and machine learning, researchers are able to harness both hypothesis-driven inquiry and data-driven discovery, enabling deeper insights into pollutant dynamics, ecological risks, and long-term environmental change [27,62,63]. Such hybrid models are likely to form the

backbone of environmental analysis in the coming decade, particularly as regulatory frameworks increasingly demand methods that are both robust and explainable [63].

6. Practical applications of statistics and machine learning in environmental chemistry and analysis

6.1. Soil analysis

In soil contamination studies, statistical methods such as ANOVA, regression, PCA and cluster analysis continue to underpin interpretation of land-use, agricultural and industrial influences on heavy-metal accumulation. More recently, machine-learning models have been brought into play. Traditional statistical tools (regression, ANOVA, PCA, and geostatistics such as kriging) remain widely used to characterize soil contamination, identify significant predictors (e.g., pH, organic matter), and provide interpretable summaries for regulators and land managers [64]. Over the last few years, machine-learning (ML) models — particularly tree ensembles such as Random Forest (RF) and gradient boosting — have been applied to predict spatial distributions of heavy metals (Cd, Pb, As, etc.), often incorporating heterogeneous predictors (remote-sensing indices, land-use, precipitation, population density), and typically outperform classical regression/geostatistical baselines in predictive accuracy while allowing variable-importance interpretation [23,64]. Hybrid workflows that apply PCA or other dimensionality reduction as preprocessing before ML (to retain interpretability while reducing computational cost) are now common in large soil datasets [23]. Moradpour et al. utilized random forest and gradient boosting algorithms to model and map heavy metal concentrations in contaminated soils, revealing superior predictive accuracy compared to traditional kriging by integrating land-use, soil-type, and topographic data [31].

Several recent studies have shown that tree-based machine-learning methods (e.g., Random Forest and gradient-boosting variants), alone or in hybrid schemes with geostatistics, often outperform classical spatial interpolation for mapping soil heavy metals — improving predictive precision and spatial detail when remote-sensing, topographic and land-use covariates are included [23,33,64,65]. Nie et al. implemented a transfer-learning and ensemble-modeling approach to predict soil heavy-metal distributions across geographical domains, effectively reducing data demands and enhancing generalization in data-poor regions. In the work, they used a Random Forest model to map soil heavy-metal distributions in a coastal city in eastern China, showing that environmental variables explained 51-63 % of variability for several metals [23]. Such ML models often outperform traditional interpolation or regression methods and provide variable-importance metrics that aid interpretability. In their work, Salaudeen et al, successfully applied multivariate statistical tools, such as Partial Least Squares Discriminant Analysis (PLS-DA) score and loading plots, Variable Importance in Projection (VIP) scores, Significance Analysis of Microarrays (SAM), and heat mapping to successfully discriminate soil samples from four distinct sites on Penang Mainland in Malaysia. PLS-DA analysis showed distinct metabolomic differences between soil samples from different sites, with certain groups showing more similarities than others [66].

6.2. Water analysis

Time-series statistical methods such as ARIMA historically provided baseline forecasts for water quality variables, but recurrent neural networks — especially LSTM and hybrid CNN-LSTM/ConvLSTM architectures — now deliver superior performance for nonlinear, seasonal, and spatially coupled data streams [67]. These deep sequence models have been applied to forecast dissolved oxygen, turbidity, and algal-bloom risk using combined in-situ sensor streams and satellite/contextual inputs; importantly, interpretable post-hoc tools (e.g., SHAP) are used to link ML drivers like temperature and nutrient inputs, preserving interpretability for management decisions [12,67]. In their work, Gao et al. used Long Short-Term Memory (LSTM) networks to multivariate water-quality time series for forecasting key parameters (including dissolved oxygen and turbidity), showing superior performance over traditional ARIMA/regression approaches for nonlinear and seasonally varying dynamics [36]. Pant et al. (2024) developed a hybrid CEEMDAN-AdaBoost-BiLSTM-LSTM model for multi-step dissolved oxygen forecasting in the River Ganga, capturing complex nonlinearities and seasonal dynamics with high accuracy [68]. Also, Kim et al. integrated partial least squares (PLS) with advanced regressors such as LightGBM and support

vector regression to build a real-time chlorophyll-a forecasting system from hyperspectral and in-situ features, achieving strong performance for bloom monitoring [69]. Izadi et al. employed a machine-learning and remote-sensing hybrid model to predict the onset of harmful algal blooms, showing that combining spectral indices and satellite-derived variables substantially enhances early detection capabilities [12].

6.3. Fish

Petrea et al. compiled literature data on 11 metals measured in turbot (*Psetta maxima*) muscle and liver and compared multiple linear regression (MLR) with non-linear Random Forest (RF) models. They used stepwise MLR and RF to predict tissue concentrations from available covariates and reported that RF models achieved >70% prediction accuracy for numerous metals (As, Cd, Cu, K, Mg, Zn in muscle; As, Ca, Cd, Mg, Fe in liver). The authors concluded that tree-based ML can reliably predict heavy-metal burdens in fish tissues and complement classical regression for monitoring and food-safety screening [70]. Bertato et al. developed and validated QSAR regression and classification models (including artificial neural networks) to predict log-BCF using an updated, large curated dataset of chemicals. Their ANN and optimized linear models showed good internal and external performance ($R^2 \sim 0.62$ – 0.70 for regression models; high classification accuracies for $BCF >$ regulatory thresholds). The study emphasized OECD-style validation and applicability domains for regulatory relevance [71]. Simionov and colleagues sampled fish and water across the lower Danube, Danube Delta, and Black Sea and combined biochemical oxidative- stress biomarkers (CAT, SOD, GPx, MDA) with measured metal concentrations. They implemented Random Forest models to predict Cd, Pb, Zn, Fe, and Cu concentrations in muscle and liver from biomarker data and environmental covariates. RF models identified MDA (malondialdehyde) and GPx as top predictors and produced high predictive accuracy across regions; the approach was proposed as a cost- effective screening/monitoring tool [72]. Lepak et al. used Random Forests to analyze predictors of mercury (Hg) concentrations in sport fish across 32 Colorado reservoirs and three fish species. They combined fish Hg concentration records with watershed, reservoir, management (e.g., fish stocking), productivity and food-web variables. RF identified different dominant predictors by species (e.g., salmonid stocking metrics for northern pike; productivity and forage for smallmouth bass and walleye) and explained substantial variance (e.g., ~55% for some species models). The authors proposed RF as a practical tool to prioritize monitoring and predict changes in fish Hg under management or environmental change scenarios [73].

6.4. Air analysis

Air quality and source apportionment. Statistical receptor models (e.g., PCA, positive matrix factorization) continue to underpin source-apportionment studies and policy reporting, but ML improves temporal forecasting and augments source-apportionment by modeling meteorological interactions and nonlinearity. Hybrid approaches that combine PMF with supervised ML (or use ML to post-process PMF outputs) reveal synergistic effects of sources and weather on $PM_{2.5}$ concentration patterns and improve short-term predictions used for warnings and mitigation planning [74]. In the work of Lin et al. a hybrid RF-XGBoost framework using MAIAC aerosol optical depth, meteorological, and land-use variables was applied to estimate $PM_{2.5}$ concentrations across Chinese urban regions, showing improved accuracy and spatial generalization compared to regression-based models [75].

deSouza et al. established calibration and transfer-learning protocols for dense networks of low-cost $PM_{2.5}$ sensors, demonstrating that nonlinear and ensemble machine-learning corrections substantially improve measurement reliability under variable ambient conditions [76]. Zhang et al. (2022) combined receptor modeling (PMF) with machine-learning classifiers to enhance $PM_{2.5}$ source apportionment and temporal prediction, thereby linking emission factors to predictive patterns in ambient data [74].

6.5. Plant analysis

In plant science applications (crop stress, nutrient status, disease detection), chemometrics and classical multivariate regression remain important for interpretable calibration from spectral indices, while ML (convolutional and other deep networks) enables high-accuracy mapping from hyperspectral or multispectral imagery and yields robust classification under varying illumination and phenology. Recent hyperspectral

and multispectral studies show ML substantially improves detection of plant stress and contaminant uptake proxies compared with simple regression models [77].

6.6. Remote sensing and environmental monitoring

Liu et al. leveraged convolutional neural networks (CNNs) to detect *Noctiluca scintillans* harmful algal blooms from multispectral satellite imagery, demonstrating that deep-learning architectures can robustly separate bloom signals from background noise [78]. Kim et al. explored hyperspectral and machine-learning integration for near-real-time water-quality estimation, emphasizing the value of dimensionality reduction (PLS) as a statistical pre-step for optimizing model interpretability [69].

6.7. Ecotoxicology and chemical modeling

Chen et al. implemented a graph-convolutional neural network (GCN) for chemical-toxicity prediction, revealing that molecular graph embeddings yield more reliable toxicity classification than descriptor-based QSAR models [79]. Ketkar et al. benchmarked multiple graph-based architectures for acute-toxicity prediction, finding that attention-based and message-passing neural networks deliver higher accuracy and interpretability than conventional statistical QSAR models [80].

6.8. Ecotoxicology and predictive toxicology

Statistical dose–response modeling (LC50, NOEC) still forms the backbone of toxicological assessment. However, modern ML — especially graph-based neural networks and other deep models trained on molecular descriptors — has substantially advanced in-silico toxicity prediction (QSAR/graph-based models), enabling prioritization of thousands of untested compounds and reducing animal testing needs. Benchmarking studies show graph models often outperform traditional QSAR regressions for many ecotoxicological endpoints, although careful external validation is essential to avoid overfitting [79,80].

6.9. Sensors and real-time monitoring

Networks of low-cost sensors (air, water, IoT) generate dense data streams but suffer from drift, bias, and environmental interferences; ML methods (transfer learning, calibration models, recurrent models for temporal drift correction) have been widely applied to calibrate, correct drift, impute missing data, and detect anomalies in real time [76,81]. Work on model transferability (global calibration models and domain-adaptation techniques) shows ML can reduce per-unit calibration burdens, enabling larger citizen-science and urban monitoring deployments while maintaining acceptable accuracy [76,82].

6.10. Remote sensing and large-scale mapping

Remote sensing combined with ML (CNNs, Transformer variants) has enabled high-resolution mapping of proxies such as chlorophyll-a, turbidity, and bloom extent from multispectral/hyperspectral imagery, often outperforming regression or index-based approaches in heterogeneous environments [12]. These ML-driven remote sensing products are increasingly used to drive near-real-time management (e.g., HAB alerts), with statistical validation against in-situ observations used to quantify and report uncertainty.

6.11. Integration, interpretability and best practice

Across domains, the current best practice couples statistical design and uncertainty quantification with ML's predictive power. That means using statistical methods for experimental design, baseline hypothesis testing, and uncertainty estimation, while applying ML for pattern detection and forecasting; explaining ML outputs via SHAP/LIME or similar tools is critical for regulatory acceptance and domain insight [67,74]. Reproducibility requires rigorous external validation, open code/data, and clear reporting of preprocessing and hyperparameters — practices increasingly advocated in the literature [76,80]. Challenges and outlook. Persistent challenges include overfitting in very high-dimensional datasets (e.g., non-target HRMS and metabolomics), model transferability across regions and sensor types, interpretability in regulatory contexts, and the environmental cost of large model training. Hybrid strategies — embedding mechanistic constraints (physics-informed ML, Bayesian priors), using dimensionality reduction for interpretability, and

enforcing FAIR data and code sharing — are promising pathways forward for robust, transparent, and usable environmental-chemistry ML applications.

7. Future perspective

The future of environmental chemistry lies at the intersection of robust statistical reasoning, powerful machine-learning algorithms, and interdisciplinary integration. The reviewed literature indicates that environmental systems are often too complex to be fully captured by any single method, and promising advances are likely to come from hybrid approaches that combine mechanistic interpretability with predictive power. One important direction is the formal integration of machine learning with mechanistic environmental models. Rather than treating machine learning as a black box, researchers are beginning to use it to emulate sub-components of process-based models, such as chemical reaction kinetics, pollutant dispersion, or sorption processes [83]. This strategy can enable faster simulations without discarding physical insight, potentially reducing computational demands while retaining mechanistic constraints. A second promising area is the integration of uncertainty quantification into machine learning. Environmental policy and risk assessment rely on transparent estimates of uncertainty, and statistical approaches such as Bayesian inference can be embedded within machine learning frameworks to provide probability distributions rather than single-point predictions. Work cited in this review shows that uncertainty quantification can be integrated into environmental ML frameworks [10]. The growing availability of high-resolution environmental data from satellites, drones, in-situ sensors, and advanced analytical chemistry platforms creates an unprecedented opportunity for scaling research. However, these data streams are vast, noisy, and heterogeneous. Future work must focus on developing machine learning models that are not only accurate but also resilient to missing data, data drift, and sensor biases. Transfer learning and domain adaptation, already demonstrated in remote-sensing and spatial environmental studies [76,82] will likely become standard practices for leveraging global data to inform local predictions, especially in data-scarce regions.

Another frontier lies in the integration of multi-omics with environmental chemistry. With advances in metabolomics, proteomics, and metagenomics, researchers can now explore the biological consequences of environmental contamination with unprecedented detail. Machine learning is already being used to integrate these high-dimensional omics datasets with traditional chemical measurements, revealing hidden pathways of toxicity and resilience [29]. The future will likely see greater emphasis on systems-level models that couple chemical data with biological and ecological indicators, allowing for holistic assessments of ecosystem health.

Ethics and sustainability will also play increasingly central roles. As environmental machine learning relies on large-scale computation, the carbon footprint of training and deploying deep models must be considered. Researchers are beginning to explore energy-efficient algorithms and green artificial intelligence, ensuring that the very tools designed to address environmental challenges do not inadvertently worsen them [84]. Equally critical is ensuring equity in data representation. Many regions most affected by pollution and climate change remain under-represented in global environmental datasets. Future work must prioritise inclusive data collection and model development that serves diverse communities, ensuring that scientific advances are not confined to well-resourced regions but benefit vulnerable populations worldwide [85]. A further perspective is the move toward regulatory acceptance of machine learning in environmental decision-making. Currently, most machine-learning studies remain at the research stage, with limited direct translation into policy. This reflects both interpretability challenges and a lack of standardised protocols. In the coming decade, further guidelines are likely to emerge for the validation and reporting of environmental machine-learning models, and domain-specific guidance is already emerging for water and environmental modeling [86]. Such guidelines could accelerate the adoption of machine learning in regulatory contexts, helping ensure that predictive tools are only scientifically sound but also legally defensible.

Finally, education and training will be critical. Future environmental chemists will require fluency not only in analytical chemistry and statistics but also in machine learning, programming, and data ethics. Graduate programmes and professional training will need to adapt to this reality, fostering a new generation of scientists capable of navigating both the laboratory bench and the computational workspace. This interdisciplinary training will ensure that the advances of machine learning are harnessed responsibly and effectively for solving pressing environmental challenges. Taken together, these perspectives indicate that the next phase of research will move beyond methodological comparison toward integration, sustainability, and

application. Statistics and machine learning will not remain separate cultures but will converge into a unified toolkit for environmental chemistry, one that is predictive, transparent, ethical, and globally relevant.

8. Conclusion

The application of statistics and machine learning to environmental chemistry and analysis has advanced rapidly over the past decade, with the period 2021–2025 marking a decisive acceleration. Statistical tools remain essential for interpretability, mechanistic understanding, and uncertainty quantification, while machine learning excels at capturing nonlinear relationships, high-dimensional datasets, and predictive tasks across air, water, and soil systems. The reviewed studies indicate that, for many complex environmental applications, neither approach alone is universally sufficient. A strong direction is therefore the use of hybrid models that combine the interpretability of statistics with the predictive power of machine learning. The challenges ahead are clear: overfitting, reproducibility, interpretability, ethical responsibility, and regulatory acceptance. Yet the opportunities are equally compelling: real-time environmental monitoring, global transferability of models, integration of multi-omics, and sustainable artificial intelligence. As the field matures, the most successful applications will be those that bridge the gap between prediction and explanation, ensuring that computational advances translate into meaningful insights and actionable policies.

In conclusion, statistics and machine learning should be viewed not as competing paradigms but as complementary partners in the advancement of environmental chemistry. By embracing both traditions, the scientific community can move toward a future in which environmental data are not only collected but also understood, predicted, and applied to safeguard ecosystems and human health.

Funding Information: The authors received no financial support for the research, authorship, and/or publication of this article.

Conflicts of Interest: The authors declare that there are no conflicts of interest or competing interests regarding the publication of this paper.

Data Availability: This article is a review paper and does not contain new experimental data. All data discussed in this study are derived from previously published sources, which are appropriately cited within the manuscript.

This review article does not involve original computational modelling or custom-developed code. Therefore, code availability is not applicable.

Author Contributions: Abdulwasii Olawale Salaudeen conceptualized the review, designed the structure and scope, conducted the comprehensive literature search, and led the drafting and critical revision of the manuscript and approved the manuscript for submission. Yemisi Ajoke Olawore contributed to literature interpretation, organization of thematic sections, and manuscript editing. Hajara Yakubu assisted in literature compilation, synthesis of key findings, and table preparation. Hajara Yakubu contributed to critical reading, citation management, and refinement of discussion sections. Habib Abdulkadiri supported the integration of statistical and machine learning perspectives and contributed to manuscript review. Akoji Godwin Mathias provided intellectual input on environmental chemistry aspects and critically reviewed the final version for accuracy and coherence. All authors read and approved the final version of the manuscript and agree to be accountable for all aspects of the work

References

- [1] Zhu, J.-J., Yang, M., & Ren, Z. J. (2023). Machine learning in environmental research: Common pitfalls and best practices. *Environmental Science & Technology*, 57(46), 17671–17689.
- [2] 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. *Environmental Science & Technology*, 55(19), 12741–12754.
- [3] Arturi, K., & Hollender, J. (2023). Machine learning-based hazard-driven prioritization of features in nontarget screening of environmental high-resolution mass spectrometry data. *Environmental Science & Technology*, 57(46), 18067–18079.
- [4] Renner, G., & Reuschenbach, M. (2023). Critical review on data processing algorithms in non-target screening: Challenges and opportunities to improve result comparability. *Analytical and Bioanalytical Chemistry*, 415(18), 4111–4123.
- [5] Hollender, J., Schymanski, E. L., Ahrens, L., Alygizakis, N., Béen, F., Bijlsma, L., Brunner, A. M., Celma, A., Fildier, A., Fu, Q., Gago-Ferrero, P., Gil-Solsona, R., Haglund, P., Hansen, M., Kaserzon, S., Krueve, A., Lamoree, M., Margoum, C., Meijer, J., ... Krauss, M. (2023). NORMAN guidance on suspect and non-target screening in environmental monitoring. *Environmental Sciences Europe*, 35, Article 75.

- [6] Arashpour, M. (2023). AI explainability framework for environmental management research. *Journal of Environmental Management*, 342, Article 118149.
- [7] Zhi, W., Appling, A. P., Golden, H. E., Podgorski, J., & Li, L. (2024). Deep learning for water quality. *Nature Water*, 2(3), 228–241.
- [8] Bush, T., Papaioannou, N., Leach, F., Pope, F. D., Singh, A., Thomas, G. N., Stacey, B., & Bartington, S. (2022). Machine learning techniques to improve the field performance of low-cost air quality sensors. *Atmospheric Measurement Techniques*, 15(10), 3261–3278.
- [9] Secci, D., Godoy, V. A., & Gómez-Hernández, J. J. (2024). Physics-informed neural networks for solving transient unconfined groundwater flow. *Computers & Geosciences*, 182, Article 105494.
- [10] Liu, S., Lu, D., Painter, S. L., Griffiths, N. A., & Pierce, E. M. (2023). Uncertainty quantification of machine learning models to improve streamflow prediction under changing climate and environmental conditions. *Frontiers in Water*, 5, Article 1150126.
- [11] Srivastava, S., Wang, W., Zhou, W., Jin, M., & Vikesland, P. J. (2024). Machine learning-assisted surface-enhanced Raman spectroscopy detection for environmental applications: A review. *Environmental Science & Technology*, 58(47), 20830–20848.
- [12] Izadi, M., Sultan, M., El Kadiri, R., Ghannadi, A., & Abdelmohsen, K. (2021). A remote sensing and machine learning-based approach to forecast the onset of harmful algal bloom. *Remote Sensing*, 13(19), Article 3863.
- [13] Zheng, H., Liu, Y., Wan, W., Zhao, J., & Xie, G. (2023). Large-scale prediction of stream water quality using an interpretable deep learning approach. *Journal of Environmental Management*, 331, Article 117309.
- [14] Nalakurthi, N. V. S. R., Abimbola, I., Ahmed, T., Anton, I., Riaz, K., Ibrahim, Q., Banerjee, A., Tiwari, A., & Gharbia, S. (2024). Challenges and opportunities in calibrating low-cost environmental sensors. *Sensors*, 24(11), Article 3650.
- [15] Liu, B., & Li, T. (2024). A machine-learning-based framework for retrieving water quality parameters in urban rivers using UAV hyperspectral images. *Remote Sensing*, 16(5), Article 905.
- [16] Joshi, N., Park, J., Zhao, K., Londo, A., & Khanal, S. (2024). Monitoring harmful algal blooms and water quality using Sentinel-3 OLCI satellite imagery with machine learning. *Remote Sensing*, 16(13), Article 2444.
- [17] Salaudeen, A. O., & Lim, V. (2026). Metabolomics approach in environmental studies: Methodologies, application and challenges. *Critical Reviews in Analytical Chemistry*, 56(5), 1331–1347.
- [18] Pang, M., Du, E., & Zheng, C. (2024). Contaminant transport modeling and source attribution with attention-based graph neural network. *Water Resources Research*, 60(6), Article e2023WR035278.
- [19] Ferreira, C. R., de Lima Gomes, P. C. F., Robison, K. M., Cooper, B. R., & Shannahan, J. H. (2024). Implementation of multiomic mass spectrometry approaches for the evaluation of human health following environmental exposure. *Molecular Omics*, 20(5), 296–321.
- [20] Luan, H. (2022). Machine learning for screening active metabolites with metabolomics in environmental science. *Environmental Science: Advances*, 1(5), 605–611.
- [21] Malm, L., Liigand, J., Aalizadeh, R., Alygizakis, N., Ng, K., Frøkjær, E. E., Nanusha, M. Y., Hansen, M., Plassmann, M., Bieber, S., Letzel, T., Balest, L., Abis, P. P., Mazzetti, M., Kasprzyk-Hordern, B., Ceolotto, N., Kumari, S., Hann, S., Kochmann, S., ... Kruve, A. (2024). Quantification approaches in non-target LC/ESI/HRMS analysis: An interlaboratory comparison. *Analytical Chemistry*, 96(41), 16215–16226.
- [22] Stanciu, A. R., Gillespie, C., & Britz-McKibbin, P. (2025). Environmental exposures and health risks: A metabolomics perspective on exposomics research. *Annual Review of Analytical Chemistry*, 18, 47–71.
- [23] Nie, S., Chen, H., Sun, X., & An, Y. (2024). Spatial distribution prediction of soil heavy metals based on random forest model. *Sustainability*, 16(11), Article 4358.
- [24] Kapoor, S., Cantrell, E. M., Peng, K., Pham, T. H., Bail, C. A., Gundersen, O. E., Hofman, J. M., Hullman, J., Lones, M. A., Malik, M. M., Nanayakkara, P., Poldrack, R. A., Raji, I. D., Roberts, M., Salganik, M. J., Serra-Garcia, M., Stewart, B. M., Vandewiele, G., & Narayanan, A. (2024). REFORMS: Consensus-based recommendations for machine-learning-based science. *Science Advances*, 10(18), Article eadk3452.
- [25] Zhang, J. (2024). Comparative analysis of water applicability predictions explained by the LightGBM model using SHAP and LIME. *Applied and Computational Engineering*, 104, 151–159.
- [26] McGovern, A., Ebert-Uphoff, I., Gagne, D. J., II, & Bostrom, A. (2022). Why we need to focus on developing ethical, responsible, and trustworthy artificial intelligence approaches for environmental science. *Environmental Data Science*, 1, Article e6.
- [27] Hsieh, W. W. (2022). Evolution of machine learning in environmental science—A perspective. *Environmental Data Science*, 1, Article e3.
- [28] Greenacre, M., Groenen, P. J. F., Hastie, T., Iodice D’Enza, A., Markos, A., & Tuzhilina, E. (2022). Principal component analysis. *Nature Reviews Methods Primers*, 2, Article 100.

- [29] Pichler, M., & Hartig, F. (2023). Machine learning and deep learning—A review for ecologists. *Methods in Ecology and Evolution*, 14(4), 994–1016.
- [30] von Borries, K., Holmquist, H., Kosnik, M., Beckwith, K. V., Jolliet, O., Goodman, J. M., & Fantke, P. (2023). Potential for machine learning to address data gaps in human toxicity and ecotoxicity characterization. *Environmental Science & Technology*, 57(46), 18259–18270.
- [31] Moradpour, S., Entezari, M., Ayoubi, S., Karimi, A., & Naimi, S. (2023). Digital exploration of selected heavy metals using random forest and a set of environmental covariates at the watershed scale. *Journal of Hazardous Materials*, 455, Article 131609.
- [32] Palansooriya, K. N., Li, J., Dissanayake, P. D., Suvarna, M., Li, L., Yuan, X., Sarkar, B., Tsang, D. C. W., Rinklebe, J., Wang, X., & Ok, Y. S. (2022). Prediction of soil heavy metal immobilization by biochar using machine learning. *Environmental Science & Technology*, 56(7), 4187–4198.
- [33] Guo, H.-N., Liu, H.-T., & Wu, S. (2022). Simulation, prediction and optimization of typical heavy metals immobilization in swine manure composting by using machine learning models and genetic algorithm. *Journal of Environmental Management*, 323, Article 116266.
- [34] Camps-Valls, G., Tuia, D., Zhu, X. X., & Reichstein, M. (Eds.). (2021). *Deep learning for the Earth sciences: A comprehensive approach to remote sensing, climate science, and geosciences*. Wiley.
- [35] Thapa, A., Horanont, T., Neupane, B., & Aryal, J. (2023). Deep learning for remote sensing image scene classification: A review and meta-analysis. *Remote Sensing*, 15(19), Article 4804.
- [36] Gao, Z., Chen, J., Wang, G., Ren, S., Fang, L., Yinglan, A., & Wang, Q. (2023). A novel multivariate time series prediction of crucial water quality parameters with long short-term memory (LSTM) networks. *Journal of Contaminant Hydrology*, 259, Article 104262.
- [37] Mahesh, N., Babu, J. J., Nithya, K., & Arunmozhi, S. A. (2024). Water quality prediction using LSTM with combined normalizer for efficient water management. *Desalination and Water Treatment*, 317, Article 100183.
- [38] Zhou, Z.-H. (2022). Open-environment machine learning. *National Science Review*, 9(8), Article nwac123.
- [39] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 30, pp. 4765–4774). Curran Associates.
- [40] Paszkiewicz, M., Godlewska, K., Lis, H., Caban, M., Białk-Bielińska, A., & Stepnowski, P. (2022). Advances in suspect screening and non-target analysis of polar emerging contaminants in the environmental monitoring. *TrAC Trends in Analytical Chemistry*, 154, Article 116671.
- [41] Brown, L. E., Maavara, T., Zhang, J., Chen, X., Klaar, M., Moshe, F. O., Ben-Zur, E., Stein, S., Grayson, R., Carter, L., Levintal, E., Gal, G., Ziv, P., Tarkowski, F., Pathak, D., Khamis, K., Barquín, J., Philamore, H., Gradilla-Hernández, M. S., & Arnon, S. (2025). Integrating sensor data and machine learning to advance the science and management of river carbon emissions. *Critical Reviews in Environmental Science and Technology*, 55(9), 600–623.
- [42] Delabriere, A., Warmer, P., Brennstainer, V., & Zamboni, N. (2021). SLAW: A scalable and self-optimizing processing workflow for untargeted LC-MS. *Analytical Chemistry*, 93(45), 15024–15032.
- [43] Pang, Z., Xu, L., Viau, C., Lu, Y., Salavati, R., Basu, N., & Xia, J. (2024). MetaboAnalystR 4.0: A unified LC-MS workflow for global metabolomics. *Nature Communications*, 15, Article 3675.
- [44] Li, S., Siddiqua, A., Thapa, M., Chi, Y., & Zheng, S. (2023). Trackable and scalable LC-MS metabolomics data processing using asari. *Nature Communications*, 14, Article 4113.
- [45] Acal, C., Aguilera, A. M., Sarra, A., Evangelista, A., Di Battista, T., & Palermi, S. (2022). Functional ANOVA approaches for detecting changes in air pollution during the COVID-19 pandemic. *Stochastic Environmental Research and Risk Assessment*, 36(4), 1083–1101.
- [46] Kisi, O., Heddam, S., Parmar, K. S., Petroselli, A., Külls, C., & Zounemat-Kermani, M. (2025). Integration of Gaussian process regression and K means clustering for enhanced short term rainfall runoff modeling. *Scientific Reports*, 15, Article 7444.
- [47] Milà, C., Ludwig, M., Pebesma, E., Tonne, C., & Meyer, H. (2024). Random forests with spatial proxies for environmental modelling: Opportunities and pitfalls. *Geoscientific Model Development*, 17(15), 6007–6033.
- [48] Simon, S. M., Glaum, P., & Valdovinos, F. S. (2023). Interpreting random forest analysis of ecological models to move from prediction to explanation. *Scientific Reports*, 13, Article 3881.
- [49] Agbehadj, I. E., & Obagbuwa, I. C. (2025a). Explainable artificial intelligence and machine learning for air pollution risk assessment and respiratory health outcomes: A systematic review. *Atmosphere*, 16(10), Article 1154.
- [50] Binetti, M. S., Massarelli, C., & Uricchio, V. F. (2024). Machine learning in geosciences: A review of complex environmental monitoring applications. *Machine Learning and Knowledge Extraction*, 6(2), 1263–1280.

- [51] Lokman, A., Ismail, W. Z. W., & Aziz, N. A. A. (2025). A review of water quality forecasting and classification using machine learning models and statistical analysis. *Water*, 17(15), Article 2243.
- [52] Madni, H. A., Umer, M., Ishaq, A., Abuzinadah, N., Saidani, O., Alsubai, S., Hamdi, M., & Ashraf, I. (2023). Water-quality prediction based on H₂O AutoML and explainable AI techniques. *Water*, 15(3), Article 475.
- [53] Osman, A. I., Zhang, Y., Lai, Z. Y., Rashwan, A. K., Farghali, M., Ahmed, A. A., Liu, Y., Fang, B., Chen, Z., Al-Fatesh, A., Rooney, D. W., Yiin, C. L., & Yap, P.-S. (2023). Machine learning and computational chemistry to improve biochar fertilizers: A review. *Environmental Chemistry Letters*, 21(6), 3159–3244.
- [54] Elsayed, A., Levison, J., Binns, A., Larocque, M., & Goel, P. (2025). Regression-based machine learning models for nitrate and chloride prediction in surface water in a small agricultural sand plain sub-watershed in Southwestern Ontario, Canada. *Frontiers in Environmental Science*, 13, Article 1543852.
- [55] Kalian, A. D., Benfenati, E., Osborne, O. J., Gott, D., Potter, C., Dorne, J.-L. C. M., Guo, M., & Hogstrand, C. (2023). Exploring dimensionality reduction techniques for deep learning driven QSAR models of mutagenicity. *Toxics*, 11(7), Article 572.
- [56] Banerjee, A., & Roy, K. (2024). ARKA: A framework of dimensionality reduction for machine-learning classification modeling, risk assessment, and data gap-filling of sparse environmental toxicity data. *Environmental Science: Processes & Impacts*, 26(6), 991–1007.
- [57] Agbehadj, I. E., & Obagbuwa, I. C. (2025b). A hybrid long short-term memory with generalized additive model and post-hoc explainable artificial intelligence with causal inference for air pollutants prediction in Kimberley, South Africa. *Frontiers in Artificial Intelligence*, 8, Article 1620019.
- [58] Abu El-Magd, S. A., Maged, A., & Farhat, H. I. (2022). Hybrid-based Bayesian algorithm and hydrologic indices for flash flood vulnerability assessment in coastal regions: Machine learning, risk prediction, and environmental impact. *Environmental Science and Pollution Research*, 29(38), 57345–57356.
- [59] Le, X.-H., Choi, C., Eu, S., Yeon, M., & Lee, G. (2024). Quantitative evaluation of uncertainty and interpretability in machine learning-based landslide susceptibility mapping through feature selection and explainable AI. *Frontiers in Environmental Science*, 12, Article 1424988.
- [60] Weller, D. L., Love, T. M. T., & Wiedmann, M. (2021). Interpretability versus accuracy: A comparison of machine learning models built using different algorithms, performance measures, and features to predict *E. coli* levels in agricultural water. *Frontiers in Artificial Intelligence*, 4, Article 628441.
- [61] Zaresefat, M., & Derakhshani, R. (2023). Revolutionizing groundwater management with hybrid AI models: A practical review. *Water*, 15(9), Article 1750.
- [62] Luan, H., & Cai, Z. (2023). Introduction to artificial intelligence and machine learning in environmental science. *Environmental Science: Advances*, 2(9), 1149–1150.
- [63] Slater, L. J., Arnal, L., Boucher, M.-A., Chang, A. Y.-Y., Moulds, S., Murphy, C., Nearing, G., Shalev, G., Shen, C., Speight, L., Villarini, G., Wilby, R. L., Wood, A., & Zappa, M. (2023). Hybrid forecasting: Blending climate predictions with AI models. *Hydrology and Earth System Sciences*, 27(9), 1865–1889.
- [64] Cao, J., Guo, Z., Lv, Y., Xu, M., Huang, C., & Liang, H. (2023). Pollution risk prediction for cadmium in soil from an abandoned mine based on random forest model. *International Journal of Environmental Research and Public Health*, 20(6), Article 5097.
- [65] Sun, Y., Chen, S., Jiang, H., Qin, B., Li, D., Jia, K., & Wang, C. (2024). Towards interpretable machine learning for observational quantification of soil heavy metal concentrations under environmental constraints. *Science of the Total Environment*, 926, Article 171931.
- [66] Salaudeen, A. O., Olawore, Y. A., Yetunde, A. A., Yakubu, H., & Dewa, A. U. (2025). Preliminary study into the application of metabolomics in soil discrimination. *Kwaghe International Journal of Sciences and Technology*, 2(2), 258–278.
- [67] Pan, D., Zhang, Y., Deng, Y., Van Griensven Thé, J., Yang, S. X., & Gharabaghi, B. (2024). Dissolved oxygen forecasting for Lake Erie's central basin using hybrid long short-term memory and gated recurrent unit networks. *Water*, 16(5), Article 707.
- [68] Pant, N., Toshniwal, D., & Gurjar, B. R. (2024). Multi-step forecasting of dissolved oxygen in River Ganga based on CEEMDAN-AdaBoost-BiLSTM-LSTM model. *Scientific Reports*, 14, Article 11199.
- [69] Kim, D., Lee, K., Jeong, S., Song, M., Kim, B., Park, J., & Heo, T.-Y. (2024). Real-time chlorophyll-a forecasting using machine learning framework with dimension reduction and hyperspectral data. *Environmental Research*, 262(Part 1), Article 119823.
- [70] Petrea, S.-M., Costache, M., Cristea, D., Strungaru, S.-A., Simionov, I.-A., Mogodan, A., Oprica, L., & Cristea, V. (2020). A machine learning approach in analyzing bioaccumulation of heavy metals in turbot tissues. *Molecules*, 25(20), Article 4696.

- [71] Bertato, L., Chirico, N., & Papa, E. (2022). Predicting the bioconcentration factor in fish from molecular structures. *Toxics*, 10(10), Article 581.
- [72] Simionov, I.-A., Cristea, D. S., Petrea, S.-M., Mogodan, A., Jijie, R., Ciornea, E., Nicoară, M., Turek Rahoveanu, M. M., & Cristea, V. (2021). Predictive innovative methods for aquatic heavy metals pollution based on bioindicators in support of blue economy in the Danube River basin. *Sustainability*, 13(16), Article 8936.
- [73] Lepak, J. M., Johnson, B. M., Hooten, M. B., Wolff, B. A., & Hansen, A. G. (2023). Predicting sport fish mercury contamination in heavily managed reservoirs: Implications for human and ecological health. *PLOS ONE*, 18(8), Article e0285890.
- [74] 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. *Environmental Research*, 212(Part B), Article 113322.
- [75] Lin, L., Liang, Y., Liu, L., Zhang, Y., Xie, D., Yin, F., & Ashraf, T. (2022). Estimating PM_{2.5} concentrations using the machine learning RF-XGBoost model in Guanzhong urban agglomeration, China. *Remote Sensing*, 14(20), Article 5239.
- [76] deSouza, P., Kahn, R., Stockman, T., Obermann, W., Crawford, B., Wang, A., Crooks, J., Li, J., & Kinney, P. (2022). Calibrating networks of low-cost air quality sensors. *Atmospheric Measurement Techniques*, 15(21), 6309–6328.
- [77] Souza, A., Rojas, M. Z., Yang, Y., Lee, L., & Hoagland, L. (2022). Classifying cadmium contaminated leafy vegetables using hyperspectral imaging and machine learning. *Heliyon*, 8(12), Article e12256.
- [78] Liu, R., Cui, B., Dong, W., Fang, X., Xiao, Y., Zhao, X., Cui, T., Ma, Y., & Wang, Q. (2024). A refined deep-learning-based algorithm for harmful-algal-bloom remote-sensing recognition using *Noctiluca scintillans* algal bloom as an example. *Journal of Hazardous Materials*, 467, Article 133721.
- [79] Chen, J., Si, Y.-W., Un, C.-W., & Siu, S. W. I. (2021). Chemical toxicity prediction based on semi-supervised learning and graph convolutional neural network. *Journal of Cheminformatics*, 13, Article 93.
- [80] Ketkar, R., Liu, Y., Wang, H., & Tian, H. (2023). A benchmark study of graph models for molecular acute toxicity prediction. *International Journal of Molecular Sciences*, 24(15), Article 11966.
- [81] Sun, Y., & Zheng, Y. (2023). A method of gas sensor drift compensation based on intrinsic characteristics of response curve. *Scientific Reports*, 13, Article 11971.
- [82] Abu-Hani, A., Chen, J., Balamurugan, V., Wenzel, A., & Bigi, A. (2024). Transferability of machine-learning-based global calibration models for NO₂ and NO low-cost sensors. *Atmospheric Measurement Techniques*, 17(13), 3917–3931.
- [83] Xu, Q., Shi, Y., Bamber, J. L., Tuo, Y., Ludwig, R., & Zhu, X. X. (2025). Physics-aware machine learning revolutionizes scientific paradigm for process-based modeling in hydrology. *Earth-Science Reviews*, 271, Article 105276.
- [84] Li, P., Yang, J., Wierman, A., & Ren, S. (2024). Towards environmentally equitable AI via geographical load balancing. In *Proceedings of the 15th ACM International Conference on Future and Sustainable Energy Systems* (pp. 291–307). Association for Computing Machinery.
- [85] Hilling, D. E., Ihaddouchen, I., Buijsman, S., Townsend, R., Gommers, D., & van Genderen, M. E. (2025). The imperative of diversity and equity for the adoption of responsible AI in healthcare. *Frontiers in Artificial Intelligence*, 8, Article 1577529.
- [86] He, M., Sandhu, P., Namadi, P., Reyes, E., Guivetchi, K., & Chung, F. (2025). Protocols for water and environmental modeling using machine learning in California. *Hydrology*, 12(3), Article 59.



© 2026 by the authors; licensee PSRP, Lahore, Pakistan. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) license (<http://creativecommons.org/licenses/by/4.0/>).