



OPEN

Machine learning-based anomaly detection in surface water quality data using ensemble models with residual analysis

Iryna Mashkina & Tetiana Nosenko

This study addresses the challenge of detecting anomalies in water-quality monitoring data, where traditional approaches often lack sufficient temporal and spatial resolution. A data-driven framework integrating ensemble machine-learning models (Random Forest and XGBoost) with residual-based anomaly detection was developed and evaluated using open surface-water monitoring data from Ukraine for 2022. The models were trained to predict key hydrochemical indicators, including nitrogen concentration and dissolved oxygen, using chemical, spatial, and seasonal features. Model robustness was assessed through 5-fold cross-validation, which demonstrated stable predictive performance for both indicators. Anomalies were identified as statistically significant deviations between observed and predicted values. Recurrent anomaly hotspots were subsequently analyzed and spatially mapped, with hotspot locations independently confirmed by both Random Forest and XGBoost models. The results demonstrate that the proposed framework can effectively identify statistically unusual observations and reveal spatially persistent anomaly hotspots across multiple river basins. However, the detected anomalies should be interpreted as statistical deviations rather than confirmed pollution events, as independent pollution-source information was not available. The study highlights the potential of combining ensemble learning, residual analysis, and spatial hotspot identification to support data-driven environmental assessment and monitoring activities.

Keywords Artificial intelligence (AI), Machine learning (ML), Anomaly detection, Water pollution monitoring, Ensemble models, Random forest, XGBoost

The increasing volume of environmental monitoring data creates significant challenges for the development of efficient analytical methods. Recent studies have demonstrated the growing applicability of machine-learning approaches for water-quality assessment, prediction, and environmental monitoring, particularly when large and heterogeneous datasets are involved^{1,2}. Therefore, there is a need for approaches that can efficiently transform environmental monitoring data into actionable information.

Monitoring and assessment of river water quality are critically important for ecosystem protection and public health^{3,4}. Artificial intelligence and machine-learning techniques offer promising opportunities for improving environmental data analysis and decision support. Recent studies have also emphasized the importance of spatial and spatiotemporal perspectives in water-quality assessment and environmental monitoring^{5,6}.

Anomaly detection has become an important component of environmental data analysis because it enables the identification of statistically unusual observations in complex datasets⁷.

In contrast to dedicated anomaly-detection approaches such as Isolation Forest and One-Class SVM^{8,9}, as well as recent machine-learning methods developed specifically for water-quality anomaly detection¹⁰, this study proposes a hybrid ensemble framework that combines the predictive capabilities of Random Forest and XGBoost with residual-based analysis to identify statistically unusual observations in heterogeneous environmental datasets. Random Forest and XGBoost are among the most widely used ensemble-learning algorithms because of their robustness and predictive performance across diverse application domains^{11,12}.

Department of Computer Science, Faculty of Information Technologies and Mathematics, Borys Grinchenko Kyiv Metropolitan University, Levka Lukyanenka St., 13-B, Room 517, Kyiv 04053, Ukraine. ✉email: t.nosenko@kubg.edu.ua

Methods

Dataset

The study is based on open environmental monitoring data of surface-water quality in Ukraine for the year 2022. The hydrochemical dataset used in this study was obtained from the publicly available Surface Water Monitoring Database of the State Water Resources Agency of Ukraine (Derzhavne ahentsvo vodnykh resursiv Ukrayiny, n.d.)¹³. The data are organized as time series with monthly observations collected at multiple monitoring sites distributed across the country.

The monitoring network covers multiple hydrographic regions of Ukraine, including the Dnipro, Danube, Southern Bug, Dniester, Western Bug, and other river basins. Such spatial coverage provides a diverse set of environmental conditions, hydrological characteristics, and anthropogenic pressures, making the dataset suitable for evaluating machine-learning approaches under heterogeneous monitoring conditions.

The original monitoring records include measurements of numerous physicochemical indicators routinely used for environmental assessment. Among the available parameters are total nitrogen, dissolved oxygen, biochemical oxygen demand (BOD5), suspended solids, sulfate ions, chloride ions, ammonium ions, nitrate ions, nitrite ions, phosphate ions, chemical oxygen demand (COD), and additional hydrochemical characteristics. These indicators reflect both natural environmental processes and anthropogenic influences affecting surface-water quality.

The original dataset contained 10,953 records collected from 611 monitoring sites. Missing values were present in the raw dataset and were handled through the removal of incomplete observations to ensure consistency of the machine-learning analysis. Records with missing values were removed separately for each target variable using the corresponding target and predictor variables. After the cleaning procedure, 2,570 records from 104 monitoring sites were available for nitrogen modeling, while 4,295 records from 225 monitoring sites were available for dissolved oxygen modeling. The difference in sample size reflects variable-specific data completeness.

The resulting dataset represents a large-scale environmental monitoring resource characterized by substantial spatial variability and heterogeneous environmental conditions. These characteristics make it particularly suitable for evaluating the robustness of machine-learning models and investigating anomaly-detection approaches in real-world water-quality monitoring applications.

Data preprocessing and feature selection

The dataset was merged from monthly files into a unified structure and converted into numerical format. Selected input features included chemical indicators (e.g., nitrate, phosphate, biochemical oxygen demand), spatial attributes (geographical coordinates), and seasonal variables representing temporal patterns. Feature selection was performed based on domain relevance and data availability. No additional normalization was applied, as tree-based models are robust to feature scaling.

The same predictor variables were used for both target models: biochemical oxygen demand (BSK5), phosphate concentration (Fosfat), nitrate concentration (Nitrat), month of observation, latitude, and longitude. Two ensemble regression models were evaluated: Random Forest and XGBoost. The Random Forest model was configured with `n_estimators = 100` and `random_state = 42`. The XGBoost model was configured with `n_estimators = 100` and `random_state = 42`. All remaining hyperparameters were retained at their default library settings.

The dataset for each target variable was divided into training and testing subsets using an 80/20 split with `random_state = 42`. Model robustness was additionally assessed using 5-fold cross-validation. Model performance was evaluated using the coefficient of determination (R^2) and root mean square error (RMSE). Residuals were calculated on the test subset as the difference between observed and predicted values. The residual standard deviation (σ) was estimated from these test-set residuals, and observations with residuals exceeding $\pm 2\sigma$ were flagged as anomalies. Thus, anomaly detection was applied only to the test subset and not to the training data.

All analyses were performed in Python 3.14.2 using pandas for data processing, scikit-learn version 1.9.0 for Random Forest modeling and evaluation, and XGBoost version 3.2.0 for gradient-boosting regression.

Machine learning models

Random Forest (RF) and XGBoost were selected because of their robustness and strong predictive performance in heterogeneous environmental datasets. Both algorithms are capable of capturing complex nonlinear relationships among hydrochemical variables while maintaining resistance to overfitting. In addition, tree-based ensemble methods can effectively handle environmental datasets characterized by mixed variable types, missing-value filtering, and heterogeneous spatial conditions without requiring extensive feature scaling.

Random Forest combines predictions from multiple decision trees constructed using bootstrap sampling and random feature selection, thereby improving generalization performance and reducing model variance¹¹. XGBoost employs gradient boosting with sequential tree construction and regularization mechanisms, enabling accurate modeling of nonlinear environmental processes while maintaining computational efficiency¹².

Compared with many deep-learning approaches, both RF and XGBoost provide relatively high interpretability and have demonstrated strong performance in environmental monitoring and water-quality prediction studies¹. For these reasons, the two algorithms were selected as complementary ensemble-learning approaches for the present analysis.

Both models were trained separately for nitrogen concentration and dissolved oxygen prediction. Model hyperparameters and implementation details are provided in Sect. “Data preprocessing and feature selection”.

Model evaluation

Model performance was evaluated using the coefficient of determination (R^2) and Root Mean Square Error (RMSE). In addition to the train–test evaluation, 5-fold cross-validation was performed to assess model

robustness and generalizability. The resulting performance metrics were subsequently used to compare model behavior and support the interpretation of anomaly-detection results.

Anomaly detection

Anomalies were identified based on residual analysis, where the residual is defined as the difference between the observed and predicted values. The residual for each observation is calculated as:

$$r_i = y_i - \hat{y}_i \quad (1)$$

where y_i is the observed value and $\hat{y}_i$ is the corresponding model prediction.

An observation is classified as an anomaly if the absolute value of the residual exceeds a predefined threshold:

$$|r_i| > k \cdot \sigma_r \quad (2)$$

where σ_r is the standard deviation of the residuals and $k = 2$ in the present study.

The $\pm 2\sigma$ threshold was adopted as a pragmatic operational screening rule for identifying statistically unusual observations. It is intended to flag potential statistical outliers rather than confirm environmental pollution events and should not be interpreted as a universally optimal anomaly-detection criterion.

Results

Residual analysis allowed for the identification of anomalous observations that significantly deviated from model predictions (Fig. 1). These deviations represent statistically unusual observations that differ substantially from expected values and therefore merit additional environmental assessment and verification.

Anomalous observations that deviated substantially from model predictions were identified through residual analysis (Fig. 1). To evaluate the robustness of the detected anomalies, model performance was assessed using R^2 , RMSE, anomaly counts, and 5-fold cross-validation (Table 1).

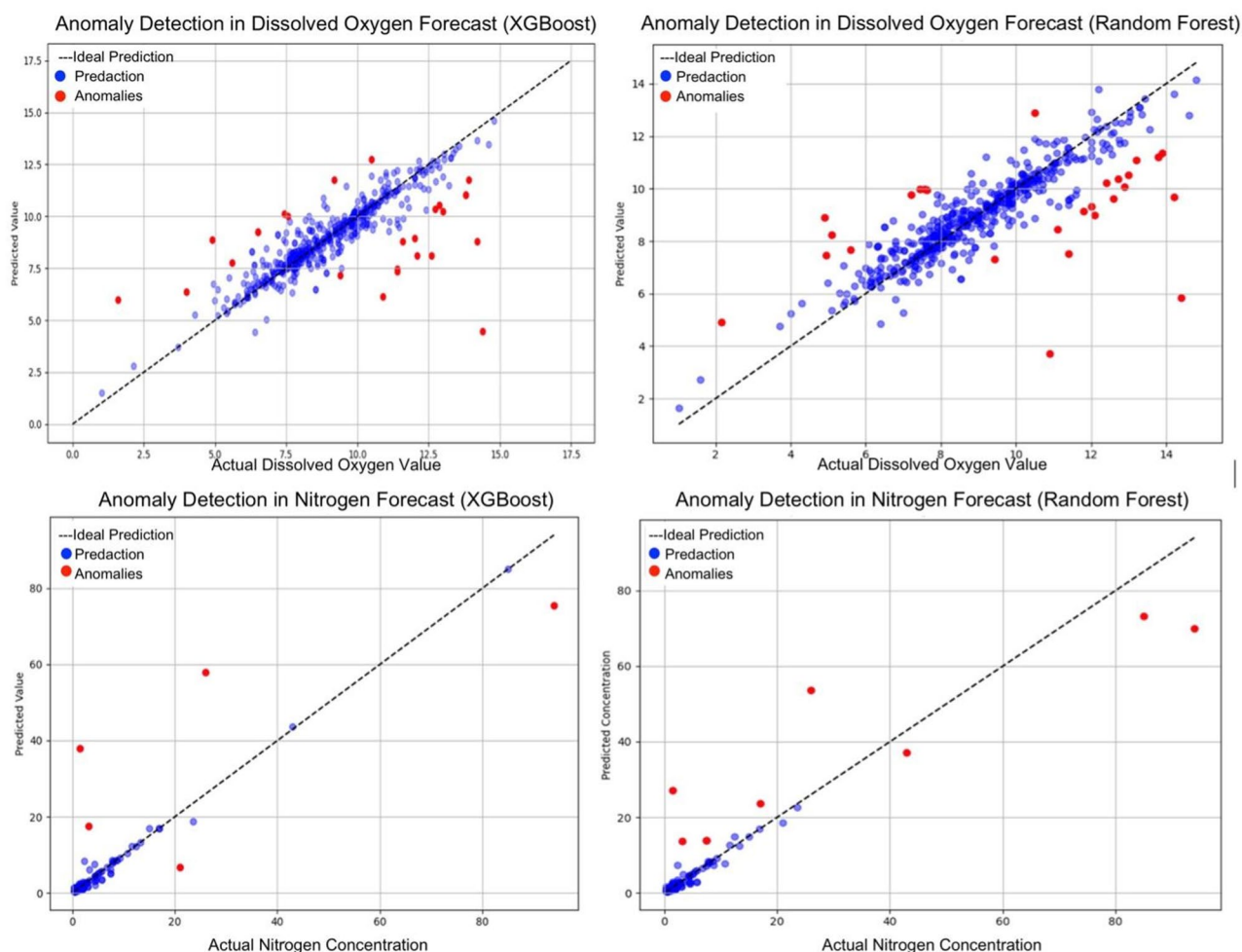


Fig. 1. Prediction results for nitrogen concentration and dissolved oxygen using Random Forest and XGBoost models. Blue points represent regular observations and red points indicate anomalies identified using the $\pm 2\sigma$ residual threshold. Model performance metrics (R^2 , RMSE) and sample sizes are reported in Table 1.

Indicator	Model	<i>n</i>	Sites	<i>R</i> ² , mean ± SD	RMSE, mean ± SD
Nitrogen/Azot	Random Forest	2570	104	0.975 ± 0.020	0.914 ± 0.383
Nitrogen/Azot	XGBoost	2570	104	0.980 ± 0.025	0.741 ± 0.618
Dissolved oxygen/Kisen	Random Forest	4295	225	0.897 ± 0.028	0.714 ± 0.119
Dissolved oxygen/Kisen	XGBoost	4295	225	0.912 ± 0.037	0.655 ± 0.161

Table 1. Results for predicting nitrogen and dissolved oxygen indicators using Random Forest and XGBoost models.

Monitoring site	RF-N	RF-O	XGB-N	XGB-O	Total detections
Yahorlyk River (Artyrivka, Moldova border)	2	2	2	2	8
Ros River (Korsun-Shevchenkivskiy drinking-water intake)	0	3	0	3	6
Southern Bug River (Ladyzhyn drinking-water intake)	0	4	0	2	6
Sarata River (Miniailivka, Moldova border)	2	2	0	2	6
Kropyvnyk River (Dombrovskiy quarry impact zone)	2	2	0	2	6
Kohyllyk River (Serpneve, Moldova border)	1	2	0	2	5
Western Bug River (Kamyanka-Buzka intake)	2	0	2	0	4
Inhulets River (Inhulets city)	0	2	0	2	4
Danube River (Kiliya drinking-water intake)	2	0	2	0	4
Khomora River (Polonne drinking-water intake)	0	2	0	2	4
Turunchuk River (Troitske, Moldova border)	0	2	0	2	4
Styr River (Zarichne, Belarus border)	0	2	0	2	4
Sluch River (Khmelnyskiy drinking-water intake)	0	2	0	2	4
Kyrhyzh-Kytai River (Malyi Yaroslavets, Moldova border)	2	0	2	0	4
Sluch River (Novohrad-Volynskiy drinking-water intake)	0	2	0	2	4
Southern Bug River (Kalynivka drinking-water intake)	0	2	0	2	4

Table 2. Recurrent anomaly hotspots ranked by cumulative detection frequency across machine-learning models.

Although the single train–test split favored Random Forest for nitrogen prediction, the 5-fold cross-validation results indicated slightly more stable overall performance for XGBoost. For nitrogen prediction, XGBoost achieved the highest performance ($R^2 = 0.980 \pm 0.025$; $RMSE = 0.741 \pm 0.618$), followed by Random Forest ($R^2 = 0.975 \pm 0.020$; $RMSE = 0.914 \pm 0.383$). For dissolved oxygen, XGBoost also slightly outperformed Random Forest, achieving $R^2 = 0.912 \pm 0.037$ and $RMSE = 0.655 \pm 0.161$.

Residual analysis was performed for each indicator to identify observations that deviated substantially from model predictions. For dissolved oxygen, both machine-learning models demonstrated comparable predictive performance, while the Random Forest model identified a larger number of anomalous observations than XGBoost. For nitrogen, a smaller set of recurrent anomaly locations was identified. The spatial distribution of anomalies differed between indicators: nitrogen hotspots were concentrated at a limited number of monitoring sites, whereas dissolved oxygen hotspots were observed across multiple river basins of Ukraine. Several monitoring locations were repeatedly identified as anomalous by different machine-learning models, increasing confidence that these locations represent robust statistical anomalies rather than model-specific artefacts.

The recurrent anomaly hotspots identified by the machine-learning models are summarized in Table 2.

The cross-validation results indicate that model performance remained stable across different folds, suggesting that the reported results are not driven by a specific train–test partition. To enhance the reliability of spatial interpretation, anomaly locations independently identified by both Random Forest and XGBoost models were retained for mapping. This approach reduces the influence of model-specific artefacts and highlights locations consistently recognized as anomalous by different machine-learning techniques.

For nitrogen concentration, four monitoring locations were simultaneously detected by both models: the Yahorlyk River, Kyrhyzh-Kytai River, Danube River (Kiliya monitoring station), and Western Bug River (Kamyanka-Buzka monitoring station). These locations represent the most robust nitrogen anomaly hotspots identified in the study (Fig. 2A).

For dissolved oxygen, thirteen monitoring locations were confirmed by both Random Forest and XGBoost models (Fig. 2B). The recurrent hotspots include sections of the Southern Bug, Ros, Sluch, Styr, Turunchuk, Yahorlyk, Kohyllyk, Sarata, Khomora, Inhulets, and Kropyvnyk rivers. Several hotspots are located within drinking-water abstraction zones, while others occur in transboundary river sections and regions with different environmental settings.

The spatial distribution of confirmed hotspots demonstrates that anomalous observations are not concentrated within a single hydrographic region but occur across multiple river basins of Ukraine.

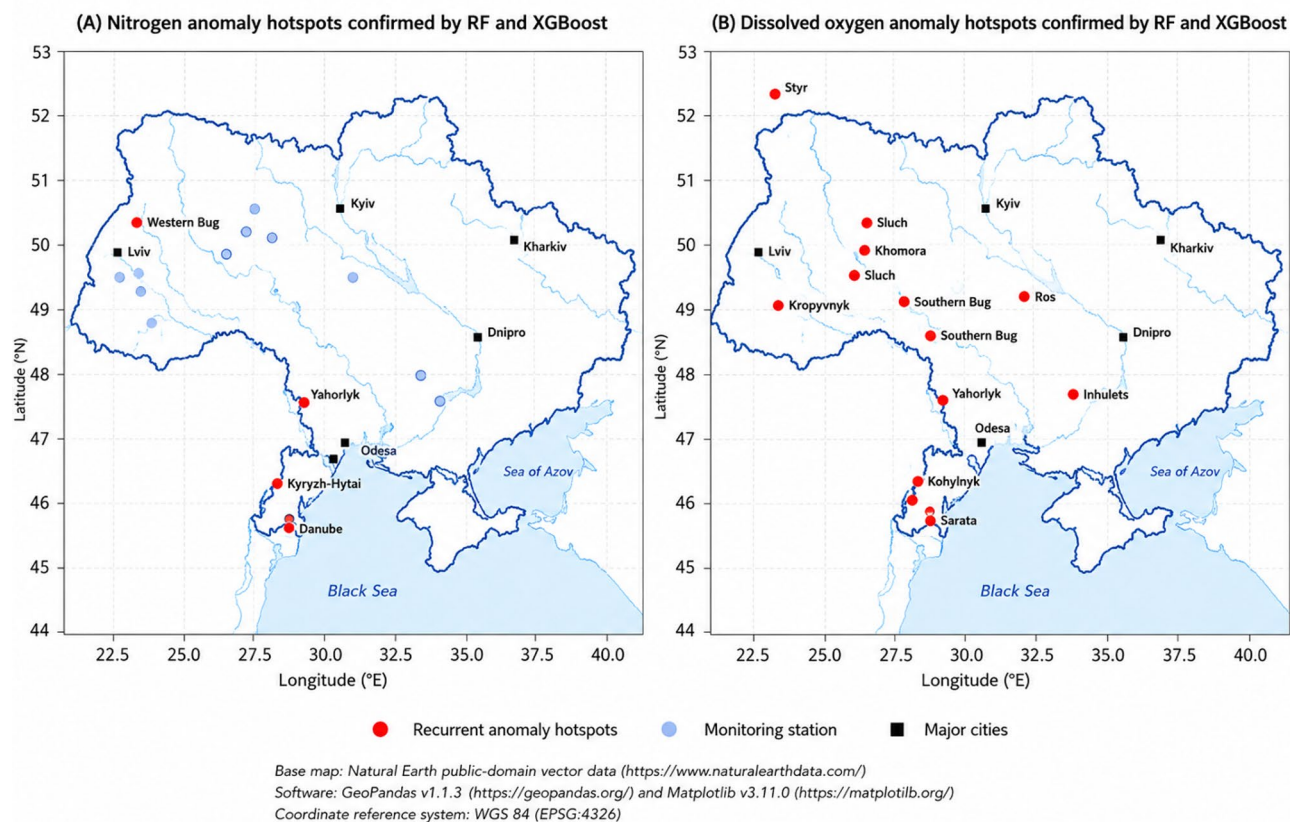


Fig. 2. Spatial distribution of anomaly hotspots confirmed by independent machine-learning models. (A) Nitrogen anomaly hotspots independently identified by both Random Forest and XGBoost models. (B) Dissolved oxygen anomaly hotspots independently identified by both Random Forest and XGBoost models. Major river networks, monitoring stations, and selected reference cities are shown to provide geographic context. Only locations confirmed by both machine-learning approaches were retained for spatial interpretation. Base map created by the authors using Natural Earth public-domain vector data (<https://www.naturalearthdata.com/>), GeoPandas v1.1.3 (<https://geopandas.org/>), and Matplotlib v3.11.0 (<https://matplotlib.org/>).

The agreement between two independent machine-learning approaches increases confidence that the detected hotspots represent robust statistical anomalies rather than artefacts of a specific predictive model. Some anomalous locations are situated near urbanized or industrialized areas. However, the present study does not include independent pollution-source data, and therefore no causal attribution can be established.

It should be emphasized that the identified hotspots indicate locations exhibiting unusual statistical behaviour in the monitored water-quality indicators. The presented analysis does not establish causal relationships with specific pollution sources and therefore requires further field investigation and environmental assessment.

The results demonstrate that combining ensemble machine-learning models (Random Forest and XGBoost) with residual-based anomaly detection provides an effective framework for identifying statistically unusual observations in environmental monitoring data. Recurrent anomaly hotspots detected by independent models indicate locations that warrant further investigation and monitoring. The proposed approach is transferable to other environmental domains, including air quality, soil monitoring, and ecological assessment, where large-scale observational datasets are available. These findings highlight the potential of machine-learning methods to support environmental monitoring and the prioritization of locations for additional investigation and monitoring.

Discussion

The results demonstrate that ensemble machine-learning models can effectively identify statistically unusual observations in large-scale environmental monitoring datasets. Both Random Forest and XGBoost achieved high predictive performance for nitrogen concentration and dissolved oxygen, while simultaneously providing a basis for residual-based anomaly detection. The agreement between independent machine-learning models increased confidence in the identified anomaly hotspots and reduced the likelihood that the detected observations were artefacts of a particular modeling approach.

The cross-validation results indicate that model performance remained stable across different folds of the dataset. This finding suggests that the reported predictive accuracy was not strongly dependent on a specific train–test partition and supports the robustness of the proposed framework. The relatively small variability of R^2 and RMSE across folds further indicates that the selected predictor variables captured a substantial portion of the variability in the target water-quality indicators.

The obtained results are consistent with recent studies demonstrating the effectiveness of machine-learning approaches for environmental monitoring and water-quality assessment. Ahmed¹, reported that ensemble learning methods frequently outperform conventional statistical approaches when complex nonlinear relationships exist among environmental variables. Similarly, recent studies have shown that machine-learning techniques can successfully model heterogeneous hydrochemical datasets characterized by spatial variability, incomplete observations, and non-stationary environmental conditions⁶.

Unlike many previous studies that focus primarily on prediction accuracy, the present work emphasizes anomaly identification through residual analysis. This distinction is important because environmental management often requires the identification of unusual observations that may indicate emerging environmental issues, measurement inconsistencies, or localized disturbances. In this context, residual-based anomaly detection provides an interpretable mechanism for prioritizing observations and monitoring locations for further investigation.

The identified anomaly hotspots were distributed across multiple river basins rather than being concentrated within a single geographic region. Such spatial dispersion suggests that anomalous observations may arise under diverse environmental conditions and monitoring contexts. The results therefore support the usefulness of nationwide monitoring datasets for identifying locations that warrant additional attention from environmental agencies and researchers.

An additional strength of the present study is the use of a nationwide monitoring dataset covering diverse hydrochemical and geographical conditions across Ukraine. Unlike many previous investigations that focus on individual river systems, local catchments, or limited regional datasets, the analyzed data encompass multiple river basins characterized by different environmental pressures, hydrological regimes, and water-quality conditions. Such diversity provides a more realistic testing environment for machine-learning models and supports the evaluation of their robustness under heterogeneous monitoring conditions. Consequently, the proposed framework may offer broader applicability for large-scale environmental monitoring programs than approaches developed and validated using geographically restricted datasets.

Recent developments in anomaly detection increasingly rely on deep-learning architectures and multivariate approaches^{7,10}. Although such methods may provide additional flexibility for modeling complex data structures, they often require substantially larger datasets, higher computational resources, and reduced model interpretability. In contrast, the combination of Random Forest, XGBoost, and residual analysis used in the present study provides a transparent and computationally efficient framework that can be readily implemented in practical environmental-monitoring workflows.

An additional strength of the proposed framework is its reproducibility. The complete analysis workflow, including the processed dataset, source code, software requirements, and validation procedures, has been made publicly available through a version-controlled repository archived with a permanent DOI. This approach supports transparent research practices and facilitates independent verification of the reported results.

Overall, the findings suggest that ensemble machine-learning models combined with residual-based anomaly detection can serve as a practical analytical tool for the exploration of large-scale environmental monitoring datasets. Rather than replacing conventional environmental assessment procedures, such methods can complement existing monitoring programs by highlighting statistically unusual observations and supporting the prioritization of locations for additional investigation.

Limitations and future work

Several limitations of the present study should be acknowledged. First, the analysis was conducted using monitoring data from a single year (2022). Although the dataset provides broad spatial coverage across Ukraine, interannual variability in hydrological conditions, climatic factors, and anthropogenic pressures was not explicitly evaluated. Future studies should investigate multi-year datasets to assess the temporal stability of the identified anomaly patterns and the long-term robustness of the proposed framework.

Second, the study focused on two water-quality indicators, namely nitrogen concentration and dissolved oxygen. While these indicators are environmentally important and widely monitored, river ecosystems are influenced by a much broader set of physical, chemical, and biological parameters. Future research could extend the proposed methodology to additional indicators, including phosphorus compounds, heavy metals, organic pollutants, microbiological parameters, and composite water-quality indices.

Third, the anomaly-detection procedure relied on a pragmatic statistical threshold ($\pm 2\sigma$) applied to model residuals. Although this approach provides a transparent and interpretable mechanism for identifying unusual observations, it should not be considered a universally optimal anomaly-detection criterion. Different environmental datasets may require alternative threshold-selection strategies. Future work could compare residual-based screening with quantile-based approaches, probabilistic anomaly-detection methods, Isolation Forest, One-Class SVM, and other modern anomaly-detection techniques.

Another limitation arises from the observational nature of the monitoring data. The identified hotspots indicate locations exhibiting statistically unusual behavior in the analyzed indicators; however, the present study does not establish causal relationships with specific pollution sources, hydrological processes, or anthropogenic activities. Additional environmental investigations, field measurements, and independent datasets would be required to determine the mechanisms underlying the detected anomalies.

The analyzed dataset corresponds to 2022, a period characterized by wartime conditions in Ukraine. Consequently, some detected anomalies may reflect temporary environmental disturbances that differ from long-term ecological patterns. Future studies incorporating post-war monitoring data will be important for evaluating the persistence and environmental significance of the identified hotspots.

Furthermore, the proposed framework treats observations independently and does not explicitly model temporal dependencies or spatial interactions among monitoring locations. Recent developments in environmental informatics increasingly employ spatiotemporal machine-learning models, graph-based approaches, and deep-learning architectures capable of capturing complex dependencies across monitoring networks. Investigating such methods represents a promising direction for future research.

Another area for future work involves uncertainty quantification. Although cross-validation was used to assess model robustness, additional uncertainty-analysis techniques could provide deeper insights into prediction reliability and anomaly confidence. Bayesian machine-learning approaches, ensemble uncertainty estimation, and probabilistic forecasting methods may offer valuable extensions of the current framework.

Finally, the present study demonstrates the applicability of ensemble machine-learning models for anomaly detection in river-water monitoring data. Future research may explore the transferability of the proposed framework to other environmental domains, including groundwater monitoring, air-quality assessment, soil contamination studies, and integrated environmental-observation systems. Such investigations would help determine the generalizability of residual-based anomaly detection across diverse environmental-monitoring applications.

Conclusions

This study demonstrated that the combination of ensemble machine-learning models (Random Forest and XGBoost) with residual-based anomaly detection provides an effective framework for identifying statistically unusual observations in hydrochemical monitoring data. Both models achieved stable predictive performance for nitrogen concentration and dissolved oxygen, with 5-fold cross-validation confirming the robustness and generalizability of the obtained results.

Residual analysis enabled the identification of anomalous observations that deviated substantially from expected patterns. The integration of anomaly detection with spatial analysis revealed a set of recurrent hotspots that were independently confirmed by both Random Forest and XGBoost models. These locations represent robust statistical anomalies that warrant further environmental investigation and monitoring.

The detected anomalies should not be interpreted as confirmed pollution events. Although several hotspots were located near urbanized, industrialized, or transboundary river sections, the present study did not include independent pollution-source information and therefore does not allow causal attribution.

The proposed approach is particularly useful in environmental monitoring scenarios where labeled anomaly data are unavailable. The combination of machine-learning prediction, residual analysis, and spatial hotspot identification can support data-driven environmental assessment and prioritization of locations for further investigation.

Future research should investigate the incorporation of spatiotemporal dependencies into the predictive modeling framework. Although the present study treated observations independently, river-water quality is inherently influenced by both temporal dynamics and spatial interactions among monitoring locations. Advanced machine-learning approaches, including recurrent neural networks, graph-based methods, and spatiotemporal modeling techniques, may provide additional insights into the evolution and propagation of water-quality anomalies across monitoring networks.

Further improvements may be achieved by expanding the feature space through the integration of additional environmental variables, such as meteorological conditions, hydrological characteristics, land-use information, and remote-sensing observations. Such extensions could improve the ability of machine-learning models to distinguish natural environmental variability from potential anthropogenic disturbances.

Future studies should also investigate alternative anomaly-detection strategies and uncertainty-quantification approaches, including probabilistic and Bayesian frameworks. In addition, independent validation using external environmental records, where available, would provide further evidence regarding the practical relevance of detected anomaly hotspots.

Finally, the transferability of the proposed framework should be evaluated in other environmental-monitoring applications, including groundwater-quality assessment, air-quality monitoring, soil-contamination studies, and integrated environmental-observation systems. Such investigations would help determine the generalizability of residual-based anomaly detection across diverse environmental domains.

Data availability

Data and Materials Availability The datasets analyzed during the current study are derived from publicly available repositories of the state monitoring of surface waters in Ukraine (Derzhavne ahentsvo vodnykh resursiv Ukrainy. (n.d.)¹³. Surface water monitoring data [Data set]. Unified State Open Data Portal. Retrieved May 12, 2025, from <https://data.gov.ua/dataset/surface-water-monitoring>). The processed dataset used for the machine learning models is available from the corresponding author upon reasonable request.

Code availability

The source code used for data preprocessing, model validation, cross-validation, and anomaly detection is publicly available in a GitHub repository archived through Zenodo: DOI: <https://doi.org/10.5281/zenodo.20582393>. The repository contains the processed dataset, source code, software requirements, and validation results necessary to reproduce the analyses reported in this study.

Received: 15 February 2026; Accepted: 23 July 2026

Published online: 29 July 2026

References

1. Ahmed, A. N. et al. Machine learning methods for better water quality prediction. *J. Hydrol.* **578**, Article 124084. <https://doi.org/10.1016/j.jhydrol.2019.124084> (2019).
2. Farouk, M. I. H. Z., Jamil, Z. & Latip, M. F. A. Towards online surface water quality monitoring technology: A review. *Environ. Res.* **238**, Article 117147. <https://doi.org/10.1016/j.envres.2023.117147> (2023).
3. Rodionov, V. et al. Combined nonthermal plasma and carbon sorption in water decontamination from persistent organic pollutants. In *9th International Conference on Carbon for Energy Storage and Environmental Protection*. (2023).
4. Zabulonov, Y. L., Kadoshnikov, V. M. & Melnychenko, T. I. Koloidni nanodispersiji dlja ocyshennja tekhnoghenno zabrudnennykh vod [Colloidal nanodispersions for the purification of technogenically contaminated waters]. In VII naučna konferenciya Fundamentalni ta prykladni doslidzhennja u sučasnij nauci (65). *Tehnologichnyj Tsent*. (2019).
5. Lin, J. et al. An extensive spatiotemporal water quality dataset covering four decades (1980–2022) in China. *Earth Syst. Sci. Data.* **16** (2), 1137–1149. <https://doi.org/10.5194/essd-16-1137-2024> (2024).
6. Wu, R. et al. Spatiotemporal variation in water quality and identification and quantification of areas sensitive to water quality in Hulun Lake, China. *Ecol. Indic.* **149**, 110176. <https://doi.org/10.1016/j.ecolind.2023.110176> (2023).
7. Pang, G., Shen, C., Cao, L. & Hengel, A. V. D. Deep learning for anomaly detection: A review. *ACM Comput. Surveys.* **54** (2), 1–38. <https://doi.org/10.1145/3439950> (2021).
8. Al Farizi, W. S., Hidayah, I. & Rizal, M. N. Isolation forest-based anomaly detection: A systematic literature review. In 2021 8th International Conference on Information Technology, Computer and Electrical Engineering (ICITACEE) (pp. 118–122). IEEE. (2021)., September <https://doi.org/10.1109/ICITACEE53184.2021.9613869>
9. Aschepkov, V. The use of the Isolation Forest model for anomaly detection in measurement data. *Innovative Technol. Sci. Solutions Industries.* **1** (27), 236–243. <https://doi.org/10.30837/ITSSI.2024.27.236> (2024).
10. Rigueira, X., Olivieri, D., Araujo, M., Saavedra, A. & Pazo, M. Multivariate functional data analysis and machine learning methods for anomaly detection in water quality sensor data. *Environ. Model. Softw.* <https://doi.org/10.1016/j.envsoft.2024.106443> (2025)., 190, Article 106443.
11. Breiman, L. Random forests. *Mach. Learn.* **45** (1), 5–32. <https://doi.org/10.1023/A:1010933404324> (2001).
12. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). ACM. (2016). August <https://doi.org/10.1145/2939672.2939785>
13. Derzhavne ahentsvo vodnykh resursiv Ukrainy. (n.d.). Surface water monitoring data [Data set]. Unified State Open Data Portal. Retrieved May 12, from 2025. <https://data.gov.ua/dataset/surface-water-monitoring>

Author contributions

Tetiana Nosenko: Methodology; Software; Data Curation; Formal analysis; Visualization; Writing – original draft. Iryna Mashkina: Conceptualization; Validation; Formal analysis; Writing – review & editing; Supervision. All authors have read and agreed to the published version of the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to T.N.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2026