Filling in missing pieces in the co-development of artificial intelligence and environmental science

Zhenyu Wang,[2] Jin Zhang,[1,*] Pei Hua,[3] Yuanzheng Cui,[4] Chunhui Lu,[1] Xiaojun Wang,[5] Qiuwen Chen,[5] and Peter Krebs[2]

[1]The National Key Laboratory of Water Disaster Prevention, Yangtze Institute for Conservation and Development, Hohai University, Nanjing 210098, China

[2]Institute of Urban and Industrial Water Management, Technische Universität Dresden, Dresden 01069, Germany

[3]MOE Key Laboratory of Theoretical Chemistry of Environment, School of Environment, South China Normal University, Guangzhou 510006, China

[4]College of Hydrology and Water Resources, Hohai University, Nanjing 210098, China

[5]Department of Hydrology and Water Resources, Nanjing Hydraulic Research Institute, Nanjing 210029, China

*Correspondence: jin.zhang@hhu.edu.cn

Received: March 27, 2023; Accepted: May 17, 2023; Published Online: May 28, 2023; https://doi.org/10.59717/j.xinn-geo.2023.100007

Citation: Wang Z., Zhang J., Hua P., et al., (2023). Filling in missing pieces in the co-development of artificial intelligence and environmental science. The Innovation Geoscience 1(1), 100007.

Environmental science is an interdisciplinary field that focuses on studying the natural environment, including its physical, chemical, and biological characteristics, as well as the interactions between humans and the natural environment. Due to the nonlinear natural systems, implicit changes in the environment, and multifaceted environmental issues, there remain significant challenges in properly deciphering the complexity and dynamics of environmental science.

Artificial intelligence (AI) is on a steep upward trajectory, and its benefits are unearthed and implemented in business, traffic, and especially scientific research. The powerful nonlinearity and flexible model architecture of AI allow it to clarify complex relationships between variables and uncover underlying patterns behind large datasets with few computing resources. As a result, it has tremendous potential to address the challenges in environmental science and facilitate a better understanding of it. Based on recent research statistics, AI techniques have infiltrated many fields of environmental science, including the collection and analysis of informing environmental data, fast prediction of environment-related parameter values, chemical screening analysis, risk assessment and management, and environmental decision-making (Figure 1A).

Despite the growing influence of AI on environmental science, the problems and challenges created by AI have been seldom discussed. The challenges, such as insufficient or poor-quality data, difficulties in model interpretation, and concerns of bias in modeling outcomes, are important but missing pieces for co-developing AI and environmental science (Figure 1B). These missing pieces could give rise to misleading scientific findings, especially when AI is applied with inexperience. As next-generation AI technologies are built today, such as ChatGPT designed by OpenAI, AI enters a new era. More user-friendly and widely-used AI tools could magnify the challenges. At this crossroads of scientific development, it is necessary to pay more concern or re-think how we respond to the impact of the new AI wave on environmental science and how we properly promote co-development in both scientific fields.

AI DATA DEPENDENCY AND ENVIRONMENTAL DATA MANAGEMENT

The first missing piece is the data. The quantity and quality of the data dominate the AI performance. These are also challenges in environmental science caused by the problems such as limitations in monitoring systems, data inconsistency among data sources and formats, and insufficient database maintenance. Information gaps within small datasets and bias or errors in poor-quality data could compromise the AI model precision and lead

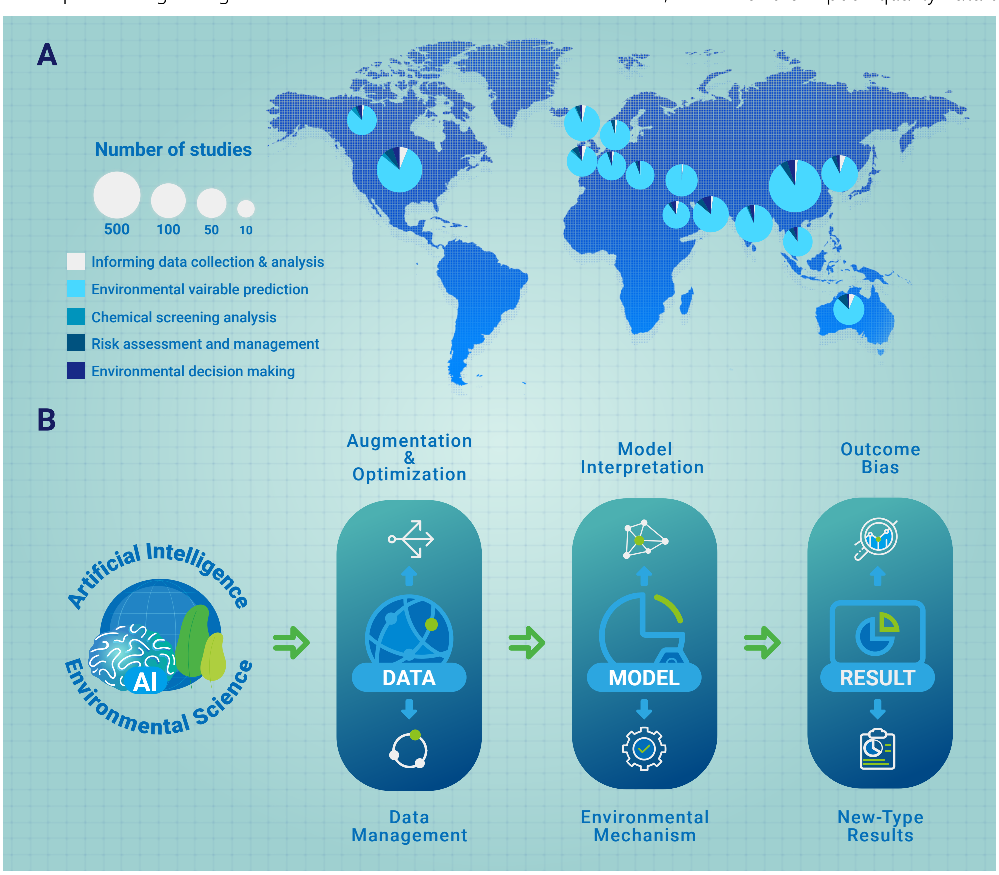

Figure 1. (A) Literature review of AI technique application in environmental science, derived from 2,106 relevant articles. The top 15 countries with the most publications are shown on the map. The categories are informing data collection and analysis, environmental variable prediction, chemical screening analysis, risk assessment and management, and environmental decision-making. (B) A schematic diagram describing missing pieces and relevant solvers in the co-development of artificial intelligence and environmental science.

to an increase in overfitting or underfitting risks, failing to represent the complexities of the environmental systems. Thus, developing adaptive AI techniques with weaker data dependency and improving environmental data management are the keys to data issues underlying the AI technology application in environmental science.

From the perspective of AI, reducing data dependency through the application of data augmentation, model validation, and learning optimization algorithms can adaptively solve the quantity and quality problems from data. With regards to data quantity, data augmentation techniques (e.g., data fuzzification, bootstrap resampling, and deep generative models) are being developed to refine the feature information or distributions in original data and artificially expand the datasets, which have achieved success in improving AI performance on small environmental datasets.[1]

Improving the utilization efficiency of available data in the model validation process can also mitigate models' reliability issues caused by small datasets. Pasini et al. applied a generalized leave-one-out approach to maximize the training data extension, successfully reconstructing the global temperature based on the limited datasets.[2] To address data quality issues, learning optimization algorithms can improve AI generalization performance and data-quality endurance. For example, regularization techniques (e.g., boosting methods, dropout, and weight decay) can encourage sparse solutions in model weights and focus on more important data features; Bayesian optimization can build an uncertainty-considering objective function and incorporate prior knowledge to handle uncertainty and bias in data and models.

In addition to reducing AI data dependency, effective environmental data management serves as the foundation for addressing data problems in the co-development of AI and environmental science. Comprehensive data management contains several steps: data documentation, data-collection standardization, database establishment, and data-sharing promotion. The former two steps are the basics of data management. These handle the issues of data transparency, consistency, harmonization, and reliability, helping avoid data debt during the AI technique application on environmental science. It is especially important for these steps in current when we have historical data, with incompletion, inaccuracy, or bias, and new data, collected by the advent of new technologies (e.g., sensors and satellites) with the non-standardized data format. The latter two steps are the growth of data management, which enhances the data accessibility for AI technology applications and foster the development of environmental research. Given the critical role of data management throughout the regulatory and scientific communities, a collaborative effort from environmental scientists is essential for improving environmental data management.

AI MODEL INTERPRETATION AND ENVIRONMENTAL MECHANISM EXPLORATION

As the second missing piece, the AI model interpretation is often neglected in the AI application due to the characteristics of AI algorithms. Understanding the interactions between natural and man-made processes remains many barriers, namely the abundance of interacting environmental and human components, unobservable chemical, physical and biological processes, as well as unpredictable ecosystemic and human behaviors. AI models with advanced nonlinear algorithms and flexible parameter systems can bypass the barriers and make a gigantic leap in the interaction analysis. However, these barriers are also mechanism information for interpreting the causality of natural environmental change, which are hidden as high-level representations in layers and parameters of black boxy AI models. Therefore, improving interpretability in AI techniques is critical for AI applications in environmental science.

Recently, a sub-field of AI, namely explainable AI (XAI), is fast developing, aiming to realize fairness, accountability, transparency, and explainability (FATE) in AI. Heuristics analysis is one of the branches of XAI development. It is a post-hoc XAI method, in which the users heuristically evaluate explanations of AI outcomes based on their prior knowledge and partial understanding of the algorithm. As it is subjective due to the dependence on human judgment and understandability, a heuristics analysis involves some methods, such as SHapley Additive exPlanations (SHAP), Partial Dependence Plots (PDP), Local Surrogate (LIME), and Layer-wise relevance propagation (LRP), to provide quantitative importance of input features and variables to AI outcomes, increasing the objectivity and quality of interpretation. For example, in the field of environmental science, the SHAP in random forest modeling has been used to interpretably identify the sensitivity of leaf area index to sub-surface soil moisture.[3] Based on the reinforced heuristic interpretation, the black-boxy nature of AI in environmental science is, to some extent, shifted into translucent-boxy, which points out the potential mechanisms underlying the environmental interactions and paves the way for the subsequent mechanism research in environmental science.

Despite the convenience and potential applicability in understanding the model's logic, the heuristics analysis may suffer from low fidelity with the presence of data bias, which is a common issue for most data-based models. Recently, scientists are trying to integrate prior knowledge directly into the AI model structure to handle data issues and enhance models' interpretability from the data level to the mechanism level. Physics-informed neural networks (PINN) are AI models in which physical laws, described by partial differential equations (PDEs), regularize the model learning process, increasing model robustness and outcomes' reliability. PINN has significant application potential in environmental science, as PDEs can describe most environmental behaviors, such as chemical kinetics, population dynamics, heat transfer, and fluid dynamics. Recently, a PINN has been successfully used in predicting the global evapotranspiration process.[4]

AI OUTCOME BIAS AND NEW-TYPE ENVIRONMENTAL RESULTS

The concern for bias in AI modeling outcomes represents the third missing piece. The reliability, robustness, and generalizability of models are contingent upon the quality of data and appropriate AI algorithms. Undoubtedly, robust datasets in quantity and quality can prevent bias in AI outcomes in environmental science. Appropriate AI algorithms are also crucial in reducing outcome bias in environmental science applications. Different AI algorithms have specific strengths and weaknesses in handling varying patterns in datasets. For instance, convolutional neural networks and recurrent neural networks have unique neural network architectures that give them advantages in processing spatial and temporal data, respectively; Fuzzy logic systems use membership functions to handle categorical data, such as environmental quality indices, that involve non-linearity and uncertainty. These highlight the need for a deep understanding and preparation of the characteristics of data and AI algorithms to minimize biases in AI outcomes during application in environmental science.

As environmental science has evolved, research has shifted from utilizing 1D monitoring data to 2D image data (e.g., sensing and satellite images). It is expected that text-type data (e.g., social media and news) with higher-dimensional information could also play a crucial role in environmental data analysis in the future. However, concerns about bias in this type of data remain challenging. For example, ChatGPT, a large language model, has gained popularity for its capabilities of text generation and information summarization in various fields. Zhu et.al. tested ChatGPT's performances in environmental science and found it outstandingly capable of summarizing knowledge in the field.[5] However, the model also produced plausible-sounding but incorrect data on chemical formulas and frontier knowledge. The confusion of information can significantly hinder even experts' ability to judge the information accuracy, resulting in bias during research or utilization. Therefore, assessing the bias of these new types of data is something we need to concern about in the future co-development of AI and environmental science.

REFERENCES

1. Bergen, K.J., Johnson, P.A., de Hoop, M.V., and Beroza, G.C. (2019). Machine learning for data-driven discovery in solid Earth geoscience. Science **363**: eaau0323. DOI: 10.1126/science.aau0323.
2. Pasini, A., Racca, P., Amendola, S., et al. (2017). Attribution of recent temperature behaviour reassessed by a neural-network method. Scientific Reports **7**: 17681. DOI: 10.1038/s41598-017-18011-8.
3. Li, W., Migliavacca, M., Forkel, M., et al. (2022). Widespread increasing vegetation sensitivity to soil moisture. Nature Communications **13**: 3959. DOI: 10.1038/s41467-022-31667-9.
4. Lian, X., Zhao, W., and Gentine, P. (2022). Recent global decline in rainfall interception loss due to altered rainfall regimes. Nature Communications **13**: 7642. DOI: 10.1038/s41467-022-35414-y.
5. Zhu, J., Jiang, J., Yang, M., et al. (2023). ChatGPT and Environmental Research. Environmental Science & Technology **57**: 17667–17670. DOI: 10.1021/acs.est.3c01818.

DECLARATION OF INTERESTS

The authors declare no competing interests.