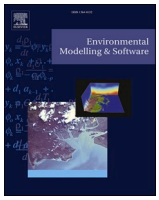
Environmental Modelling and Software 194 (2025) 106700

Contents lists available at ScienceDirect

Environmental Modelling and Software

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



Pitfalls of XAI interpretation in environmental modeling: A warning on model bias in air quality data analysis

ARTICLE INFO

Keywords:
Ozone analysis
Machine learning
Explainable AI (XAI)
Feature importance
Interpretation reliability
Statistical validation

ABSTRACT

Jung et al. (2025) achieved high predictive accuracy in interpolating missing ozone data using graph machine learning (ML) and conducted feature importance analysis with explainable AI (XAI). This correspondence acknowledges their significant contribution but discusses the limitations and biases inherent in ML models and XAI methods (e.g., Random Forest/Bootstrap Test, SHapley Additive exPlanations (SHAP)) and their impact on the reliability of derived feature importance. High predictive accuracy does not necessarily guarantee trustworthy interpretation of feature relevance, as evidenced by inconsistent importance rankings across models and XAI techniques. To enhance interpretability and scientific reliability, we advocate a validation strategy integrating ML with rigorous statistical analysis. It combines model-driven insights with statistical measures such as Spearman's rho and Kendall's tau, and information-theoretic metrics like Mutual Information and Total Correlation to capture complex, non-linear dependencies. Such integration improves the robustness of feature importance assessments and supports more reliable interpretations in environmental modeling.

Letter to the Editor

The recent paper by Jung et al. (2025), "Interpolation of missing ozone data using graph machine learning and parameter analysis through explainable artificial intelligence comparison," makes a valuable contribution to interpolating missing ozone data and analyzing parameter influence. Their study evaluates the potential of a graph-based machine learning (GML) model that integrates statistical interpolation methods and explores feature importance using explainable AI (XAI) methods, specifically the Bootstrap Test (BT) and SHapley Additive exPlanations (SHAP). However, the reliance on complex machine learning (ML) models and the interpretation of feature importance warrants further discussion.

Jung et al. employed a GML model incorporating methods such as Spatial Mean (SM), Spatiotemporal Mean (STM), Nearest Neighbor Hybrid (NNH), and Random Forest (RF), enhanced by a Correct and Smooth (CaS) process for interpolating missing ozone data. They reported robust performance for this interpolation task, with the model effectively simulating O_3 variations with R^2 of up to 0.96 and RMSE of 3.60 ppbv (for RF with CaS in Seoul, >7d missing interval). Beyond evaluating the performance of the interpolation model, a key aspect of their work involved feature importance analysis using XAI. For this analysis, they specifically applied the Bootstrap Test (BT) to the Random Forest (RF) model and SHapley Additive exPlanations (SHAP) to the XGBoost (XGB) model, revealing influential predictors such as NO2, Day of Year (DOY), Hour of Day (HOD), and Temperature. This raises critical concerns about potential bias in the ranked features derived from these specific models and XAI methods.

A growing body of recent research has underscored the complex challenges associated with interpreting feature importance in environmental modeling. Gu et al. (2025) introduced a vision-based model that integrates local and global information through a self-adaptive multiscale transform domain for air pollution monitoring. This approach complements Jung et al.'s graph-based method by demonstrating the potential of lightweight image-based techniques for efficient and scalable air quality analysis. Wu et al. (2024) comprehensively review the utilization of geospatial AI in air pollution prediction, summarizing current model statuses and proposing future research directions. Their insights support the need for caution when interpreting feature importance in complex models, particularly in the presence of spatial heterogeneity and data-driven biases. Rabbani et al. (2024) applied chemometric and ML techniques to assess environmental stress on regional plants from thermal power plant emissions. Their findings revealed that fly ash increased soil copper and iron, impairing plant growth and chlorophyll content, while also demonstrating the high accuracy of satellite data and ML in predicting air quality and vegetation health (NDVI). Fang et al. (2025) emphasize the importance of temporal and meteorological parameters in environmental monitoring through their remote sensing-based analysis of phycocyanin. This highlights how environmental variables, such as DOY and temperature, can be identified as influential in ozone prediction.

While Jung et al. present a valuable method achieving high ozone interpolation accuracy, this letter addresses a critical interpretational issue: the reliability of feature importance derived from their RF/BT and XGBoost/SHAP analysis. While their GML model demonstrated impressive predictive performance, this risks inadvertently implying reliability in the subsequent feature importance interpretation. However, it is crucial to emphasize that high predictive accuracy does not inherently guarantee reliable feature importance, a limitation widely recognized in literature. As demonstrated by numerous previous studies, strong predictive performance does not guarantee dependable feature importance interpretation (Lipton, 2018; Fisher et al., 2019; Lenhof et al., 2024; Mandler and Weigand, 2024; Potharlanka and Bhat, 2024; Wood et al., 2024). This issue is perhaps best exemplified by two

https://doi.org/10.1016/j.envsoft.2025.106700
Received 16 May 2025; Received in revised form 15 August 2025; Accepted 19 September 2025
Available online 19 September 2025

influential works. As Lipton (2018) cautions, "While the machine-learning objective might be to reduce error, the real-world purpose is to provide useful information." This highlights the risk of conflating predictive success with explanatory validity. Fisher et al. (2019) further demonstrate that feature importance is not uniquely defined across models with similar performance. They introduce the concept of Model Class Reliance (MCR), noting that: "Our central goal is to understand how much, or how little, models may rely on covariates of interest while still predicting well." More details and supporting literature can be found in the supplementary material.

Section 2.4 of Jung et al. (2025) clearly acknowledges this concern, stating: "it remains unclear whether these methods produce consistent results, as they are based on different ML models and estimation techniques for feature influence." This issue has been further examined in simulation-based research (Oka and Takefuji, 2025), which demonstrates that SHAP-based rankings from XGBoost can diverge substantially from statistically grounded benchmarks such as Spearman's rho and Kendall's tau. Drawing insights from Jung et al.'s valuable research, we advocate for integrating complementary statistical methods to strengthen feature importance analysis in environmental modeling. This need for contextual and multidisciplinary validation is echoed in Wang et al. (2025), who demonstrate how life-cycle assessment (LCA) combined with economic valuation (WTP) can reveal hidden health risks associated with construction waste. Their approach underscores the importance of integrating domain-specific knowledge when interpreting model outputs, especially in environmental health contexts.

Interpreting complex ML models like RF and XGBoost, particularly when understanding feature importance, presents methodological complexities. These models, while offering strong predictive power, can generate feature importance scores influenced by their internal structure and operation, such as splitting logic and handling interactions. RF and XGBoost tend to overemphasize features used in earlier tree splits, potentially leading to skewed assessments (Huti et al., 2023; Salles et al., 2021; Touw et al., 2013; Ugirumurera et al., 2024; Adler and Painsky, 2022; Alaimo Di Loro et al., 2023). Collinearity among features can also distort importance estimates. When predictors are highly correlated, the model may choose one and ignore the others, underestimating their shared influence. Tree-based models do not account for joint variance or mutual contributions.

The analysis by Jung et al. relies on RF, BT, and XGBoost combined with SHAP. As a result, both model-specific biases and limitations of SHAP affect the interpretation. SHAP values are calculated by averaging marginal contributions across all feature permutations. In cases of strong collinearity, this can dilute importance scores because correlated features split their contributions. SHAP may also produce unrealistic feature combinations that do not exist in actual data, which can mislead the interpretation (Bilodeau et al., 2024; Huang and Marques-Silva, 2024; Kumar et al., 2021; Lones, 2024; Molnar et al., 2022). Therefore, the ranked features reflect what is important to the model and its explanation method, rather than the true causal factors behind ozone concentration.

Fundamentally, validating feature importance is challenging due to the lack of ground truth. Different models can yield varying rankings because of their distinct methodologies and inherent biases. Indeed, Jung et al. themselves observed that "An interesting observation in Province regions is that the order of influence of HOD, DOY, and T differs between the BT and SHAP results" in Section 3.3, highlighting this critical issue. In environmental studies, complex feature sets with potential collinearity can further complicate interpretation and amplify these issues, making it difficult to confidently identify the actual determinants using only model-dependent analysis. These factors highlight that high predictive accuracy does not guarantee reliable feature importance interpretation from biased models and XAI methods. Although simulation-based validation can offer insights under controlled conditions, it cannot resolve the fundamental absence of ground truth in observational data.

Addressing the identified limitations in interpreting feature importance from complex ML models and XAI methods requires a robust analytical strategy. Such a framework must begin with a thorough understanding of the data characteristics and the underlying environmental processes governing ozone formation and distribution. Moving beyond model-dependent metrics, it is crucial to explore the statistical relationships between variables, including complex, non-monotonic associations, by utilizing appropriate non-parametric methods. Furthermore, rigorous statistical validation, incorporating techniques like hypothesis testing, is essential to ensure that findings are not merely artifacts of the modeling process but reflect genuine relationships.

Rather than relying solely on ML models and their inherent interpretability tools for feature selection and understanding model behavior, we advocate for a synergistic approach. This involves integrating the predictive power of ML with impartial and robust statistical methodologies, such as Spearman's rho and Kendall's tau, which are particularly adept at characterizing monotonic relationships (Okoye and Hosseini, 2024; Yu and Hutson, 2024). For more complex dependencies, including non-monotonic interactions among variables, alternative non-parametric methods like Mutual Information and Total Correlation offer valuable insights (Gibson, 2025; Kerby et al., 2024; Shi et al., 2024; Tserkis et al., 2025). By prioritizing fundamental statistical principles and employing methods capable of revealing diverse types of variable relationships, researchers can significantly enhance the credibility and dependability of feature importance assessments and model interpretations within environmental modeling domains.

To further enhance the statistical validity of feature importance assessments, it is essential to consider structural issues in the data. Collinearity among predictors can distort importance estimates by distributing influence arbitrarily among correlated features. In ML models, collinearity often leads to unstable and inconsistent feature rankings, as influence is arbitrarily split among correlated variables. In statistical methods, shared variance can exaggerate the apparent importance of features, resulting in misleading interpretations. Thus, whether using ML or statistical validation, careful attention to collinearity is crucial. To address this, we suggest feature agglomeration (FA), which clusters correlated features into unified groups, reducing redundancy while preserving interpretability. Although Principal Component Analysis (PCA) is widely used for dimensionality reduction, it may introduce bias specific to linear models and obscure interpretability when components lack clear domain relevance.

In conclusion, while Jung et al.'s valuable GML model achieves high ozone interpolation accuracy, this letter highlights a critical challenge for environmental modeling. The inherent limitations of ML models and XAI methods, such as RF/BT and XGBoost/SHAP, mean that the identified features cannot be automatically interpreted as the definitive drivers of ozone variability. High predictive accuracy in prediction tasks does not, in itself, guarantee dependable interpretation of feature importance. Solely relying on such interpretations risks misrepresenting the true mechanisms of complex environmental phenomena. Gaining truly trustworthy scientific insights therefore requires a robust approach that seamlessly integrates the predictive power of ML with the rigor and objectivity of complementary statistical methodologies. To advance reliable scientific insights, it is essential to integrate the predictive strengths of ML with the rigor of complementary statistical approaches. This perspective is reinforced by Li and Li (2023), who show that AI-based methods can reduce emissions and improve environmental performance when embedded within broader sustainability frameworks. Together, these insights point to the need for multifaceted, context-aware methodologies that support robust and unbiased environmental decision-making.

CRediT authorship contribution statement

Souichi Oka: Writing – original draft, Conceptualization. **Takuma Yamazaki:** Investigation. **Yoshiyasu Takefuji:** Writing – review &

editing, Supervision, Project administration.

Funding sources

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

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

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envsoft.2025.106700.

Data availability

No data was used for the research described in the article.

References

Adler, A.I., Painsky, A., 2022. Feature importance in gradient boosting trees with cross-validation feature selection. Entropy 24 (5), 687. https://doi.org/10.3390/e24050687.

Alaimo Di Loro, P., Scacciatelli, D., Tagliaferri, G., 2023. 2-step Gradient Boosting approach to selectivity bias correction in tax audit: an application to the VAT gap in Italy. Stat. Methods Appl. 32, 237–270. https://doi.org/10.1007/s10260-022-00643-4.

Bilodeau, B., Jaques, N., Koh, P.W., Kim, B., 2024. Impossibility theorems for feature attribution. Proc. Natl. Acad. Sci. U. S. A 121 (2), e2304406120. https://doi.org/10.1073/pnas.2304406120.

Fang, C., Song, K., Yan, Z., Liu, G., 2025. Monitoring phycocyanin in global inland waters by remote sensing: progress and future developments. Water Res. 275, 123176. https://doi.org/10.1016/j.watres.2025.123176.

Fisher, A., Rudin, C., Dominici, F., 2019. All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. J. Mach. Learn. Res. 20, 177. https://doi.org/10.48550/arXiv.1801.01489.

Gibson, J.D., 2025. Entropy and mutual information. In: Information Theoretic Principles for Agent Learning. Springer, Cham. https://doi.org/10.1007/978-3-031-65388-9_2.

Gu, K., Liu, Y., Liu, H., Liu, B., Qiao, J., Lin, W., Zhang, W., 2025. Air pollution monitoring by integrating local and global information in self-adaptive multiscale transform domain. IEEE Trans. Multimed. 27, 3716–3728. https://doi.org/10.1109/TMM.2025.3535351.

Huang, X., Marques-Silva, J., 2024. On the failings of Shapley values for explainability. Int. J. Approx. Reason. 171, 109112. https://doi.org/10.1016/j.ijar.2023.109112.

Huti, M., Lee, T., Sawyer, E., King, A.P., 2023. An investigation into race bias in random forest models based on breast DCE-MRI derived radiomics features. Clinical Image Based Procedure Fairness AI Med Imaging Ethical Philos Issues Med Imaging 14242, 225–234. https://doi.org/10.1007/978-3-031-45249-9_22.

Jung, S., Gil, J., Lee, M., Betancourt, C., Schultz, M., Choi, Y., Joo, T., Kim, D., 2025. Interpolation of missing ozone data using graph machine learning and parameter analysis through eXplainable artificial intelligence comparison. Environ. Model. Software 190, 106466. https://doi.org/10.1016/j.envsoft.2025.106466.

Kerby, T., White, T., Moon, K.R., 2024. Learning local higher-order interactions with total correlation. Proc. 2024 IEEE 34th Int. Workshop Mach. Learn. Signal Process. (MLSP), pp. 1–6. https://doi.org/10.1109/MLSP58920.2024.10734758.

Kumar, I., Scheidegger, C., Venkatasubramanian, S., Friedler, S., 2021. Shapley residuals: quantifying the limits of the Shapley value for explanations. Adv. Neural Inf. Process. Syst. 34, 26598–26608.

Lenhof, K., Eckhart, L., Rolli, L.M., Lenhof, H.P., 2024. Trust me if you can: a survey on reliability and interpretability of machine learning approaches for drug sensitivity prediction in cancer. Briefings Bioinf. 25 (5), bbae379. https://doi.org/10.1093/bib/bbae379.

Li, J., Li, X., 2023. Artificial intelligence for reducing the carbon emissions of 5G networks in China. Nat. Sustain. 6, 1522–1523. https://doi.org/10.1038/s41893-023-01208-3.

Lipton, Z.C., 2018. The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery. ACM Queue 16 (3), 31–57. https://doi.org/10.1145/3236386.3241340.

Lones, M.A., 2024. Avoiding common machine learning pitfalls. Patterns 5 (10), 101046. https://doi.org/10.1016/j.patter.2024.101046.

Mandler, H., Weigand, B., 2024. A review and benchmark of feature importance methods for neural networks. ACM Comput. Surv. 56. https://doi.org/10.1145/3679012. Article 318.

Molnar, C., König, G., Herbinger, J., Freiesleben, T., Dandl, S., Scholbeck, C.A., Casalicchio, G., Grosse-Wentrup, M., Bischl, B., 2022. General pitfalls of model-agnostic interpretation methods for machine learning models. In: Holzinger, A., Goebel, R., Fong, R., Moon, T., Müller, K.R., Samek, W. (Eds.), xxAI - beyond Explainable AI. Springer, Cham, p. 4. https://doi.org/10.1007/978-3-031-04083-2_4.

Oka, S., Takefuji, Y., 2025. Comments on "Dialogue between algorithms and soil: Machine learning unravels the mystery of phthalates pollution in soil" by Pan et al. (2025). J. Hazard. Mater. 493, 138366. https://doi.org/10.1016/j.jhazmat.2025.138366.

Okoye, K., Hosseini, S., 2024. Correlation tests in R: pearson cor, Kendall's tau, and Spearman's rho. In: Okoye, K., Hosseini, S. (Eds.), R Programming: Statistical Data Analysis in Research. Springer Nature, pp. 247–277. https://doi.org/10.1007/978-981-97-3385-9_12.

Potharlanka, J.L., Bhat, M.N., 2024. Feature importance feedback with Deep Q process in ensemble-based metaheuristic feature selection algorithms. Sci. Rep. 14 (1), 2923. https://doi.org/10.1038/s41598-024-53141-w.

Rabbani, M., Hossain, M.S., Islam, S.S., Roy, S.K., Islam, A., Mondal, I., Imam Saadi, S.M. A., 2024. Assessing thermal power effluent-induced air quality and associated environmental stress on Blumea lacera and Phyla nodiflora using chemometric, remote sensing and machine learning approach. Geol. Ecol. Landscapes. 1–19. https://doi.org/10.1080/24749508.2024.2430042.

Salles, T., Rocha, L., Gonçalves, M., 2021. A bias-variance analysis of state-of-the-art random forest text classifiers. Adv. Data Anal. Classif. 15, 379–405. https://doi.org/10.1007/s11634-020-00409-4.

Shi, Y., Golestanian, R., Vilfan, A., 2024. Mutual information as a measure of mixing efficiency in viscous fluids. Phys. Rev. Res. 6, L022050. https://doi.org/10.1103/PhysRevResearch.6.L022050.

Touw, W.G., Bayjanov, J.R., Overmars, L., Backus, L., Boekhorst, J., Wels, M., van Hijum, S.A.F.T., 2013. Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle? Briefings Bioinf. 14 (3), 315–326. https://doi.org/10.1093/bib/bbs034.

Tserkis, S., Assad, S.M., Lam, P.K., Narang, P., 2025. Quantifying total correlations in quantum systems through the Pearson correlation coefficient. Phys. Lett. 543, 130432. https://doi.org/10.1016/j.physleta.2025.130432.

Ugirumurera, J., Bensen, E.A., Severino, J., Sanyal, J., 2024. Addressing bias in bagging and boosting regression models. Sci. Rep. 14, 18452. https://doi.org/10.1038/s41598-024-68907-5.

Wang, L., Lv, Y., Wang, T., Wan, S., Ye, Y., 2025. Assessment of the impacts of the life cycle of construction waste on human health: lessons from developing countries. Eng. Construct. Architect. Manag. 32 (2), 1348–1369. https://doi.org/10.1108/ECAM-06-2023-0610.

Wood, D., Papamarkou, T., Benatan, M., Allmendinger, R., 2024. Model-agnostic variable importance for predictive uncertainty: an entropy-based approach. Data Min. Knowl. Discov. 38, 4184–4216. https://doi.org/10.1007/s10618-024-01070-7.

Wu, C., Lu, S., Tian, J., Yin, L., Wang, L., Zheng, W., 2024. Current situation and prospect of geospatial AI in air pollution prediction. Atmosphere 15, 1411. https://doi.org/10.3390/atmos15121411.

Yu, H., Hutson, A.D., 2024. A robust Spearman correlation coefficient permutation test. Commun. Stat. Theory Methods 53 (6), 2141–2153. https://doi.org/10.1080/03610926.2022.21211414.

Souichi Oka[a,*], Takuma Yamazaki[a], Yoshiyasu Takefuji[b]

[a] *Science Park Corporation, 3-24-9 Iriya-Nishi, Zama-shi, Kanagawa, 252-0029, Japan*

[b] *Faculty of Data Science, Musashino University, 3-3-3 Ariake Koto-ku, Tokyo, 135-8181, Japan*

[*] Corresponding author.

E-mail addresses: souichi.oka@sciencepark.co.jp (S. Oka), tyamazaki@sciencepark.co.jp (T. Yamazaki), takefuji@keio.jp (Y. Takefuji).