Machine learning-based in-situ detection of toxic petroleum hydrocarbons in groundwater

Wu, C. L.R.; Wagterveld, R. M.; Rietveld, L. C.; van Breukelen, B. M.

DOI
10.1016/j.jconhyd.2025.104771

Publication date
2025

Document Version
Final published version

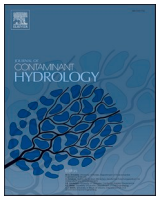
Published in
Journal of Contaminant Hydrology

Journal of Contaminant Hydrology 276 (2026) 104771

Contents lists available at ScienceDirect

Journal of Contaminant Hydrology

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



Machine learning-based in-situ detection of toxic petroleum hydrocarbons in groundwater

C.L.R. Wu [a,b,*], R.M. Wagterveld [a], L.C. Rietveld [b], B.M. van Breukelen [b]

[a] Wetsus, European Centre of Excellence for Sustainable Water Technology, Oostergoweg 9, 8911 MA Leeuwarden, the Netherlands
[b] Department of Water Management, Faculty of Civil Engineering and Geosciences, Delft University of Technology, 2628 CN Delft, the Netherlands

ARTICLE INFO

Keywords:
BTEX
Real-time monitoring
Sensor fusion
Groundwater pollution
Early warning system
Data-driven modeling

ABSTRACT

Monitored natural attenuation is commonly used to manage petroleum hydrocarbon-contaminated groundwater. However, it requires periodic, costly grab sampling. We propose a cost-effective, real-time groundwater monitoring proof-of-concept machine learning (ML) framework using in-situ sensors—pH, dissolved oxygen, electrical conductivity, and redox potential—to detect benzene, ethylbenzene, and xylenes (BEX). We built upon the established correlations between hydrocarbon concentrations and in-situ water quality parameters (iWQPs). Due to limited field data, we validated the framework using datasets at virtual wells within a simulated aquifer from our previously developed reactive transport model. In this application, we detected the spreading of pollution downstream of the established pollution plume. The used framework is a binary classification system that flags contamination at virtual downstream wells. We compared five ML classifiers, i.e. Logistic Regression, Random Forest, XGBoost, Multi-layer Perceptron, and Support Vector Classifier, for early warning when BEX reached or exceeded the regulatory threshold of 5 μg/L. The models were trained on virtual wells at and near the source zone and predicted contamination before BEX reached the threshold at downstream virtual wells. This reflects the spatial variability in flow and reaction dynamics that altered BEX-iWQP relationships. Scenario analyses revealed the ML models' sensitivity to aquifer properties, i.e., hydraulic conductivity, electrical conductivity, and electron acceptor availability. We also assessed the impact of sensor noise and seasonal fluctuations on iWQPs. We found that even moderate levels of noise (10–20 %) can significantly affect model accuracy, particularly when the noise was introduced into the test data. Therefore, we recommended to combine hardware stabilization with adaptive smoothing techniques. With these approaches, our proposed framework remains promising for providing early warnings of plume migration toward sensitive receptors.

1. Introduction

Petroleum hydrocarbons (PHCs) are widely used across various sectors, including residential, agricultural, and transportation industries. Among these compounds, aromatic hydrocarbons such as benzene, toluene, ethylbenzene, and xylenes (BTEX), are particularly hazardous due to their toxicity and persistence in groundwater (Li et al., 2021). These contaminants often infiltrate groundwater through landfill leachate, leaks from underground storage tanks, and industrial discharges (Haider et al., 2021). To safeguard water resources, agencies such as the U.S. Environmental Protection Agency (EPA) established strict guidelines for permissible levels of organic contaminants in water systems (U.S. Environmental Protection Agency, 2024).

Conventional PHC monitoring in groundwater uses manual grab sampling and laboratory analysis. While accurate, this approach is costly, infrequent, and lacks the capability to provide real-time data. These limitations are especially crucial when Monitored Natural Attenuation (MNA) is the remediation strategy. MNA relies on naturally occurring physical, chemical, and biological processes within the soil and groundwater to reduce contamination levels, offering a cost-effective and environment-friendly alternative to immediate full-scale remediation (Beck and Mann, 2010).

Under MNA, contaminant plumes are expected to shrink over time, resulting in relatively stable conditions with respect to receptor exposure. Contaminant concentrations should not exceed the compliance limits at designated warning wells, located between the contaminant source and the receptor areas. Receptors encompass both human populations and ecological systems dependent on groundwater resources

* Corresponding author at: Wetsus, European Centre of Excellence for Sustainable Water Technology, Oostergoweg 9, 8911 MA Leeuwarden, the Netherlands.
E-mail address: c.l.wu@tudelft.nl (C.L.R. Wu).

https://doi.org/10.1016/j.jconhyd.2025.104771
Received 28 July 2025; Received in revised form 10 October 2025; Accepted 1 November 2025
Available online 3 November 2025

(McKnight et al., 2010). Exceeding contaminant thresholds at warning wells may signal an unaccounted pollution source, requiring early warning systems and prompt remediation to protect receptors.

Consequently, there is a need for cost-effective, continuous monitoring systems capable of real-time PHC detection in groundwater. Advances in sensor technology and machine learning (ML) offer promising solutions. Emerging portable sensors are being developed for field-based detection of PHCs (Cova et al., 2022). But these remain costly, require complex sample preparation, and have slow response times.

ML models and sensor fusion have predicted contamination in both surface and groundwater using environmental and hydrogeological variables. LightGBM and XGBoost (XGB) effectively predicted *E. coli* levels on three Lake Erie beach sites (Li et al., 2022). Logistic regression (LR), random forest (RF), and support vector machines have estimated contamination probabilities of specific pollutants in groundwater. These include using static well measurements to predict nitrate and arsenic contamination (Singh et al., 2021). Linear and RF regressors were also used to predict and identify wells with high PFAS levels in California groundwater using co-contaminant fingerprints, hydrological and soil properties, proximity to pollution sources, and geospatial data (George and Dixit, 2021). Additionally, continuous sensor data integrated with Kalman filter had estimated tritium and uranium concentrations, enabling real-time plume tracking (Schmidt et al., 2018a).

Despite these advancements, a significant gap remains in using continuous data from conventional, low-cost sensors to detect PHCs at warning wells. Although Li et al. (2017) monitored groundwater for potential oil and gas contamination using sensor data, PHCs were not directly detected; instead, anomalies were identified by comparing new data to historical records, without converting data to PHC levels. Furthermore, integrating sensor data with ML models for real-time PHC detection has not been explored. For MNA, early warning of contaminant migration toward the receptors is essential for triggering active remediation measures.

Indirect detection of PHCs is possible by monitoring their degradation processes, which influence groundwater quality through various terminal electron-accepting processes (Wu et al., 2024). Studies have shown that following an inland oil spill, these degradation reactions drive measurable changes in geochemical conditions. To better understand these processes at contaminated sites, reactive transport models (RTMs) have been developed (e.g., Ng et al., 2015).

In an earlier study, we developed an RTM to examine how PHC degradation affects groundwater quality under hypothetical yet representative conditions. We simulated a stationary oil source dissolving at the top of a heterogeneous, shallow sandy aquifer in two dimensions. The model showed the spatiotemporal evolution of dissolved PHCs under various realistic conditions and revealed correlations between PHC concentrations and in-situ water quality parameters (iWQPs), namely pH, dissolved oxygen (DO), electrical conductivity (EC), and redox potential (ORP) (Wu et al., 2024). ORP in water measures the tendency of a chemical species to gain (reduction) or lose (oxidation) electrons (Copeland and Lytle, 2014). These findings suggested that conventional water quality sensors could detect PHCs in groundwater.

Qiao et al. (2025) have recently presented a study that closely parallels the methods and frameworks introduced in our earlier work (Wu et al., 2024). They made minor modifications to our original Python-based RTM to generate the datasets required for the ML-based integration of iWQPs in estimating PHC concentrations in groundwater. While their stated objective was to "predict the spatiotemporal distribution of dissolved-phase NAPL plumes," our study focused on the early detection of benzene, ethylbenzene, and xylenes (BEX) compounds at warning wells across a range of aquifer conditions. Toluene was excluded in the BTEX group due to its tendency to degrade directly from the pollution source without dissolving (Ng et al., 2015; Wu et al., 2024).

We present a proof-of-concept ML framework that integrates data from affordable in-situ sensors to detect BEX in groundwater. BEX compounds were grouped based on their shared migration and degradation properties (Ng et al., 2015; Wu et al., 2024). The sensors include pH, DO, EC, and ORP. Due to limited field data, we validated the framework using RTM-generated datasets from virtual observation wells. We framed the task as a binary classification problem, determining whether a virtual well is contaminated ($\geq$5 μg/L BEX, based on U.S. EPA standards) or uncontaminated. To our knowledge, this is the first study to explore the use of low-cost sensor data in combination with ML to provide early warnings of PHC contamination at downstream wells. Our approach differs from prior work by directly predicting exceedances of regulatory thresholds, rather than mapping plumes or identifying anomalies. In essence, our ML framework was trained to detect contamination at warning wells and provide a timely warning of contaminant migration before it reaches sensitive receptor areas.

Additionally, we evaluated the ML models across diverse hydrogeological conditions by training and testing separate models for each RTM scenario. In our scenario simulations, we varied the amplitude of the water table fluctuations, hydraulic conductivity, and background salinity. We also modified the electron acceptor availability by removing or increasing DO, nitrate, and sulfate concentrations, and adjusted mineralogical controls via calcite presence, $Fe(OH)_3$ reduction, and cation exchange capacity. Details are provided in Tables S1, S2, and in Wu et al. (2024).

To further evaluate our ML models under realistic conditions, we introduced noise to iWQPs from our previous RTM (Wu et al., 2024). We incorporated (1) Gaussian sensor noise based on commercial specifications, and (2) sinusoidal fluctuations to simulate seasonal groundwater variations. These were introduced only to the RTM outputs. We then evaluated the models' performance on these noisy datasets by comparing BEX arrival times from both the RTM and ML predictions at virtual warning wells. To improve detection accuracy under noisy conditions, we implemented a moving-average filter with varying window sizes.

2. Methods

2.1. Reactive transport modeling as a data source

The previously developed RTM (Wu et al., 2024) simulates the dissolution, advection, dispersion, and biodegradation of BEX and non-volatile dissolved organic carbon (NVDOC) from a light non-aqueous phase liquid (LNAPL) source in a porous media aquifer. Furthermore, the model accounts for the direct degradation of other components within the source zone. In addition to organic processes, the RTM incorporates inorganic reactions such as mineral dissolution and precipitation, cation exchange, and redox-related outgassing (Wu et al., 2024). We based the model's domain and parameter values on existing studies on PHC contamination, particularly the extensively studied crude oil spill site in Bemidji, Minnesota (Ng et al., 2015). In our previous study, we examined the correlation between BEX and various in-situ water quality parameters (iWQPs), specifically pH, DO, ORP, and EC. Based on these findings, we developed ML models to detect the arrival of dissolved BEX using iWQPs at virtual warning wells located at various depths and distances from the oil source.

While synthetic datasets generated by RTM offer a controlled and informative environment for proof-of-concept development, they cannot fully capture the complexity and variability of real-world aquifers including microbial variability, contamination source changes, and sensor fouling. As such, the applicability of our ML framework to field conditions remains to be validated. Still, the use of RTM facilitates generalizability by allowing testing with diverse hydrogeochemical scenarios without the constraints of cost, safety, or other physical limitations typically associated with field studies.

2.2. Machine learning model development

In this study, we focused on a binary classification problem:

determining whether a virtual observation well is contaminated or uncontaminated. Classification problems are a type of supervised learning task where each data point is associated with a class label (Sen et al., 2020). The goal is to develop a function or a classifier that can accurately categorize data points into the correct classes based on their features. To be effective, this classifier must demonstrate both high predictive accuracy—meaning it correctly classifies most of the data it is tested on—and a low generalization error, which refers to its ability to maintain that accuracy when applied to new, unseen data (Wang and Shen, 2006). Rather than estimating exact BEX concentrations, we aimed to demonstrate how an early-warning system based on inexpensive water quality sensors can support timely field investigations during contamination events. For stakeholders, a binary output indicating the presence or absence of contamination offers a more practical and interpretable first step in risk management. This simplification, in which 5 μg/L and 9.9 mg/L are treated as the same class, was designed to prioritize rapid risk identification over accurate concentration estimation.

We implemented several of the most widely used ML classifiers, which have demonstrated their effectiveness across various domains (Ahsan et al., 2021; Li et al., 2022). We used LR, Support Vector Classifier (SVC), RF, XGB, and Multi-layer Perceptron (MLP) (Text S1). We opted for classification ML models over regression since classification models can provide probability estimates, offering deeper insights into the certainty of the predictions. Furthermore, classification algorithms can generally handle imbalanced datasets. This is particularly relevant to our virtual dataset produced with RTM given the underrepresentation of the uncontaminated class. To classify contamination levels, we used the U.S. EPA's maximum benzene concentration limit of 5 μg/L in drinking water as a threshold during ML model development; benzene is recognized as the most toxic constituent within the BEX group. In comparison, ethylbenzene and xylenes have higher allowable concentrations in drinking water, which are 700 μg/L and 10,000 μg/L, respectively (U.S. EPA, 2024).

The ML model development process starts with creating a database that captures the relationship between BEX concentration and the iWQPs. Due to limited high-resolution temporal data from contaminated sites, including the well-studied Bemidji site, we principally used our RTM to generate a comprehensive virtual dataset for model training and evaluation. We focused on the initial 5 years of simulation, when BEX

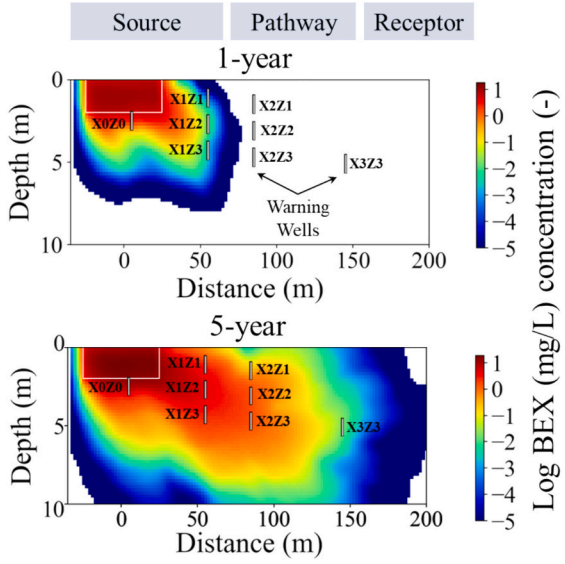


Fig. 1. BEX plume distribution at 1 and 5 years of simulation. BEX concentration is converted to base 10 log form with 1 mg/L as reference. The white square indicates the oil source zone, and the eight vertical white lines represent virtual observation wells. BEX concentration reached a 5 μg/L threshold at the farthest virtual warning well approximately 4 years into simulation.

was the primary electron donor. Fig. 1 illustrates the simulated cross-sectional BEX plume after 5 years (Wu et al., 2024). DO, pH, EC, and ORP were selected as input variables based on strong correlations with PHCs (Fig. S1). To capture spatial variability, we also included the depth and horizontal distance of each virtual observation well from the contamination source (Table S3) in both training and test datasets.

Model training and evaluation followed a two-layer approach combining Tree-Structured Parzen Estimator for hyperparameter optimization and leave-one-group-out cross-validation across four virtual training wells. In each iteration, one well was used for validation while the remaining three were used for training. A custom scoring function was used to prioritize timely detection of BEX contamination, balancing prediction accuracy (F1 score) and delay in alarm timing. Although we did not apply explicit techniques to address class imbalance, we relied on the inherent robustness of the selected classifiers to handle imbalanced datasets. Additionally, we incorporated penalty weighting for missed contamination events and required at least two consecutive detections to trigger an alarm, reducing the impact of isolated false positives. Details of the model training and hyperparameter tuning process are provided in Text S2.

To evaluate the real-world applicability, we conducted sensitivity analyses using a realistic base case and additional simulations reflecting varied but plausible hydrogeological conditions. The base case, based on typical field settings (Wu et al., 2024), served as a benchmark. Meanwhile, scenario simulations assessed the model performance under different boundary conditions, aquifer characteristics, and groundwater chemistries. We trained and tested the ML models across these diverse conditions to assess their robustness and adaptability to field-relevant scenarios.

2.3. Data preprocessing

A fundamental concept in ML is the separation of training and testing data to ensure unbiased performance evaluation (Xu and Goodacre, 2018). The training dataset is used to calibrate the models, allowing it to learn the relationships and patterns inherent in the data. The testing dataset is then used to evaluate the models and assess the model's generalization capabilities and predictive accuracy on unseen data. This process is crucial in preventing overfitting, a phenomenon where the model performs exceptionally well on the training data but poorly on new data.

Unlike random shuffling approaches that mix all available data (Zhang, 2025), we implemented a spatially structured training-testing split that reflects the natural time progression of contamination (i.e., plume migrates from upstream to downstream). This method avoids potential biases that arise from mixing data across different locations and time periods. Our models were trained on 5-year data from virtual wells near the source zone (X0Z0, X1Z1, X1Z2, X1Z3) where BEX concentrations exceed regulatory thresholds. The models were then tested on downstream virtual wells—which we identify as warning wells (X2Z1, X2Z2, X2Z3, X3Z3)—to simulate early detection in uncontaminated areas. The receptor boundary was defined beyond the furthest virtual well (X3Z3). This approach achieves our practical objective of predicting contamination migration using in-situ sensor data.

Since the iWQPs vary in scale and units, we standardized the data using the StandardScaler function from the scikit-learn library, transforming each parameter to have a mean of 0 and a standard deviation of 1. This ensures that all parameters contribute equally to model training, preventing any single feature from disproportionately influencing the results (Ahsan et al., 2021).

We categorized the data into two groups: non-contaminated (assigned a value of 0) for all data with BEX levels of 0 to below 5 μg/L, and contaminated (assigned a value of 1) for all data with BEX levels of 5 μg/L and above. This binary labeling approach (assigning a label of 0 or 1) is widely used in binary classification tasks (Li and Tong, 2020). The highest BEX level in the training data from the base case is 9.9 mg/L,

observed at the virtual observation well X0Z0.

2.4. Model evaluation

As a proof-of-concept, the performance of the ML models was assessed based on their ability to detect when BEX reached the downstream virtual observation wells (X2Z1, X2Z2, X2Z3, and X3Z3) using the virtual iWQPs. We quantified this by comparing the BEX arrival time predicted by the ML models with the actual arrival time determined by the RTM. A key criterion in our evaluation was whether the ML model could trigger an alarm when BEX contamination occurred in a virtual observation well, specifically when it reached the threshold of 5 μg/L.

When the ML model triggered an alarm, we compared its prediction to the RTM-generated data. We defined a threshold breach as the moment when BEX concentrations reached or exceeded the regulatory limit of 5 μg/L. If the RTM showed BEX concentrations above 0 μg/L (not detectable or ND) but still below 5 μg/L, we identified the event as a premature alarm, meaning the ML model predicted a potential risk before the threshold was breached. In these cases, we recorded how many days the alarm occurred before the threshold was breached. If the RTM showed BEX concentrations as ND, we considered the ML alarm a false alarm, meaning the ML model predicted contamination when none was present. If BEX concentrations from the RTM reached or exceeded 5 μg/L but the ML model did not trigger an alarm on the same day, we labeled it a missed alarm. This means that contamination was present, but the ML model failed to detect it. We evaluated missed alarms by calculating the delay in the model's response, defined as the number of days between the actual threshold breach and the ML model's alarm trigger.

2.5. Introduction of sensor noise and seasonal fluctuations in groundwater chemistry

Water quality sensors are inherently prone to measurement errors (de Winter et al., 2019). Therefore, we integrated measurement noise into our virtual data to accurately represent real-world conditions. We added zero-mean Gaussian noise to the RTM output data using fixed standard deviations that match the accuracy of commercial sensors (Table S4). This measurement noise was introduced only at the RTM output stage, not during the RTM simulation. Our approach for adding noise follows the method outlined by Zhu and Wu (2004), in which noise is injected into the training data, test data, or both to simulate measurement uncertainty (Fig. S2). We increased the noise levels in a stepwise manner, using 10 % increments, where 100 % noise corresponds to the standard deviation specified by the sensor's accuracy.

We observed the changes in iWQPs induced by BEX degradation, which is the signal of interest. While changes in iWQPs may also influence BEX degradation, we assumed first-order biodegradation kinetics in our RTM that are temporally and spatially uniform and independent of redox conditions (Wu et al., 2024). This simplification allowed us to isolate the effect of BEX degradation on iWQPs. We determined that the aquifer's initial baseline iWQP values do not significantly influence the magnitude of this signal, as the observed changes are governed primarily by BEX degradation kinetics. We therefore assumed that our noise augmentation approach—introducing noise specifically to iWQPs—adequately captures the real-world noisy measurements.

Whereas the RTM assumed that the groundwater flowing into the pollution zone had a constant chemistry, groundwater composition can vary spatially and temporally due to anthropogenic exploitation, natural mixing processes, and varying recharge conditions (Yao et al., 2024). While bulk groundwater chemistry responds slowly to these changes, depending on aquifer characteristics (Khatri and Tyagi, 2015), iWQPs are generally more sensitive to environmental fluctuations. Additionally, iWQPs are measured using sensors characteristically sensitive to (seasonally changing) temperature. While automatic temperature compensation (ATC) is typically applied, temperature fluctuations can still introduce variability in sensor readings. To simulate the effect of seasonal variations in the composition of clean upstream groundwater, we superimposed a sinusoidal annual fluctuation on the RTM-derived iWQPs. We introduced fluctuations exclusively at the RTM output stage while maintaining noise-free RTM simulations. These fluctuations were introduced solely to the training data, in contrast to sensor noise, which was added to both training and testing datasets.

To introduce sinusoidal fluctuations into the iWQPs, we first needed to determine the signal amplitude (A) to generate the synthetic sine wave for each parameter. We began by calculating the standard deviation of field measurements ($\sigma_{measurement}$) using continuous measurement data obtained from the U.S. Geological Survey (n.d.) and from a published study (Lyons et al., 2023) (Table S5). For a pure sine wave with zero mean, A is related to the standard deviation of the signal (σ_{signal}) through the root-mean-square relationship (Irvine, n.d.): $\sigma_{signal} = 1/\sqrt{2}A$.

Since $\sigma_{measurement}$ includes sensor noise, we estimated σ_{signal} by removing the noise component using the following expression: $\sigma_{signal} = \sqrt{\sigma_{measurement}{}^2 - \sigma_{sensor}{}^2}$, where the standard deviation of the sensor, σ_{sensor} corresponds to the technical accuracy of the sensor. Because the actual accuracy of the USGS sensors is unknown, we subtracted the variance based on the specifications of sensors used in our noise simulations (Table S4). For EC, the sensor accuracy is expressed as a percentage of the measured value, which complicates amplitude calculation because the resulting σ_{sensor} varies with each reading. To simplify, we used the mean EC value from the USGS data (approximately 520 μS/cm). Following the same procedure used for sensor noise, we increased the amplitude of the sinusoidal fluctuation in 10 % increments, where 100 % corresponds to the calculated amplitude A.

Thus, we evaluated the performance of our ML models under three different scenarios: 1) data with added sensor noise, 2) data with added sinusoidal annual fluctuations in clean groundwater composition, and 3) data combining both factors.

Finally, we explored various methods for filtering the data with added noise and fluctuations to enhance the predictive performance of our ML models. These included applying smoothing techniques using 3-day, 5-day, and 10-day averages. Fig. S3a, b, and c illustrate the iWQPs at virtual training well X0Z0 under conditions of 10 %, 50 %, and 100 % added sensor noise and sinusoidal fluctuations, along with the corresponding 5-day moving average smoothing.

3. Results and discussion

In this section, we present the performance evaluation of our ML models through four key assessments. First, we established the baseline accuracy by testing the models using noise-free RTM-generated base case data. The goal is to trigger a contamination alarm when BEX concentration reached 5 μg/L. While detection earlier than this threshold is generally acceptable, our models were trained to prioritize timely rather than excessively premature alarms to avoid unnecessary field responses. For example, an alarm triggered 5 days before the threshold breach is considered more desirable than one triggered 30 days early. However, there is no standardized optimal time for premature alarms, as alarms triggered far in advance may resemble false positives and can be costly in terms of operational impact.

Next, we assessed their adaptability to varying conditions using data from scenario simulations (Wu et al., 2024). We then tested their robustness by introducing artificial noise at multiple intensity levels to the base case data and quantifying the corresponding performance degradation. Finally, we evaluated the effectiveness of noise-reduction techniques in restoring the timely detection accuracy of the ML models.

To reflect the practical goal of early warning, we evaluated model performance using a custom scoring function that prioritizes timely detection of BEX contamination. The score penalizes premature alarms and missed contamination events more heavily than overall

classification accuracy. Details of this scoring approach are provided in Text S2. These assessments demonstrate the capabilities and limitations of the ML models, and the implications for real-world groundwater monitoring application.

3.1. Baseline performance evaluation

The five ML models trained to detect BEX performed differently in triggering the contamination alarm when tested on virtual warning wells. LR, RF, XGB, and MLP models triggered the alarms in all virtual warning wells before BEX concentrations reached the 5 μg/L threshold (Fig. 2). In virtual warning wells X2Z1, X2Z2, X2Z3 (located 85 m from the source zone's center), these concentrations ranged from 3.4 μg/L to 4.7 μg/L. These values correspond to predictions that predated the actual concentration threshold breach by 25 to 4 days. In contrast, SVC triggered alarms at the X2 virtual warning wells when BEX concentrations exceeded the threshold (6.4 μg/L to 7.1 μg/L), resulting in a delayed alarm trigger of 12 to 19 days.

At the farthest virtual warning well, X3Z3 (located 145 m from the source zone's center), most models triggered alarms earlier than expected. The LR model triggered the alarm 189 days earlier at a concentration of 0.6 μg/L, while the MLP model triggered the alarm 126 days earlier at 1.2 μg/L. The RF and XGB models showed moderately better performance, triggering alarms 54 days (at 2.6 μg/L) and 61 days (at 2.4 μg/L) earlier, respectively. However, the SVC performed best at this virtual warning well, triggering the alarm 18 days earlier at a concentration of 4.1 μg/L.

Most ML models predicted contamination prior to the actual event. This discrepancy likely stems from differences between the training and test datasets: the models were trained on data from source-zone virtual wells (X0Z0) and nearby virtual wells (55 m from the source center) but tested on downstream virtual wells farther from the source (warning wells). Due to varying flow and reaction dynamics, the relationships between BEX and iWQPs were not consistent across distances. Specifically, BEX arrived more quickly at closer virtual wells with higher concentrations, whereas geochemical reactions had more time to influence the iQWPs at farther virtual wells, leading to different parameter relationships (Fig. 3). This mismatch in geochemical conditions across distances affected model generalization, as seen in the premature alarms at X3Z3. The results suggest that the initial ML framework was less robust at distant wells, and that either warning wells should be placed closer to the source zone or additional training data from farther locations is needed to improve performance.

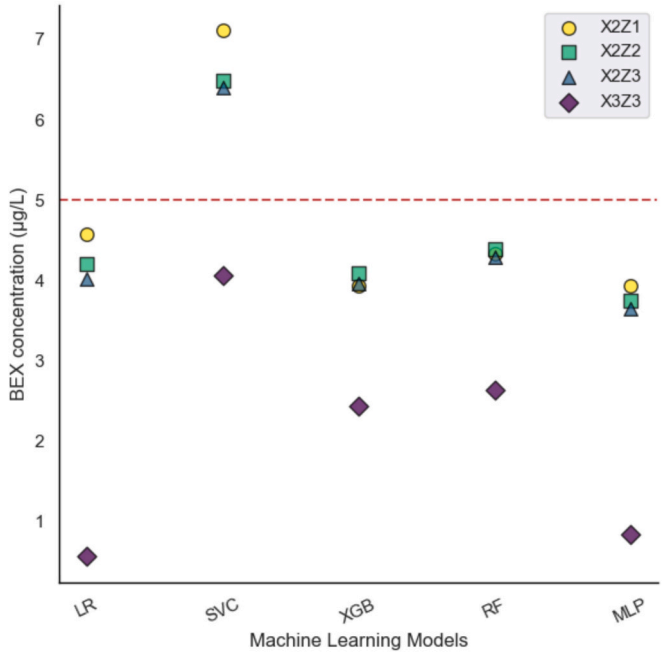


Fig. 2. Concentration of BEX when the different ML models triggered the contamination alarm at the virtual warning wells. Red dashed line represents the US EPA threshold of 5 μg/L. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Furthermore, the iWQPs were responsive to changes in BEX due to degradation rather than to its absolute concentration. As a result, the models still tended to trigger contamination alarms earlier than the set thresholds even when the training threshold was increased from 5 μg/L to 10, 50, 100, and 500 μg/L (Fig. S4).

RF and XGB performed best at all virtual warning wells. This can be attributed to their ensemble nature, which enables them to effectively capture complex, non-linear relationships in the data (Lin et al., 2022). Ensemble methods can also improve predictive performance and robustness against overfitting by combining multiple models or learners, allowing them to generalize better across different datasets.

Although LR and SVC are both linear models (with a linear kernel used for SVC), SVC was more conservative in its predictions, requiring a higher BEX concentration and thus greater certainty in exceeding the threshold before it triggered the alarm. The LR models the probability of a sample belonging to a class (i.e., contaminated vs. uncontaminated). However, it assumes a linear relationship between the independent variables and the log odds of the outcome, which may not be valid in all scenarios (Bisong, 2019). The training data may have caused the decision boundary to be close to the threshold, leading to more samples being classified as contaminated, even if their concentrations were below but near the threshold. On the other hand, SVC aims to find the hyperplane that maximizes the margin between the two classes (Vapnik, 1982). SVC focuses on samples closest to the decision boundary (support vectors), making it less sensitive to the overall data distribution, which results in the SVC being more conservative when classifying samples as contaminated.

While MLP can capture complex patterns due to its multi-layer architecture, it is also prone to overfitting (Rynkiewicz, 2019). MLP triggered an alarm at a BEX concentration of 1.2 μg/L at the virtual well X3Z3, suggesting that it did not generalize well to unseen data or underperformed with data that were outside the range of the training data. This lack of generalization is often a consequence of the model's reliance on specific training data characteristics that may not be representative of the unseen conditions.

Furthermore, due to the stochastic initialization of weights in MLP, as in other neural networks (Narkhede et al., 2022), the BEX concentration that triggered the alarm varied between runs. To account for this variability, we ran the prediction 1000 times and recorded the results. Fig. S5 shows the histogram of the BEX concentration when MLP triggered the alarm at the virtual warning wells. Most of the alarms were triggered at BEX concentration of approximately 3.3 μg/L to 3.8 μg/L for all virtual warning wells, except for the farthest virtual warning well X3Z3. At this virtual well, MLP was prone to false alarms, triggering alarms even when the BEX concentration was ND.

It is important to note that small variations in sensor data may be difficult to detect in real-life applications, and that the technical limitations of available sensors must be considered when interpreting the practical implications of these findings. Sensors typically have accuracies and resolutions of ±0.1 and ± 0.01 for pH, ±5 mV and ± 0.1 mV for ORP, ±0.5 % (of reading) and ± 0.1 μS/cm for EC, and ± 0.1 mg/L and ± 0.01 mg/L for DO (Wu et al., 2024). Despite these limitations, our results suggest that the combined response of multiple sensor parameters—even when individual deviations are small—still supported our proof-of-concept framework for an early warning system.

3.2. Scenario simulations performances

In this section, we present the results of our scenario simulations,

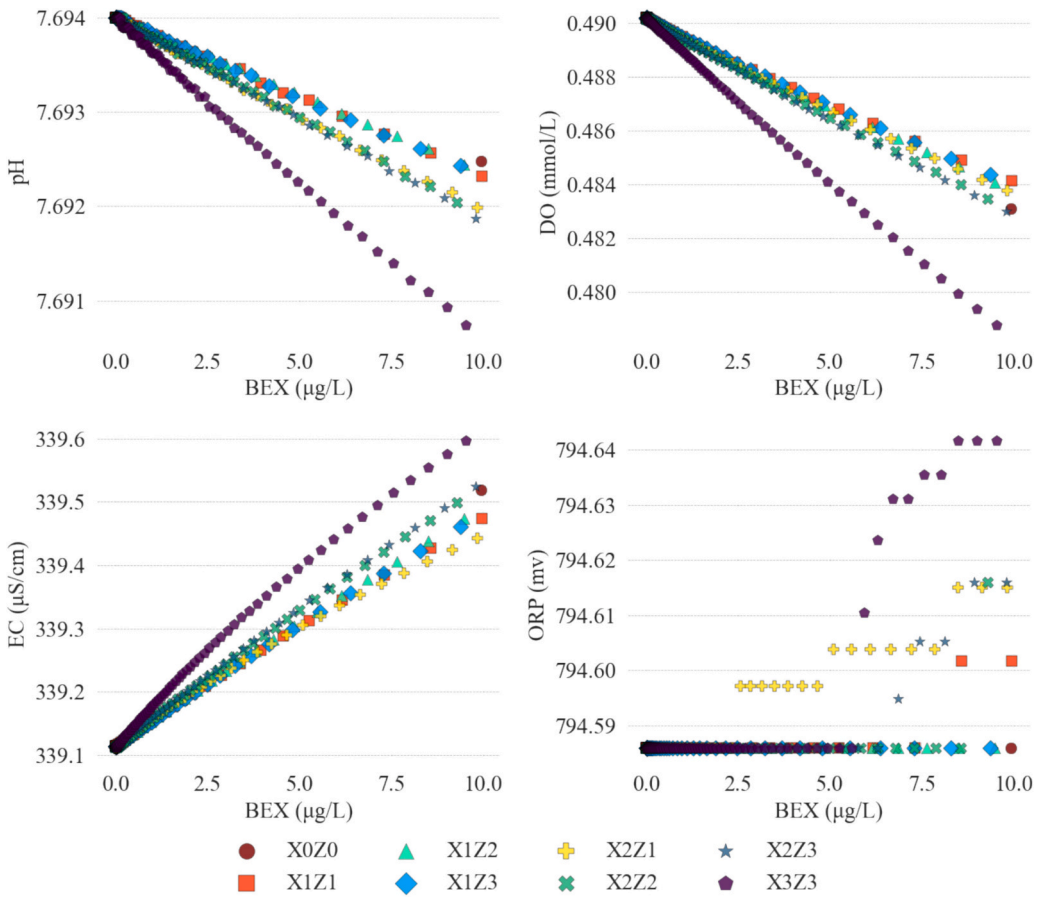


Fig. 3. Scatter plots of four in-situ water quality parameters versus BEX concentration (µg/L) based on RTM results. Data points are plotted at 5-point intervals to reduce visual clutter while preserving trend resolution. Note that variations in sensor parameter values were minimal across the ND–10 µg/L range.

which compare the performance of ML classifiers under different aquifer conditions (Fig. 4). Each scenario involves a specific modification to the aquifer's properties in our previous RTM (Wu et al., 2024), including hydraulic conductivity, background EC, nitrate and sulfate concentrations, presence/absence of calcite, and cation exchange capacity (CEC). As in the base case, we trained the ML models using data from virtual wells X0Z0, X1Z1, X1Z2, and X1Z3 within a given scenario and tested them on downstream virtual warning wells X2Z1, X2Z2, X2Z3, and X3Z3 under the same conditions.

Our scenario simulations revealed consistent patterns in model performance relative to the base case. Most models triggered detections earlier than the actual 5 µg/L threshold, except for the SVC model, which showed a more conservative behavior. However, these general trends of the models' performances were significantly affected by three hydrogeochemical changes: reduced hydraulic conductivity, elevated background salinity, and the availability of electron acceptors.

Reduced hydraulic conductivity slowed down advective transport, increasing residence time and enhancing redox-driven degradation of BEX. This condition altered the geochemical gradients of the aquifer, weakening the iWQP-BEX correlations on which the ML models relied for detection. The LR model became particularly conservative under these conditions, triggering alarm at elevated concentration of 17.4 µg/L at virtual warning well X2Z1 (Fig. 4). In contrast, the SVC model triggered the alarm at 4.3 µg/L in virtual warning well X2Z1 and falsely detecting contamination in virtual warning well X3Z3 (ND). These responses reflect how reduced flow velocities fundamentally shift chemical relationships in the aquifer, making parameters like pH and EC more responsive to NVDOC than to BEX compounds (Wu et al., 2024).

Changes in background water chemistry also affected the performance of the ML models. The elevated EC, induced by increased salinity, reduced the signal-to-noise ratio of BEX contamination (Wu et al., 2024). High baseline EC masked the subtle changes caused by BEX degradation, resulting in SVC to generate false positives even in uncontaminated virtual warning wells (ND at X3Z3).

Scenarios with elevated nitrate or absent oxygen/sulfate disrupted the expected BEX degradation pathways, causing the ML models to trigger false or delayed contamination alarms. For instance, LR incorrectly flagged contamination at all virtual warning wells (ND) when dissolved oxygen was absent at the RTM simulation. Similarly, both RF and XGB triggered alarms during the simulation scenario with elevated nitrate concentrations at BEX levels above 30 µg/L at the virtual warning wells X2Z2, X2Z3, and X3Z3 (Fig. 4). This scenario likely shifted the redox balance, delayed BEX degradation, and thus delayed the alarm until the concentration was significantly higher than the 5 µg/L threshold.

The kernel density estimates in Fig. S6 clearly illustrate these overall trends, showing that most model predictions clustered below the 5 µg/L threshold (i.e., at the peak of the curves). These KDE plots provide a smoothed visualization of the data distribution, offering intuitive insights into prediction patterns while masking exact frequencies (Węglarczyk, 2018).

While our scenario simulations explored hydrogeochemical variability, they did not yet account for potential measurement uncertainty in real-world field deployments. To address this limitation and assess the robustness of our ML models under more practical conditions, we introduced Gaussian noise and sinusoidal fluctuations to the iWQPs.

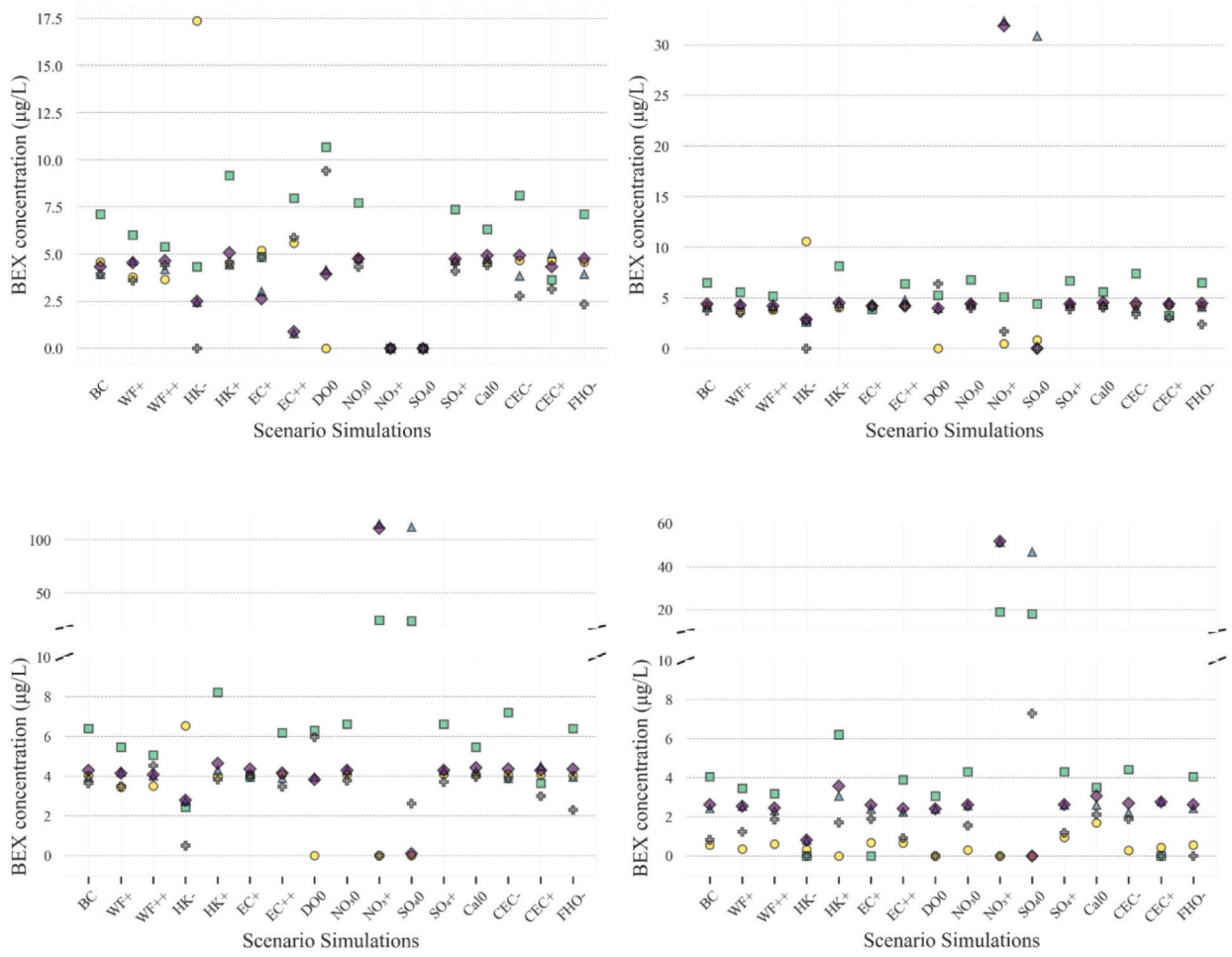


Fig. 4. Model prediction results showing the BEX concentration at which the contamination alarm was triggered at four virtual monitoring wells (i.e., X2Z1, X2Z2, X2Z3, X3Z3) under different scenario simulations. We used a broken y-axis to clearly display BEX concentrations between ND and 10 μg/L at X2Z3 and X3Z3.

3.3. Sensitivity to noise and groundwater variability

3.3.1. Impact of sensor noise on model performance

Adding Gaussian noise to the test data significantly degraded the performance of all ML models. Even at relatively low noise levels (10–20 %), false alarms were consistently triggered, particularly at virtual warning well X2Z1 (Fig. 5). This is presumably because typical sensor noise levels were comparable to the subtle signal variations of iWQPS at BEX concentration threshold of 5 μg/L. As the threshold increased, the impact of sensor noise diminished, since higher BEX levels were associated with stronger signals (Fig. S7). False alarms occurred at 10–30 % noise levels for the 10 μg/L threshold, while premature alarms—approximately 30 to 100 μg/L lower than the set threshold—were observed at 80–100 % noise levels for thresholds at 50 to 500 μg/L (Fig. S8). Among the tested models, MLP was the most sensitive, leading to inconsistent alarm triggering. This suggests a loss of generalization and stability under noisy conditions, highlighting the need for enhanced regularization techniques (Dey et al., 2018).

When noise was added to the training data, its effect on model performance was more gradual compared to test data noise. As noise levels increased from 10 % to 100 %, the BEX concentration at which the models triggered alarms also increased (Fig. S9a to c), except for MLP. This trend aligns with findings from other studies, which have shown that the type and level of noise significantly affect classifier accuracy; for example, experiments with Gaussian noise demonstrated a more gradual decline in accuracy, particularly when noise was added to the training data (Schooltink, 2020).

However, the addition of training noise caused LR and SVC to generate false positives at the farthest virtual warning well X3Z3 (Fig. S9c). This points to a possible interaction between noise and signal strength at the fringe of the plume, where BEX concentrations are already low. Under these conditions, even small fluctuations in input features were enough to push the model toward incorrect classifications. Furthermore, as noted by Schmidt et al. (2018b), even small input perturbations that are often imperceptible to humans can cause state-of-the-art classifiers to make incorrect predictions with high confidence.

Compared to other models, LR demonstrated greater robustness. While noise introduced a slight delay in alarm triggering (from 5 μg/L to about 8 μg/L) at virtual warning well X2Z1 (Fig. 5), LR maintained relatively stable performance across noise levels. This can be attributed to its linear decision boundary, which is inherently less sensitive to small perturbations in the input data, allowing it to better filter out minor fluctuations (Hasan and Chu, 2022).

3.3.2. Impact of seasonal fluctuation on model performance

Sinusoidal fluctuations in the training data had a limited impact on the performance of linear models such as LR and SVC, especially when compared to sensor noise (Fig. 6). These simpler models lack the flexibility to model non-linear patterns, which in this context proved advantageous. Because they could not overfit to the fluctuations, LR and SVC maintained stable alarm-triggering behavior across varying fluctuation levels.

In contrast, complex models such as XGB, RF, and MLP were more sensitive to these fluctuations. As the amplitude of the fluctuations

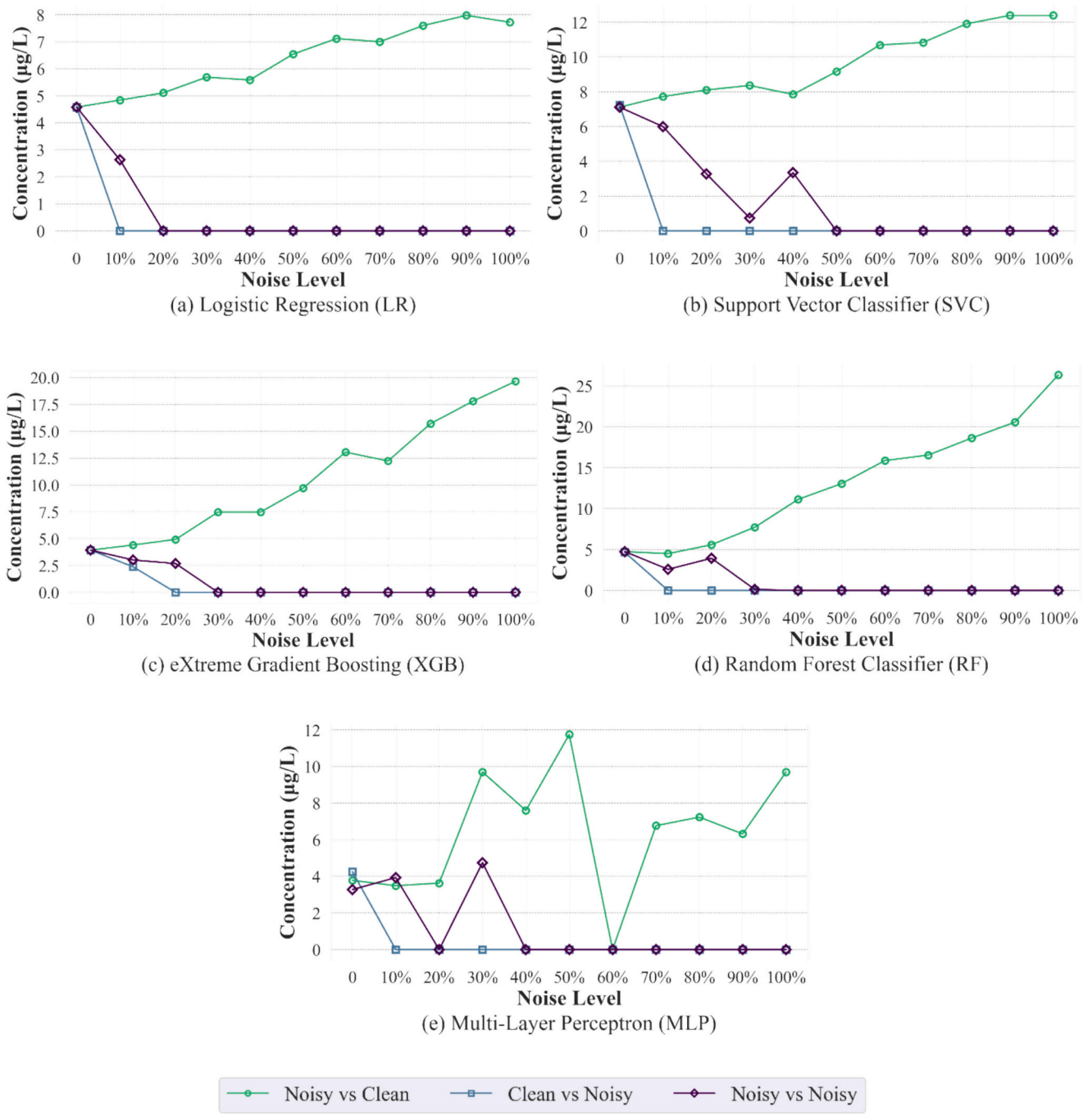


Fig. 5. BEX concentrations (µg/L) at which the five ML models triggered the contamination alarm under varying sensor noise levels at virtual warning well X2Z1: Noisy vs Clean (noisy training, clean test data), Clean vs Noisy (clean training, noisy test data), and Noisy vs Noisy (noisy training and test data).

increased, XGB and RF became increasingly conservative, delaying alarms until higher BEX concentrations were reached. RF, in particular, only triggered alarms at concentrations exceeding 60 µg/L under extreme (100 %) fluctuation conditions. The MLP model exhibited the most erratic behavior, likely due to the MLP's complex internal representations and susceptibility to overfitting, particularly when the hyperparameters were tuned on clean data (Rynkiewicz, 2019).

These findings suggest that fluctuations not representative of the true underlying signal can degrade the performance of complex models. Such models are prone to fitting high-frequency patterns in the training data, mistaking noise for meaningful trends (Hakkal and Lahcen, 2024). In contrast, simpler linear models such as LR and SVC cannot capture complex relationships such as these fluctuations. However, if the fluctuations were part of the actual signal, the inability of linear models to learn non-linear patterns would likely degrade their performance.

To account for possible seasonal trends in iWQPs, we also conducted additional experiments by shifting the sinusoidal peaks to represent different months. These seasonal adjustments did not significantly alter model behavior, and the general conclusions described above held across all scenarios.

3.3.3. *Influence of combined noise and seasonal fluctuation*

When both sensor noise and sinusoidal annual fluctuations were introduced into the training data, model performance degraded more noticeably than when only one type of variability was present (Fig. 7). For instance, at 100 % combined noise and fluctuation levels, LR

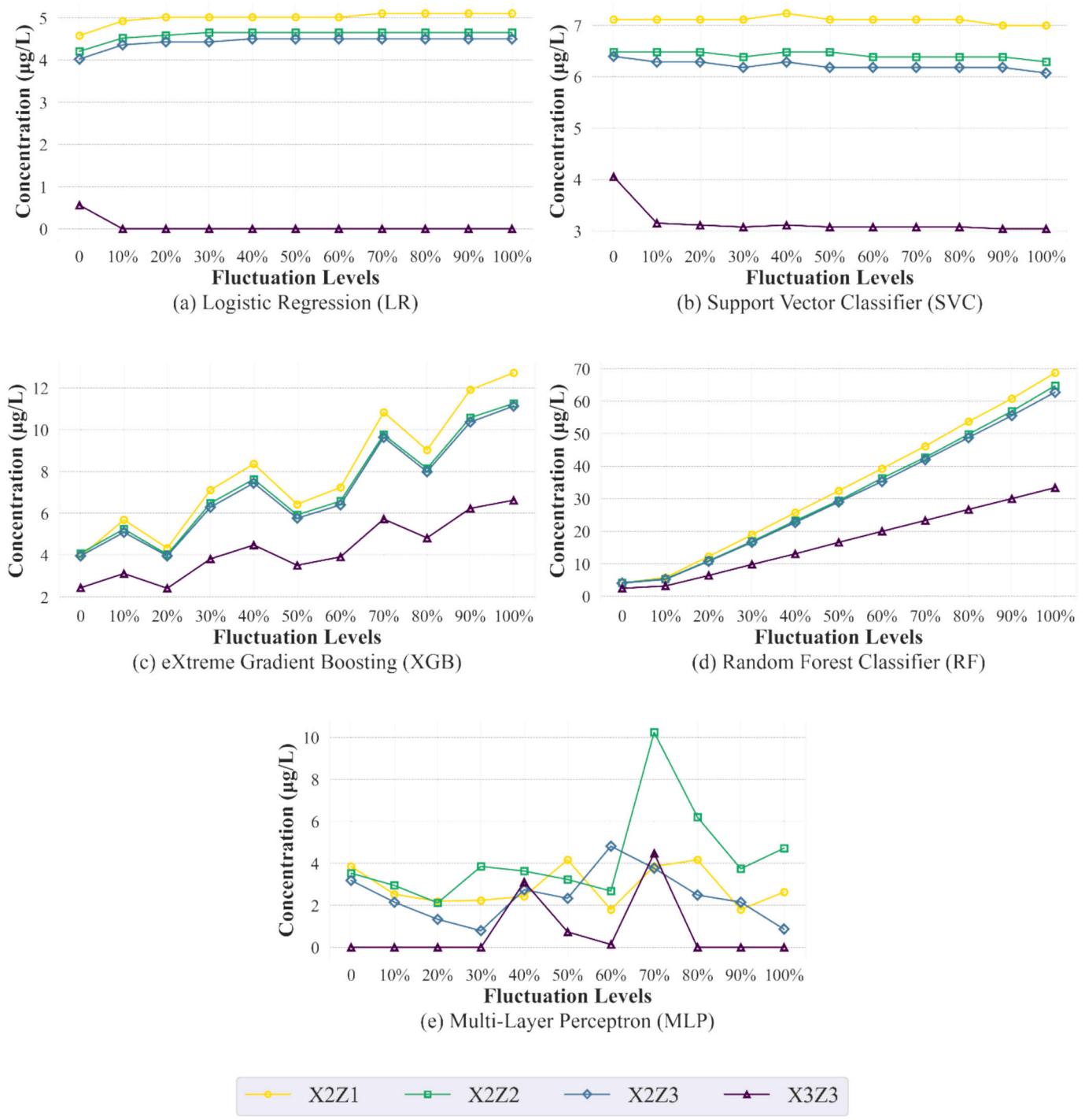


Fig. 6. BEX concentrations (µg/L) at which the five ML models triggered the contamination alarm under varying fluctuation levels added to the training data at all virtual warning wells.

delayed its alarm at virtual warning well X2Z1. The alarm was triggered only until BEX concentrations reached 14 µg/L which is well above the intended 5 µg/L threshold. In comparison, LR triggered the alarm at around 8 µg/L with only sensor noise, and correctly at 5 µg/L with only sinusoidal fluctuations. Similarly, XGB triggered the alarm at X2Z1 at 50 µg/L under combined noise and fluctuations, compared to 20 µg/L with only sensor noise and 13 µg/L with only sinusoidal fluctuations.

Interestingly, model performance did not always degrade linearly with increasing noise levels. In some cases, such as XGB at the four virtual warning wells, the model performed slightly better at 100 % combined noise and fluctuation than at 90 %. At 90 %, the alarm was triggered at approximately 60 µg/L, compared to around 50 µg/L at 100 %. This counterintuitive result may be due to instances where the sensor noise and the sinusoidal fluctuations partially canceled each other out.

3.4. *Improving model performance*

To mitigate the effects of sensor noise and seasonal fluctuations, we applied 3-day, 5-day, and 10-day moving average smoothing to the training data after introducing Gaussian noise and sinusoidal fluctuations. Among these, the 5-day moving average yielded the best model performance (Fig. 8). The results for 3-day and 10-day smoothing are

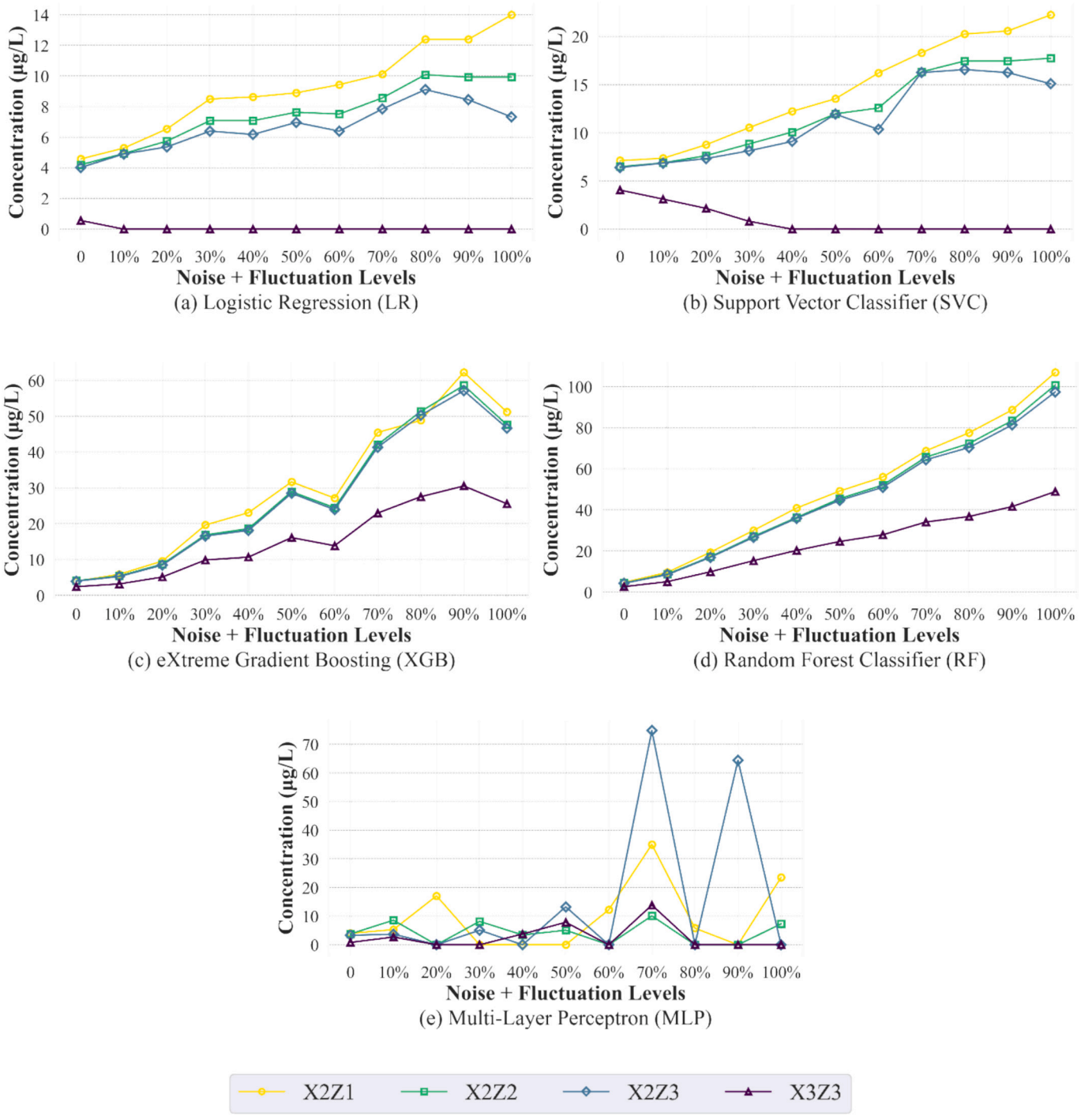


Fig. 7. BEX concentrations (µg/L) at which the five ML models triggered the contamination alarm with combined sensor noise and sinusoidal fluctuation levels added to the training data at all virtual warning wells.

provided in Fig. S10a and b. LR initially triggered the contamination alarm at ~10 µg/L at 100 % noise and fluctuation levels in virtual warning well X2Z2 (Fig. 7). However, LR detected BEX contamination at ~7 µg/L when 5-day smoothing was applied (Fig. 8), which is closer to the 5 µg/L regulatory threshold. Similarly, SVC showed improved sensitivity, with alarms triggered at ~10 µg/L (smoothed) compared to ~17 µg/L (unsmoothed).

While smoothing can generally help reduce the impact of sensor noise in the ML model performance (Xiao et al., 2022), the choice of window size is critical: Too few days (e.g., 3-day) makes smoothing overly sensitive to noise, as short-term fluctuations disproportionately influence the average. On the other hand, too many days (e.g., 10-day) leads to over-smoothing and can delay the detection by masking short-term contamination spikes. The 5-day window in our case served as an optimal balance, effectively dampening random noise while still preserving meaningful sensor signals. However, seasonal fluctuations require a different approach such as Fourier decomposition or wavelet transforms (Bi et al., 2023) to separate periodic trends from contamination signals.

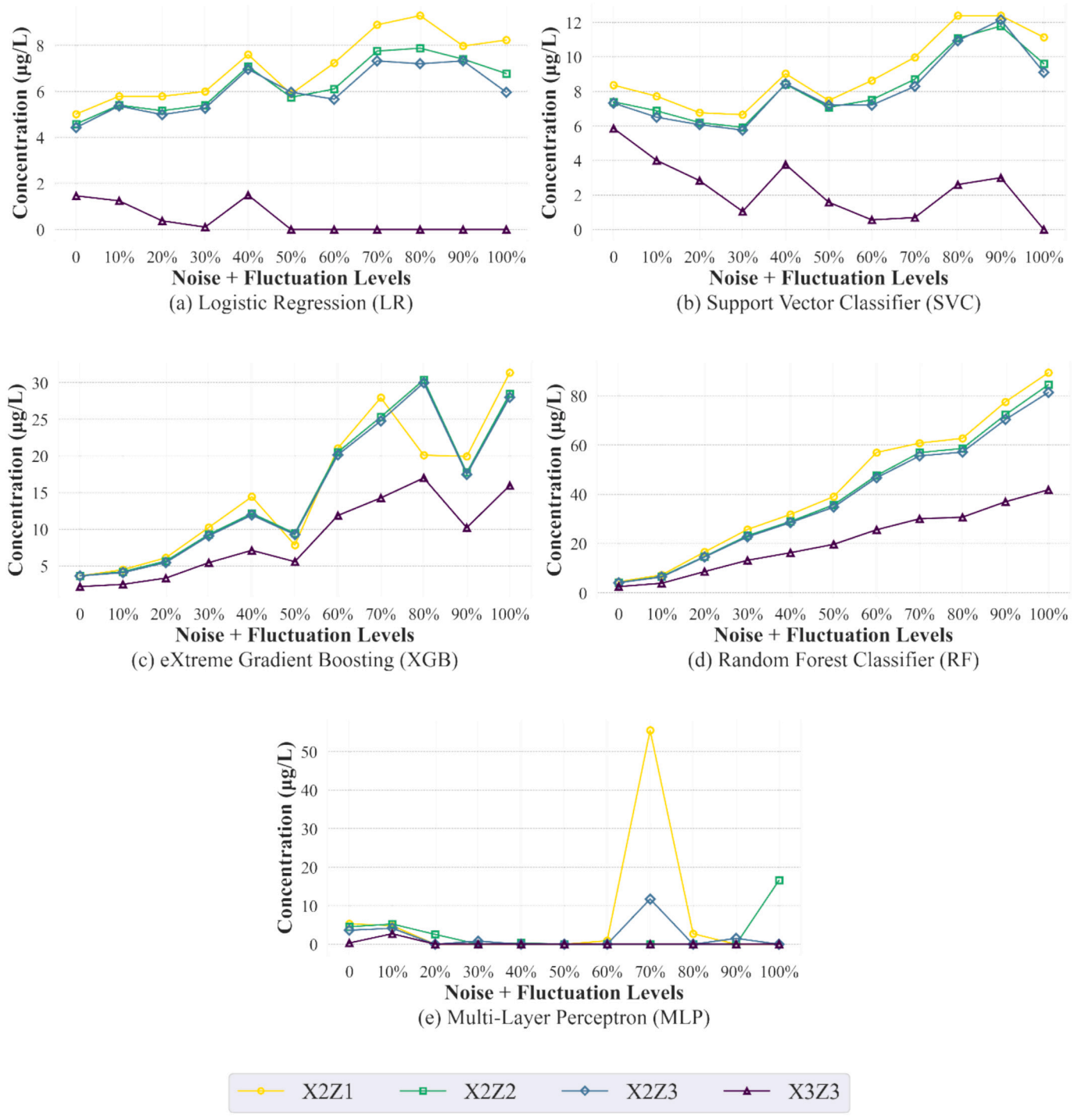


Fig. 8. BEX concentrations (µg/L) at which the five machine learning models triggered contamination alarms. A 5-day moving average was applied for smoothing after adding different levels of combined sensor noise and sinusoidal fluctuations in the training data at all virtual warning wells.

4. Conclusion

In this study, we present a proof-of-concept ML framework that integrates data from affordable in-situ sensors to detect BEX in groundwater. The used virtual sensors in the presented case include pH, DO, EC, and ORP. Our ML framework was trained to detect contamination at warning wells and provide timely alerts of contaminant migration before it reaches sensitive receptor areas.

Ensemble models such as RF and XGB consistently demonstrated reliable performance across diverse aquifer scenarios. This robustness suggests their potential suitability for future field deployment after further field validation. However, successful real-world implementation still requires careful consideration of site-specific hydrogeochemical properties. Our findings showed that scenario-specific deviations consistently emerged under three conditions: when altered transport times changed the degradation signatures, when background chemistry obscured the contamination signals, or when terminal electron acceptors were absent. These factors affect the relationships between iWQPs and BEX concentrations.

The integration of sensor data into ML systems raises challenges related to data quality and signal stability. Even modest levels of noise, such as 10–20 % Gaussian variation, considerably impacted the model

performance, especially for complex models such as the MLP. This highlights the need for both hardware and software-based solutions to stabilize data inputs. For example, hardware modifications such as flow-stabilization chambers in monitoring wells can help reduce turbulence and provide more consistent sensor readings. On the software side, preprocessing techniques to remove seasonal patterns and smooth noise can also enhance data quality. Our results indicate that fixed-window smoothing can improve model reliability; however, static and adaptive smoothing methods may mask short-term concentration spikes or fail to adjust to seasonal variability.

To maintain model accuracy over time, all deployed systems will require continuous learning mechanisms capable of incorporating new sensor data and adapting to evolving aquifer conditions. Without such updates, model predictions may degrade as site conditions shift. Moreover, since class imbalance can affect model performance, resampling or weighting techniques can be explored for future work to improve contaminant detection sensitivity. While our study demonstrates the feasibility of this approach in a controlled virtual environment, further field-based research is needed to validate its practical applicability.

For field validation and practical implementation, low-cost in-situ sensors (e.g., pH, DO, EC, and ORP) could be installed at monitoring wells within contaminated area to collect continuous water quality data. To train the ML models, high-resolution spatiotemporal data would be needed to establish a reliable and site-specific baseline. Monthly BTEX measurements via laboratory analysis are essential during the initial years of contamination. More frequent sampling, such as weekly or daily, can further improve model robustness; direct BTEX sensors may be installed at contaminated wells for daily measurements. Once trained, the models can be deployed at downstream warning wells where similar sensors would be installed. The warning wells are positioned before sensitive receptors such as drinking water wells. Sensor data from these wells would be streamed to a central system for ML-based analysis, triggering alarms when predicted BEX concentrations exceed the contamination threshold. Manual sampling is then required to check for actual contamination.

Initial costs include sensor installation and calibration, while ongoing costs involve maintenance and periodic model retraining to account for seasonal or anthropogenic changes in aquifer chemistry. Compared to manual grab sampling, which costs approximately €130 per BTEX analysis (ALS Global, n.d.), sensors with a combined cost of around €400 (Atlas Scientific LLC, 2025a, 2025b, 2025c, 2025d) offer continuous monitoring and faster detection, potentially reducing long-term costs and improving response times. Importantly, this framework is intended to complement, not replace, existing manual sampling strategies. Nevertheless, the integration of low-cost sensor networks, adaptive ML models, and robust validation strategies offers a promising path toward real-time, scalable, and continuous groundwater quality monitoring.

CRediT authorship contribution statement

C.L.R. Wu: Writing – original draft, Visualization, Validation, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **R.M. Wagterveld:** Writing – review & editing, Supervision, Project administration, Conceptualization. **L.C. Rietveld:** Writing – review & editing, Supervision, Conceptualization. **B.M. van Breukelen:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization.

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.

Acknowledgements

This work has been performed within the cooperation framework of Wetsus, the European Centre of Excellence for Sustainable Water Technology (wetsus.nl), and Technische Universiteit Delft (tudelft.nl). Wetsus is co-funded by the Dutch Ministry of Economic Affairs, the Ministry of Infrastructure and Environment, the European Union Regional Development Fund, the Northern Netherlands Provinces, and the Province of Fryslân. We would like to thank the participants of the "Monitoring and Quality" theme for the informative discussions and financial support, and Nadia van Pelt for her thorough proof reading and valuable suggestions on the flow and structure of this paper. This research has been financially supported by the Dutch Research Council (NWO; Sustainable Water Technology Call 2018; contract number: ALWET.2019.003).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jconhyd.2025.104771.

Data availability

Data will be made available on request.

References

Ahsan, M.M., Mahmud, M.A.P., Saha, P.K., Gupta, K.D., Siddique, Z., 2021. Effect of data scaling methods on machine learning algorithms and model performance. Technologies 9 (3), 52. https://doi.org/10.3390/TECHNOLOGIES9030052.

ALS Global. (n.d.) Petrol Pack Basic: BTEX and TPH [OV-20h] in Water. Retrieved from https://www.alsglobal.no/en/package/environment_1/water_2/combination-packages_8/petrol-pack-basic-btex-and-tph-ov-20h-in-water_44640 (on 28 September 2025).

Atlas Scientific LLC, 2025a. Lab Grade pH Probe: #ENV-40-pH. Retrieved from. https://atlas-scientific.com/probes/ph-probe/ (on 15 September 2025).

Atlas Scientific LLC, 2025b. Mini Conductivity Probe K 1.0: #ENV-20-EC-K1.0. Retrieved from. https://atlas-scientific.com/probes/mini-e-c-probe-k-1-0/. on 15 September 2025.

Atlas Scientific LLC, 2025c. Mini Lab Grade Dissolved Oxygen Probe: #ENV-20-DOX. Retrieved from. https://atlas-scientific.com/probes/mini-d-o-probe/. on 15 September 2025.

Atlas Scientific LLC, 2025d. Mini Lab Grade ORP Probe: #ENV-20-ORP. Retrieved from. https://atlas-scientific.com/probes/mini-orp-probe/. on 15 September 2025.

Beck, P., Mann, B., 2010. A Technical Guide for Demonstrating Monitored Natural Attenuation of Petroleum Hydrocarbons in Groundwater. CRC CARE Technical Report no. 15,. CRC for Contamination Assessment and Remediation of the Environment, Adelaide, Australia.

Bi, J., Li, Y., Chang, X., Yuan, H., Qiao, J., 2023. Hybrid Water Quality Prediction with Frequency Domain Conversion Enhancement and Seasonal Decomposition. Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics, 5200–5205. https://doi.org/10.1109/SMC53992.2023.10394421.

Bisong, E., 2019. Logistic regression. In: Building Machine Learning and Deep Learning Models on Google Cloud Platform. Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4842-4470-8_20.

Copeland, A., Lytle, D.A., 2014. Measuring the Oxidation–reduction Potential of Important Oxidants in Drinking Water, 106. American Water Works Association, pp. E10–E20. https://doi.org/10.5942/jawwa.2014.106.0002.

Cova, C.M., Rincón, E.M., Espinosa, E., Serrano, L., Zuliani, A., Zuliani, A., 2022. Paving the way for a green transition in the design of sensors and biosensors for the detection of volatile organic compounds (VOCs). Biosensors 12 (2), 51. https://doi.org/10.3390/BIOS12020051.

de Winter, C., Palleti, V.R., Worm, D., Kooij, R., 2019. Measuring imperfections of water quality sensors in water distribution networks. Meas. Sci. Technol. 30 (9), 095101. https://doi.org/10.1088/1361-6501/AB1EEB.

Dey, P., Nag, K., Pal, T., Pal, N.R., 2018. Regularizing multilayer perceptron for robustness. IEEE Trans Syst Man Cybern Syst 48 (8), 1255–1266. https://doi.org/10.1109/TSMC.2017.2664143.

George, S., Dixit, A., 2021. A machine learning approach for prioritizing groundwater testing for per-and polyfluoroalkyl substances (PFAS). J. Environ. Manage. 295, 113359. ISSN 0301-4797. https://doi.org/10.1016/j.jenvman.2021.113359.

Haider, F.U., Ejaz, M., Cheema, S.A., Khan, M.I., Zhao, B., Liqun, C., Salim, M.A., Naveed, M., Khan, N., Núñez-Delgado, A., Mustafa, A., 2021. Phytotoxicity of petroleum hydrocarbons: sources, impacts and remediation strategies. Environ. Res. 197, 111031. https://doi.org/10.1016/J.ENVRES.2021.111031.

Hakkal, S., Lahcen, A.A., 2024. XGBoost to enhance learner performance prediction. Comput. Educ. 7, 100254. https://doi.org/10.1016/J.CAEAI.2024.100254.

Hasan, R., Chu, H., 2022. Noise in Datasets: What Are the Impacts on Classification Performance? 11th International Conference on Pattern Recognition Applications and Methods. https://doi.org/10.5220/0010782200003122.
Irvine, T. (n.d.). Shock and Vibration Response Spectra Course Unit 2A. Sine Vibration Characteristics.
Khatri, N., Tyagi, S., 2015. Influences of natural and anthropogenic factors on surface and groundwater quality in rural and urban areas. Front. Life Sci. 8 (1), 23–39. https://doi.org/10.1080/21553769.2014.933716.
Li, J.J., Tong, X., 2020. Statistical hypothesis testing versus machine learning binary classification: distinctions and guidelines. Patterns 1 (7), 100115. https://doi.org/10.1016/J.PATTER.2020.100115.
Li, H., Son, J.-H., Hanif, A., Gu, J., Dhanasekar, A., Carlson, K., 2017. Colorado water watch: real-time groundwater monitoring for possible contamination from oil and gas activities. J. Water Resour. Protect. 9. https://doi.org/10.4236/jwarp.2017.913104.
Li, P., Karunanidhi, D., Subramani, T., Srinivasamoorthy, K., 2021. Sources and consequences of groundwater contamination. Arch. Environ. Contam. Toxicol. 80 (1), 1–10. https://doi.org/10.1007/s00244-020-00805-z.
Li, L., Qiao, J., Yu, G., Wang, L., Li, H.Y., Liao, C., Zhu, Z., 2022. Interpretable tree-based ensemble model for predicting beach water quality. Water Res. 211, 118078. https://doi.org/10.1016/J.WATRES.2022.118078.
Lin, S., Zheng, H., Han, B., Li, Y., Han, C., Li, W., 2022. Comparative performance of eight ensemble learning approaches for the development of models of slope stability prediction. Acta Geotech. 17 (4), 1477–1502. https://doi.org/10.1007/s11440-021-01440-1.
Lyons, K.J., Ikonen, J., Hokajärvi, A.M., Räsänen, T., Pitkänen, T., Kauppinen, A., Kujala, K., Rossi, P.M., Miettinen, I.T., 2023. Monitoring groundwater quality with real-time data, stable water isotopes, and microbial community analysis: a comparison with conventional methods. Sci. Total Environ. 864, 161199. https://doi.org/10.1016/J.SCITOTENV.2022.161199.
McKnight, U.S., Funder, S.G., Rasmussen, J.J., Finkel, M., Binning, P.J., Bjerg, P.L., 2010. An integrated model for assessing the risk of TCE groundwater contamination to human receptors and surface water ecosystems. Ecol. Eng. 36 (9), 1126–1137. https://doi.org/10.1016/J.ECOLENG.2010.01.004.
Narkhede, M.V., Bartakke, P.P., Sutaone, M.S., 2022. A review on weight initialization strategies for neural networks. Artif. Intell. Rev. 55 (1), 291–322. https://doi.org/10.1007/S10462-021-10033-Z/METRICS.
Ng, G.H.C., Bekins, B.A., Cozzarelli, I.M., Baedecker, M.J., Bennett, P.C., Amos, R.T., Herkelrath, W.N., 2015. Reactive transport modeling of geochemical controls on secondary water quality impacts at a crude oil spill site near Bemidji, MN. Water Resour. Res. 51 (6), 4156–4183. https://doi.org/10.1002/2015WR016964.
Qiao, F., Wang, J., Song, J., Chen, Z., Kwaw, A.K., Zhao, Y., Zheng, S., 2025. The spatiotemporal evolution of dissolved-phase NAPL plumes revealed by the integrated groundwater quality and machine learning models. Water Res. 280, 123535. https://doi.org/10.1016/j.watres.2025.123535.
Rynkiewicz, J. (2019). On overfitting of multilayer perceptrons for classification. In J. Rynkiewicz (ed.), Computational Intelligence and Machine Learning. http://www.i6doc.com/en/.
Schmidt, F., Wainwright, H.M., Faybishenko, B., Denham, M., Eddy-Dilek, C., 2018a. In situ monitoring of groundwater contamination using the kalman filter. Environ. Sci. Tech. 52 (13), 7418–7425. https://doi.org/10.1021/ACS.EST.8B00017.
Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., Madry, A., 2018b. Adversarially robust generalization requires more data. Adv. Neural Inf. Proces. Syst. 31. https://doi.org/10.48550/arXiv.1804.11285.
Schooltink, W.T., 2020. Testing the Sensitivity of Machine Learning Classifiers to Attribute Noise in Training Data. http://deeplearning.net/datasets/.
Sen, P.C., Hajra, M., Ghosh, M., 2020. Supervised classification algorithms in machine learning: a survey and review. Adv. Intell. Syst. Comput. 937, 99–111. https://doi.org/10.1007/978-981-13-7403-6_11.
Singh, S.K., Shirzadi, A., Pham, B.T., 2021. Application of artificial intelligence in predicting groundwater contaminants. Water Pollut. Manag. Pract. 71–105. https://doi.org/10.1007/978-981-15-8358-2_4.
U.S. Environmental Protection Agency, 2024. National Primary Drinking Water Regulations. https://www.epa.gov/ground-water-and-drinking-water/national-primary-drinking-water-regulations.
U.S. Geological Survey. (n.d.). USGS Water Data for the Nation. U.S. Department of the Interior. Retrieved August 6, 2024, from https://waterdata.usgs.gov/nwis.
Vapnik, V., 1982. Estimation of Dependences Based on Empirical Data (Springer Series in Statistics). Springer-Verlag, New York (ISBN 3-540-90733-5).
Wang, J., Shen, X., 2006. Estimation of generalization error: random and fixed inputs. Stat. Sin. 16, 569–588.
Węglarczyk, S., 2018. Kernel density estimation and its application. In: ITM Web of Conferences, 23. https://doi.org/10.1051/ITMCONF/20182300037, 00037.
Wu, C.L.R., Wagterveld, R.M., van Breukelen, B.M., 2024. Reactive transport modeling for exploring the potential of water quality sensors to estimate hydrocarbon levels in groundwater. Water Resour. Res. 60, e2023WR036644. https://doi.org/10.1029/2023WR036644.
Xiao, Z., Gang, W., Yuan, J., Chen, Z., Li, J., Wang, X., Feng, X., 2022. Impacts of data preprocessing and selection on energy consumption prediction model of HVAC systems based on deep learning. Energ. Buildings 258, 111832. https://doi.org/10.1016/J.ENBUILD.2022.111832.
Xu, Y., Goodacre, R., 2018. On splitting training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. J. Anal. Test. 2, 249–262. https://doi.org/10.1007/s41664-018-0068-2.
Yao, Y., Tu, C., Hu, G., Zhang, Y., Cao, H., Wang, W., Wang, W., 2024. Groundwater hydrochemistry and recharge process impacted by human activities in an Oasis–Desert in Central Asia. Water 16 (5), 763. https://doi.org/10.3390/W16050763.
Zhang, Z., 2025. Enhancing distributed machine learning through data shuffling: techniques, challenges, and implications. In: ITM Web of Conferences, 73. https://doi.org/10.1051/ITMCONF/20257303018, 03018.
Zhu, X., Wu, X., 2004. Class noise vs. attribute noise: a quantitative study. Artif. Intell. Rev. 22 (3), 177–210. https://doi.org/10.1007/S10462-004-0751-8.