Artificial Intelligence & Environment

https://doi.org/10.66178/aie-0026-0015

Viewpoint

Open Access

From R^2 to W^2: rethinking predictive metrics for AI models in environmental health

Bin Wang[1,2,3,4*], Fei Xia[5*], Bo Pan[6*], Xiqing Li[4], Guofeng Shen[4], Yichao Huang[7], Tiantian Li[8], Mengyuan Ren[9], Le Zhang[10], Michael S Bloom[11], Pablo Gago-Ferrero[12] and Mingliang Fang[13]

Citation: Wang B; Xia F; Pan B; et al. From R^2 to W^2: rethinking predictive metrics for AI models in environmental health. *AI Environ.* 2026, 1(3): 135-138.

Received: 13 March 2026

Revised: 1 July 2026

Accepted: 15 July 2026

Published online: 28 July 2026

Abstract

Artificial intelligence (AI) has been increasingly used in environmental health research to predict pollutant concentrations, human exposure levels, and disease risks. However, the evaluation of predictive models depends largely on traditional metrics such as the coefficient of determination (R^2) or predicted R^2, which is frequently misinterpreted in predictive modeling contexts. In many studies, R^2 is used interchangeably with predictive performance metrics, potentially leading to overestimation of model reliability. Moreover, conventional evaluation indicators primarily quantify correlation or error magnitude and may fail to detect systematic prediction bias across exposure gradients. As a result, models with high apparent accuracy can still exhibit discriminatory prediction patterns, such as consistent overestimation or underestimation across the prediction range. Here, we clarify the distinction between the R^2 and predictive squared correlation coefficient (Q^2), and highlight the overlooked issue of prediction discrimination in AI-based environmental modeling. To address this limitation, we propose an integrated evaluation metric, W^2, which combines predictive accuracy with a penalty for systematic deviation from the 1:1 prediction line, as demonstrated with real environmental health data. This provides an auxiliary tool for assessing predictive reliability and improving the methodological rigor of AI-driven environmental health studies.

Keywords: Predictive squared correlation coefficient, Coefficient of determination, AI, Model performance, Risk assessment

Highlights

- Rapid integration of AI into environmental health research holds transformative potential.
- Predictive R^2 is usually misused in previously published AI studies.
- We proposed that W^2 and F values as the auxiliary perspective to differentiate model prediction accuracy and discrimination.

Artificial intelligence (AI) is rapidly transforming environmental health science by enabling large-scale predictions of pollutant concentrations, human exposure levels, and disease risks[1]. AI models have been increasingly used to reconstruct high-resolution air pollution fields, estimate chemical exposure in biomonitoring studies, and prioritize hazardous chemicals for regulatory decision-making[2,3]. As environmental health research moves toward data-intensive modeling frameworks, the evaluation of predictive performance has become a central methodological challenge. Among commonly used performance metrics, the coefficient of determination (R^2) or predicted R^2, remains one of the most widely reported indicators for regression-based prediction models[4]. However, R^2 is often misinterpreted or misapplied in the context of predictive modeling. In parallel, increasing application of AI models in environmental health has introduced additional methodological concerns. Environmental exposure data often span multiple orders of magnitude, exhibit strong spatial heterogeneity, and involve nonlinear relationships between environmental factors and health outcomes. For example, in the "HExpPredict" AI model, which predicts human blood exposome for chemical risk prioritization, chemical concentrations span eight orders of magnitude[1]. Under these conditions, conventional evaluation metrics may fail to capture systematic deviations between predicted and observed values. In

* Correspondence: Bin Wang (binwang@pku.edu.cn); Fei Xia (thuxia@gmail.com); Bo Pan (panbocai@gmail.com)

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

https://doi.org/10.66178/aie-0026-0015

particular, models with high R^2 values may still produce biased predictions across exposure gradients or population groups. These issues raise an important methodological question that how should predictive performance be evaluated in AI-based environmental health models to ensure both prediction accuracy and balance of the prediction discrimination? In this study, we clarify the distinction between the R^2 and the predictive squared correlation coefficient (Q^2), which are frequently conflated in the literature. We then highlight a less recognized limitation of conventional evaluation metrics, the inability to capture prediction discrimination, defined as systematic deviation between predicted and observed values. Finally, we propose an integrated evaluation metric, W^2, which combines predictive accuracy with a penalty for systematic prediction bias. This approach aims to improve the reliability and interpretability of AI-based environmental health predictions.

Misinterpretation of R^2 in predictive modeling

R^2 is traditionally used to evaluate how well a regression model fits observed data. Mathematically, R^2 quantifies the proportion of variance in the response variable that is explained by the model. In classical statistical modeling, this metric serves as an indicator of model goodness-of-fit within the dataset used to construct the regression model. However, predictive modeling requires a different evaluation perspective. When models are applied to new, unseen data, the primary concern is not how well the model fits the training dataset, but how accurately it predicts external observations. In this context, the Q^2 provides a more appropriate evaluation on predictive performance by comparing predicted values with observed values in a validation dataset. Our previously published mathematical proof and empirical calculations revealed that using the training dataset activity mean for computing Q^2 introduces systematic bias into the results, leading to an overestimation of a model's predictive power[4]. Thus, the validation dataset activity mean should be used as the reference as Eq. (1):

$$Q^2 = 1 - \frac{PRESS}{SS} = 1 - \frac{\Sigma(y_{\text{pred-}} y_{\text{obs}})^2}{\Sigma(y_{\text{obs-}} \bar{y})^2} \quad (1)$$

where *PRESS* denotes predictive sum of the squared differences between predicted and observed values; *SS* denotes the sum of squares of the differences between observed values and their mean ($\bar{y}$). Hence, Q^2 reflects the degree of aggregation of the predicted values and the observed values around the 1:1 line ranging $(-\infty, 1]$ for evaluating the predictive power of a model (in red, Figure 1a). We emphasized that $\bar{y}$ should come from the corresponding validation set (i.e., new, unseen data). For example, this approach helps ensure that quantitative structure-activity relationship (QSAR) models maintain scientific credibility and reliability, and was used to reshape the guidance document of Organization for Economic Co-operation and Development (OECD) for toxicity prediction[4]. For convenience, researchers tend to refer to Q^2 as R^2 to measure the predicting ability, which can be calculated by functions from the most popular software, e.g., "r2_score()" in the package "scikit-learn" (Python software) and "R2()" in the package "caret" (R software).

However, a key issue is that some studies do not clearly distinguish Q^2 in Eq. (1) from the R^2 (coefficient of determination) as defined in Eq. (2):

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\Sigma(y_{\text{fit-}} y_{\text{obs}})^2}{\Sigma(y_{\text{obs-}} \bar{y})^2} \quad (2)$$

where y_{fit} is the (possibly newly) fitted value of the target property using the training data, as opposed to the (unmodified) predicted value of the model y_{pred}, when employing the validation data to calculate Q^2 (Eq. 1); SSE denotes sum of squared error; and SST denotes sum of squares total. R^2 is used for evaluating the "goodness-of-fit" of a regression model with the range of $[0, 1]$. In this case, $\bar{y}$ refers to the data set under investigation for deriving the model (or training), not for the validation (or test) data set. For the training dataset, the values of $\bar{y}$ for Eq. (1) and Eq. (2) are equal, and $y_{\text{pred}} = y_{\text{fit}}$, resulting in $Q^2 = R^2$. Still, misuse of these terms is common. For example, a study introduced R^2 as the "coefficient of determination" and used this term for both training and validation data sets[5]. This practice is inappropriate according to our previous explanation that R^2 for the validation data set should be "predictive squared correlation coefficient" [4]. Another study fitted the predictive and observed values using Eq. (2) to obtain the correlation coefficient ($\sqrt{R^2}$) for both training and validation data sets[6], analogous to the model fitting line (in blue) in Figure 1a. These examples highlight the importance of clearly differentiating R^2 from Q^2 to ensure accurate model evaluation. Therefore, clarifying the conceptual differences between R^2 and Q^2 is an important step toward improving the methodological rigor of AI-based environmental health studies.

Prediction discrimination in environmental health models

Beyond the distinction between R^2 and Q^2, prediction discrimination of conventional evaluation metrics has received relatively little attention. It refers to systematic deviations between predicted and observed values that remain hidden when models are evaluated solely based on correlation-based metrics. For example, a model may consistently overestimate low concentrations while underestimating high concentrations. In such cases, predicted values may still correlate strongly with observed values, resulting in high R^2 or Q^2 values, even though the predictions exhibit systematic bias. This issue is particularly relevant for environmental health modeling. Environmental exposure data often cover several orders of magnitude and may involve complex nonlinear relationships between environmental drivers and pollutant concentrations[1]. Under these conditions, a model may achieve high correlation with observed data while still failing to reproduce the correct scale of environmental exposures.

To address these limitations, we propose a new evaluation metric, W^2, which integrates predictive accuracy with an explicit penalty for systematic prediction deviation. The central idea is to combine the Q^2 with an angular deviation term that quantifies the difference between the regression line of predicted versus observed values and the ideal 1:1 line. The angular deviation reflects the degree to which predictions deviate from perfect agreement in scale. To formalize this concept, we define the loss function *J*, which incorporates both Q^2 and a penalty for angular deviation of *F* in Eqs. (3)–(5):

$$J = Q^2 \cdot \exp\left(-\gamma \frac{F}{90}\right) \quad (3)$$

$$F = \tan^{-1}\left|\frac{\beta - 1}{\beta + 1}\right| \cdot \frac{180}{\pi} \quad (4)$$

$$\beta = \frac{\Sigma(y_{\text{obs-}} \bar{y}_{\text{obs}})(y_{\text{pred-}} \bar{y}_{\text{pred}})}{\Sigma(y_{\text{obs-}} \bar{y}_{\text{obs}})^2} \quad (5)$$

where β is the slope of the regression fit for predicted vs. observed values; the term (*F*/90) normalizes *F* (since *F* is in degrees ranging [0–90°]; and gamma (γ) is a tuning hyperparameter controlling the strength of the penalty and ranging $[0, +\infty]$. Using the exponential term $\exp\left(-\gamma \frac{F}{90}\right)$ can seamlessly integrate both prediction accuracy (through Q^2) and slope fidelity (through *F*), creating a balanced metric

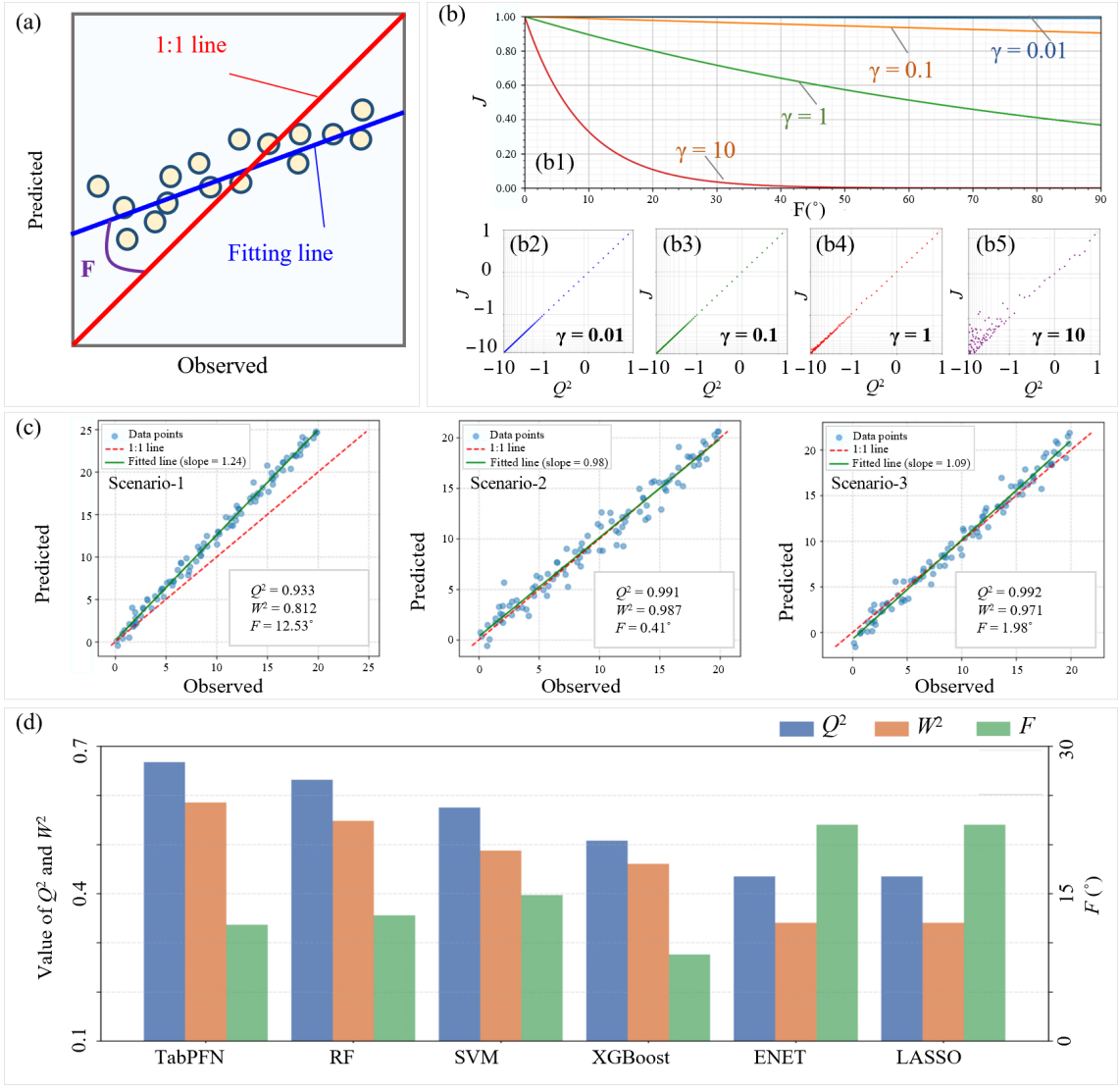

Fig. 1 Conceptual framework and performance comparison of the proposed predictive evaluation metric W^2. (**a**) Conceptual illustration of prediction discrimination in regression models. Predicted values are plotted against observed values, where the dashed red line represents the ideal 1:1 agreement. The blue line indicates the fitted regression line between predicted and observed values. The angular deviation (F) between these two lines quantifies systematic prediction bias, reflecting the extent to which model predictions deviate from scale-consistent agreement. (**b**) Performance analysis of the loss function J under different penalty strengths (γ). Panel (b1) shows how the penalty parameter γ modulates the influence of angular deviation F on J. Panels (b2–b5) illustrate the relationship between J and predictive squared correlation (Q^2) across simulated prediction scenarios with different γ values, demonstrating how increasing γ strengthens the penalty for prediction discrimination. (**c**) Model comparison using Q^2, F, and the integrated metric W^2. (**d**) Application of W^2 to evaluate an environmental health prediction task across the six representative models of tabular prior-data fitted network (TabPFN), random forest (RF), support vector machine (SVM), eXtreme Gradient Boosting (XGBoost), Elastic Net (ENET), and Least Absolute Shrinkage and Selection Operator (LASSO).

that encourages models to be both accurate and consistent in their scale of predictions with several advantages. As F increases, typical monotonic decreasing patterns are observed when γ varies from 0.01 to 10 (Figure 1b1). For simulated data sets of the predicted vs. observed values with Q^2 from −10 to 1, there is a positive relationship between J and Q^2, of which the curve shape is greatly modulated by γ value (Figure 1b1–b5). Based on our experience, the model performance can be well evaluated when the value of γ ranging [0.1, 1]. For the tuned model with the optimal γ (γ_{opt}), we define the new evaluation metric (W^2) in Eq. (6), which is the final form of Eq. (3),

$$W^2 = Q^2 \cdot \exp\left(-\gamma_{opt}\frac{F}{90}\right) \quad (6)$$

where Q^2 and F are calculated using the predicted vs. observed values in the validation data set.

To illustrate the utility of W^2, consider two predictive models that produce similar Q^2 values when evaluated against validation data. Although both models may appear equally accurate based on correlation metrics, their prediction behaviors may differ substantially. If the predicted values closely follow the 1:1 line, it suggests accurate reproduction of the magnitude of observed values. In another case, predicted values exhibit systematic slope deviation, indicating of consistent overprediction or underprediction across the exposure range. Taking the three assumed model results by setting $\gamma = 1$ for example (see Figure 1c), we can see that Scenario-1 have relatively good Q^2 (= 0.933), but great discriminatory prediction with F =

https://doi.org/10.66178/aie-0026-0015

12.53°, as also indicated by W^2 = 0.812. Scenario-2 and Scenario-3 have equivalent Q^2 values (i.e., 0.991 *vs*. 0.992), but greater discriminatory prediction was found for Scenario-2 than Scenario-3, reflected by W^2 (i.e., 0.987 *vs*. 0.971, respectively). By incorporating this angular deviation into the evaluation metric, W^2 captures both correlation strength and prediction consistency. For a more practically relevant evaluation, we applied $W^2$to the HExpPredict human blood exposome prediction dataset[1] (see Figure 1d). Although all models achieved high Q^2 values (i.e., prediction R^2), their deviations from the 1:1 prediction line differed. Tabular prior-data fitted network (TabPFN) and random forest (RF) showed the highest W^2, whereas Elastic Net (ENET) and Least Absolute Shrinkage and Selection Operator (LASSO) exhibited larger *F*, revealing calibration limitations despite moderate predictive correlations. Their integration in W2 therefore distinguishes models that merely reproduce relative trends from those that also generate quantitatively reliable predictions.

In summary, we made a comprehensive summary of R^2, Q^2, and propose an optimization strategy. Through theoretical derivations and empirical validation, we illustrated that W^2 and *F* can differentiate models based on both prediction accuracy and discrimination as auxiliary indices. However, W^2 is a complementary metric, not a standalone solution. Overfitting should still be controlled through appropriate procedures such as cross-validation, external validation, and strict separation of training and test data, and W^2 should be interpreted together with root mean squared error (RMSE), mean absolute error (MAE), and other relevant performance indicators.

Ethical statements

Not applicable.

Author contributions

Bin Wang, Fei Xia, and Bo Pan: Conceptualization; Bin Wang and Fei Xia: Methodology; Bin Wang, Fei Xia, and Bo Pan: Investigation; Bin Wang: Visualization; Bin Wang and Mingliang Fang: Funding acquisition; Bin Wang: Project administration; Bin Wang and Mingliang Fang: Supervision; Bin Wang, Fei Xia, and Bo Pan: Writing – original draft; Bin Wang, Mingliang Fang, Fei Xia, Bo Pan, Xiqing Li, Guofeng Shen, Yichao Huang, Tiantian Li, Mengyuan Ren, Le Zhang, Michael S Bloom, and Pablo Gago-Ferrero: Writing – review & editing.

Data availability

The data and code used to reproduce the demonstration results are available at https://github.com/ExposomeX/Data-and-Code-for-W2-Manuscript.git. We additionally provide the demo script "*w2_calculation_function.py*" to compute R^2, Q^2, *F*, and W^2 for readers' convenience.

Acknowledgments

We would like to express our gratitude to the working group of environmental exposure and human health of the China Cohort Consortium (https://cohortconsortium.com/index).

Funding

This research was supported by the National Natural Science Foundation of China (Grant No. 42477455), Beijing Municipal Ecology and Environment Bureau (No. BJST20250209), and the Strategy Priority Research Program (Category B) of the Chinese Academy of Sciences (XDB0750300).

Competing interests

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.

Author details

[1]Department of Epidemiology and Biostatistics, School of Public Health, Peking University, Beijing 100191, China; [2]Institute of Reproductive and Child Health, Key Laboratory of Reproductive Health, National Health and Family Planning Commission of the People's Republic of China, Beijing 100191, China; [3]Key Laboratory of Epidemiology of Major Diseases (Peking University), Ministry of Education, Beijing 100191, China; [4]Laboratory for Earth Surface Processes, College of Urban and Environmental Science, Peking University, Beijing 100871, China; [5]Independent researcher, New York, United States; [6]Yunnan Provincial Key Lab of Soil Carbon Sequestration and Pollution Control, Faculty of Environmental Science & Engineering, Kunming University of Science & Technology, Kunming 650500, China; [7]Department of Toxicology, Anhui Medical University, Hefei, Anhui 230032, China; [8]China CDC Key Laboratory of Environment and Population Health, National Institute of Environmental Health, Chinese Center for Disease Control and Prevention, Beijing 100088, China; [9]Gangarosa Department of Environmental Health, Rollins School of Public Health, Emory University, Atlanta, GA 30322, USA; [10]School of Information and Communication Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu 611731, China; [11]Department of Global and Community Health, College of Health and Human Services, George Mason University, Fairfax, Virginia 22030, USA; [12]Institute of Environmental Assessment and Water Research (IDAEA-CSIC), Barcelona 08034, Spain; [13]Department of Environmental Science and Engineering, Fudan University, Shanghai, 200433, China

References

[1] Zhao FR, Li L, Lin PH, Chen Y, Xing SP, et al. 2023. HExpPredict: In vivo exposure prediction of human blood exposome using a random forest model and its application in chemical risk prioritization. *Environmental Health Perspectives* 131:37009

[2] Gui K, Zhang XT, Che HZ, Li L, Zheng Y, et al. 2026. Advancing operational global aerosol forecasting with machine learning. *Nature* 651:658–665

[3] Bodnar C, Bruinsma WP, Lucic A, Stanley M, Allen A, et al. 2025. A foundation model for the Earth system. *Nature* 641:1180–1187

[4] Schüürmann G, Ebert RU, Chen JW, Wang B, Kühne R. 2008. External validation and prediction employing the predictive squared correlation coefficient - test set activity mean vs training set activity mean. *Journal of Chemical Information and Modeling* 48:2140–2145

[5] Wen YF, Zhang SJ, Wang Y, Yang JN, He LY, et al. 2024. Dynamic traffic data in machine-learning air quality mapping improves environmental justice assessment. *Environmental Science & Technology* 58:3118–3128

[6] Stelzer IA, Ghaemi MS, Han XY, Ando K, Hédou JJ, et al. 2021. Integrated trajectories of the maternal metabolome, proteome, and immunome predict labor onset. *Science Translational Medicine* 13:eabd9898

Copyright: © 2026 by the author(s). Published by NEW HORIZON PRESS LIMITED, TSIM SHA TSUI, HK, China. This is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY 4.0) License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.