Artificial Intelligence & Environment

https://doi.org/10.66178/aie-0026-0004

Perspective

Open Access

Artificial intelligence-aided new paradigm of environmental research

Ziyu Chen, Jinhui Yuan, Jianing Liu, Dirong Zhang, Hou Guo, Peirong Wu and Shulin Zhuang*

Citation: Chen ZY; Yuan JH; Liu JN; et al. Artificial intelligence-aided new paradigm of environmental research. *AI Environ.* 2026, 1(1): 23-32.

Received: 10 December 2025

Revised: 11 January 2026

Accepted: 20 January 2026

Published online: 11 February 2026

Abstract

Artificial intelligence (AI) is revolutionizing a paradigm shift in environmental research from exploratory investigation to predictive study, facilitating the development of precision-guided research to tackle global environmental challenges at an unprecedented scale and speed. Against this backdrop, this perspective introduces the transformative impact of AI across various interconnected environmental domains. These technologies, including machine learning (ML), deep learning (DL), and large language models (LLMs), uncover latent environmental research patterns, integrate cross-scale data, and generate novel scientific insights. Systematically, we review advances in AI applications across water, soil, and atmospheric environments, as well as solid waste management. To further facilitate the construction and application of AI models, we summarize key considerations for AI model development, including data curation, data preprocessing, model training, validation, and interpretability. Significant challenges regarding AI application, including data-related issues, model selection, and ethical considerations are finally addressed. Surmounting these obstacles is a prerequisite for transitioning AI from passive analytical tools to autonomous research partners. Consequently, more innovative AI scientific agents are expected to promote cross-disciplinary collaboration and innovation to facilitate high-throughput, large-scale, and precision-guided composite investigations that integrate multi-source data across spatial-temporal scales to enable predictive and systemic insights, and to support holistic environmental governance.

Keywords: AI for science, Research paradigm, Environmental science, Engineering applications, Functional materials, Environmental management

Introduction

Artificial intelligence (AI) has rapidly progressed from early rule-based systems to today's data-driven and foundation-model-based architectures, marking a shift from specialized problem-solving patterns toward more general and adaptive intelligence. Emerging AI technologies have increasingly transformed our interactions with both the digital and physical worlds[1]. The emergence of AI for Science (AI4S) has advanced AI techniques into the core logic of scientific inquiry[2]. AI4S generally leverages machine learning, deep learning and large language models (LLMs) to empower scientific discovery with robust capabilities in reasoning, generalization, and multimodal comprehension. The AI-driven scientific discovery process in the research field of environment (AI for environment) forms a complete closed loop, spanning from hypothesis generation to practical validation (Fig. 1). It is driving a paradigm shift across environmental science, engineering applications, and functional materials, transitioning these domains from discrete, observation-based approaches toward integrated, intelligent, and predictive frameworks[3]. Within this framework, New Artificial Intelligence-based Methods (NAMs) represent a pivotal advancement, fusing high-throughput in vitro screening with AI to achieve robust toxicity prediction while minimizing reliance on in vivo testing. Transcending static data correlations and computationally intractable physical models, this paradigm integrates physical principles with foundation models to autonomously decode complex, cross-scale interactions across heterogeneous systems.

* Correspondence: Shulin Zhuang (shulin@zju.edu.cn)

Full list of author information is available at the end of the article.

https://doi.org/10.66178/aie-0026-0004

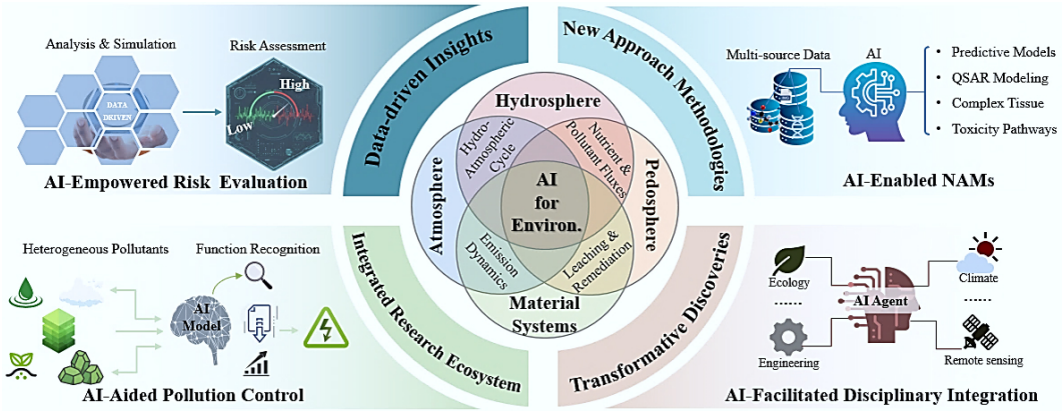

Fig. 1 AI for environment consists of four key areas, including data-driven insights, new approach methodologies, integrated research ecosystem and transformative discoveries.

Considering that many traditional methods relying on isolated data and physics-based models are often limited in fully capturing the cross-scale interactions and the coupled environmental processes, AI offers a unified framework to augment both analytical depth and operational efficacy[4], thereby reorienting the scientific lens from compartmentalized analysis toward a holistic understanding of synergistic effects[5,6]. AI facilitates a transition from static, point-based monitoring to dynamic, multi-scale perception, allowing continuous tracking of pollutants across atmospheric, aquatic, and terrestrial systems[7]. By integrating physical principles with data-driven modeling, AI enhances predictive accuracy and quantifies uncertainty, thereby informing risk mitigation and decision-making under complex, nonlinear conditions[8]. This capability underpins a shift from reactive measures to proactive, adaptive, and precision-oriented management—enabling early warning systems, targeted ecological restoration, and circular resource strategies that collectively promote a more resilient and sustainable relationship between human activities and environmental systems. Specifically in environmental analysis, AI drives a shift from discrete laboratory measurements to continuous, intelligent sensing. By leveraging computer vision, spectral analysis, and multi-sensor fusion, AI enables rapid, high-precision quantification and non-target screening of contaminants. This transition not only enhances data resolution but also captures dynamic pollution events in real-time across aquatic, terrestrial, and atmospheric systems.

By integrating data, algorithms, and computing power, AI4S is developing comprehensive solutions across environmental science and engineering. These solutions support a wide range of environmental applications, including energy and climate change, advanced early warning and forecasting, ecosystem assessments, circular resource strategies, and sustainable development planning. We focus particularly on its transformative impact in water, soil, air, and solid waste systems. Synthesizing these domains enables us to transcend the boundaries of discrete environmental media, orchestrating a synergy where functional material innovations and engineering interventions are deployed to sever complex cross-media pollutant pathways, thereby minimizing health risks and empowering holistic environmental management.

Water environment: a digital water system from early warning to intelligent operation

The acceleration of urbanization and the escalating pressure on global water resources have caused unprecedented challenges for the management of water environments. These challenges highlight that traditional methods for water quality monitoring and treatment have limitations in real-time responsiveness, efficiency, and accuracy, driving the need for digital transformation. With the rapid advancement of machine learning and deep learning technologies, the digital transformation of water environment systems is accelerating and has become a core driver in the field. These technologies enhance data processing and predictive capabilities, effectively overcoming the limitations of traditional methods in real-time monitoring and response to emergencies.

Water quality monitoring systems integrate multi-source data from hydrological, meteorological, and water quality sensors, significantly improving predictive performance. For instance, AI-based multispectral sensor systems have been effectively used for real-time monitoring of inland water bodies, showing strengths in pollutant detection and supporting efforts to address eutrophication assessment[9]. The innovative application of these technologies has shifted water environment management from traditional passive responses to proactive, intelligent decision support. Water quality early warning systems leverage deep learning and soft detection strategies to provide real-time risk predictions and early alerts. By utilizing advanced attention mechanisms and time-series analysis models, these systems can identify sudden shock loads at an early stage, offering accurate warnings to decision-makers and driving the shift in water environment management from traditional passive responses to proactive, intelligent management. The feature-temporal attention-based long short-term memory (FTA-LSTM) model is employed to process multidimensional sensor data, enabling efficient detection of anomalous water quality patterns, thereby significantly enhancing real-time response capabilities[10].

Furthermore, image analysis supported by artificial intelligence technologies is replacing traditional manual interventions, overcoming subjectivity and inefficiency in monitoring, and improving the accuracy and efficiency of surveillance. For example, by combining computer vision and deep learning[11], researchers utilize image recognition technologies to monitor surface water changes in real-time. In the event of anomalies such as increased phytoplankton concentration or surface pollutants, early warnings are issued, providing opportunities for preemptive responses to pollutant spread and water quality deterioration[12]. By providing timely alerts, these systems facilitate early intervention, curbing the spread of pollutants and improving the cost-effectiveness of environmental management.

The digital transformation extends beyond monitoring and early warning systems, opening new development pathways in water

treatment processes. For instance, in the challenging domain of high-saline organic wastewater treatment, deep learning has catalyzed a paradigm shift for Fenton-like reactions. Models significantly enhance the accuracy of degradation predictions while effectively reducing reagent consumption and treatment costs by tackling critical bottlenecks such as catalyst deactivation and complex reaction pathways. Specifically, Transformers utilize attention mechanisms to decipher chloride-hydroxyl radical interactions for accurate quenching prediction. Furthermore, reinforcement learning agents (e.g., Deep Q-Networks) autonomously regulate pH and dosage in real-time to stabilize efficiency against salinity fluctuations[13]. Complementing these advances in pollutant degradation, deep learning benefits water and wastewater disinfection by navigating non-linear water matrix dynamics to optimize dosages for agents like peracetic acid. These models surmount conventional limitations by precisely predicting by-product formation and residues, securing the optimal equilibrium between pathogen inactivation and chemical safety[14].

With the introduction of LLMs and other technologies, digital water systems are achieving intelligent optimization through more sophisticated data reasoning and processing. For example, LLMs can alleviate the scarcity of real hydrological data by generating synthetic datasets[15]. Furthermore, these technologies not only enhance real-time optimization of water quality, energy consumption, and operational costs but also drive the development of fully autonomous and self-optimizing water infrastructure[16]. In the future, as these technologies continue to evolve and be applied, water management will become more intelligent and precise, providing robust support for sustainable water resource utilization and environmental protection. Crucially, the integrative capability of AI extends beyond the water body itself. By assimilating data from atmospheric deposition and soil leaching, AI models enable the tracking of cross-media pollutant fluxes, bridging the gap between aquatic health and terrestrial or atmospheric inputs. This interconnected perspective is essential for realizing the holistic watershed governance proposed in the new paradigm.

Soil environment: precise monitoring and restoration strategy development

Accurately monitoring the concentration, spatial distribution, and temporal dynamics of soil pollutants is fundamental to soil quality assessment and pollution control. AI-driven models can integrate multi-source soil data that are typically characterized by high dimensionality, heterogeneity, and strong nonlinearity and extract key features from complex interactions, thereby supporting the prediction of contaminant concentrations, simulation of transport and transformation processes, and risk assessment[17]. Machine-learning-based approaches can identify key influencing factors from high-dimensional monitoring data, thereby improving model robustness and predictive accuracy[18]. To address the complexity of soil systems, a methodological suite for data processing and model optimization has evolved and been refined. This approach synergistically integrates dimensionality-reduction strategies, regularization techniques, and interpretability tools. Its application strengthens the predictive modeling of soil pollutants and enhances the actionable insights derived from model outputs for monitoring and remediation planning.

Algorithms such as random forests, decision trees, and convolutional neural networks can extract key patterns from high-dimensional and nonlinear soil datasets, facilitating pollution source identification, elucidation of heavy-metal and organic pollutant formation mechanisms, and significant improvement in concentration and spatial distribution prediction[19]. Demonstrating the capacity of these models to synthesize vast, heterogeneous datasets, a recent global analysis aggregated 796,084 sampling points from 1,493 regional studies. By employing the Extremely Randomized Trees (ERT) algorithm, researchers successfully modeled the non-linear relationships between environmental drivers and metal abundance, precisely mapping global zones where toxic metals exceed agricultural and human health thresholds[20]. For instance, theory-guided machine learning models have recently revealed that soil carbon sequestration strategies may inadvertently heighten metal(loid) mobility risks, thereby exposing the critical spatial divergence between total pollutant content and actual bioavailability[21]. In addition to quantitatively predicting the total concentrations of heavy metals, AI models are increasingly leveraged to predict the chemical speciation. By analyzing complex interactions between soil physicochemical properties (e.g., pH, redox potential, and organic matter), machine learning algorithms can accurately estimate the partitioning of metals among different fractions (e.g., exchangeable vs. residual forms). For soil property inference, neural networks and ensemble learning models consistently outperform traditional linear regression or interpolation methods in predicting indicators such as soil organic carbon, nitrogen content, pH, and enzyme activity, offering more accurate tools for soil health assessment and carbon-cycle studies. Integrated ensemble learning frameworks (e.g., Lasso-Stacking) have demonstrated superior capability in capturing high-dimensional environmental heterogeneity, thereby significantly improving the accuracy of predicting soil heavy metal(loid) concentrations and their spatial distributions for precise risk assessment[22]. Meanwhile, algorithm-driven intelligent control systems are accelerating the automation of soil remediation equipment, enabling autonomous material loading, precise reagent dosing, and uniform mixing through automated recognition, quantitative analysis, and real-time feedback, thereby improving remediation efficiency and reducing secondary pollution risks. For instance, recent advancements include robotic systems equipped with integrated sensors that autonomously collect soil samples and analyze properties in real-time, significantly reducing manual labor[23]. At broader spatial scales, the integration of remote-sensing inversion algorithms, deep-learning image models, and multi-sensor fusion technologies enables the construction of high spatiotemporal resolution soil monitoring systems capable of continuous soil-condition tracking, dynamic pollution-spread warning, and decision support for remediation strategies[24]. Deep integration of LLMs with soil monitoring, remediation, and management practices, completes a crucial feedback loop within intelligent environmental governance. This deep integration of artificial intelligence redefines soil pollution management. By refining soil-specific analytical frameworks and generating interoperable models, it enhances remediation precision, promotes sustainable land use, and strengthens ecological resilience. Furthermore, these intelligent systems facilitate the assessment of pollutant migration risks from soil to groundwater and crops, effectively linking soil remediation with water safety and food security within a unified management framework(Fig. 2).

Atmospheric environment: spatiotemporal analysis and process-oriented insight

The atmospheric environment presents a unique governance challenge due to its high spatiotemporal variability, rapid pollutant transformation, and the transboundary nature of pollutant transport[25]. AI technologies are reshaping the study and

https://doi.org/10.66178/aie-0026-0004

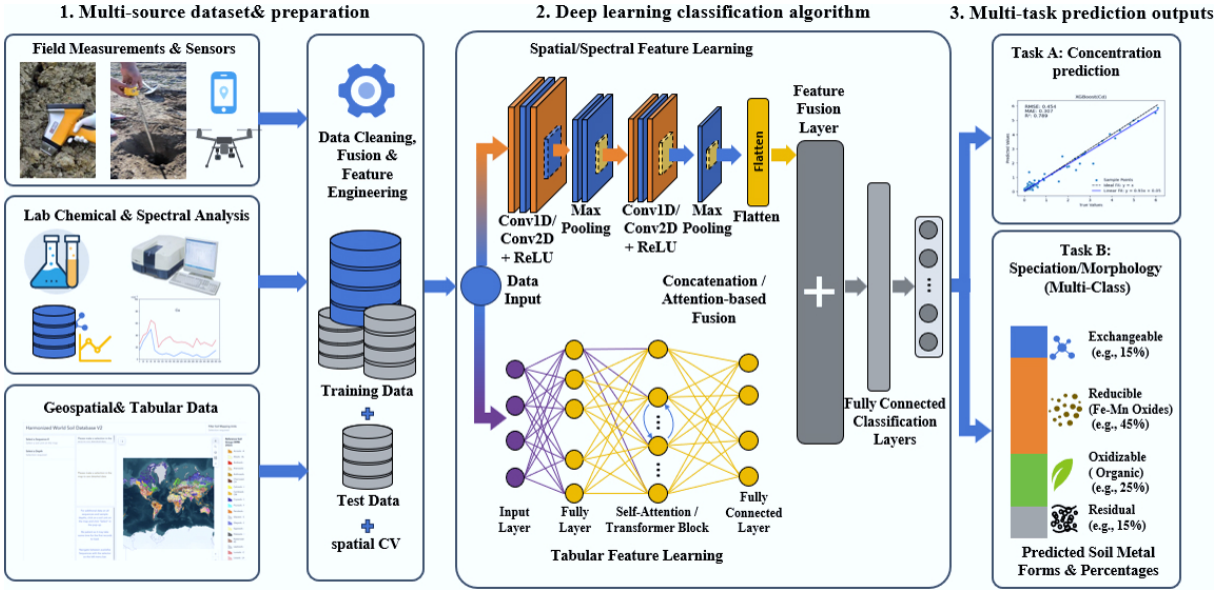

Fig. 2 AI-driven soil heavy metal prediction framework: from dataset to multi-task classification

management of the atmospheric environment by enabling a shift from fragmented observations and mechanistic simulations to data-driven, integrated, and predictive approaches. With the rapid expansion of ground-based monitoring networks, mobile sensing systems, satellite observations, and meteorological reanalysis products, atmospheric science is in a landscape characterized by massive, heterogeneous data streams. AI allows multi-source and multi-scale datasets to be fused, interpreted, and transformed into coherent spatiotemporal representations, greatly enhancing the capability to monitor pollution dynamics, identify emergent patterns, and diagnose environmental anomalies. AI also advances the understanding of atmospheric processes by uncovering nonlinear relationships and latent structures in pollutant interactions, providing insights essential for responsive and scientifically grounded air-quality management[26].

Substantial progress has been realized where AI is applied to resolve intractable analytical challenges in atmospheric environmental research, particularly those associated with heterogeneity, nonlinearity, and the incomplete nature of observational datasets. Within the realm of data synthesis, machine learning models have proven effective at reconciling discrepancies among satellite retrievals, ground-based measurements, and low-cost sensor observations, thereby enabling the construction of continuous, high-resolution pollutant fields. For example, a hybrid approach that combined graph-based learning with random forests was shown to substantially reduce spatial discontinuities in tropospheric ozone monitoring by exploiting the relational structure among monitoring sites[27]. Advanced machine learning models can precisely estimate daily concentrations of landscape fire PM2.5 by synthesizing outputs from chemical transport models, meteorological data, and ground-level observations. Extending these data-driven capabilities to high-fidelity Urban Air Pollution Mapping, researchers have further deployed fleets of vehicles equipped with low-cost sensors to serve as dynamic mobile monitoring networks. By leveraging decision tree algorithms such as XGBoost to fuse voluminous mobile sensing data with sparse stationary monitor readings, road network topology, and traffic flow metrics, these systems effectively overcome the spatial limitations of traditional infrastructure, inferring pollutant distribution at granular scales to pinpoint localized contamination hotspots[28,29]. Beyond improvements to the observational foundation, AI has also advanced the investigation of complex atmospheric chemical processes by identifying nonlinear dependencies that govern secondary pollutant formation, multiphase reactions, and new particle events—dependencies that conventional mechanistic models often struggle to capture with sufficient flexibility. A notable example lies in source apportionment, where data-driven models trained on extensive organic carbon measurements increased the volume of analyzable organic aerosol source information in Europe by more than a factor of four, enabling more refined delineation of pollution hotspots and improved assessments of population exposure. AI-driven emulators of chemical transport and meteorological models further contribute to rapid sensitivity analysis and scenario evaluation, offering a computationally efficient complement to full-physics simulations. Together, these developments illustrate how AI serves as a methodological extension of atmospheric science, enhancing its capacity to extract mechanistic insight and evaluate environmental interventions with greater speed and resolution.

The integration of AI into atmospheric environmental science is poised to synergistically enhance observation, interpretation, and decision support. Future progress will likely be driven by advances in multimodal data assimilation, physics-informed learning[30], and high-resolution predictive systems capable of capturing the complex interactions among emissions, chemistry, and meteorology. As LLMs evolve toward more domain-adapted and physically constrained forms, they hold particular promise for integrating dispersed atmospheric knowledge, facilitating model-data integration, and enabling more interpretable assessments of pollution episodes and regulatory scenarios. The convergence of these developments is anticipated to support increasingly adaptive, transparent, and system-level air-quality management frameworks, marking a significant step toward intelligent atmospheric governance[31].

Solid waste: intelligent sorting and circular economy optimization

The escalating challenge of Municipal Solid Waste (MSW) management, fueled by global trends in urbanization and

consumption, underscores the unsustainability of the traditional linear economic model and its attendant environmental risks[32–34]. In response, the Circular Economy framework emerges as a transformative paradigm, advocating for a system redesign where waste is eliminated, and materials are perpetually circulated[33–34]. However, the inherent challenge[35–37] of implementing Circular Economy principles in complex urban systems creates an imperative for a smarter approach, positioning AI as the cornerstone of a new paradigm. This AI-driven paradigm is being realized through capabilities that address two interconnected pillars of sustainable waste management: intelligent sorting and circular economy optimization.

AI overcomes critical limitations in waste segregation. Deep learning models (e.g., CNNs, YOLO) provide the perceptual foundation for high-precision identification and classification of complex waste streams, from municipal recyclables such as plastics and metals[38,39] to disaster debris[40,41]. This capability extends to distinguishing organic from inorganic fractions[42,43] and detecting specific waste types like construction debris and illegal dump sites[44,45]. Such perceptual intelligence is subsequently embodied in dexterous robotic action, where systems equipped with advanced control architectures perform precise manipulation and sorting tasks[46].

AI plays an equally pivotal role in circular economy optimization by enabling predictive analytics and system-level orchestration. Machine learning techniques, such as Random Forest (RF) and Long Short-Term Memory (LSTM) networks, enable accurate forecasting of municipal solid waste generation across scales, effectively reducing collection routing costs and optimizing collection logistics and resource allocation[47,48]. AI further enhances resource recovery by predicting outputs from processes such as plastic pyrolysis[49] and improving sorting efficiency for food and agricultural waste[50,51]. At the system level, this intelligence culminates in systemic orchestration through Digital Twin frameworks powered by AI and multi-agent systems, which simulate, predict, and optimize the waste lifecycle in real time, transforming linear flows into dynamic, closed-loop material flows[52]. Collectively, these developments illustrate AI's progression from a tool for automation to a foundational technology that underpins a smart, resilient, and truly circular waste management system.

Building on proven operational intelligence in waste management, future studies will focus on advancing the synergy between AI and complementary digital infrastructures. The convergence of machine learning with technologies such as digital twins and blockchain is critical for creating a transparent, verifiable, and adaptive ecosystem. Subsequent research will prioritize the development of robust algorithms capable of generalizing across heterogeneous waste types and facility operations, linking model innovation to measurable circular economy outcomes. Through this integrated approach, AI will solidify its role as the foundational layer of waste management systems—transforming them from linear cost centers into predictive, resilient, and inherently sustainable pillars of a circular economy.

Key considerations for environmental AI model development

Advanced application of AI requires a deliberate integration of three foundational pillars of AI (data, algorithms, and computational power) according to the real application scenarios and the practical needs. Especially, the inherent nonlinear relationships among multiple factors and the presence of multidimensional stresses within environmental systems further underscore the need for rigorous model development. This section outlines key considerations and provides references for constructing AI models (Fig. 3).

Dataset construction

High-quality data are essential for building robust and highly accurate models. Environmental datasets are inherently complex and heterogeneous, spanning diverse spatial and temporal scales, measurement resolutions, and data modalities. These datasets encompass structured data (e.g., physicochemical parameters, pollutant toxicity) and unstructured data (e.g., images, videos, acoustic signals). The quality and utility of such data are further shaped by measurement uncertainty, sensor variability, missing records, and inconsistencies across monitoring networks. Consequently, constructing a comprehensive and reliable dataset requires the integration of heterogeneous data sources, including long-term monitoring records, controlled experimental measurements, and open-source platforms guided by the specific objectives and constraints of the research problem[53].

The representativeness of a dataset fundamentally determines model reliability, often outweighing sheer data volume. In environmental research, representativeness entails comprehensive coverage across relevant spatiotemporal scales, environmental gradients (e.g., pollution levels, ecosystem states), and data modalities. Given the intrinsic heterogeneity of environmental systems, datasets that fail to capture key sources of variability, such as seasonal dynamics, extreme events, or anthropogenic disturbances, inevitably limit the transferability and robustness of model predictions. For example, river water-quality modeling depends on long-term, multi-site observations of meteorological, hydrological, and physicochemical parameters to capture both baseline conditions and episodic

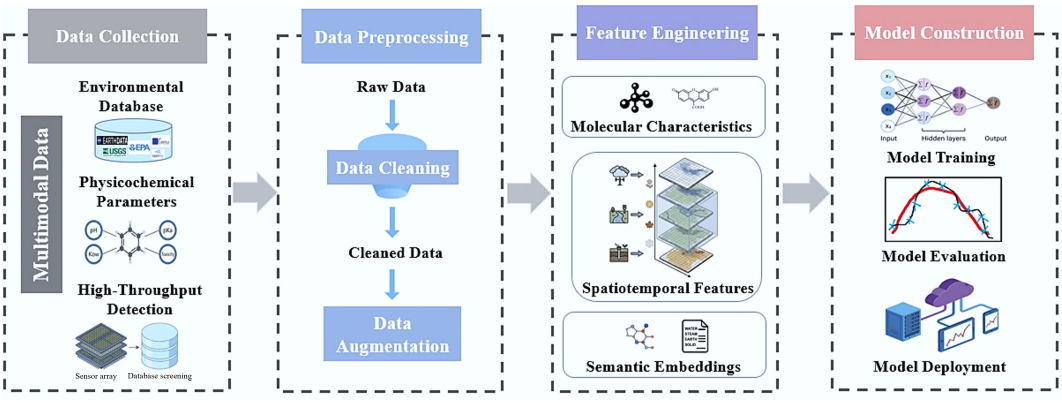

Fig. 3 Key components and workflow in environmental AI model development

anomalies such as seasonal shifts, storm-driven runoff, and pollution incidents. Predictive toxicology modeling similarly requires datasets that span diverse biological systems, exposure durations, and concentration ranges to ensure reliable characterization of chemical hazards. Therefore, ensuring dataset representativeness is a fundamental requirement for developing environmental AI models that can generalize reliably, support policy-relevant decision-making, and remain robust under real-world variability[54]. At the same time, strict data standardization is required—unified measurement units, sensor calibration protocols, and metadata formats. Without this, massive amounts of data will become noise rather than valuable resources.

Data preprocessing

Data preprocessing is a critical stage for ensuring data quality, enhancing model robustness, and aligning the dataset with the requirements of downstream modeling tasks. Common preprocessing procedures include the following components:

(1) Data curation constitutes a foundational component of data preprocessing and is essential for establishing the integrity and analytical reliability of a dataset. This process generally encompasses the systematic evaluation of data completeness, coherence, and internal consistency, with attention to the handling of missing information, the assessment of atypical or extreme observations, and the remediation of structural or formatting discrepancies that may undermine downstream analyses. Particularly in public toxicology databases (e.g., ToxCast), handling chemical structures constitutes a major bottleneck known as structural curation. Raw data often contain inorganic salts, indeterminate stereochemistry, or mixtures that confound predictive modeling. Therefore, specific chemical standardization steps—including the removal of inorganic salts, tautomer normalization, and the canonicalization of SMILES strings—are indispensable. These measures ensure that the digital representation of a molecule corresponds accurately to its bioactive form, preventing errors where identical compounds are misidentified as distinct entities due to formatting variations. Importantly, putative outliers are not treated merely as interference items; in many cases, they may be key signals that indicate environmental anomalies and therefore warrant careful contextual evaluation rather than automatic removal[55]. Through these quality-assurance measures, data curation provides a rigorously vetted dataset that supports more robust feature engineering and modeling practices.

(2) Feature engineering aims to refine the representation of raw data and improve model applicability, typically through feature selection, dimensionality reduction, and basic transformation procedures. Feature selection seeks to retain variables that are both informative and non-redundant, drawing on statistical relevance assessments, multicollinearity diagnostics, and model-based criteria to identify features that contribute meaningfully to prediction. Dimensionality reduction further streamlines the feature space by capturing dominant variance structures through linear techniques such as Principal Component Analysis (PCA), while nonlinear manifold-learning approaches, including t-distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP), are commonly used to visualize complex high-dimensional patterns. In practice, these core procedures are complemented by routine steps, such as constructing derived variables, encoding categorical attributes, and applying appropriate scaling, all of which help ensure coherent and analytically robust data representations across studies. Specifically for molecular characteristics (Fig. 3), a shift is occurring from traditional 1D/2D descriptors (e.g., ECFP) to deep learning-based representation learning. Unlike fixed descriptors, techniques like Graph Neural Networks (GNNs) extract latent high-dimensional features directly from molecular structures, capturing complex topological dependencies often overlooked by manual feature engineering.

(3) Imbalanced datasets, where certain classes or feature distributions are underrepresented, may lead to biased model learning, resulting in overfitting to majority classes and poor minority-class recognition. Typical imbalance types include binary imbalance, multiclass imbalance, within-class imbalance, and small-sample rare-event scenarios. To mitigate these issues, multiple strategies are employed. Resampling and synthetic data generation (e.g., SMOTE) help rebalance class distributions at the data level[56]. It is worth noting that SMOTE needs to run strictly within the abstract feature space. In the context of environmental chemometrics and toxicology, this creates a significant risk of generating chemical noise—synthetic feature vectors that mathematically exist but do not correspond to physically valid or synthesizable molecular structures. Consequently, when handling molecular data, such methods must be coupled with rigorous chemical validity filters, or replaced by structure-aware generative deep learning models (e.g., Graph Neural Networks or VAEs) that respect valence and topological constraints. At the algorithmic level, cost-sensitive learning incorporates asymmetric misclassification costs directly into model training, minimizing overall expected cost rather than error rate, an approach shown to be effective in neural networks and other classifiers[57]. Ensemble methods, including bagging, boosting, and hybrid frameworks, offer a complementary approach by combining multiple learners to emphasize difficult or minority-class instances, thereby improving predictive robustness[58]. Collectively, such approaches help counteract class-distribution bias and support more reliable model performance across underrepresented groups.

(4) Datasets should be partitioned into training, validation, and test subsets (or training and test sets combined with cross-validation strategies), with commonly used ratios of 7:2:1 or 8:1:1, depending on dataset size and variability. To ensure unbiased model evaluation, preprocessing steps that rely on model parameters (e.g., scaling, imputation, or dimensionality reduction) should be fitted only on the training set, with the resulting transformations subsequently applied to the validation and test sets to avoid data leakage[59]. In studies involving external validation, it is also important to examine whether the external dataset is distributionally comparable to the training data, as substantial shifts in data characteristics may compromise the reliability of generalization assessments.

Model architecture selection

The exponential growth of environmental data has catalyzed a paradigm shift toward AI-driven environmental modeling. Machine learning and deep learning methods offer unprecedented opportunities to decipher complex, nonlinear relationships and alleviate computational burdens in scenarios where underlying processes remain partially obscure or theoretically intractable. For instance, variational autoencoders (VAEs) effectively reduce dimensionality and extract salient features from vast environmental datasets, while isolation forests and clustering models robustly detect anomalies such as pollutant level deviations[60]. These methods can be complemented by traditional statistical techniques, like threshold-based methods, to ensure model outputs remain physically plausible and theoretically grounded. However, traditional machine learning models are typically designed for specific tasks, limiting their ability to capture the interconnectedness of various environmental processes. To overcome the limitations of the aforementioned single-paradigm approaches, process-guided or knowledge-guided machine learning

integrates mechanistic insights into data-driven models[61]. This hybrid paradigm embeds scientific laws and domain knowledge into machine learning workflows to improve accuracy, generalization, and consistency with fundamental principles such as conservation laws. For instance, a modeling framework for predicting vertical temperature variations in lakes has employed the combination of process-based components and recurrent neural networks. This yields better performance for long-term trend predictions by constraining outputs with ecological principles[62,63].

The use of ensemble methods in environmental science integrates outputs from multiple predictive models to improve the accuracy and robustness of predictions. This technique leverages the strengths of diverse models through approaches such as model blending, which combines various predictions, and post-processing, which refines outputs using observational data[64]. Ensembling is widely applied in environmental disciplines, including weather forecasting, where it aggregates predictions from multiple numerical weather models to enhance forecast precision[65,66].

Designing suitable model architectures for environmental science involves addressing challenges related to scientific data complexity, data heterogeneity, and scalability across regions and timescales[67]. Transfer learning enables the development of general-purpose models that can leverage pre-trained knowledge. Building on this, foundation models such as BERT and GPT have expanded transfer learning by training on massive datasets, allowing them to generalize across tasks with minimal tuning[68]. For instance, ChemBERTa, a Transformer-based model pre-trained on millions of chemical structures (SMILES), serves as a general-purpose encoder that can be fine-tuned for specific downstream tasks like toxicity prediction with limited labeled data. Similarly, in remote sensing, foundation models such as Prithvi leverage transfer learning from vast satellite imagery archives to generalize across diverse environmental monitoring tasks—ranging from flood detection to fire scar mapping—without requiring training from scratch. This capability enables scalability and adaptability from well-studied systems to data-scarce or unseen environments through effective knowledge transfer. Furthermore, developing LLMs based on environmental domain expertise and fine-tuned with data is also a research topic. They can automatically extract key information, such as compound names, their properties, and environmental effects, from extensive literature. Additionally, they can offer recommendations for environmental planning and management[67,69]. LLMs demonstrate significant potential in organizing unstructured toxicological literature for constructing Adverse Outcome Pathways (AOPs). By automating the extraction of molecular initiating events and key events, LLMs can accelerate the mechanistic understanding required for modern risk assessment. These models are revolutionizing domains and transforming environmental science by enabling holistic, scalable, and integrated modeling approaches.

Model evaluation and selection

Effective evaluation of AI models focuses on their ability to address unique challenges, such as multi-modal data integration, scalability, and generalization across diverse environmental tasks and datasets. Data splitting and cross-validation: Time-based and region-based splits ensure models account for temporal dynamics and spatial heterogeneity without introducing data leakage. K-fold cross-validation assesses transferability across varied datasets, while domain-specific methods, such as leave-one-region-out validation, evaluate performance on unobserved geographic areas. For classification tasks, common evaluation metrics include accuracy, precision, recall, F1-score, and AUC-ROC curve. For regression tasks, MSE, MAE, and R-squared can be used, combined with residual analysis to test model fitting effectiveness[70]. Beyond traditional regression metrics like root mean square error (RMSE), multi-modal alignment scores, such as contrastive loss in joint embeddings, evaluate the model's ability to combine text, images, and numerical data. Furthermore, feature importance tools help interpret the contribution of variables across multi-modal inputs. They can reveal whether vegetation indices or precipitation data dominate predictions in drought forecasting tasks. Complementing performance metrics, defining the Applicability Domain (AD) is a fundamental requirement for regulatory acceptance (e.g., under REACH or OECD principles). The AD delineates the specific physicochemical, structural, or biological space within which the model makes reliable predictions. Without a strictly defined AD, models risk 'blind extrapolation' to new chemical spaces, leading to potentially hazardous errors. Therefore, evaluation protocols must incorporate distance-based methods (e.g., Euclidean or Mahalanobis distance), probability density estimation, or bounding box approaches to establish these boundaries. Predictions falling outside the AD should be explicitly flagged as low-confidence, ensuring that the model is applied only within its validated limits. The generalization capabilities of different baseline models should be compared on the test set to select the model most suitable for the problem for in-depth research.

In the context of AI for environmental research, the ultimate goal is to facilitate practical application in industrial production and scientific research. Therefore, during the model evaluation and selection stage, emphasis should extend beyond traditional metrics to include interpretability and deployability[71]. The model with high performance is not necessarily the most suitable. For instance, while deep learning models might excel in prediction accuracy, their complex structures often lack interpretability, and their substantial computational requirements can complicate deployment on devices. In summary, model selection should holistically balance performance, explainability, and real-world constraints.

Challenges

The integration of AI into environmental science offers unprecedented opportunities; meanwhile, it also faces significant challenges. These challenges span data-related issues, model selection, and ethical considerations, all of which need to be addressed for AI to fully realize its potential in environmental research[72]. These challenges not only impact the effective application of AI in the environmental field, but also determine the key directions for future development and technological breakthroughs.

Data challenges

The success of AI in environmental applications is heavily dependent on the quality of input data. However, environmental data are often heterogeneous, sparse, and noisy, complicating data preprocessing and feature engineering. These datasets are sourced from various domains, including remote sensing, sensor data, and laboratory-generated data, each with unique structures and formats, presenting a challenge to data uniformity and consistency[73]. For example, remote sensing data may suffer from low resolution and high noise levels, while sensor data may contain missing or erroneous values due to equipment malfunctions. These data issues directly impact model training, reducing the performance of AI models and limiting their predictive accuracy. A particularly persistent challenge, especially with the proliferation of low-cost sensors, is sensor drift over time. To mitigate this, AI-based auto-calibration is emerging as a crucial solution, utilizing algorithms to dynamically detect and rectify signal biases against reference nodes without the need for frequent manual intervention.

https://doi.org/10.66178/aie-0026-0004

Researchers should develop techniques capable of handling heterogeneous data. In particular, tackling the scarcity of multi-source data in specific practical scenarios requires innovative methods to integrate data from various origins[74]. This could involve exploring cross-disciplinary data fusion and multi-modal learning approaches[75], which could more effectively capture the complexity and diversity of environmental datasets. For instance, fusing high-resolution satellite imagery (spatial data) with point-source ground sensor time-series (temporal data) enables a more holistic monitoring framework. Modern architectures, particularly Transformers, facilitate the alignment of these disparate modalities through cross-attention mechanisms, which map heterogeneous inputs into a shared latent space to uncover deep spatiotemporal correlations. Additionally, synthetic data generation, while useful in filling data gaps, introduces the risk of non-physical artifacts that, if not rigorously validated, can lead to model collapse and undermine the scientific validity of results[76].

Model challenges

Apart from data-related issues, challenges also arise in the selection and application of AI models themselves. In environmental research, there is often a mismatch between the complexity of models and the actual application requirements. While more complex models may provide better fitting capabilities in theory, they often lead to excessive computational costs and reduced interpretability in practice. Crucially, this interpretability encompasses two distinct layers: statistical interpretability (e.g., identifying influential features via SHAP values) and mechanistic interpretability (e.g., aligning model logic with established toxicological pathways or physical laws). For effective risk assessment, statistical correlations must be corroborated by mechanistic plausibility to ensure that decision-making is scientifically sound and not merely artifact-driven. Given the importance of transparency and explainability in environmental studies—where the results inform policy-making and resource allocation—it is crucial to ensure that AI models are not only accurate but also understandable[73].

To mitigate this challenge, researchers should prioritize model selection based on the specific needs of the application rather than the inherent complexity of the model. Overly intricate models should be avoided unless their superior performance is clearly justified by the problem at hand. Simple models, such as linear regression or decision trees, may often be sufficient for many environmental problems, while more advanced deep learning models can be leveraged when handling large-scale data. Future research should focus on developing context-aware models that align with specific research objectives and operational constraints, balancing predictive accuracy with computational efficiency[77].

Ethical and social challenges

The widespread use of AI in environmental research also brings about significant ethical considerations. AI's application in environmental monitoring and assessment often involves large amounts of sensitive data, such as personal health information and land ownership records. This raises concerns regarding privacy and data security, especially when such data are used for predictive modeling and decision-making. For example, the collection of environmental data may involve sensitive personal health information, necessitating strict adherence to ethical guidelines and legal frameworks to protect individuals' privacy rights and ensure data security.

AI's broad application could exacerbate data inequities, particularly between developed and developing regions. In many developing countries, environmental data quality is insufficient compared to that in more technologically advanced nations, leading to unequal access to environmental monitoring and governance tools. AI technologies should be deployed with a focus on equitable access to data and technology, particularly for under-resourced and remote areas, to ensure that global environmental management efforts are inclusive and fair[78]. Future research should not only tackle technical challenges but also consider the social implications of AI, ensuring that these technologies contribute to sustainable and globally accessible solutions for environmental protection.

Future directions

With the continuous advancement of AI technologies, more innovative methods and techniques are expected to emerge, enabling better responses to environmental challenges. Future environmental research warrants the integration of AI techniques with a focus on improving data acquisition and processing technologies to enhance model accuracy and adaptability, as well as designing AI solutions that meet ethical standards and ensure the sustainability of their applications. By integrating data generation technology, remote sensing technologies, cloud computing, Internet of Things (IoT) devices, and blockchain, AI will be able to monitor global environmental changes in real time, predict future environmental risks, and provide timely and accurate feedback. These technological breakthroughs will bring profound changes to environmental research, elevating scientific precision whilst harmonizing the complex interplay among engineering applications, functional material discovery, and ecological health protection, thereby delivering more collaborative and effective solutions for global environmental governance.

Ethical statements

Not applicable.

Author contributions

The authors confirm their contributions to the paper as follows: Shulin Zhuang designed the idea for this paper and revised the manuscript critically. Ziyu Chen, Jinhui Yuan, Jianing Liu, Dirong Zhang, Hou Guo and Peirong Wu contributed to collection of research articles and prepared the manuscript. All authors reviewed and approved the paper.

Data availability

Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

Funding

This research was supported by the National Key Research & Developmental Program of China (No. 2022YFC3701303), and the Zhejiang Provincial Natural Science Foundation of China (No. LZ26B070001).

Competing interests

The authors declare no competing financial interest.

Author details

State Key Laboratory of Soil Pollution Control and Safety, and Women's Hospital, School of Medicine, Zhejiang University, Hangzhou 310058, China

References

[1] Nielsen MW, Gissi E, Heidari S, Horton R, Nadeau KC, et al. 2025. Intersectional analysis for science and technology. *Nature* 640(8058):329–337

[2] Wang H, Fu T, Du Y, Gao W, Huang K, et al. 2023. Scientific discovery in the age of artificial intelligence. *Nature* 620(7972):47–60

[3] Reichstein M, Camps-Valls G, Stevens B, Jung M, Denzler J, et al. 2019. Deep learning and process understanding for data-driven earth system science. *Nature* 566(7743):195–204

[4] Shen C, Appling AP, Gentine P, Bandai T, Gupta H, et al. 2023. Differentiable modelling to unify machine learning and physical models for geosciences. *Nature Reviews Earth & Environment* 4(8):552–567

[5] Ren ZJ, Lowry GV, Boehm AB, Brooks BW, Gago-Ferrero P, et al. 2023. Data science for advancing environmental science, engineering, and technology. *Environmental Science & Technology* 57(46):17661–17662

[6] Runge J, Bathiany S, Bollt E, Camps-Valls G, Coumou D, et al. 2019. Inferring causation from time series in Earth system sciences. *Nature Communications* 10(1):2553

[7] Yuan Q, Shen H, Li T, Li Z, Li S, et al. 2020. Deep learning in environmental remote sensing: Achievements and challenges. *Remote Sensing of Environment* 241:111716

[8] Immorlano F, Eyring V, le Monnier De Gouville T, Accarino G, Elia D, et al. 2025. Transferring climate change physical knowledge. *Proceedings of the National Academy of Sciences of the United States of America* 122(15):e2413503122

[9] Sun Y, Wang D, Li L, Ning R, Yu S, et al. 2024. Application of remote sensing technology in water quality monitoring: From traditional approaches to artificial intelligence. *Water Research* 267:122546

[10] Zhang Y, Wang J, Li C, Duan H, Wang W. 2025. Attention-based deep learning models for predicting anomalous shock of wastewater treatment plants. *Water Research* 275:123192

[11] Zhi W, Appling AP, Golden HE, Podgorski J, Li L. 2024. Deep learning for water quality. *Nature Water* 2:228–241

[12] Amieva JF, Oxoli D, Brovelli MA. 2023. Machine and deep learning regression of chlorophyll-a concentrations in lakes using PRISMA Satellite hyperspectral imagery. *Remote Sensing* 15(22):5385

[13] Chen D, Guo Y, Ren Y, Yuan D, Niu Q, et al. 2025. When deep learning encounters Fenton like reactions: A paradigm shift in intelligent degradation of organic pollutants in high salt matrices. *Industrial Crops and Products* 234:121558

[14] Ding Y, Sun Q, Lin Y, Ping Q, Peng N, et al. 2024. Application of artificial intelligence in (waste)water disinfection: Emphasizing the regulation of disinfection by-products formation and residues prediction. *Water Research* 253:121267

[15] Wala J, Rüppel U. 2025. Leveraging large language models for synthetic data generation for AI applications for water infrastructure management. *EG-ICE International Workshop on Intelligent Computing in Engineering 2025*

[16] Vekaria D, Sinha S. 2024. Ai waters: an artificial intelligence framework for the water sector. *AI in Civil Engineering* 3(1):6

[17] Anifowose B, Anifowose F. 2024. Artificial intelligence and machine learning in environmental impact prediction for soil pollution management – case for EIA process. *Environmental Advances* 17:100554

[18] Nguyen TT. 2021. Predicting agricultural soil carbon using machine learning. *Nature Reviews Earth & Environment* 2(12):825

[19] Zhang H, Yin S, Chen Y, Shao S, Wu J, et al. 2020. Machine learning-based source identification and spatial prediction of heavy metals in soil in a rapid urbanization area, eastern China. *Journal of Cleaner Production* 273:122858

[20] Hou D, Jia X, Wang L, McGrath SP, Zhu Y, et al. 2025. Global soil pollution by toxic metals threatens agriculture and human health. *Science* 388(6744):316–321

[21] Qi C, Hu T, Zheng Y, Wu M, Tang FHM, et al. 2025. Global and regional patterns of soil metal(loid) mobility and associated risks. *Nature Communications* 16(1):2947

[22] Xu Y, Li P, Zhang Z, Gu Y, Xiao L, et al. 2025. Integrating machine learning for enhanced spatial prediction and risk assessment of soil heavy metal(loid)s. *Environmental Pollution* 383:126919

[23] Norby J, Wang S, Wang H, Deng S, Jones N, et al. 2024. Path to autonomous soil sampling and analysis by ground-based robots. *Journal of Environmental Management* 360:121130

[24] Fan Y, Wang X, Funk T, Rashid I, Herman B, et al. 2022. A critical review for real-time continuous soil monitoring: advantages, challenges, and perspectives. *Environmental Science & Technology* 56(19):13546–13564

[25] Zhang Q, Zheng Y, Tong D, Shao M, Wang S, et al. 2019. Drivers of improved PM2.5 air quality in China from 2013 to 2017. *Proceedings of the National Academy of Sciences of the United States of America* 116(49):24463–24469

[26] Gagne II DJ, Haupt SE, Nychka DW, Thompson G. 2019. Interpretable deep learning for spatial analysis of severe hailstorms. *Monthly Weather Review* 147(8):2827–2845

[27] Betancourt C, Li CWY, Kleinert F, Schultz MG. 2023. Graph machine learning for improved imputation of missing tropospheric ozone data. *Environmental Science & Technology* 57(46):18246–18258

[28] Tian Y, Ma Y, Xu R, Wu Y, Li S, et al. 2024. Landscape fire PM 2.5 and hospital admissions for cause-specific cardiovascular disease in urban China. *Nature Communications* 15(1):8

[29] Zhao B, Yu L, Wang C, Shuai C, Zhu J, et al. 2021. Urban air pollution mapping using fleet vehicles as mobile monitors and machine learning. *Environmental Science & Technology* 55(8):5579–5588

[30] Karniadakis GE, Kevrekidis IG, Lu L, Perdikaris P, Wang S, et al. 2021. Physics-informed machine learning. *Nature Reviews Physics* 3(6):422–440

[31] Rolnick D, Donti PL, Kaack LH, Kochanski K, Lacoste A, et al. 2023. Tackling climate change with machine learning. *ACM Computing Surveys* 55(2):1–96

[32] Kulisz M, Kujawska J, Cioch M, Cel W, Pizoń J. 2024. Comparative analysis of machine learning methods for predicting energy recovery from waste. *Applied Sciences* 14(7):2997

[33] Sesay REV, Fang P. 2025. Circular economy in municipal solid waste management: Innovations and challenges for urban sustainability. *Journal of Environmental Protection* 16(02):35–65

[34] Temirbekova M, Wojcik W, Stoyak V. 2021. Motor fuels and energy – producing fuels generation based on the processing of municipal solid waste organic components. *Journal of Ecological Engineering* 22(7):73–79

[35] Sarc R, Curtis A, Kandlbauer L, Khodier K, Lorber KE, et al. 2019. Digitalisation and intelligent robotics in value chain of circular economy oriented waste management - A review. *Waste Management* 95:476–492

[36] Ranjbar S, Malcata FX. 2023. Hydrothermal Liquefaction: How the holistic approach by nature will help solve the environmental conundrum. *Molecules* 28(24):8127

[37] Priya A, Dutta K, Daverey A. 2022. A comprehensive biotechnological and molecular insight into plastic degradation by microbial community. *Journal of Chemical Technology & Biotechnology* 97(2):381–390

[38] Shreyas Madhav AV, Rajaraman R, Harini S, Kiliroor CC. 2022. Application of artificial intelligence to enhance collection of E-waste: A potential solution for household WEEE collection and segregation in India. *Waste Management & Research* 40(7):1047–1053

[39] Zhang Y, Kim Y, Cha G. 2023. Assessment of deep learning-based image analysis for disaster waste identification. *Journal of Cleaner Production* 428:139351

[40] Mao W, Chen W, Wang C, Lin Y. 2021. Recycling waste classification using optimized convolutional neural network. *Resources, Conservation and Recycling* 164:105132

[41] Abdallah M, Talib MA, Feroz S, Nasir Q, Abdalla H, et al. 2020. Artificial intelligence applications in solid waste management: A systematic research review. *Waste Management* 109:231–246
[42] Pitakaso R, Srichok T, Khonjun S, Golinska-Dawson P, Gonwirat S, et al. 2024. Artificial intelligence in enhancing sustainable practices for infectious municipal waste classification. *Waste Management* 183:87–100
[43] Tao J, Gu Y, Hao X, Liang R, Wang B, et al. 2023. Combination of hyperspectral imaging and machine learning models for fast characterization and classification of municipal solid waste. *Resources, Conservation and Recycling* 188:106731
[44] Cha G, Moon HJ, Kim Y, Hong W, Hwang J, et al. 2020. Development of a prediction model for demolition waste generation using a random forest algorithm based on small dataSets. *International Journal of Environmental Research and Public Health* 17(19):6997
[45] Youme O, Bayet T, Dembele JM, Cambier C. 2021. Deep learning and remote sensing: Detection of dumping waste using UAV. *Big Data, IoT, and AI for a Smarter Future* 185:361–369
[46] Gal I, Ciocîrlan A, Mărgăritescu M. 2021. State machine-based hybrid position/force control architecture for a waste management mobile robot with 5DOF manipulator. *Applied Sciences* 11(9):4222
[47] Cubillos M. 2020. Multi-site household waste generation forecasting using a deep learning approach. *Waste Management* 115:8–14
[48] Nguyen XC, Nguyen T, La DD, Kumar G, Rene ER, et al. 2021. Development of machine learning - based models to forecast solid waste generation in residential areas: A case study from vietnam. *Resources, Conservation and Recycling* 167:105381
[49] Cheng Y, Ekici E, Yildiz G, Yang Y, Coward B, et al. 2023. Applied machine learning for prediction of waste plastic pyrolysis towards valuable fuel and chemicals production. *Journal of Analytical and Applied Pyrolysis* 169:105857
[50] Sousa J, Rebelo A, Cardoso JS. 2019. Automation of waste sorting with deep learning. *2019 XV Workshop de Visão Computacional (WVC)* 43–48
[51] Jahanbakhshi A, Momeny M, Mahmoudi M, Radeva P. 2021. Waste management using an automatic sorting system for carrot fruit based on image processing technique and improved deep neural networks. *Energy Reports* 7:5248–5256
[52] Liu Z, Zhang Z, Zhang Q, Zhao L. 2024. Digital twin modeling method for environmental governance of abandoned landfills based on multi-agent systems. *Digital Twin* 4:2530328
[53] Kim J, Jung W, An J, Oh HJ, Park J. 2023. Self-optimization of training dataset improves forecasting of cyanobacterial bloom by machine learning. *Science of The Total Environment* 866:161398
[54] Cui H, Gao S, Wang H, Zhang L, Luo Y, et al. 2025. Big data integration for environmental risk assessment of emerging contaminants. *Nature Sustainability* 1–11
[55] Brokamp C, Jandarov R, Hossain M, Ryan P. 2018. Predicting daily urban fine particulate matter concentrations using a random forest model. *Environmental Science & Technology* 52(7):4173–4179
[56] Ciranni M, Murino V, Odone F, Pastore VP. 2024. Computer vision and deep learning meet plankton: Milestones and future directions. *Image and Vision Computing* 143:104934
[57] Zhou Z, Liu X. 2006. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on Knowledge and Data Engineering* 18(1):63–77
[58] Galar M, Fernandez A, Barrenechea E, Bustince H, Herrera F. 2012. A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42(4):463–484
[59] Zhong S, Zhang K, Bagheri M, Burken JG, Gu A, et al. 2021. Machine learning: New ideas and tools in environmental science and engineering. *Environmental Science & Technology* 55(19):12741–12754
[60] Ma J, Ding Y, Cheng JCP, Jiang F, Tan Y, et al. 2020. Identification of high impact factors of air quality on a national scale using big data and machine learning techniques. *Journal of Cleaner Production* 244:118955
[61] Karpatne A, Atluri G, Faghmous JH, Steinbach M, Banerjee A, et al. 2017. Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on Knowledge and Data Engineering* 29(10):2318–2331
[62] Xia D, Chen J, Fu Z, Xu T, Wang Z, et al. 2022. Potential application of machine-learning-based quantum chemical methods in environmental chemistry. *Environmental Science & Technology* 56(4):2115–2123
[63] Read JS, Jia X, Willard J, Appling AP, Zwart JA, et al. 2019. Process-guided deep learning predictions of lake water temperature. *Water Resources Research* 55(11):9173–9190
[64] Slater LJ, Arnal L, Boucher M, Chang AYY, Moulds S, et al. 2023. Hybrid forecasting: blending climate predictions with AI models. *Hydrology and Earth System Sciences* 27(9):1865–1889
[65] Schevenhoven F, Selten F, Carrassi A, Keenlyside N. 2019. Improving weather and climate predictions by training of supermodels. *Earth System Dynamics* 10(4):789–807
[66] Bauer P, Thorpe A, Brunet G. 2015. The quiet revolution of numerical weather prediction. *Nature* 525(7567):47–55
[67] Kang Y, Lee W, Bae T, Han S, Jang H, et al. 2025. Harnessing large language models to collect and analyze metal–organic framework property data set. *Journal of the American Chemical Society* 147(5):3943–3958
[68] Zhu J, Jiang J, Yang M, Ren ZJ. 2023. ChatGPT and environmental research. *Environmental Science & Technology* 57(46):17667–17670
[69] Zhang Y, Lin S, Xiong Y, Li N, Zhong L, et al. 2025. Fine-tuning large language models for interdisciplinary environmental challenges. *Environmental Science & Ecotechnology* 27:100608
[70] Alvi M, Batstone D, Mbamba CK, Keymer P, French T, et al. 2023. Deep learning in wastewater treatment: a critical review. *Water Research* 245:120518
[71] Irwan D, Ibrahim SL, Latif SD, Winston CA, Ahmed AN, et al. 2025. River water quality monitoring using machine learning with multiple possible in-situ scenarios. *Environmental and Sustainability Indicators* 26:100620
[72] Bao Q, Yu N, Jiang Q, Wang X, Jiao Z, et al. 2025. Recent advances in implementation of machine learning for environmental nontarget identification. *Environmental Health* 3(11):1286–1298
[73] Liu W, Chen J, Wang H, Fu Z, Peijnenburg WJGM, et al. 2024. Perspectives on advancing multimodal learning in environmental science and engineering studies. *Environmental Science & Technology* 58(38):16690–16703
[74] Rong J, Zeng J, Chen K, Ma H, Shi P, et al. 2025. Synergizing machine learning and interpolation methods: A Stacking framework for global-scale satellite soil moisture gap filling. *Remote Sensing of Environment* 331:115040
[75] Yu N, Ma Z, Shao Q, Li L, Wang X, et al. 2025. Pseudodata-based molecular structure generator to reveal unknown chemicals. *Nature Machine Intelligence* 7(11):1879–1887
[76] Yan X, Yue T, Winkler DA, Yin Y, Zhu H, et al. 2023. Converting nanotoxicity data to information using artificial intelligence and simulation. *Chemical Reviews* 123(13):8575–8637
[77] von Rueden L, Mayer S, Beckh K, Georgiev B, Giesselbach S, et al. 2021. Informed machine learning - A taxonomy and survey of integrating prior knowledge into learning systems. *IEEE Transactions on Knowledge and Data Engineering* 35(1):614–633
[78] Roscher R, Bohn B, Duarte MF, Garcke J. 2020. Explainable machine learning for scientific insights and discoveries. *IEEE Access* 8:42200–42216

Copyright: © 2026 by the author(s). Published by NEW PRESS LIMITED, TSIM SHA TSUI, HK, China. This article is an open access article distributed under Creative Commons Attribution License (CC BY 4.0), visit https://creativecommons.org/licenses/by/4.0.