Supplementary Material for

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

C. L. R. Wu[1, 2], R. M. Wagterveld[1], L.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.

Contents of this file

Equation S1. Average of pumping flow-rate-weighted mean concentration equation

$$C_{\text{sample}} = \frac{\sum C_i K_i}{\sum K_i},$$

where:

C_{sample} = concentration at the virtual observation well;

C_{i} = species concentration; and

K_i = hydraulic conductivity of each grid cell *i*.

Text S1. LSTM Architecture and Hyperparameters

The LSTM model was constructed with two layers of 256 units each to capture complex temporal dependencies. A dropout rate of 0.1 was applied between layers, regularized by L2 regularization (λ=0.001), to mitigate overfitting. The learned sequences were passed to a dense feedforward layer with 32 units for final output processing. The model was compiled with the Adam optimizer using a learning rate of 0.001.

Text S2. Hyperparameter Search Space of Classical Regression Models

The search spaces for the classical regression models were defined as follows: for support vector regression (SVR), we tuned the kernel function (i.e., radial basis function, polynomial and, sigmoid), the regularization parameter C from e^{1} to e^{5}, epsilon ε from e^{-10} to e^{10}, and kernel coefficient γ from e^{-1} to e^{10}; for random forest (RF), we optimized the number of trees from 2 to 500, the maximum depth from 1 to 100, min_samples_split from 0.001 to 0.2, min_samples_leaf from 0.001 to 0.1, and the feature selection strategy (i.e., sqrt, log2, or none); for XGBoost, we tuned the number of estimators from 2 to 500, max depth from 1 to 100, the learning rate log-uniformly from e^{-5} to $e^{-0.7}$, the subsampling ratio from 0.5 to 1.0, the column sampling from 0.5 to 1.0, and gamma from 0 to 7.

Text S3. Hybrid Kalman Filter Architecture

The hybrid predictive framework integrates Kalman filtering with machine learning to estimate BEX concentrations using both direct BEX and iWQPs values from the RTM (**Figure S2**). The system architecture consists of four key components: data preparation, effective attenuation rate estimation, ensemble observation modeling, and Kalman filtering with rolling window validation. Daily time-series from the RTM were compiled for BEX concentrations and four iWQPs at each observation well.

To model the attenuation of BEX following source removal, we assumed a first-order kinetic model governed by the differential equation,

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

where C_{bex} is the BEX concentration and λ is the effective attenuation rate constant. Because of time scale decomposition, the effective attenuation rate was estimated per 5-year training window using constrained nonlinear least-squares fitting of the exponential model:

$$C_{bex,t} = (C_{bex,t-1}) \cdot e^{-\lambda_{t-1}\Delta t},$$

implemented via SciPy's curve fit function with bounds to prevent non-physical negative values. We assumed that λ approximates the combined effects of advective transport and biodegradation.

Residuals from the fitting process were used to derive the process noise covariance matrix Q, which incorporated both concentration variance and λ uncertainty. An off-diagonal term was included to reflect the expected negative correlation between these variables.

For observation modeling, we employed an ensemble approach combining two SVR models with an RF regressor. One of the SVR had a radial basis function kernel (kernel='rbf', C=100, epsilon=0.01) while the other has a second-degree polynomial kernel (kernel='poly', degree=2, C=100). The RF regressor has n_estimators=100. These models were trained to estimate BEX concentrations from iWQPs when direct BEX measurements were unavailable. Prior to training, all input features were standardized using StandardScaler() from Scikit-learn. The prediction uncertainty of the ensemble, quantified as the standard deviation of the training residuals, was used to dynamically define the uncertainty for the BEX estimate in the observation covariance matrix R. For the physically measured iWQPs, the diagonal elements of R were set to the reported accuracies of commercial in-situ water quality sensors (**Table S2**).

The KF tracked both BEX concentration and effective attenuation rate in a two-dimensional state vector [C_{bex}, λ]. The state transition matrix encoded the first-order kinetic model using the analytical solution of the governing equation, while allowing sensitivity to variations in λ. Unlike classical KFs that rely on a fixed observation model H, our implementation uses an ensemble of ML models to estimate BEX from iWQPs when direct measurements are missing. In this hybrid configuration, the ensemble ML model predictions are treated as surrogate observations with dynamically computed uncertainty, which is incorporated into the measurement noise covariance matrix R. A sparse observation matrix H_k is defined to link the predicted BEX to the state vector, enabling computation of the Kalman gain even in the absence of direct physical mapping. This approach allows the filter to adaptively balance model predictions and ensemble ML model estimates based on their respective uncertainties. Additional enhancements included adaptive scaling of process noise based on the time elapsed since the last measurement, Mahalanobis distance-based outlier detection (Chang, 2014), and numerical stabilization of the covariance matrix through eigen decomposition and regularization.

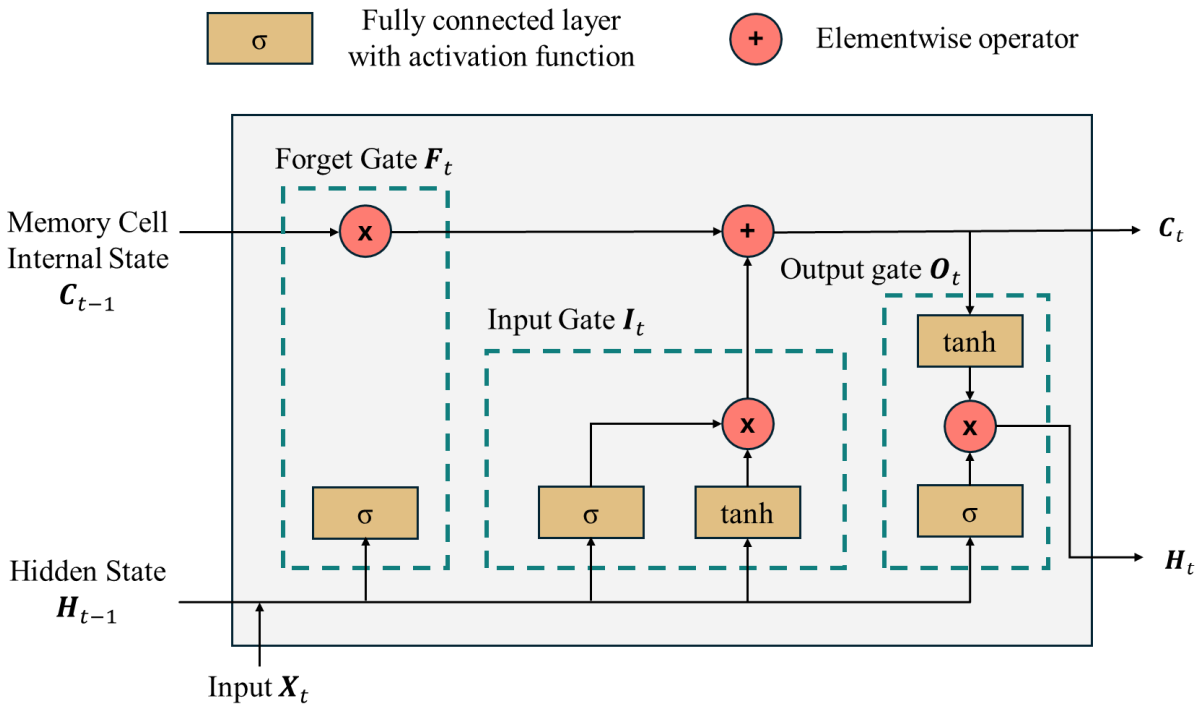

Figure S1. Framework of the LSTM model showing the fully connected layers with activation functions, gated mechanisms, and the computation of the hidden state. Modified from Zhang et al., 2023.

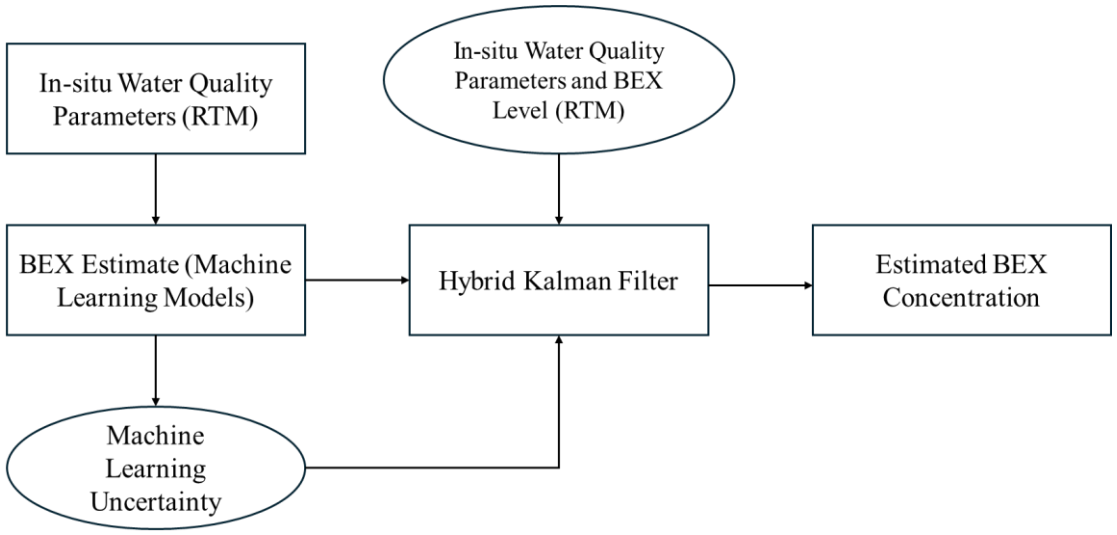

Figure S2. Schematic diagram of the hybrid Kalman filter framework. Monthly BEX levels were considered as input for the filter.

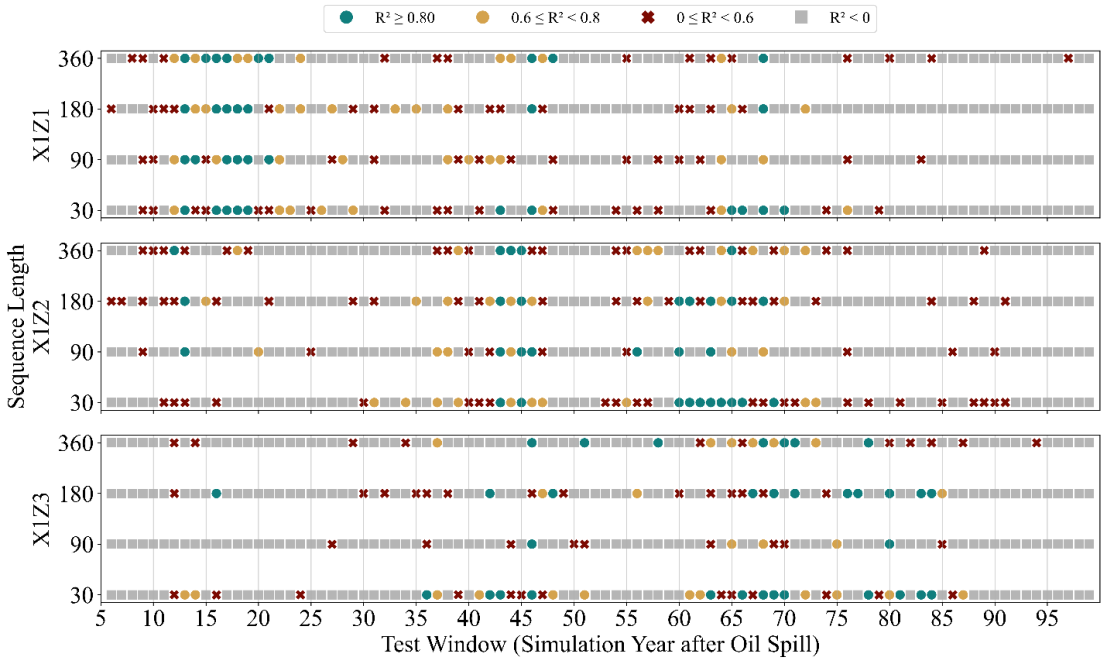

Figure S3. LSTM model predictions for observation wells X1Z1, X1Z2, and X1Z3 across time windows and sequence lengths of 30, 90, 180, and 360 days. Predictions with $R^2 \geq 0.80$ are shown as green circles, those with $0.60 \leq R^2 < 0.80$ as golden circles, and those with $0 \leq R^2 < 0.60$ as red crosses. Negative R^2 values were set to 0 and are shown as grey squares. **Figures S4** and **S5** reuse the legend from this figure for consistency across panels.

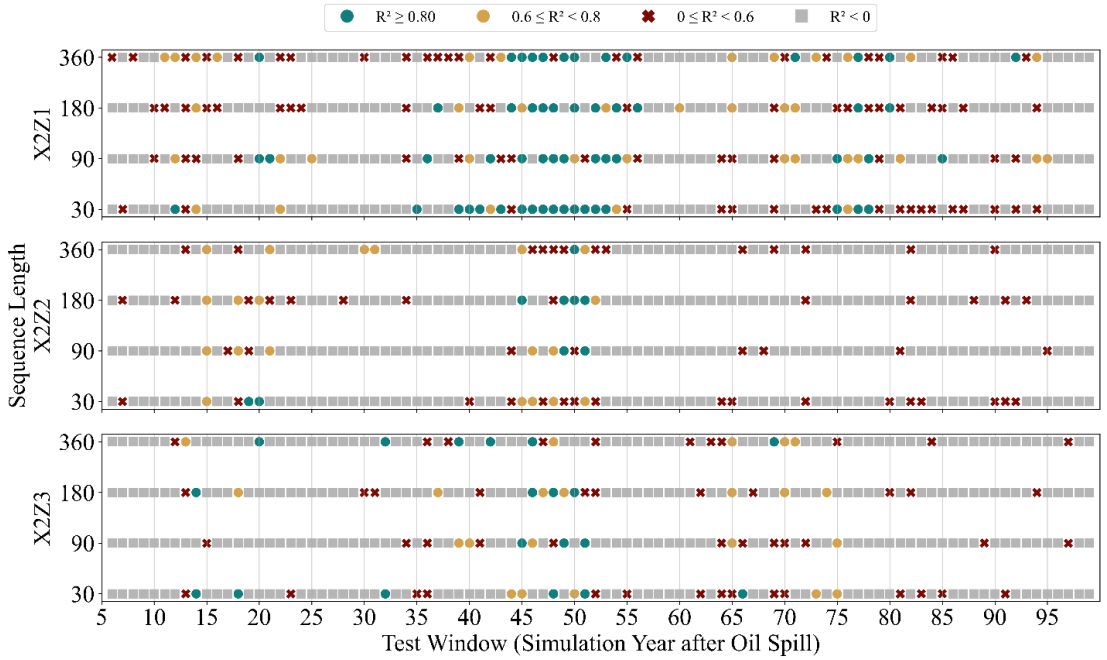

Figure S4. LSTM model predictions for observation wells X2Z1, X2Z2, and X2Z3 across time windows and sequence lengths of 30, 90, 180, and 360 days.

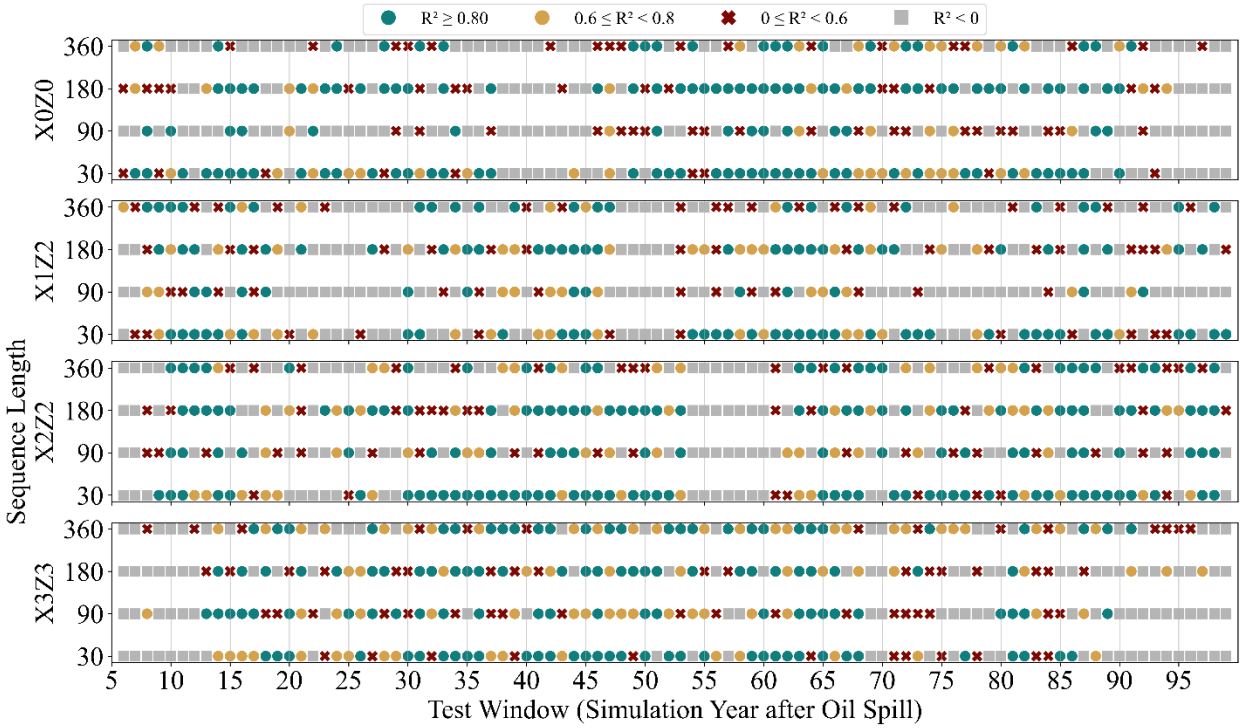

Figure S5. LSTM model predictions with hydraulic head as additional input parameter for observation wells X0Z0, X1Z2, X2Z2, and X3Z3 across time windows and sequence lengths of 30, 90, 180, and 360 days.

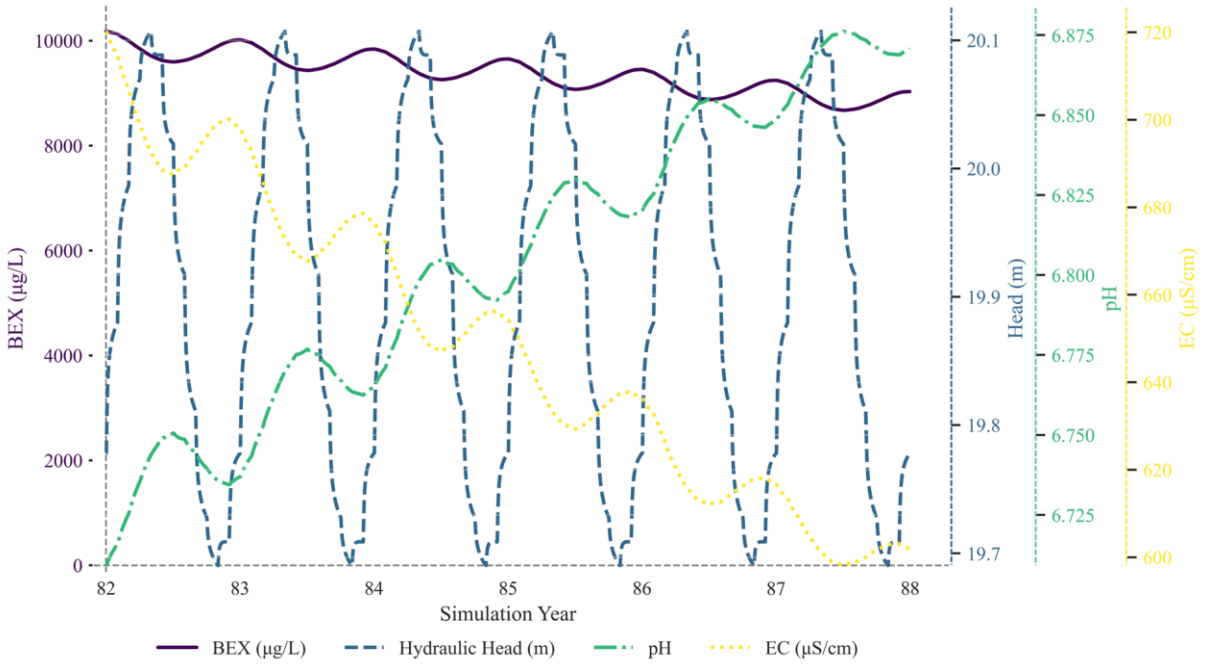

Figure S6. BEX concentrations with hydraulic head, pH, and EC at well X0Z0 during simulation years 82–88.

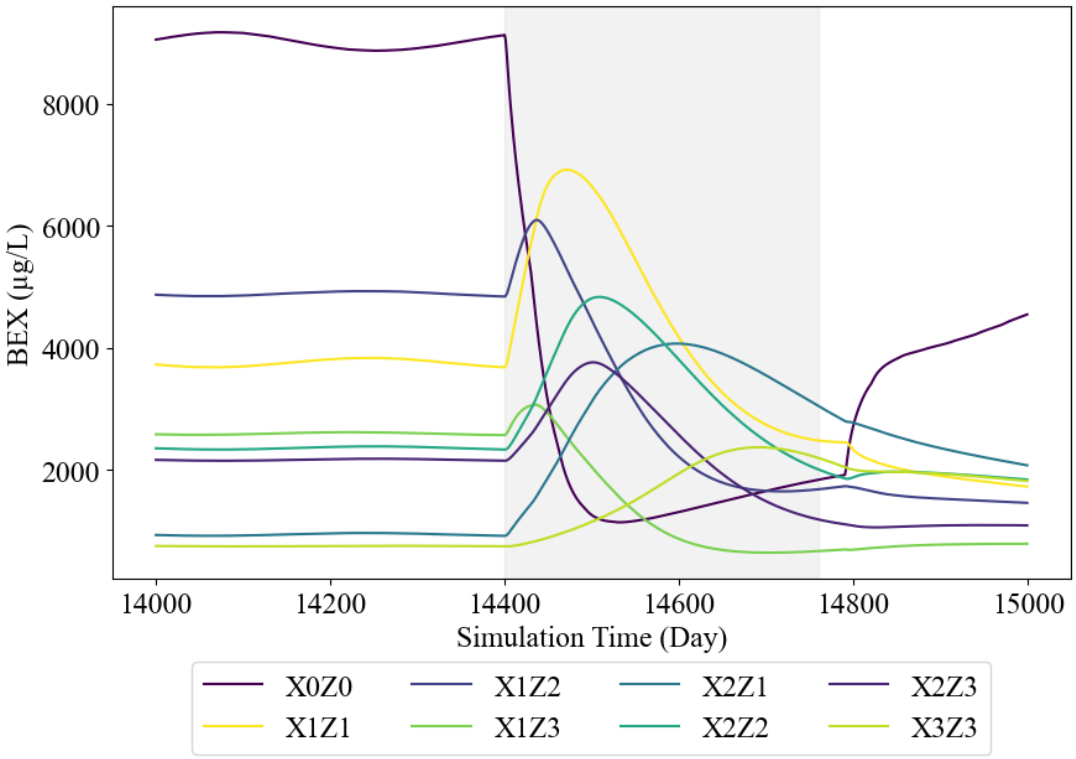

Figure S7. BEX concentration across eight observation wells with a year of increased hydraulic gradient starting from simulation year 40 (shaded region).

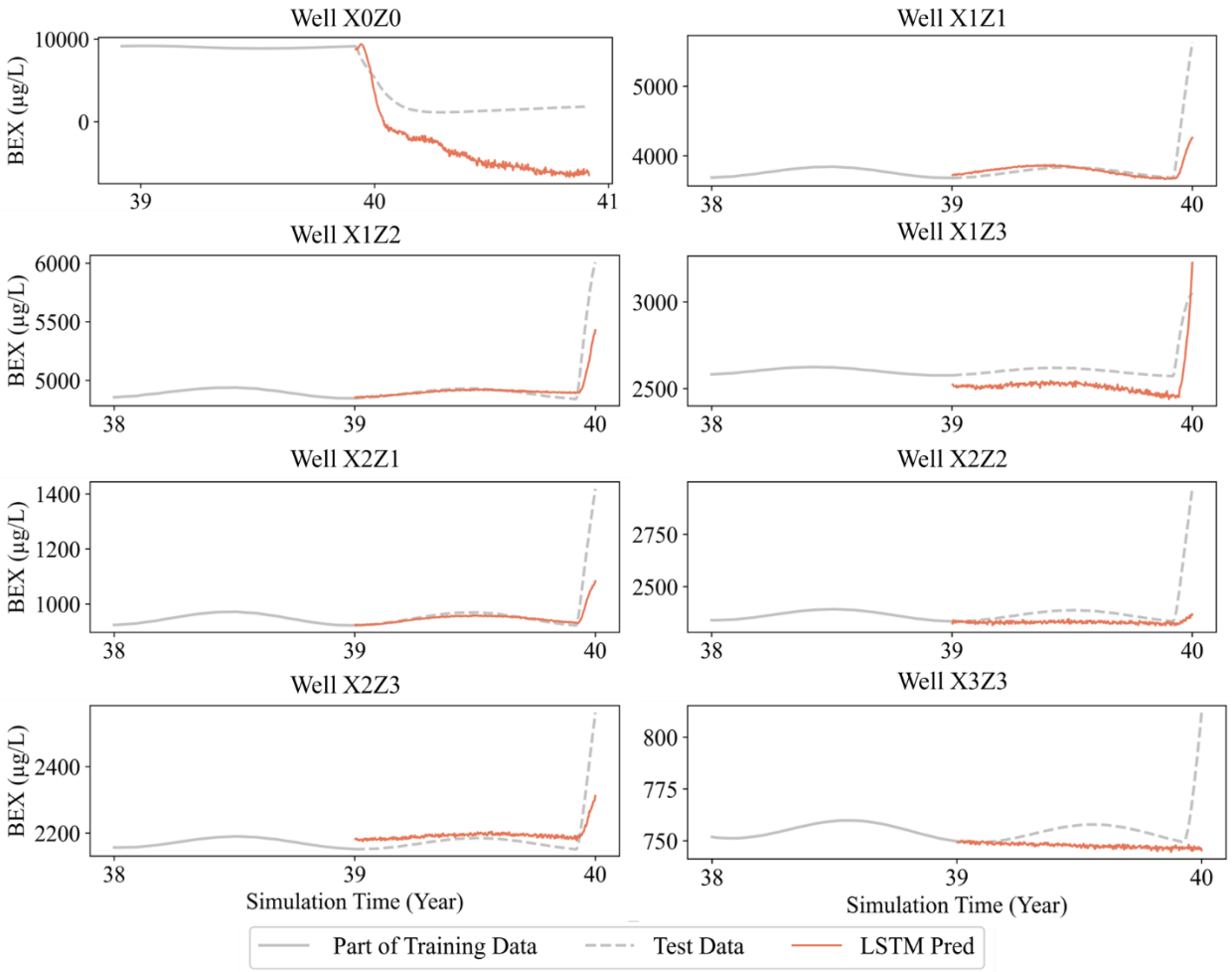

Figure S8. LSTM-predicted BEX concentrations under increased hydraulic gradient, showing one year of training data prior to the gradient increase and the subsequent one-year test period across eight observation wells.

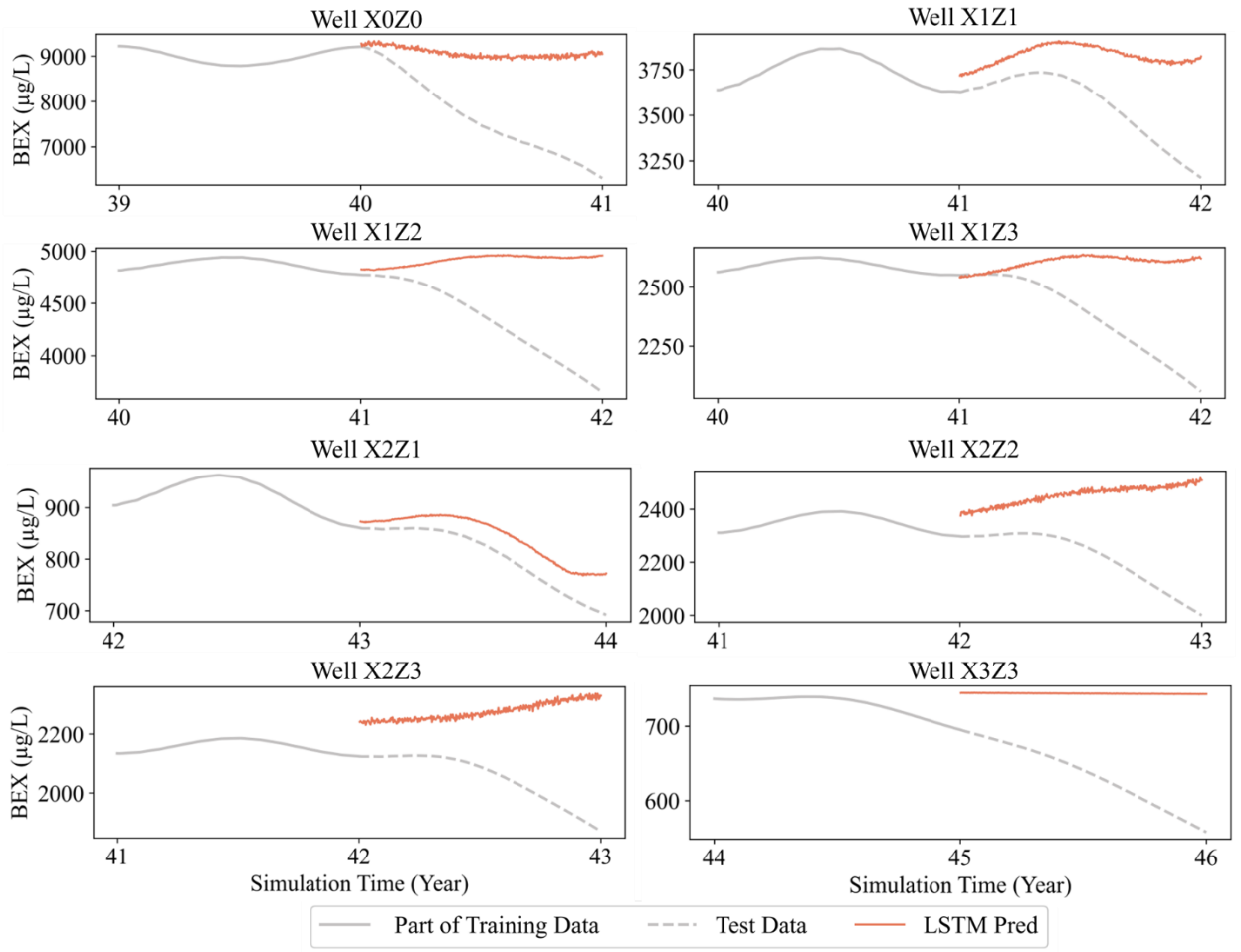

Figure S9. LSTM-predicted BEX concentrations after LNAPL source removal, showing one year of training data prior to the source removal and the subsequent one-year test period across eight observation wells.

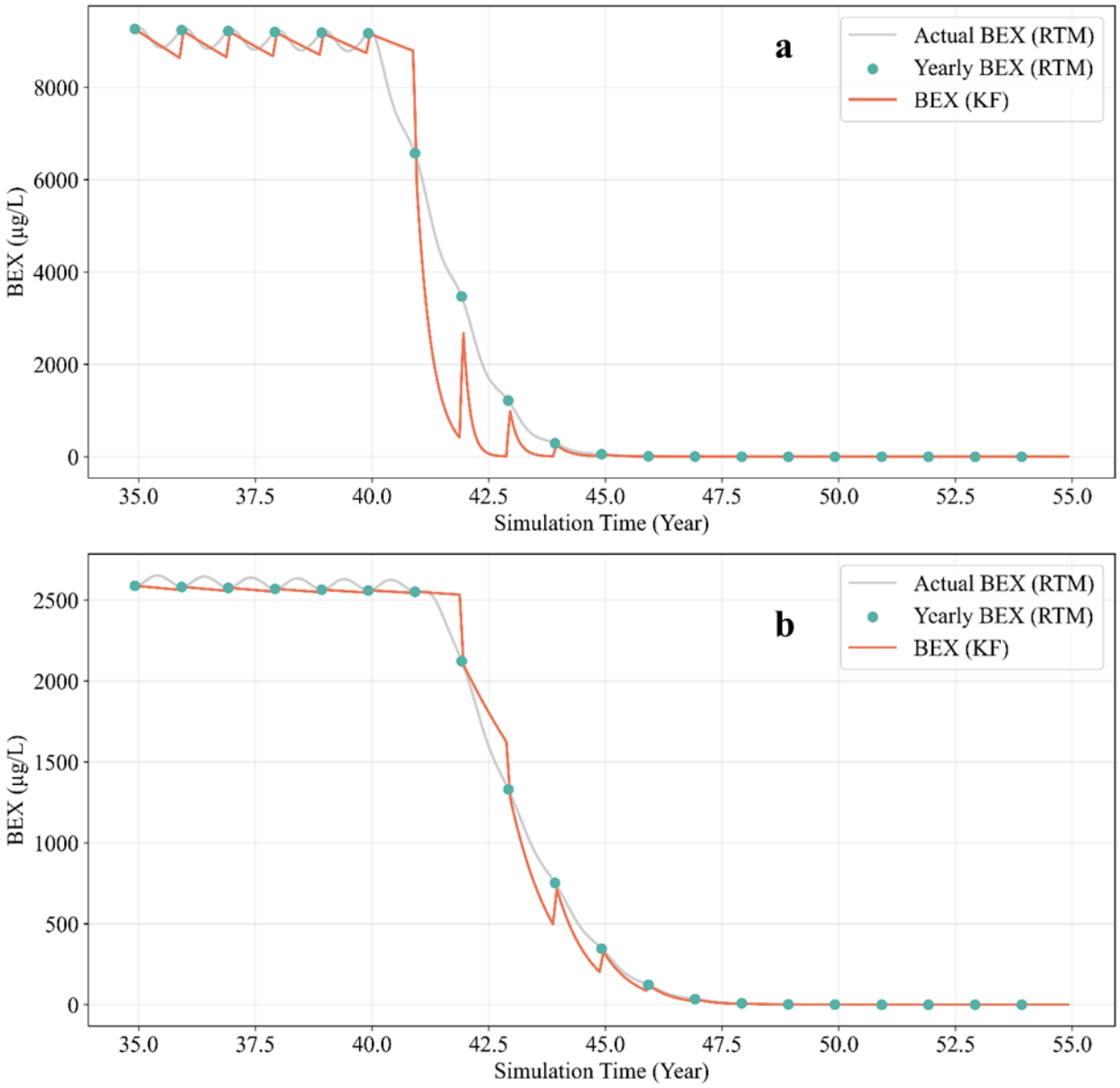

Figure S10. Sample Kalman filter results at observation wells X0Z0 (a) and X1Z3 (b) after LNAPL source removal.

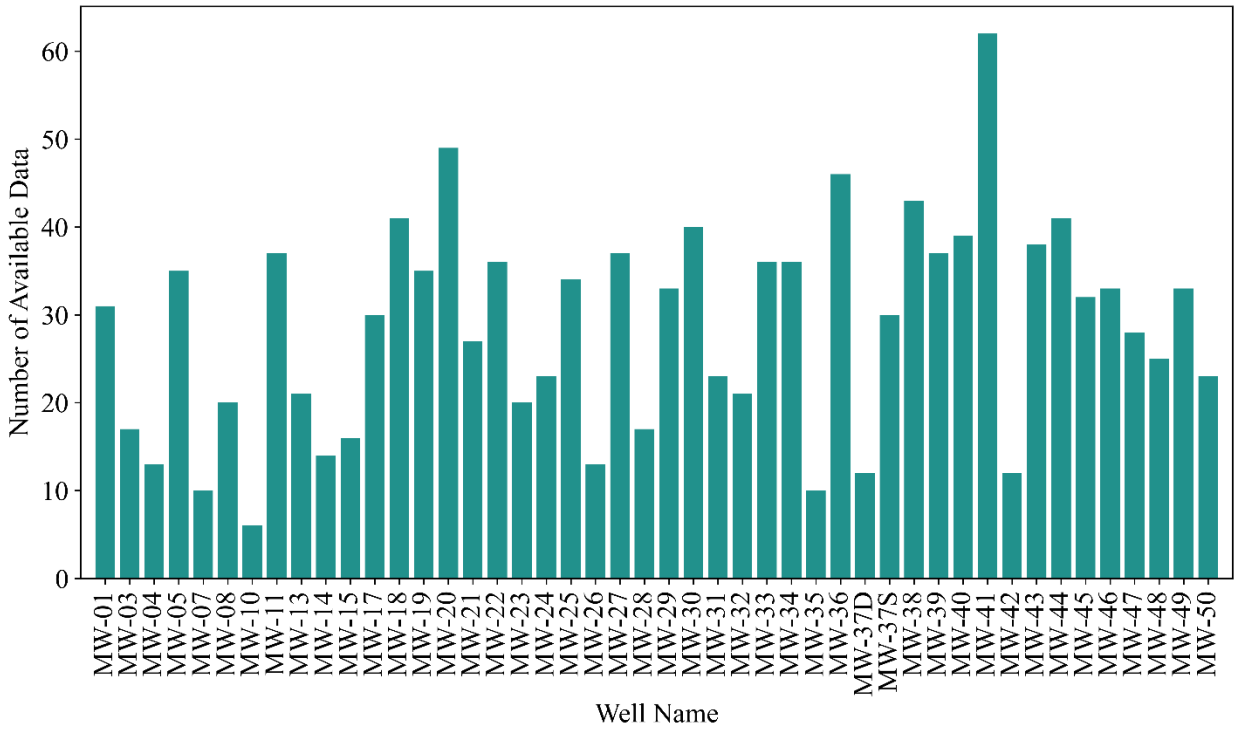

Figure S11. Number of available field data from an oil company containing both BTEX and in-situ water quality measurements at multiple monitoring wells from 2002 to 2020.

Table S1. Locations of eight observation wells in the reactive transport model, showing distance from the source zone center and depth below ground surface.

Well	Distance from Source Zone Center	Depth from Ground Surface
X0Z0	5 m	2 m – 3 m
X1Z1	55 m	0.6 m – 1.6 m
X1Z2	55 m	2.2 m – 3.2 m
X1Z3	55 m	3.8 m – 4.8
X2Z1	85 m	1 m – 2 m
X2Z2	85 m	2.6 m – 3.6 m
X2Z3	85 m	4.2 m – 5.2 m
X3Z3	145 m	4.6 m – 5.6 m

Table S2. Accuracy and price of commercial sensors based on manufacturer's specifications (Atlas Scientific LLC 2025a; 2025b; 2025c; 2025d; van Essen Instruments 2025a; 2025b).

Sensor	Accuracy	Estimated Price
pH	0.002	€72
Dissolved oxygen	0.2 mg/l	€170
Electrical conductivity	2% of reading	€94
Redox potential	1 mV	€72
TD-Diver (water level)	±0.5 cm H_2O	€558
Baro-Diver (atmospheric pressure)	±0.5 cm H_2O	€470

References

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

Chang, G. Robust Kalman filtering based on Mahalanobis distance as outlier judging criterion. J Geod 88, 391–401 (2014). https://doi.org/10.1007/s00190-013-0690-8

Van Essen Instruments. (2025a). TD-Diver. Accessed on 14 October 2025 from https://www.vanessen.com/products/data-loggers/td-diver/

Van Essen Instruments. (2025b). Baro-Diver. Accessed on 14 October 2025 from https://www.vanessen.com/products/data-loggers/baro-diver/

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