https://doi.org/10.5194/egusphere-2025-5842
Preprint. Discussion started: 4 January 2026

Real-time Monitoring of Petroleum Hydrocarbons in Groundwater using Hybrid Machine Learning Architectures

Chen Lester R. Wu[1,2], R. Martijn Wagterveld[1], Luuk C. Rietveld[2], B. M. van Breukelen[2]

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

Correspondence to: Chen Lester R. Wu (crwu2@alum.up.edu.ph)

Abstract. Monitoring petroleum hydrocarbon (PHC) plumes in groundwater is essential for managing oil contamination but is often hindered by high costs. We evaluated machine learning (ML) frameworks that estimate concentrations of benzene, ethylbenzene, and xylenes (BEX), using affordable, in situ water quality parameters (iWQPs) as inputs: pH, dissolved oxygen, electrical conductivity, and oxidation-reduction potential. Due to a scarcity of field data, we trained and tested models on high-resolution virtual data generated by a reactive transport model. We compared a long short-term memory (LSTM) network against classical algorithms (multiple linear regression, random forest, support vector regression, XGBoost) and an LSTM-XGBoost hybrid. Model performance depended on the underlying geochemical relationship between iWQPs and BEX. Accurate predictions ($R^2 \geq 0.80$, MAPE < 2.3 %) were achieved when iWQPs were strongly correlated with BEX degradation (e.g., as a primary electron donor); the LSTM model yielded predictions within a 5 % error margin for 70 % of the test cases. Performance declined sharply ($R^2 < 0$) during periods where iWQPs were correlated with non-volatile dissolved organic carbon, another component of dissolved PHC. Incorporating hydraulic head data improved accuracy by informing the model of groundwater flow dynamics. While the LSTM model struggled to extrapolate beyond its training data (e.g., during extreme flow events), it reliably detected the direction of concentration trends, providing a valuable trigger for adaptive monitoring. We also demonstrated how a hybrid Kalman filter could successfully capture concentration trends after source removal through recursive updating. Our proposed ML framework provides BEX level estimation for improved groundwater monitoring.

1 Introduction

Groundwater contamination by petroleum hydrocarbons (PHCs) remains an environmental challenge, particularly in areas affected by historical spills or leaks. Compounds such as benzene, toluene, ethylbenzene, and xylenes (BTEX) pose serious risks to ecosystems and human health due to their toxicity and persistence (Zanello et al., 2021). Monitored natural attenuation (MNA) is a widely adopted remediation strategy that relies on natural processes such as biodegradation, dilution, and sorption to reduce contaminant concentrations over time (Chiu et al., 2013). However, the effectiveness of MNA depends on monitoring contaminant levels accurately and continuously to ensure that attenuation is proceeding as expected (Beck and Mann, 2010). Traditional monitoring methods, which rely on periodic manual sampling and laboratory analysis, are time-consuming, costly, and spatially limited, often resulting in data gaps that hinder timely decision-making.

Existing technologies for continuous BTEX monitoring face limitations. In situ optical sensors provide near-real-time data but are susceptible to interference from other dissolved organics and require frequent calibration (Buerck et al., 2001; Larsson and Dasgupta, 2003, Wong et al., 2024). Membrane interface probes offer high-resolution vertical contaminant profiling but are invasive and expensive, making them impractical for long-term, large-scale monitoring (Industrial Economics, Incorporated, 2023). Portable gas chromatography systems, while accurate, require specialized operators and regular maintenance (Ji et al., 2006). Even with these technologies, a gap remains in developing scalable, low-maintenance solutions for real-time monitoring of BTEX contamination in groundwater.

Machine learning (ML) has emerged as a promising alternative by leveraging easily measurable in situ water quality parameters (iWQPs) to infer contaminant concentrations. Using a reactive transport model (RTM) with a 2D cross-section and a spatial resolution of 0.2 m (vertical) × 2 m (horizontal), Wu et al. (2024) demonstrated that iWQPs, namely pH, dissolved oxygen (DO), electrical conductivity (EC), and oxidation-reduction potential (ORP), are correlated with dissolved PHCs. The RTM was used as the data source due to the lack of an available dataset on groundwater-contaminated sites for a robust statistical analysis. Building on this, Wu et al. (2026) developed a binary classification model to flag benzene, ethylbenzene, and xylenes (BEX) contamination for early warning systems.

Qiao et al. (2025) also proposed a two-stage predictive framework for non-aqueous phase liquid (NAPL) plumes, using the RTM from Wu et al. (2024) to generate 30 years of simulated daily data. They used year 30 as the prediction target and the preceding years as training data. The training and test datasets were generated for each of the 7,500 grid cells (150 × 50), resulting in thousands of synthetic data observations. In the first stage, a sliding-window random forest (RF) algorithm predicted iWQPs for the target year. In the second stage, these forecasted parameters were then combined with sparse NAPL measurements. ML models, including long short-term memory (LSTM) networks and extreme gradient boosting (XGBoost), estimated the future NAPL plume extent and concentration distributions in the simulated 2D cross-section.

Despite these advances, current ML applications for hydrocarbon monitoring have limitations. While existing frameworks can detect PHC presence and predict NAPL plume extent based on historical virtual data, they lack real-time capability for estimating PHC concentrations in groundwater. Moreover, these models are often poorly suited to dynamic field conditions, such as active remediation or sudden shifts in hydraulic gradients. This stems from their feed-forward design, which processes data linearly without recursive updates from new measurements. These gaps illustrate the need for more responsive systems capable of continuous operation under variable site conditions.

Therefore, we evaluated several ML regression models that address these limitations by estimating BEX concentrations in real time through sensor data fusion (SDF). SDF involves integrating multiple sensor data to indirectly estimate unknown parameters that are otherwise difficult or costly to measure (Mitchell, 2007). Our objective was to quantify the prediction accuracy of ML models in estimating BEX concentrations using iWQPs from cheap sensors (i.e., pH, DO, EC, and ORP)

https://doi.org/10.5194/egusphere-2025-5842
Preprint. Discussion started: 4 January 2026

under controlled conditions. Due to a lack of high-resolution temporal and spatial field data from contaminated sites, we used virtual data generated from our previous RTM (Wu et al., 2024).

To capture temporal dependencies, we implemented LSTM networks with variable sequence lengths (30–360 days). The sequence length determines how many successive daily data points are considered when making each prediction. This architecture was selected for its demonstrated effectiveness in modeling environmental time-series data (Arsenault et al., 2023). The main advantage is its ability to learn long-range dependencies while overcoming the vanishing gradient problem inherent in simpler recurrent neural networks. We also developed a hybrid architecture that combines LSTM-based sequence modeling with XGBoost regression. This configuration takes advantage of the LSTM's temporal pattern recognition capabilities alongside XGBoost's strengths in modeling nonlinear relationships and feature importance weighting (Karimi et al., 2025).

For comparison, we also implemented traditional regression models for their interpretability and established use in water quality monitoring (Singh et al., 2021): multiple linear regression (MLR) to provide a benchmark for linear relationships (Sani Gaya et al., 2020); support vector regression (SVR) to capture nonlinear patterns (Banadkooki et al., 2020); and RF and XGBoost to handle complex feature interactions (Szomolányi and Clement, 2023). These models treated daily iWQP measurements as independent observations, without incorporating temporal context.

To enhance the reliability of BEX concentration estimates, we also explored the use of Kalman filtering. The Kalman filter (KF) is a recursive algorithm that integrates noisy sensor measurements over time to produce optimal state estimates (Simon, 2006). It is particularly well-suited for dynamic systems where measurements are uncertain and temporally correlated, making it a valuable addition to our SDF framework. By incorporating Kalman filtering, we could continuously refine BEX concentration estimates as new iWQP data and BEX measurements became available.

We evaluated the robustness of these models through stress-testing scenarios, including source removal and increased hydraulic gradient. To further improve model accuracy, we incorporated hydraulic head as an additional input parameter. This parameter contains information on the advective transport and residence times, both of which are critical for capturing the influence of flow dynamics on solute distribution and reactive processes.

2 Materials and Methods

2.1 Data Generation

The RTM used in this study is based on the model developed in Wu et al. (2024), implemented in Python using FloPy within a Jupyter Notebook environment. Groundwater flow was simulated using MODFLOW 2005, and contaminant transport was modeled with MT3DMS, incorporating advection and dispersion processes. Geochemical reactions were coupled through PHT3D and PHREEQC-2 to account for biodegradation, mineral-phase reactions, cation exchange, and outgassing. The model also incorporated transient water table fluctuations and spatial heterogeneity in hydraulic conductivity. The synthetic aquifer domain represents a saturated porous medium with a length of 300 m and a thickness of 10 m, inspired by an existing study on the Bemidji crude oil spill site (Ng et al., 2015). It was discretized into 0.2 m (vertical) × 2 m (horizontal) grid cells (150 × 50 grids). Detailed descriptions of the model specifications and reaction processes are provided in Wu et al. (2024).

To evaluate the robustness of our machine learning models under realistic stress-test conditions, we simulated three scenarios: (1) a base case scenario representing general conditions at the Bemidji site but with more generic parameterization (Wu et al., 2024); (2) a realistic remediation case in which the light non-aqueous phase liquid (LNAPL) source zone is fully excavated; and (3) a controlled hydraulic stress to represent a potential extreme anthropogenic influence, such as the use of an injection well for artificial aquifer recharge upstream or an excessive downstream pumping to simulate accelerated groundwater flow.

The second scenario involved completely suppressing PHC dissolution from the oil source zone in simulation year 40. This setup is particularly relevant for evaluating post-remediation monitoring strategies and plume tailing behavior, and to test whether the ML model can still accurately estimate BEX concentrations. The third scenario involved an injection well that continuously injected water for one year, starting in simulation year 40. The injected water, simulated in a larger groundwater flow model, increased the hydraulic gradient by approximately 10 m above the initial conditions. This resulting gradient was used as the boundary condition for the RTM simulation. We designed the third scenario to evaluate the ML model's performance under extreme groundwater flow velocity. Although we simulated this scenario via upstream artificial recharge, extreme flow conditions could also result from excessive downstream pumping or extreme natural groundwater table fluctuations; these conditions can accelerate contaminant advection and dispersion (Ahmadi et al., 2021).

The third scenario was designed to evaluate the ML model's performance under conditions of extreme groundwater flow velocity, induced through upstream artificial recharge. Such extreme hydraulic gradients, which can also arise from excessive downstream pumping or natural water table fluctuations, are critical to study as they promote accelerated contaminant advection and dispersion (Ahmadi et al., 2021).

We analyzed time-series data from a 100-year simulation with daily timesteps, capturing the long-term persistence of PHCs in groundwater systems. Eight virtual observation wells were placed at depths ranging from 1.5 to 5.2 m and distances of 0 to 165 m from the source zone (Fig. 1) to monitor plume dynamics (Table S1). Each well was modeled with a 1-meter screened interval. Groundwater chemistry was calculated as a flow-rate-weighted average concentration; the weights were proportional to the hydraulic conductivity of each RTM grid cell (Höyng et al., 2015; Thouement & Van Breukelen, 2020) (Eq S1 in the Supporting Information).

https://doi.org/10.5194/egusphere-2025-5842
Preprint. Discussion started: 4 January 2026

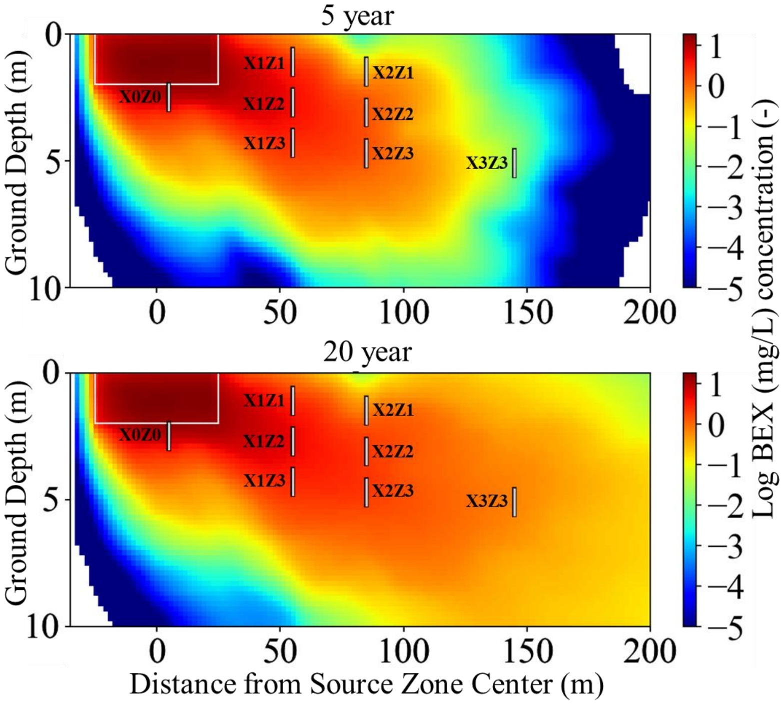


Figure 1: BEX plume distribution at 5 and 20 years of simulation. BEX concentration is converted to base 10 log form with 1 mg/L as reference. The white square indicates the oil source zone, and the eight vertical white lines represent virtual observation wells.

2.2 Data Preprocessing, Handling, and Modeling Approach

Our goal is to predict daily BEX concentrations using iWQPs at eight virtual observation wells. Hereafter, all observation wells and time-series data correspond to virtual datasets obtained from the RTM. For brevity, these were referred to as observation wells and time-series data throughout, without explicitly repeating the qualifier "virtual". All analyses used daily time-series data, with input features standardized using Scikit-learn's StandardScaler (Ahsan et al., 2021) to ensure consistent scaling, while BEX concentrations served as the target variable. Separate models were developed for each observation well to account for spatial variability across the aquifer and allow localized calibration and performance assessment.

In a subsequent modeling phase, we evaluated whether incorporating hydraulic head measurements would improve the predictive accuracy of our ML model. The average hydraulic head in an observation well was analyzed together with the four iWQPs to assess its contribution to model performance.

To account for seasonal variability while ensuring model adaptability, we employed a rolling five-year training window approach. Models were trained on daily data from each five-year window and subsequently tested on the following year's data (Fig. 2). The window was then advanced by one year, with the model retrained on the new five-year segment and tested on the subsequent year. We chose this one-year interval to capture a full hydrological cycle for a robust performance assessment. While more frequent retraining is possible in practice, the annual update provides a computationally efficient and conservative benchmark for our evaluation framework. If the ML model performs well with annual updates, it establishes a baseline from which we can infer that more frequent updates would perform at least as well as, if not better than, yearly updates. This approach enabled us to evaluate performance across different time periods.

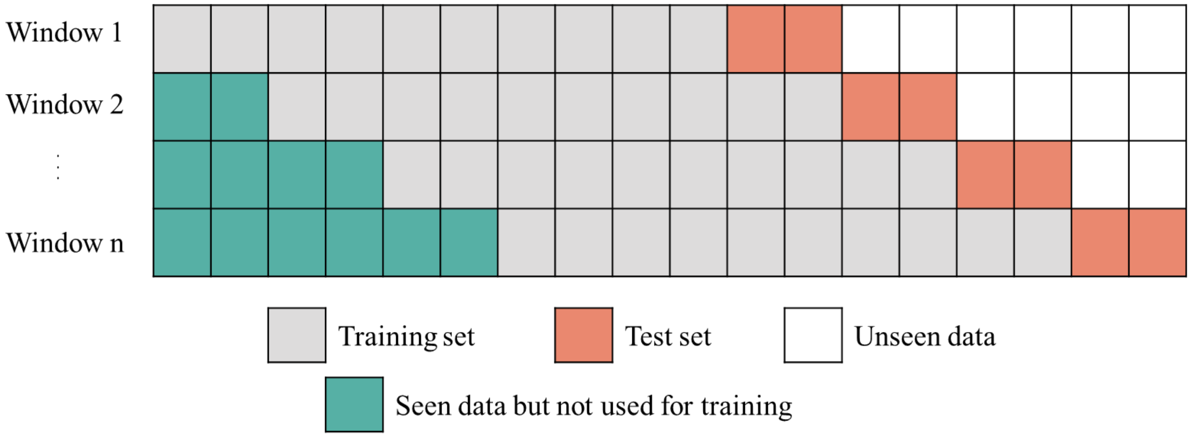


Figure 2: Illustration of the rolling window approach for model training and testing. The model is trained on a fixed-size window of past observations, which progressively moves forward through the dataset over time. The model is tested on the immediately succeeding data points to evaluate its performance (Modified from Amat Rodrigo and Escobar Ortiz, 2024).

The five-year training window duration was selected to balance model stability and responsiveness. While longer windows provide more data for learning complex relationships and reduce overfitting to short-term fluctuations, they may incorporate outdated patterns. By advancing the window yearly, the model adapts to recent data while maintaining sufficient historical context for robust generalization.

2.3 Regression Model Architectures and Hyperparameter Tuning

We implemented an LSTM network (Text S1) to predict BEX concentrations from timeseries iWQP data (i.e., pH, DO, EC, ORP). This architecture was selected to address a critical limitation of classical regression approaches: their inability to process temporal relationships between sequential measurements (Zhang et al., 2023). Unlike regression models that treat daily inputs as independent observations, LSTMs maintain an internal memory state through their gated architecture (input, forget, and output gates) (Hochreiter and Schmidhuber, 1997), allowing them to process data sequentially and selectively retain information over time (Fig. S1 in the Supporting Information).

An important aspect of the LSTM is its sequence-based input. The hyperparameter governing this is the sequence length, which defines the number of consecutive daily observations used to make a single prediction. Hyperparameters are configuration settings for an ML algorithm that are established prior to the training process and govern how the model learns from data. Unlike internal model parameters that are learned during training, hyperparameters must be predefined and significantly impact predictive performance (Wu et al., 2019).

The LSTM model was trained on overlapping sequences to predict the next value in the sequence. For example, if the sequence length is 30, the model uses data from 30 consecutive days to predict the value for the next day. As shown in Fig. 2, for each five-year training window, we generated all possible overlapping sequences of the chosen length. For instance, with a 30-day sequence length, the first training sample consists of days 1–30 of iWQP values, and the target output is the BEX concentration on day 31. The next sample uses days 2–31 to predict day 32, and so on, until the entire five-year period is covered. We tested four sequence lengths (i.e., 30, 90, 180, and 360 days) to capture different temporal contexts. Shorter sequences (30 days) emphasize recent changes, while longer sequences (360 days) incorporate seasonal and long-term trends. These values provide a balance between capturing relevant patterns and maintaining computational feasibility.

For comparative assessment, we implemented four classical machine learning models trained on the same 5-year rolling windows. As these models require single-day input, the iWQPs from a single day were used to predict that day's BEX value. MLR served as a simple, interpretable baseline that captures linear relationships without tunable hyperparameters. We also implemented SVR, RF, and XGBoost to handle non-linear relationships. To optimize the performance of these non-linear models, we tuned their hyperparameters (Text S2) using Tree-structured Parzen Estimator (TPE) from the Hyperopt library, a Bayesian optimization method (Bergstra et al., 2013). TPE efficiently navigates the hyperparameter space by building a probabilistic model from previous trials to intelligently select the most promising configurations for subsequent evaluation, thus finding high-performing settings with fewer evaluations than exhaustive methods.

We also tested an LSTM–XGBoost hybrid model that combined the strengths of temporal feature learning with robust feature-based decision making. This architecture processed sequential iWQPs through the LSTM layer and then fed the transformed features into the XGBoost component for final prediction. Furthermore, we quantified prediction uncertainty for the LSTM using Monte Carlo dropout to generate 95 % confidence intervals for the estimates.

All modeling was conducted in Python. We employed Scikit-learn for the regression models (Pedregosa et al., 2011), TensorFlow and Keras for the LSTM (Abadi et al., 2016; Chollet et al., 2015), Hyperopt for tuning (Bergstra et al., 2013), Pandas for data manipulation (McKinney, 2010), NumPy for numerical computing (Harris et al., 2020), and Matplotlib for visualization (Caswell et al., 2020).

2.4 Hybrid Kalman Filtering for Source Removal Scenario

To improve model responsiveness during abrupt system changes, we implemented a hybrid KF (Text S3, Fig. S2 in the Supporting Information) specifically for the source removal scenario. This scenario simulates the complete excavation of the LNAPL source zone at year 40, resulting in a sudden drop in PHC dissolution. The KF recursively updated BEX concentration estimates by integrating incoming iWQP measurements with direct BEX measurements (from RTM output), enabling real-time adaptation to these transient conditions.

KF operates through two primary steps: prediction and update. In the prediction step, the state vector is projected forward in time using a state transition model. This model is defined by the state transition matrix F and contains the system's variables of interest. Thus, it predicts the BEX concentration at the next time step based on its current value and the system's dynamics. This step also incorporates the inherent uncertainty of the model itself, characterized as process noise with covariance matrix Q (Simon, 2006).

For the subsequent update step, this prediction is corrected using new observations typically obtained from sensors or laboratory measurements. An observation model maps the state vector to the measurement domain and is defined by the observation matrix. The Kalman gain is then computed to optimally balance the uncertainty between the model's prediction and the new measurement by weighting their respective covariances. The measurement noise, representing sensor inaccuracy, is defined by its covariance matrix. Both the process and measurement noise are assumed to be Gaussian.

In our KF implementation, we considered the BEX concentration and effective attenuation rate as the state vectors in the prediction step. We assumed a first-order kinetic model governed by the differential Eq (1):

$$dC_{bex}/dt = -\lambda C_{bex} \ ,$$

where C_{bex} is the BEX concentration, and λ is the effective attenuation rate constant. We assumed that λ approximates the combined effects of advective transport and biodegradation after the LNAPL source was removed.

For the update step, we employed an ensemble approach combining SVR with radial basis and polynomial kernels, and RF regression. These models were trained to estimate BEX concentrations from iWQPs when direct BEX measurements were unavailable.

The hybrid KF model performance was evaluated using a rolling window validation strategy, with five-year training periods followed by 1-year testing windows. During testing, we assumed that BEX measurements were only available yearly. Thus, we masked the BEX concentration data to simulate yearly sampling frequency, while iWQPs remained available daily.

2.5 Performance Metrics

The predictive performance of machine learning models was evaluated by comparing the estimated BEX concentrations against values simulated by the RTM, considered as the ground truth. We used standard metrics for evaluation: mean absolute error (MAE), representing the arithmetic mean of absolute differences between predicted and actual values; coefficient of determination (R^2), indicating the proportion of variance in the observed data explained by the model; and mean absolute percentage error (MAPE), providing normalized error measurement independent of concentration magnitude.

3 Results and Discussion

We aimed to estimate daily BEX concentrations at observation wells using iWQPs by training separate LSTM models for each well and each test window. We defined the test window as a one-year period following the 5-year training data, which was used to evaluate model performance on unseen data (i.e., the first test window corresponds to year 6 of the RTM simulation after an oil spill).

3.1 Sequence Length Selection

A 30-day sequence length for the LSTM model consistently produced the highest number of test windows with accurate estimates of BEX concentration from iWQPs ($R^2 \geq 0.80$), across all observation wells when compared to longer sequence lengths. For instance, at well X0Z0, 34 out of 94 test windows had $R^2 \geq 0.80$ using the 30-day sequence (green circles in Fig. 3), compared to only 11, 32, and 21 test windows with 90-, 180-, and 360-day sequence lengths, respectively. While the 30-day model did not estimate BEX concentration accurately at all test windows, its better relative performance indicates that a month of historical iWQPs data (in our case, the virtual training data) provided the most effective temporal pattern for BEX estimation at the current timestep.

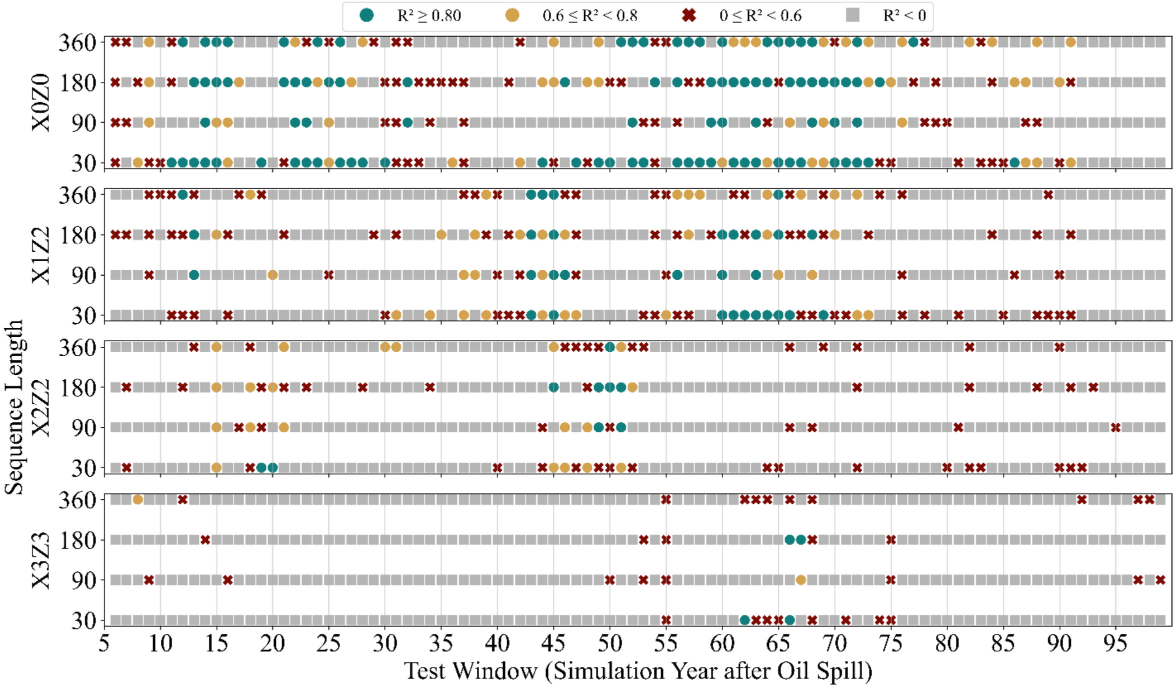


Figure 3: LSTM model predictions for observation wells X0Z0, X1Z2, X2Z2, and X3Z3 across time windows and sequence lengths of 30, 90, 180, and 360 days. Predictions with $R^2 \geq 0.80$ are shown as green circles, those with $0.60 \leq R^2 < 0.80$ as golden circles, and those with $0 \leq R^2 < 0.60$ as red crosses. Negative R2 values were set to 0 and are shown as grey squares.

While the high R^2 value of 0.80 was not met across all test windows, the model demonstrated a potential for practical application in general contaminant monitoring; more than 70 % of test windows at each observation well (68 out of 94) had a MAPE value below 5 %, while over 87 % (82 out of 94 test windows) had less than 10 % MAPE value at each observation well (Fig. 4). A 5 % error is widely considered as an acceptable margin of error for general monitoring of groundwater quality (Hounslow, 1995). Thus, R^2 values within the 0.6 to 0.8 range can still be considered as acceptable. For field applications, the model performance could be further improved through strategies such as more frequent retraining cycles and by limiting prediction horizons to shorter timeframes (e.g., one month instead of one year).

Furthermore, Fig. 3 shows suboptimal results as indicated by the predominance of red crosses and grey squares ($R^2 < 0.60$). At well X3Z3, for example, 98 % of test windows had $R^2 < 0.80$. However, this scenario served as a baseline; its main purpose was to illustrate the model's limitations before introducing additional input parameters that could improve performance, as presented in the subsequent sections.

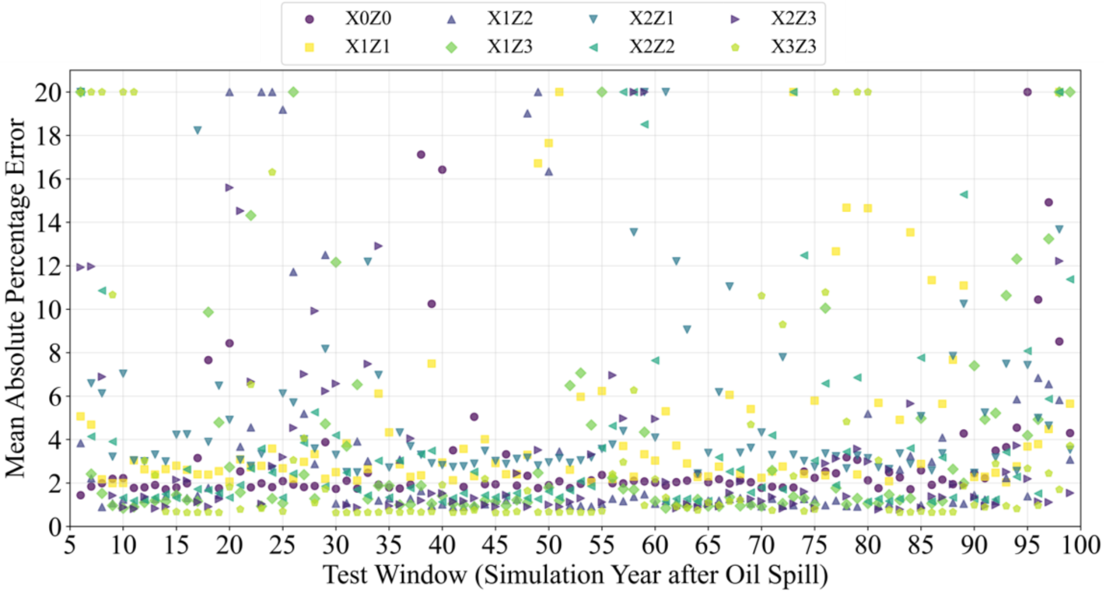


Figure 4. MAPE values for LSTM-predicted BEX concentrations across observation wells. Values above 20 % were capped at 20 % for visualization.

Results for other wells at different distances and depths relative to the source zone are included in the supporting information (Fig. S3 and Fig. S4). The exception was observation well X2Z2, where four test windows had $R^2 \geq 0.80$ for the 180-day sequence and only two test windows for the 30-day sequence. Although the improved performance with a longer sequence could be due to its ability to capture hydrogeochemical processes, such as plume migration or degradation rates that are specific to this well's location and depth, the inconsistency across only a few test windows is more likely attributable to stochastic variations in model weight initialization during the LSTM's training phase than to a fundamentally different hydrogeochemical condition.

3.2 Baseline Model Performance in Well X0Z0

Figure 5 illustrates the different performances of the LSTM model at well X0Z0 by comparing estimated versus actual (RTM-output) BEX concentrations across three representative test windows. These examples capture the range of observed outcomes: poor performance (R² ≈ -28.6, MAE ≈ 567 µg/L, MAPE ≈ 5.85 %), good performance (R² ≈ 0.80, MAE ≈ 71 µg/L, MAPE ≈ 1.84 %), and excellent performance (R² ≈ 0.95, MAE ≈ 32 µg/L, MAPE ≈ 1.96 %). In our study, we focused on R² when comparing model performance due to its interpretability and more truthful nature (Chicco et al., 2021), while acknowledging that MAE and MAPE also provide complementary insights into the model's accuracy.

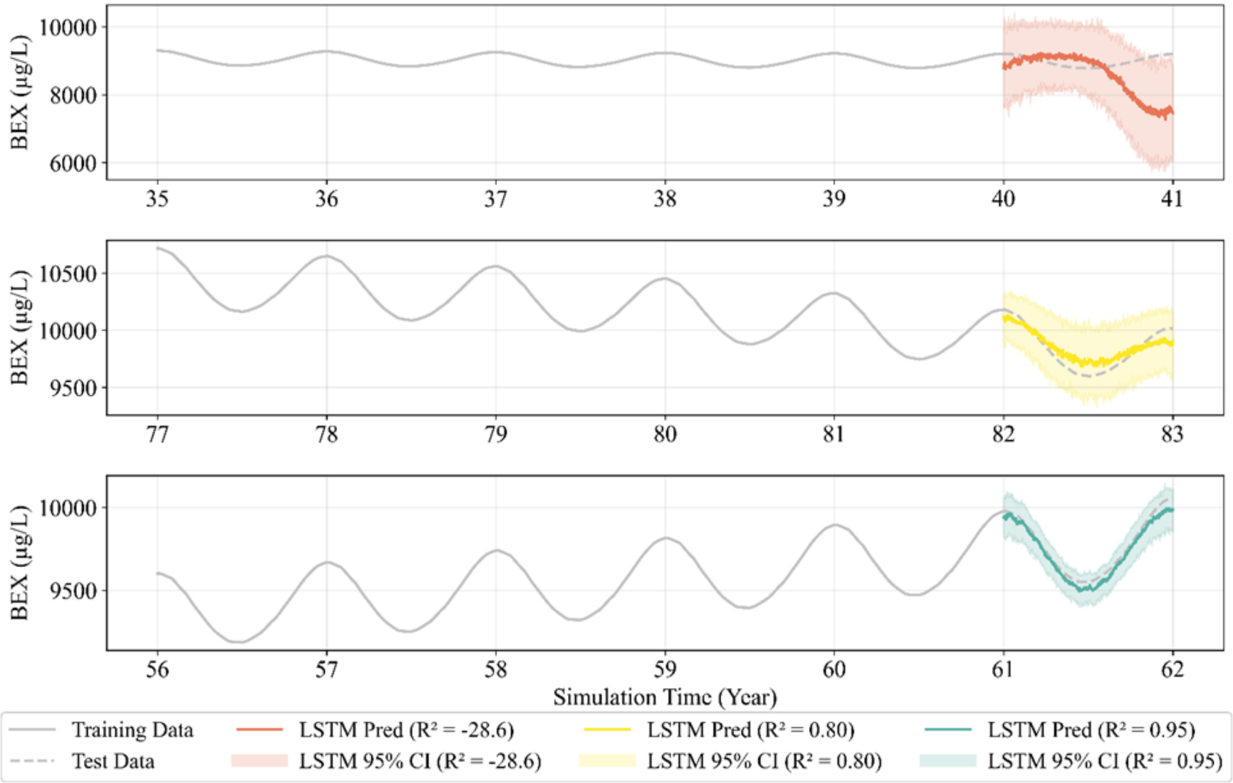


Figure 5: Sample training and testing windows with LSTM predictions using a 30-day input sequence at observation well X0Z0. Three predictions with different R² values are illustrated: approximately -28.6 (red), 0.80 (yellow), and 0.95 (green), corresponding to mean absolute error values of about 567 µg/L, 71 µg/L, and 32 µg/L, respectively. Shaded areas represent 95 % confidence intervals estimated using Monte Carlo dropout.

A negative R² indicates that the model performs worse than a simple mean predictor by failing to capture the underlying trend in the data (Chicco et al., 2021). Although its magnitude reflects the degree of deviation from the test data, we treated any negative R² as an indication of poor performance and truncated these values to zero.

This variability in predictive accuracy, as indicated by metric values such as R², MAE, and MAPE, is linked to the evolving hydrogeochemical conditions within the aquifer. The observed fluctuations in the LSTM model's performance align with the five identified hydrogeochemical periods defined in our previous study (Wu et al., 2024), where the relationship between iWQPs and BEX concentration changed in time.

The window with the poor fit ($R^2 < 0$) coincided with period 3 (around year 20 to year 60), where the BEX compounds reached a state of quasi-equilibrium. In this period, the rolling Spearman's correlation between iWQPs and BEX dropped to approximately zero from around year 25 to year 40. Instead, iWQPs became highly correlated with non-volatile dissolved organic carbon (NVDOC), a fraction of the dissolving light non-aqueous phase liquid (LNAPL). NVDOC became the primary electron donor, and thus the dominant driver of geochemical change. This results in the poor performance of the LSTM model for BEX concentration estimation.

Conversely, the window with a high R^2 of 0.95 coincided with the transition between period 3 and period 4 (around year 55 to year 65). During this time, iWQPs were strongly correlated with both BEX and NVDOC concentrations. The model successfully learned this combined, strong signal of the relationship between BEX and iWQPs, resulting in highly accurate predictions.

The window with good performance ($R^2 \approx 0.80$) corresponded to period 4 (around year 60 to year 90). Here, the dissolved NVDOC concentration peaked and began to decline. This reduction in NVDOC's influence allowed the correlation between iWQPs and the persistent BEX plume to re-strengthen to a moderate level. Even though the relationship was not as strong as in the transitional period, it provided a sufficient signal for the model to achieve reliable estimation accuracy. Also, while the specific R^2 values varied, the general performance trend tied to previously identified hydrogeochemical periods (Wu et al., 2024) was consistently observed at other test windows.

A key assumption in our analysis was that the models were not retrained during the one-year test period. In actual model deployment, models can and should be retrained more frequently, especially under highly variable field conditions. For test windows where the LSTM model performance declines, implementing a more frequent retraining schedule (e.g., weekly, monthly, or quarterly depending on data availability) would likely mitigate performance decay and maintain model accuracy.

3.3 Comparison Across Observation Wells

Figure 3 also shows that more test windows had accurate predictions at observation well X0Z0 (34 windows with $R^2 \geq 0.80$) than at other wells (fewer than 18 windows) for a 30-day sequence length. This spatial difference in model performance aligns with our previous finding that the relationship between BEX and iWQPs is not only temporally variable but also spatially heterogeneous. The better performance at X0Z0 is likely due to two factors, as established in our prior work (Wu et al., 2024).

First, the identified periods 1 to 4 began and ended earlier at well X0Z0 (closer to the source zone), than at distant wells such as X3Z3 (see Fig. 3 and Fig. 5 in Wu et al., 2024). For instance, the shift from BEX to NVDOC as the primary electron donor occurred after approximately 20 years at X0Z0 but was delayed until around 31 years at X3Z3 . Consequently, there were more test windows at X0Z0 where iWQPs were strongly correlated with BEX (e.g., the transition between Periods 3 and 4). At more distant wells, these predictive phases were less distinct, and their signal was weaker within any given five-year window, resulting in fewer windows with good accuracy ($R^2 \geq 0.80$).

Second, the signal (i.e., the change in iWQPs due to PHC degradation) was attenuated with distance from the source zone due to aquifer heterogeneity. From Wu et al. (2024) it can be concluded that the amplitude of water table fluctuations and their consequent impact on iWQP signals (e.g., pH and EC) were dampened at farther wells like X3Z3. This attenuation weakened the strength of the BEX-iWQP relationship, which the LSTM model relies upon. The stronger, less attenuated signal recorded at X0Z0 provides a clearer input pattern for the model to learn, thereby producing more consistently accurate predictions across multiple test windows. Thus, the location of the observation well was a key factor in determining whether BEX concentration can be accurately estimated from iWQPs, depending on travel times and signal strength.

3.4 Hydraulic Head as an Additional Input

Figure 6 presents heatmaps of R^2 values for LSTM predictions (30-day sequence length) across all observation wells, comparing model performance (a) without and (b) with hydraulic head included as an input feature alongside the four iWQPs. To avoid visual clutter, the figure illustrates a subset of the results, showing only every fifth test window starting from year 6 after the oil spill simulation. The addition of hydraulic head, which exhibited a high correlation with BEX concentration, markedly improved overall model performance.

This improvement was particularly evident at specific wells. At X0Z0, the number of test windows achieving $R^2 \geq 0.80$ increased from 34 to 43 out of 94 test windows. The highest improvement occurred at the farthest well, X3Z3, where the number of windows with $R^2 \geq 0.80$ increased from only 2 (~2 %) to 35 (~37 %) out of 94 test windows (Fig. S5 in the Supporting Information). This demonstrates that hydraulic head was a critical feature for estimating BEX concentration. Its importance is mechanistic: hydraulic head contains information on groundwater flow direction and advective transport of contaminants. In our RTM, simulated sinusoidal water table fluctuations were a key driver of concentration changes. By providing this signal explicitly, the LSTM could learn to recognize the fluctuation pattern as a coherent driver of BEX variability, rather than treating it as noise to be overcome using iWQPs alone.

On the other hand, a model trained only on hydraulic head performed poorly (yielding negative R^2 values; results not shown). While this model learned the sinusoidal temporal pattern, it failed to predict the correct magnitude of BEX concentration, showing that the geochemical information from iWQPs was essential for quantifying concentration levels.

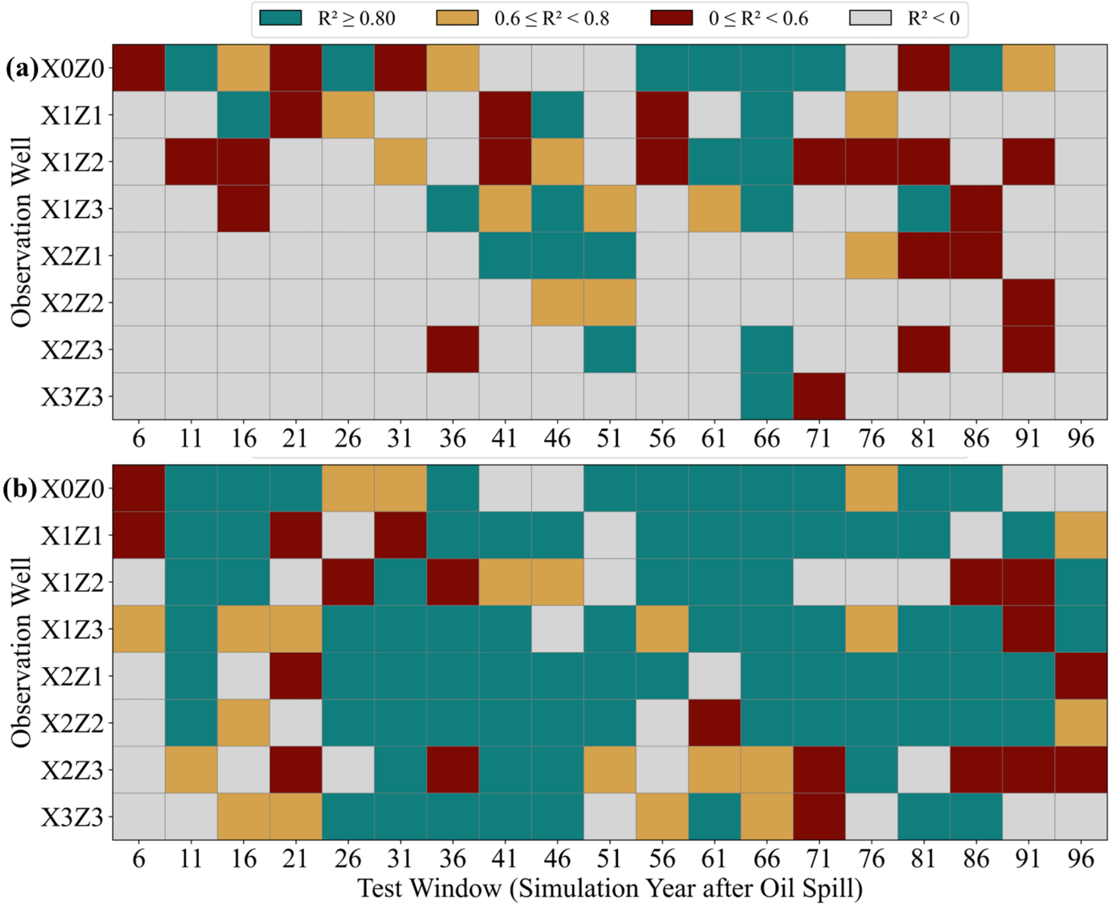


Figure 6: Comparison of LSTM model performance with and without hydraulic head as an input feature. Heatmaps show the R² values for estimating BEX concentrations across all observation wells and test windows for the 30-day sequence length LSTM model using (a) only iWQPs and (b) iWQPs and hydraulic head. Results are presented at every fifth test window for visual clarity.

345 However, performance gains were not the only observed results. Figure 7 illustrates the per-window change in R² across all eight wells after adding hydraulic head. While performance generally increased, especially at X3Z3, at well X0Z0 it decreased in more than 20 test windows. It must be noted that for Fig. 7, the test windows with a negative R² both before and after adding hydraulic head were categorized as showing "no change" in LSTM model performance (marked by yellow boxes). This classification was applied regardless of any variations in their mean absolute error (MAE).

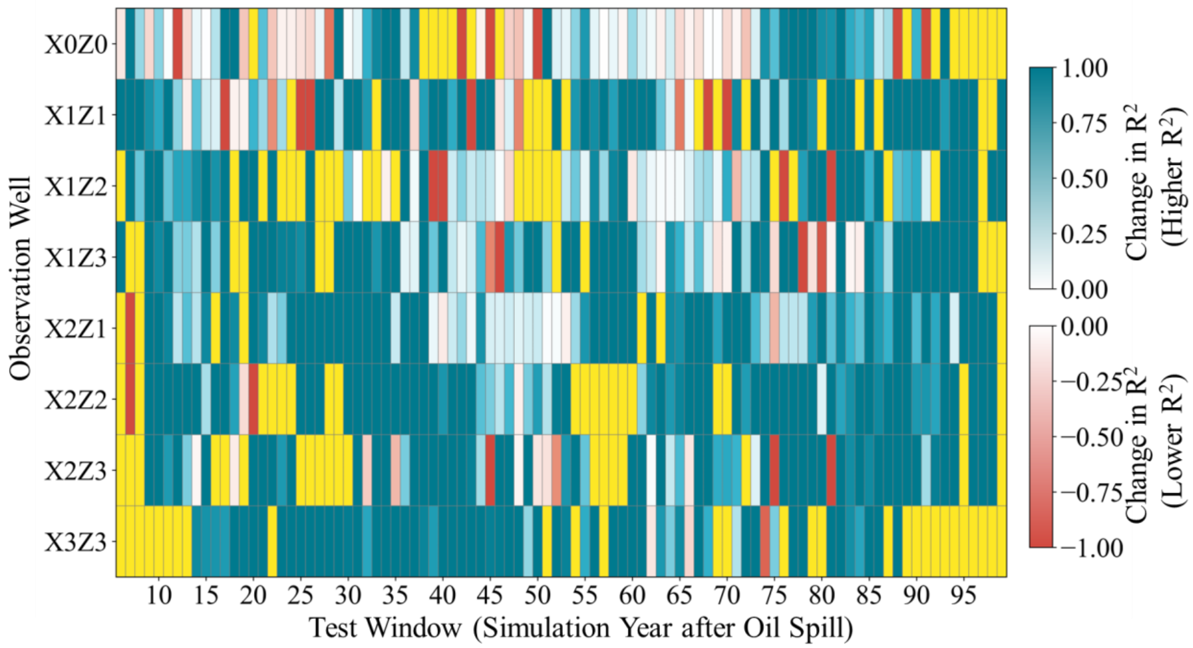
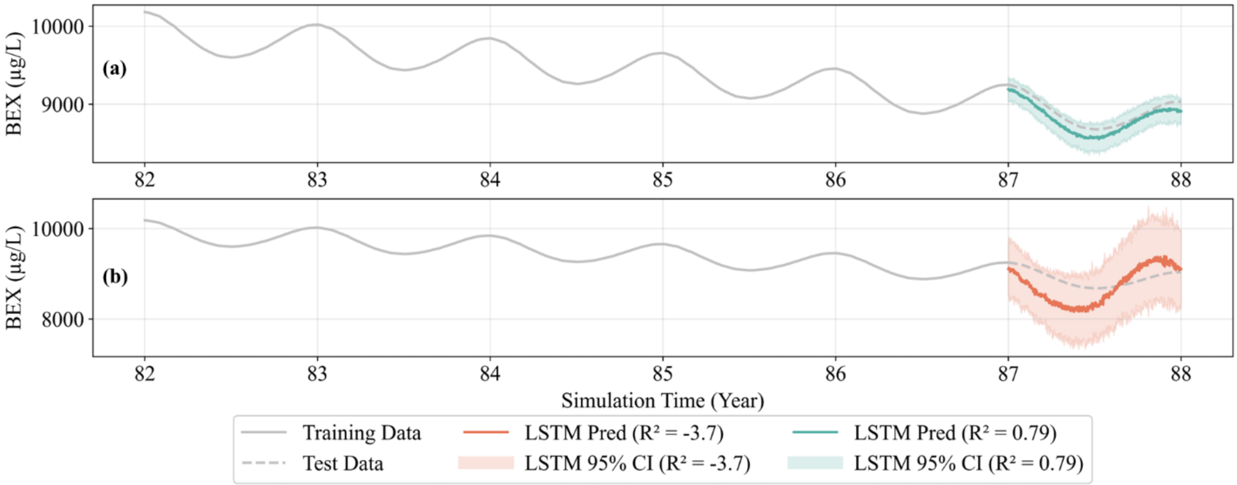


Figure 7: Changes in R^2 across test windows at eight observation wells after adding hydraulic head as an input feature alongside iWQPs in LSTM models (30-day sequence length). Yellow boxes denote windows with negative R^2 in both models (with and without hydraulic head). R^2 differences were clipped to the range [-1, 1] for visualization.

For example, during the test window from simulation year 87 to 88, a gradual decrease in BEX concentration induced changes in the iWQPs, whereas hydraulic head fluctuations remained stable. The model trained on the four iWQPs correctly captured this decreasing trend (Fig. 8a). However, when hydraulic head was added as an input parameter, the model's performance decreased (Fig. 8b). Although the hydraulic head was a useful predictive feature during the training period, its behavior in this specific test window created a conflicting signal: the average hydraulic head remained stable while the average BEX concentration decreased and other iWQPs changed (Fig. S6 in the Supporting Information). The ML model incorrectly attributed a portion of the BEX behavior to the stable hydraulic head signal it had learned during training. This reduced the influence of the more predictive iWQPs, leading to less accurate estimates and a lower R^2. This suggests that while the hydraulic head can be beneficial, it can potentially introduce noise and reduce model accuracy in periods where its behavior deviates from the target contaminant's behavior. Dynamically adjusting the importance or weight of features could potentially solve this. For example, if iWQPs show strong variation but the average hydraulic head is flat, the model could learn to reduce the weight of hydraulic head (Yao and Ge, 2021).

Beyond these hydrogeochemical explanations, the stochastic nature of model training is thus, again, a contributing factor. Similar to other neural networks, the random initialization of model weights at the start of training influences which patterns the LSTM prioritizes and its convergence path (Narkhede et al., 2022). This randomness means that the exact number of test windows with an $R^2 \geq 0.8$ can vary between independent training runs. For example, in a separate LSTM model run (training and testing) under identical conditions, high-accuracy windows ($R^2 \geq 0.80$) at well X0Z0 for the 30-day sequence length were 27. This was lower than the 34 test windows observed in the initial run. Therefore, while the reported counts indicate a strong positive trend, they must be interpreted within the context of this inherent variability.



375 **Figure 8: Comparison of LSTM model performance in estimating BEX concentration (µg/L) at observation well X0Z0 before (a) and after (b) including hydraulic head as an additional input.**

3.5 Comparison between Regression Models

Our hypothesis that the LSTM model would outperform classical regression models such as MLR, SVR, RF, and XGB, was supported in several, but not all, test windows. The LSTM's ability to learn temporal patterns from sequences of iWQP data provided an advantage for estimation, particularly in scenarios where the correlation between iWQPs and BEX evolved over time. This better performance was especially noticeable at observation well X0Z0, where classical models had fewer than 20 test windows with $R^2 \geq 0.8$.

A representative example of the LSTM's better performance is shown in the test window at simulation years 33–34 at well X0Z0 (Fig. 9a). Here, the LSTM achieved a high accuracy ($R^2 \approx 0.84$), while all classical models produced negative R^2 values with high error (MAE > 150 µg/L). Even an LSTM-GBR hybrid model performed poorly ($R^2 \approx -0.86$) in this instance. A feasible explanation for the LSTM's success is that, while the instantaneous correlation (Spearman's rolling correlation) between iWQPs and BEX was low during the test year, the model leveraged its memory of stronger correlations in the preceding months within its 30-day input sequence. This historical context allowed it to make accurate predictions even as real-time correlations weakened, a capability the static models inherently lack. The subsequent rise in MAE after approximately 60 days suggests a limit to this predictive persistence as the system dynamics eventually deviated too far from the learned historical patterns.

However, the LSTM's reliance on temporal context could also be a disadvantage. In test window 13 at well X0Z0 (Fig. 9b), the LSTM performed very poorly ($R^2 \approx -16.6$), while MLR and the LSTM-GBR hybrid excelled ($R^2 \approx 0.88$ and 0.93, respectively). In this scenario, the relationship between iWQPs and BEX was likely strong and consistent within the specific

395 test window but may have been dissimilar to patterns in the immediate historical data used by the LSTM's sequence. The classical models, which treat each day independently, could directly map this strong contemporaneous correlation without the "confusion" of irrelevant historical data. The superior performance of the LSTM-GBR model here suggests its architecture successfully fused the LSTM's feature extraction with the gradient boosting's ability to leverage strong, static feature-target relationships, making it the most robust and adaptable model overall.

400 These observations confirm that model efficacy is not universal but is instead determined by the nature of the BEX-iWQP relationship at a given time. Our results showed that the optimal model depends on whether the predictive signal is embedded in temporal patterns (favoring LSTM) or in strong instantaneous correlations (favoring models like MLR or the LSTM-GBR hybrid). Temporal patterns were crucial when the geochemical processes were more complex, for example, a change in the dominant electron acceptor. However, the existence of windows where static correlations dominate, such as 405 when the geochemical reactions reach a quasi-equilibrium state, justified the use of a hybrid modeling approach or a model selection framework.

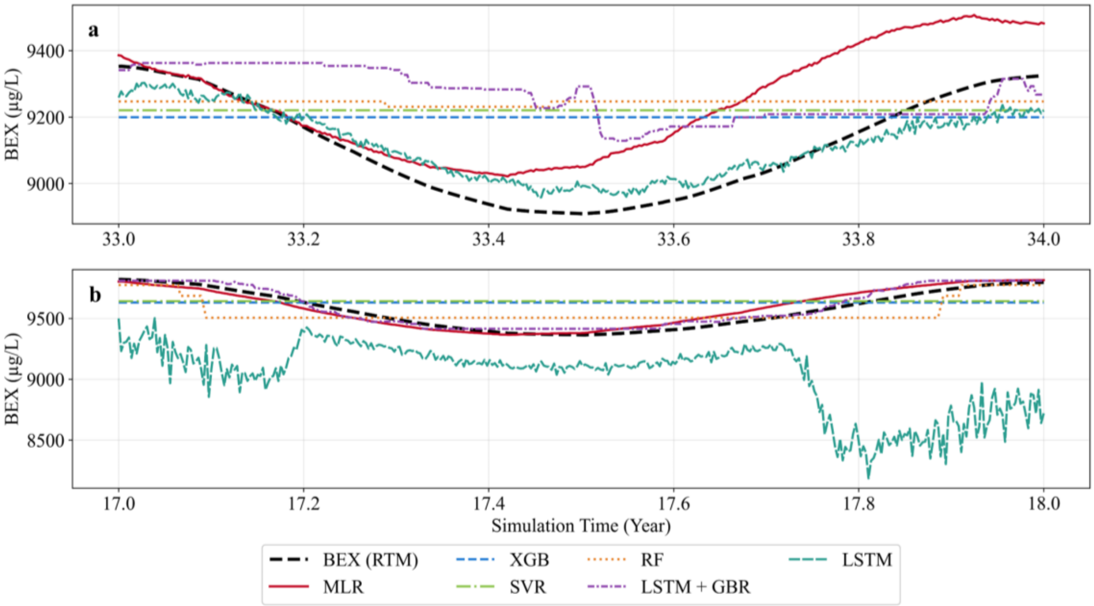


Figure 9: Comparison of BEX concentration estimates from six regression models at two test windows. The LSTM model (a) 410 outperforms and (b) underperforms the other five models.

3.6 Evaluation under Increased Hydraulic Gradient and Source Removal

Increased Hydraulic Gradient

A 10x increase in the hydraulic gradient degraded the LSTM performance across the observation wells. Probably, the sudden increase in flow velocity enhanced the dilution and "flushing" of the BEX plume, causing an immediate drop in BEX concentration at well X0Z0 (Fig. S7 in the Supporting Information). In other wells farther from the source zone, the faster groundwater flow initially pushed a concentrated BEX slug forward (causing a spike) before subsequently diluting it (causing the decline). The trained LSTM model failed to estimate BEX concentration and struggled to generalize under this hydrological condition (Fig. S8 in the Supporting Information). This behavior confirms a known limitation of deep learning models. LSTMs and other ML models often fail to extrapolate accurately when faced with conditions outside their training range (Baste et al., 2025). Improving the training strategy could address these limitations: adding more training data with a wider range or adjusting the model configurations.

Despite the poor estimation of BEX concentration, the LSTM correctly predicted the direction of concentration trends (i.e., increasing or decreasing) in six of the eight wells (X0Z0, X1Z1, X1Z2, X1Z3, X2Z1, and X2Z3); the model failed to capture the increasing BEX concentration at wells X2Z2 and X3Z3. This ability to detect a sudden change in trend, particularly in the six wells where directional shifts were correctly predicted even when the magnitude was inaccurate, serves as an effective indicator of an anomaly in groundwater processes. Thus, governing bodies could use these predictions to trigger a manual sampling campaign, changing a model's limitation into a practical tool for adaptive monitoring.

LNAPL Source Removal

The LSTM model's performance after source removal was notably poor, failing to capture the rapid decline in BEX concentration. The model predicted that concentrations would remain at an equilibrium state, despite the actual decrease observed in the RTM output (Fig. S9 in the Supporting Information). This failure probably occurred because the relationship between BEX concentration and the iWQPs was altered by the source removal; the existing BEX plume continued to undergo biodegradation without additional source, as observed from the RTM results. This sustained the geochemical reactions that affect the iWQPs. However, the sudden drop in BEX concentration due to the removed source was much faster than the decline from biodegradation. This created a discrepancy between the actual BEX concentration and the concentration effectively governing the iWQP signal. The LSTM model, trained on a system with an active source, lacked the necessary information to detect this change. It therefore interpreted the stable geochemical conditions as evidence of a persistent, steady-state plume, leading to the model's inaccurate predictions.

An exception was observed at well X2Z1, where the LSTM model correctly predicted a decreasing trend in BEX concentration following source removal (Fig. S9 in the Supporting Information). This response, however, may be attributed to a pre-existing decline in concentration at that location rather than an accurate prediction of the source removal's impact.

The LSTM's inability to predict the BEX concentration decline after source removal emphasizes an important limitation of the model that could not extrapolate beyond the conditions seen in its training data. This demonstrates the necessity for a recursive updating method like the Kalman Filter (KF), which can continuously adjust its estimates as new information becomes available (Simon, 2006).

In the subsequent KF experiment (Fig. S10 in the Supporting Information), the BEX concentration measurements were assumed to be available only annually, while iWQP measurements were available daily. Our KF implementation used a first-order attenuation process model for BEX, with an effective attenuation rate (λ) updated every five years. This parameter intended to capture the combined effects of advection and biodegradation. However, the KF did not immediately capture the initial decrease in BEX concentration (Fig. S10 in the Supporting Information**)** either. Its estimation only aligned with the RTM output after incorporating the first annual BEX measurement following source removal. While the filter did not estimate the exact concentration values well, it successfully captured the general decreasing trend following the update.

The KF framework thus served as a foundation for applying data assimilation to dynamic systems like source removal. Future improvements could involve developing a more complex process model that better represents subsurface physics, potentially even integrating the RTM itself. Additionally, the observation model could be enhanced by creating a more robust statistical translator between iWQP sensor data and BEX concentration. Finally, more frequent BEX measurements, such as quarterly or monthly, would significantly improve the accuracy and timeliness of the KF's estimations.

3.7 Final Discussion and Recommendations

Metrics commonly used in ML model evaluation provide statistical benchmarks for model performance. However, these metrics are still subject to individual interpretations especially in terms of practical applications. The required level of model accuracy is not absolute but must be evaluated against the objectives of the groundwater monitoring system. For instance, at a site with concentrations around 10,000 µg/L, a MAPE value of 5 % may be negligible for tracking general contamination concentration trends. However, this same error becomes highly significant if the goal is to ensure compliance with a strict regulatory standard. Therefore, the definition of good and bad model performance in terms of accuracy is dependent on the goal of the contaminant monitoring.

Incorporating hydraulic head data significantly enhanced BEX concentration estimates. This improvement was most pronounced in observation wells located farther from the source zone. In heterogeneous aquifers, additional information on groundwater flow and contaminant transport pathways becomes critical in these locations. The inclusion of hydraulic head offers a practical and cost-effective strategy to boost model performance, particularly at sites with strong hydrological seasonality. It must be considered that the hydraulic head adds a cost for the divers, along with the four in situ water quality sensors (**Table S2**). Moreover, the model reliability could be further improved by implementing more frequent retraining cycles and reducing prediction horizons (e.g., monthly instead of yearly).

Selecting the most suitable regression model for field application remains challenging due to deviations between real-world conditions and RTM simulations. Nevertheless, our proposed monitoring strategies were effective for estimating concentration trends and identifying hotspot areas. This enables the development of adaptive monitoring systems that can issue alerts when regulatory thresholds are exceeded, and trigger increased measurement frequency to identify the causes and characterize sudden concentration changes. It should be emphasized that for compliance monitoring affecting critical areas such as drinking water wells, more intensive and costly direct measurements remain necessary.

There is also potential to enhance the KF framework. The process model could be refined by incorporating more complex mechanisms such as dissolution, advection, and biodegradation, bringing it closer to a full RTM. Similarly, the measurement model could account for chemical reactions or be updated with more frequent PHC measurements to increase accuracy. Furthermore, we recommend using hydraulic head data from both upstream and downstream wells relative to the target point. This configuration provides a direct measurement of the hydraulic gradient, allowing the model to better infer groundwater flow velocity and the timing of contaminant arrival; better information on hydraulic gradient could improve the model's predictive performance.

A primary limitation of our study is the lack of available extensive field data. There is a scarcity of BEX concentration and iWQP measurements for real-world model validation (Fig. S11 in the Supporting Information). Thus, future research should prioritize collecting high-resolution temporal data to capture contaminant plume dynamics more accurately and allow for model evaluation under field conditions.

4 Conclusion

We evaluated the prediction accuracy of LSTM and classical regression models in estimating BEX concentrations using iWQPs, using virtual data from an RTM under controlled conditions. We found that:

- While LSTM model performance was poor, based on R^2 values ($R^2 < 0.80$ in 98 % of test windows at well X3Z3), it is useful for practical applications; over 70 % of test windows fall within a 5 % error margin (MAPE), which is acceptable for general water quality monitoring.
- Incorporating hydraulic head considerably improved LSTM accuracy, increasing the number of test windows with $R^2 \geq 0.80$ at well X3Z3 from 2 % to 37 %. However, while hydraulic head contained information on contaminant transport, it can introduce noise and reduce accuracy when contaminant behavior differs from groundwater flow (e.g., when biodegradation dominates over advection).
- The optimal model in terms of accuracy for predicting BEX concentrations depends on the nature of the BEX-iWQP relationship. LSTMs were superior for capturing temporal patterns, while the MLR and LSTM-GBR hybrid models performed better when strong instantaneous correlations exist.
- During an extreme hydraulic gradient increase, the LSTM's predictive accuracy declined, confirming its inability to extrapolate beyond training data. Nonetheless, it correctly predicted the direction of concentration trends in 75 % of the observation wells. Thus, the model remained valuable as an anomaly detection system to trigger manual sampling.
- Following source removal, the LSTM model failed to capture the rapid concentration decline, incorrectly predicting a steady-state equilibrium. In contrast, a recursive method like the KF successfully captured the decreasing trend after source removal by continuously updating its estimates with new measurements.

Future work should prioritize collecting high-resolution temporal field data from PHC-contaminated aquifers to evaluate and refine these models under real-world conditions.

Interactive computing environment

The source of time-series data in this paper was from the reactive transport model built in Jupyter Notebook using the Python package Flopy. The Jupyter Notebook for the reactive transport model is preserved at 4TU.ResearchData repository https://doi.org/10.4121/f7742f02-ee3a-4a84-adf1-625b4a9fd703 (Wu, 2024). The Jupyter Notebook and the python module created to visualize the simulation results are also available at this repository. The software/code repository for Flopy is available at https://github.com/modflowpy/flopy (Bakker et al., 2024). The Jupyter Notebook used for processing data, training, and evaluating machine learning models in this paper can be accessed at 4TU.ResearchData repository https://doi.org/10.4121/0a23147e-ba85-4ba2-a058-ba199c65d711 (Wu et al., 2025).

Author Contribution

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.

Acknowledgement

This work was performed in the cooperation framework of Wetsus, European Centre of Excellence for Sustainable Water Technology (www.wetsus.nl). Wetsus is co-funded by the European Union (Horizon Europe, LIFE, Interreg and EDRF), the Province of Fryslân and the Dutch Government: Ministry of Economic Affairs (TTT, SBO & PPS-I/TKI Water Technology), Ministry of Education, Culture and Science (TTT & SBO) and Ministry of Infrastructure and Water Management (National Growth Fund - UPPWATER). This research received funding from the Netherlands Organisation for Scientific Research (NWO) in the framework of the "Sustainable water technology Partnership Programme NWO-Wetsus on the Impact of Water Research on Energy, Industry, Health, Economy and Environment 2019. The authors like to thank the participants of the research theme "Monitoring & Quality" for the fruitful discussions and their financial support.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., … Brain, G.: TensorFlow: A System for Large-Scale Machine Learning. In Business Opp (Vol. 10, Issue July). https://www.usenix.org/conference/osdi16/technical-sessions/presentation/abadi, 2016.

Ahmadi, H., Kilanehei, F., Nazari-Sharabian, M.: Impact of Pumping Rate on Contaminant Transport in Groundwater—A Numerical Study. Hydrology, 8, 103. https://doi.org/10.3390/hydrology8030103, 2021.

Ahsan, M.M., Mahmud, M.A.P., Saha, P.K., Gupta, K.D., & Siddique, Z.: Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance. Technologies 2021, Vol. 9, Page 52, 9(3), 52. https://doi.org/10.3390/TECHNOLOGIES9030052, 2021.

Amat Rodrigo, J., & Escobar Ortiz, J.: skforecast (v0.17.0). Retrieved from https://skforecast.org/0.17.0/introduction-forecasting/introduction-forecasting (on 10 August 2025), 2024.

Arsenault, R., Martel, J.L., Brunet, F., Brissette, F., & Mai, J.: Continuous streamflow prediction in ungauged basins: Long short-term memory neural networks clearly outperform traditional hydrological models. Hydrology and Earth System Sciences, 27(1), 139–157. https://doi.org/10.5194/hess-27-139-2023, 2023.

Bakker, M., Post, V., Hughes, J. D., Langevin, C. D., White, J. T., Leaf, A. T., et al.: FloPy v3.7.0.dev0 (preliminary): U.S. Geological Survey software release, 08 February 2024, GitHub [code], https://doi.org/10.5066/F7BK19FH, 2024.

Banadkooki, F.B., Ehteram, M., Panahi, F., Sh. Sammen, S., Othman, F.B., & EL-Shafie, A.: Estimation of total dissolved solids (TDS) using new hybrid machine learning models. Journal of Hydrology, 587, 124989. https://doi.org/10.1016/J.JHYDROL.2020.124989, 2020.

Baste, S., Klotz, D., Espinoza, E.A., Bardossy, A., & Loritz, R.: Unveiling the Limits of Deep Learning Models in Hydrological Extrapolation Tasks. EGUsphere [preprint], https://doi.org/10.5194/egusphere-2025-425, 2025.

Beck, P. & Mann, B.: 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. ISBN: 978-1-921431-25-8, 2010.

Bergstra, J., Yamins, D. & Cox, D.: Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. Proceedings of the 30th International Conference on Machine Learning, in Proceedings of Machine Learning Research 28(1):115-123. Retrieved from https://proceedings.mlr.press/v28/bergstra13.html (on 30 June 2025), 2013.

Buerck, J., Roth, S., Kraemer, K., Scholz, M., & Klaas, N.: Application of a fiber-optic NIR-EFA sensor system for in situ monitoring of aromatic hydrocarbons in contaminated groundwater. Journal of Hazardous Materials, 83(1–2), 11–28. https://doi.org/10.1016/S0304-3894(00)00335-6, 2001.

Caswell, T.A., Droettboom, M., Lee, A., Hunter, J., Firing, E., Sales De Andrade, E., & Ivanov, P.: matplotlib/matplotlib: Rel: v3. 3.1 [Software]. Zenodo. https://matplotlib.org/, 2020.

Chiu, H.Y., Hong, A., Lin, S.L., Surampalli, R.Y., & Kao, C.M.: Application of natural attenuation for the control of petroleum hydrocarbon plume: Mechanisms and effectiveness evaluation. Journal of Hydrology, 505, 126–137. https://doi.org/10.1016/J.JHYDROL.2013.09.027, 2013.

Chollet, F., & others.: Keras. Retrieved from https://keras.io (on 20 June 2025), 2015.

Harris, C. R., Millman, K.J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... & Oliphant, T. E.: Array programming with NumPy. Nature, 585(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2, 2020.

Hochreiter, S., & Schmidhuber, J.: Long Short-Term Memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/NECO.1997.9.8.1735, 1997.

Hounslow, A.: Water Quality Data: Analysis and Interpretation (1st ed.). CRC Press. https://doi.org/10.1201/9780203734117, 1995.

Höyng, D., Prommer, H., Blum, P., Grathwohl, P., & Mazo D'Affonseca, F.: Evolution of carbon isotope signatures during reactive transport of hydrocarbons in heterogeneous aquifers. Journal of Contaminant Hydrology, 174, 10–27. https://doi.org/10.1016/J.JCONHYD.2014.12.005, 2015.

Industrial Economics, Incorporated.: High Resolution Site Characterization at Petroleum Underground Storage Tank Release Sites - Applicability, Benefits, and Costs. Final Report. Retrieved from https://www.epa.gov/system/files/documents/2023-04/High%20Resolution%20Site%20Characterization%20Study%20Report.pdf (on 20 August 2025), 2023.

Ji, J., Deng, C., Shen, W., & Zhang, X.: Field analysis of benzene, toluene, ethylbenzene and xylene in water by portable gas chromatography–microflame ionization detector combined with headspace solid-phase microextraction. Talanta, 69(4), 894–899. https://doi.org/10.1016/J.TALANTA.2005.11.032, 2006.

Karimi, H., Sahour, S., Khanbeyki, M., Gholami, V., Sahour, H., Shahabi-Ghahfarokhi, S., & Mohammadi, M.: Enhancing groundwater quality prediction through ensemble machine learning techniques. Environmental Monitoring and Assessment, 197(1), 1–25. https://doi.org/10.1007/S10661-024-13506-0, 2025.

Larsson, H., & Dasgupta, P.K.: Liquid core waveguide-based optical spectrometry for field estimation of dissolved BTEX compounds in groundwater: A feasibility study. Analytica Chimica Acta, 485(2), 155–167. https://doi.org/10.1016/S0003-2670(03)00423-9, 2003.

McKinney, W.: Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference, 56–61. https://doi.org/10.25080/MAJORA-92BF1922-00A, 2010.

Mitchell, H.B.: Multi-sensor data fusion: An introduction. In Multi-Sensor Data Fusion: An Introduction. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-71559-7, 2007.

Narkhede, M.V., Bartakke, P.P., & Sutaone, M.S.: A review on weight initialization strategies for neural networks. Artificial Intelligence Review, 55(1), 291–322. https://doi.org/10.1007/S10462-021-10033-Z, 2022.

Ng, G.H.C., Bekins, B.A., Cozzarelli, I.M., Baedecker, M.J., Bennett, P.C., Amos, R.T., & Herkelrath, W.N.: Reactive transport modeling of geochemical controls on secondary water quality impacts at a crude oil spill site near Bemidji, MN. Water Resources Research, 51(6), 4156–4183. https://doi.org/10.1002/2015WR016964, 2015.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830, 2011.

Qiao, F., Wang, J., Song, J., Chen, Z., Kwaw, A.K., Zhao, Y., & Zheng, S.: The spatiotemporal evolution of dissolved-phase NAPL plumes revealed by the integrated groundwater quality and machine learning models. Water Research, 280. https://doi.org/10.1016/j.watres.2025.123535, 2025.

Sani Gaya, M., Isah Abba, S., Abdu, A.M., Tukur, A.I., Saleh, M.A., Esmaili, P., Wahab, N.A., & Gaya, M.S.: Estimation of water quality index using artificial intelligence approaches and multi-linear regression. IAES International Journal of Artificial Intelligence (IJ-AI, 9(1), 126–134. https://doi.org/10.11591/ijai.v9.i1.pp126-134, 2020.

Simon, D.: Optimal state estimation: Kalman, H∞, and nonlinear approaches. In D. Simon (Ed.), Optimal State Estimation: Kalman, H∞, and Nonlinear Approaches. Wiley Blackwell. https://doi.org/10.1002/0470045345, 2006.

Singh, S.K., Shirzadi, A., & Pham, B.T.: Application of Artificial Intelligence in Predicting Groundwater Contaminants. Water Pollution and Management Practices, 71–105. https://doi.org/10.1007/978-981-15-8358-2_4, 2021.

Szomolányi, O., & Clement, A.: Use of random forest for assessing the effect of water quality parameters on the biological status of surface waters. GEM - International Journal on Geomathematics, 14(1), 1–29. https://doi.org/10.1007/S13137-023-00229-6/TABLES/3, 2023.

Thouement, H.A.A., & Van Breukelen, B.M.: Virtual experiments to assess opportunities and pitfalls of CSIA in physical-chemical heterogeneous aquifers. https://doi.org/10.1016/j.jconhyd.2020.103638, 2020.

Wong, K.T., Jang, S.B., Yoon, S. Y., Ryu, B., Valiyaveettil Basheer, R., Abd Rahman, N., Choong, C.E., Jung, J., & Jang, M.: Comparative Analysis of Reliability in On-Time Monitoring Data for NO3-N, BTEX, and TOC: Commercialized Sensors versus Spectroscopic Methods. ACS ES and T Water. https://doi.org/10.1021/acsestwater.4c00487, 2024.

Wu, C. L. R.: Virtual experiments with reactive transport modelling using FloPy: Transport and degradation of dissolved petroleum hydrocarbons in groundwater, 4TU.ResearchData [code], https://doi.org/10.4121/f7742f02-ee3a-4a84-adf1-625b4a9fd703, 2024.

Wu, C.L.R., Wagterveld, R.M., & van Breukelen, B.M.: Reactive Transport Modeling for Exploring the Potential of Water Quality Sensors to Estimate Hydrocarbon Levels in Groundwater. Water Resources Research, 60(4). https://doi.org/10.1029/2023WR036644, 2024.

Wu, C.L.R., Wagterveld, R.M., Rietveld, L.C., van Breukelen, B.M.: Hybrid Machine Learning Models for Estimating Petroleum Hydrocarbon Concentration in Groundwater, 4TU.ResearchData [code], https://doi.org/10.4121/0a23147e-ba85-4ba2-a058-ba199c65d711, 2025.

Wu, C.L.R., Wagterveld, R.M., Rietveld, L.C., van Breukelen, B.M.: Machine learning-based in-situ detection of toxic petroleum hydrocarbons in groundwater. Journal of Contaminant Hydrology. 104771, ISSN 0169-7722. https://doi.org/10.1016/j.jconhyd.2025.104771, 2026.

Wu, J., Chen, X.Y., Zhang, H., Xiong, L.D., Lei, H., & Deng, S.H.: Hyperparameter Optimization for Machine Learning Models Based on Bayesian Optimization. Journal of Electronic Science and Technology, 17(1), 26–40. DOI: 10.11989/JEST.1674-862X.80904120, 2019.

Yao, L. & Ge, Z.: Dynamic Features Incorporated Locally Weighted Deep Learning Model for Soft Sensor Development. IEEE Transactions on Instrumentation and Measurement, vol. 70, pp. 1-11, 2021, Art no. 2508011. DOI: 10.1109/TIM.2021.3073702, 2021.

Zanello, V., Scherger, L.E., & Lexow, C.: Assessment of groundwater contamination risk by BTEX from residual fuel soil phase. SN Applied Sciences, 3(3), 1–20. https://doi.org/10.1007/S42452-021-04325-W, 2021.

Zhang, A., Lipton, Z.C., Li, M., & Smola, A.J.: Dive into deep learning. Cambridge University Press. https://d2l.ai, 2023.