

Interpretable machine learning for predicting the fate and transport of pentachlorophenol in groundwater[☆]

Mehran Rad^{a,b,*}, Azra Abtahi^{c,1}, Ronny Berndtsson^{b,d}, Ursula S. McKnight^e, Amir Aminifar^c

^a Department of Agriculture and Food, Research Institutes of Sweden (RISE), Box 5401, SE-402 29, Göteborg, Sweden

^b Division of Water Resources Engineering, Department of Building and Environmental Technology, Lund University, Box 118, SE-221 00, Lund, Sweden

^c Department of Electrical and Information Technology, Lund University, Box 118, SE-221 00 Lund, Sweden

^d Centre for Advanced Middle Eastern Studies, Lund University, Box 201, SE-221 00, Lund, Sweden

^e Swedish Meteorological and Hydrological Institute, SE-601 76, Norrköping, Sweden

ARTICLE INFO

Keywords:

Explainable artificial intelligence
SHAP value
Tree-based regression
Contaminated sites
Sustainable remediation

ABSTRACT

Pentachlorophenol (PCP) is a commonly found recalcitrant and toxic groundwater contaminant that resists degradation, bioaccumulates, and has a potential for long-range environmental transport. Taking proper actions to deal with the pollutant accounting for the life cycle consequences requires a better understanding of its behavior in the subsurface. We recognize the huge potential for enhancing decision-making at contaminated groundwater sites with the arrival of machine learning (ML) techniques in environmental applications. We used ML to enhance the understanding of the dynamics of PCP transport properties in the subsurface, and to determine key hydrochemical and hydrogeological drivers affecting its transport and fate. We demonstrate how this complementary knowledge, provided by data-driven methods, may enable a more targeted planning of monitoring and remediation at two highly contaminated Swedish groundwater sites, where the method was validated. We evaluated 6 interpretable ML methods, 3 linear regressors and 3 non-linear (i.e., tree-based) regressors, to predict PCP concentration in the groundwater. The modeling results indicate that simple linear ML models were found to be useful in the prediction of observations for datasets without any missing values, while tree-based regressors were more suitable for datasets containing missing values. Considering that missing values are common in datasets collected during contaminated site investigations, this could be of significant importance for contaminated site planners and managers, ultimately reducing site investigation and monitoring costs. Furthermore, we interpreted the proposed models using the SHAP (SHapley Additive exPlanations) approach to decipher the importance of different drivers in the prediction and simulation of critical hydrogeochemical variables. Among these, sum of chlorophenols is of highest significance in the analyses. Setting that aside from the model, tetra chlorophenols, dissolved organic carbon, and conductivity found to be of highest importance. Accordingly, ML methods could potentially be used to improve the understanding of groundwater contamination transport dynamics, filling gaps in knowledge that remain when using more sophisticated deterministic modeling approaches.

1. Introduction

Groundwater is a vital resource, supporting a wide range of human and ecological functions. Despite its important role, groundwater quality is increasingly threatened by the presence of a wide range of contaminants (e.g., Lemaire et al., 2022; Sonne et al., 2017). Among these, long-lasting contaminants such as chlorinated solvents, petroleum

hydrocarbons, pesticides, and persistent organic pollutants (POPs) are the most serious threat to the environment (O'Connor et al., 2018). Also known as recalcitrant contaminants, these are persistent in the environment and can remain in soil and groundwater for decades, posing a significant risk to human health and the environment. Despite significant efforts to prevent and remediate groundwater contamination, recalcitrant contaminants continue to be a major challenge for

[☆] This paper has been recommended for acceptance by Hocheol Song.

* Corresponding author. Division of Water Resources Engineering, Faculty of Engineering, P.O. Box 118, SE-221 00 Lund, Sweden.

E-mail address: mehran.naseri.rad@ri.se (M. Rad).

¹ Main authors with equal contributions to the paper.

<https://doi.org/10.1016/j.envpol.2024.123449>

Received 24 November 2023; Received in revised form 23 January 2024; Accepted 24 January 2024

Available online 24 January 2024

0269-7491/© 2024 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

environmental scientists and managers. The challenge is twofold: first, these contaminants bioaccumulate and can be transported long distances. Secondly, they resist conventional remediation measures. Estimating their potential to spread from the source and their longevity in the subsurface after release, constitute complex research problems delaying adequate and/or sustainable remedial actions for these sites.

Dealing with contaminated groundwater requires understanding the hidden subsurface hydrogeochemical media and is associated with high variability and uncertainty (Naseri-Rad et al., 2020). Sophisticated process-based reactive contaminant transport models have been developed in the last decades to help in this regard. These models depict the interaction between physical and chemical processes, influenced by spatially and temporally varying initial and boundary conditions (e.g., Yu et al., 2020). HYDRUS 1D & 3D, and MODFLOW combined with MT3DMS, demonstrate significant efficacy in assessing the distribution of contaminants in the unsaturated and saturated zones, respectively (Zhang et al., 2019). But such models remain too complex, typically require lots of data, and are thus too resource intensive to be applied to the large number of existing contaminated sites (Naseri-Rad et al., 2021). Simplistic analytic and semi-analytic models, on the other hand, are not expensive to run and have been used to highlight the effectiveness of this method, particularly in unsteady-state conditions (Chowdhury and Rahnuma, 2023; Milašinović et al., 2019). However, these might not be as reliable for more complex geologic media and contaminant characteristics, especially when using datasets with missing values, which is common in field observations (Radelyuk et al., 2021). Here, recalcitrant contaminants are a greater challenge for site managers as their interaction with the hydrogeochemical media is largely unknown.

Traditional statistical models (e.g., linear regression) have been widely used to estimate solute transport in groundwater systems. However, they still have limitations in capturing the complex relationships between environmental variables and contaminant fate. Recently, non-linear machine learning (ML) approaches have emerged as a promising alternative to traditional statistical models for groundwater contaminant transport estimation (Motevalli et al., 2019) and vulnerability mapping (Elzain et al., 2023, 2022). Leveraging deep neural network (DNN) emulators for comprehensive groundwater flow and contamination transport models has demonstrated its efficacy as a robust basis for assessing the risk of groundwater contamination (Yu et al., 2020). These applications are currently recommended solely to alleviate the computational load of process-based transport models. This enhancement aims to increase their applicability, particularly when a substantial number of model runs are necessary for tasks such as data verification, sensitivity analysis, uncertainty analysis, and optimization. Still, this does not reduce the complexity of the entire modeling process. Non-linear ML algorithms (e.g., random forest and support vector machines) can be used to analyze large and complex datasets and identify patterns and relationships between variables, allowing for more accurate predictions of contaminant fate in the subsurface.

It is widely recognized that hydrogeochemical environmental conditions play a key role in the degradation and mobility processes of recalcitrant contaminants in the subsurface. For example, the prevailing hydrogeochemical conditions may alter these processes by many orders of magnitude over the life cycle of the contaminant. Furthermore, seasonal fluctuations may alter the existing geochemical properties of the subsurface (e.g., pH, redox), which can significantly affect contaminant biodegradation and mobility in both soil and groundwater. For this reason, long-term monitoring is becoming increasingly important at contaminated soil and groundwater sites (Libera et al., 2019; Meray et al., 2022). There is, however, a growing recognition that current remediation technologies have limited effectiveness and that recalcitrant contaminants are difficult to treat. Sustainable remediation paradigm (Hou, 2020) promotes the transition from intense soil removal and pump-and-treat alternatives to more passive remediation approaches (Naseri-Rad et al., 2022). Such techniques, especially monitored natural

attenuation (MNA), are often referred to as the most sustainable remediation option (Hou, 2020), but they normally require monitoring for long time spans, possibly decades (Meray et al., 2022).

Numerous ML methods have been developed to discover hidden patterns and key factors in subsurface contamination dynamics. Supervised learning methods to estimate the distributions of contaminant concentrations, groundwater levels, and geochemical parameters are more common (McLean et al., 2019; Motevalli et al., 2019). Unsupervised learning approaches have also been used to identify the correlation among contaminant concentrations and/or hydrogeochemical parameters (Meray et al., 2022; Schmidt et al., 2018). However, ML methods may be considered data intensive (Jordan and Mitchell, 2015) and not the best approach in dealing with the large number of contaminated sites.

In the current study, we aimed at developing a methodology to improve the understanding of the complexity of contaminated groundwater movement and unveil hidden interactions of hydrogeochemical variables that drive contaminant dynamics in the subsurface. We applied the methodology in two case studies where the groundwater is contaminated with pentachlorophenol (PCP). However, the application of the methods developed here is not limited to a specific type of contaminant. PCP sites were chosen due to its widespread presence in Swedish groundwater around old sawmills, across the country (Johansson, 2020). This is mainly because it was widely utilized for treating wood against microbial degradation, before it was banned in the 1970s, and for its physiochemical properties that makes it persistent in the subsurface, and very hard to remove through conventional remediation techniques. The latter means that PCP will remain in the subsurface for long periods, making its monitoring and prediction of great importance for the screening, i.e. investigation and remediation of contaminated sites. Moreover, a particular focus here was to extract critical information from existing site investigation datasets, which are not normally sampled at frequent intervals, leading to datasets with missing values, and so are not typically considered useful for extracting trends and drivers of governing dynamics. These datasets exist, however, because MNA strategies often build upon a large quantity of historical monitoring and characterization data, deal with unknown values, and atypical time spans of sampling are common in field projects at these sites which endeavor to provide additional information for decision-makers.

The novel aspects of our approach include (a) improved visualization of critical hydrogeochemical parameters and their impacts on contaminant transport in the subsurface based on groundwater data; (b) integration of publicly available data, such as surface elevation and groundwater table depths, with hydrogeochemical and contaminant concentration data for creating predictors in ML; (c) interpolation to integrate data-driven relationships for estimating contaminant concentrations using six different ML algorithms; (d) comparison/selection of multiple ML algorithms for this interpolation; and (e) predicting contaminant concentrations based on identified interrelationships among involved components and most influential drivers using only a limited dataset, as is common in field studies.

Accordingly, the objective of this paper was to apply ML to improve the understanding of drivers in the subsurface at sites contaminated by recalcitrant contaminants such as PCP. A variety of interpretable ML methods, specifically 6 interpretable regressor models, were used for this purpose to not only determine critical hydrogeochemical factors and optimal environmental conditions controlling the plume, but also make it possible to predict the plume's spread in the subsurface. In other words, we used 6 different regressor models for the prediction of PCP concentrations, and then interpreted the proposed models using the SHAP (SHapley Additive exPlanations) approach (Lundberg and Lee, 2017; Sundararajan and Najmi, 2020) to determine the importance of different hydrochemicals for PCP prediction. To ensure practical applicability, we applied the developed methodology at two sites in Sweden highly contaminated by PCP, which is widespread because of wood

treatment, with vast investigation data collected by the Swedish Geological Survey (SGU). All codes are supplied as open access material.

2. Methods and materials

2.1. PCP prediction using machine learning

For PCP prediction, different regressors can be used. A typical regressor predicts a dependent variable based on one or several independent variables. The independent variables are also known as the features of the regressor model. Here the dependent variable, or target variable to be predicted, is the PCP concentration and the independent variables are other constituents with the potential to influence PCP concentration in groundwater and which may include any daughter products of PCP degradation. Table 1 presents the list of independent variables,

Table 1
Independent variables used to predict PCP concentration.

Dataset 1	Dataset 2	Dataset 3
2-Monochlorophenol	2-Monochlorophenol	2,3,4,5-Tetrachlorophenol
3-Monochlorophenol	3-Monochlorophenol	2,3,4,6-Tetrachlorophenol
4-Monochlorophenol	4-Monochlorophenol	2,3,4-Trichlorophenol
2,6-Dichlorophenol	2,6-Dichlorophenol	2,3,5,6-Tetrachlorophenol
2,4- and 2,5 - Dichlorophenol	2,4- and 2,5- Dichlorophenol	2,3,5-Trichlorophenol
2,3-Dichlorophenol	2,3-Dichlorophenol	2,3,6-Trichlorophenol
3,4-Dichlorophenol	3,4-Dichlorophenol	2,3-Dichlorophenol
3,5-Dichlorophenol	3,5-Dichlorophenol	2,4- and 2,5- Dichlorophenol
2,4,6-Trichlorophenol	2,4,6-Trichlorophenol	2,4,5-Trichlorophenol
2,3,6-Trichlorophenol	2,3,6-Trichlorophenol	2,4,6-Trichlorophenol
2,3,5-Trichlorophenol	2,3,5-Trichlorophenol	2,4-Dimethyl Phenol
2,4,5-Trichlorophenol	2,4,5-Trichlorophenol	2,5-Dimethyl Phenol
2,3,4-Trichlorophenol	2,3,4-Trichlorophenol	2,6-Dichlorophenol
3,4,5-Trichlorophenol	3,4,5-Trichlorophenol	2,6-Dimethylphenol
2,3,4,6- Tetrachlorophenol	2,3,4,6- Tetrachlorophenol	2-Monochlorophenol
2,3,4,5- Tetrachlorophenol	2,3,4,5- Tetrachlorophenol	3,4,5-Trichlorophenol
2,3,5,6- Tetrachlorophenol	2,3,5,6- Tetrachlorophenol	3,4-Dichlorophenol
Chlorophenols. Sum	Chlorophenols. Sum	3,4-Dimethylphenol
PCP	PCP	3,5-Dichlorophenol
		3-Monochlorophenol
		4-Ethyl/2,3-/3,5- Dimethylphenol
		4-Monochlorophenol
		Alkalinity
		COD Cr
		Dichlorophenols. Sum
		DOC
		Fe2+
		Phenol
		Chlorophenols. Sum
		Chloride
		Conductivity
		Cresols. Sum
		m-Ethylphenol
		m-Cresol
		Mn2+
		Monochlorophenols. Sum
		Nitrate nitrogen (NO3-N)
		O-Ethylphenol
		o- Cresol
		pH
		p- Cresol
		Redox
		Sulphate
		Acid content
		Temperature
		Tetrachlorophenols. Sum
		Trichlorophenoler. Sum
		Turbidity
		Thymol

determined from the field campaigns conducted at the field sites contaminated with PCP. Notably, Dataset 3, belonging to one site, contained missing values for different constituents, while Dataset 1 and Dataset 2, which belong to another site, did not include missing values, i. e., complete datasets are defined as having had all groundwater monitoring wells sampled for the same constituents at the same times over the years.

These constituents may influence PCP concentrations in different ways (L.M. Schmidt et al., 1999). As discussed in the Introduction, linear ML models have limitations in capturing the complex relationships between environmental variables and contaminant fate. However, non-linear ML approaches, such as tree-based regressors, can identify patterns and relationships between variables more precisely. These are more advanced and might be capable of predicting contaminant fate in the subsurface more accurately. In this paper, we aimed to test the applicability of linear vs. non-linear ML approaches, on both datasets with and without missing values, in order to ascertain whether these novel approaches can be used to support and streamline future contaminated sites investigations. Hence, in this paper, we applied three widely-used linear regressors (standard linear regressor, ridge regressor, and LASSO) and three less-investigated tree-based regressors (random forest, extra trees, and XGB regressors) for PCP concentration prediction to compare the models and investigate if simpler models can identify these complex relationships or if more complex, e.g., tree-based models are needed to determine these interrelationships. We investigated if the models can be used for PCP concentration prediction. Below, we briefly explain the exploited regressors.

2.1.1. Standard linear regressor

In linear regression, the relationship between the independent and dependent variables is assumed to be linear. In other words, the dependent variable $Y = [y_1, y_2, \dots, y_n]^T$ (where T denotes the transpose of the matrix) is considered as the weighted combination of the independent variables $x_1, x_2, \dots, x_p$ plus a random error, i.e., $y_i = \epsilon_i + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}$ for $i = 1, \dots, n$, where $\beta_1, \beta_2, \dots, \beta_p$ are regression coefficients and $\epsilon = [\epsilon_1, \epsilon_2, \dots, \epsilon_n]^T$ is the random error or disturbance term.

If Gauss-Markov conditions are met (i.e., ϵ_i is assumed to be an independent and identically distributed random variable with zero mean and finite variance), then the Ordinary Least Squares (OLS) estimation is thought to provide the best linear unbiased estimator (Jobson, 2012), which is used here. In standard linear regressors, the regression coefficients are determined by using OLS, i.e., minimizing the sum of squared differences between the dependent variable and the weighted combination of the independent variables as according to Freedman (2009):

$$\hat{\beta} = \arg \min_{\beta} \|Y - \beta X\|^2 \quad (1)$$

where $\beta = [\beta_1, \beta_2, \dots, \beta_p]$ and $X = [x_1^T, x_2^T, \dots, x_p^T]^T$.

2.1.2. Ridge regressor

Ridge regressor, also known as Tikhonov regularization, is a linear regression model in which the coefficients are estimated by the ridge estimation instead of OLS. Ridge estimation adds a penalty to the size of the coefficients to the minimization problem (Gruber, 1998):

$$\hat{\beta} = \arg \min_{\beta} \|Y - \beta X\|^2 + \lambda \|\beta\|^2 \quad (2)$$

where λ is a hyperparameter that controls the amount of shrinkage of $\|\beta\|^2$. As in this problem, parameters are shrinking, the ridge regressor can be used to handle multicollinearity, where several independent variables in a model are highly correlated. Furthermore, the model complexity is reduced by the coefficient shrinkage.

2.1.3. Least absolute shrinkage and selection operator (LASSO)

LASSO is a linear regressor that estimates sparse coefficients. The feature used in LASSO is similar to the ridge regression; LASSO adds a penalty term to the OLS problem equal to the summation of the absolute values of the coefficients:

$$\hat{\beta} = \arg \min_{\beta} \|Y - \beta X\|^2 + \lambda \|\beta\|_1 \quad (3)$$

where $\|\beta\|_1 = \sum_{p=1}^p |\beta_p|$ (the norm-1 of β) (Tibshirani, 1996). Hence, LASSO can result in sparse models with few coefficients, i.e., many coefficients become zero and are eliminated from the model. LASSO is generally used when there is a greater number of independent variables as it automatically does feature selection and results in a low complexity model.

2.1.4. Random forest

Random forest is a tree-based regressor. Different types of tree-based regressors are proposed which are all based on decision trees. A decision tree is a flowchart-like or tree-like structure in which each internal node checks a constraint on the independent variables/features (this constraint can be a yes/no question or checks being smaller or larger than a specific cut-point) and the branches represent the outcomes of fulfilling or not fulfilling the constraints (Quinlan, 1986).

In random forest regression, an ensemble learning method is used which combines predictions from several machine learning algorithms to achieve a more accurate prediction. Through random forest, multiple decision trees are constructed during the training time exploiting randomly subsampled input data with replacement (bootstrapping). Then, the average prediction of the individual trees is returned as the final output (Breiman, 2001).

2.1.5. Extra trees

Extra trees or extremely randomized trees is also a tree-based regressor and an ensemble learning algorithm like random forest with two main differences: first, in extra trees, each tree is trained using the entire training data in contrast to random forest, which uses bootstrapping. Second, Extra Trees uses random cut-points for the constraints of the decision trees, which is not the case in random forest (random forest uses measures such as information gain and gini impurity for finding optimal cut-points (Breiman, 2001)). The trees in extra trees are more variable and less correlated than the trees in the random forest regressor. Hence, extra trees is faster than random forests and can sometimes outperform it in gaining better results (Geurts et al., 2006).

2.1.6. Extreme Gradient Boosting (XGB)

XGB regressors are tree-based regressors which use boosting ensemble learning based on the decision trees. In boosting ensemble learning, instead of simultaneously training independent trees all at once, trees are added one at a time to the ensemble and are trained to correct the prediction errors of the previous model. As a gradient descent optimization algorithm is used for training the new trees, the algorithm refers to Extreme Gradient Boosting (Chen and Guestrin, 2016). XGB is faster than random forest, and for giving an insight of the results, it does not need to wait until all the trees are trained, in contrast to random forest and extra trees. However, it has more hyperparameters to be set.

2.2. Interpretability for predominant contaminants

For the interpretation of PCP prediction, the main questions encompass determining the order of importance for each constituent included in the model and which constituents are most important.

2.2.1. Linear regression

The coefficients of the linear regressor give us a good insight into how the output changes due to a feature change. However, the

magnitude of these coefficients are not necessarily good measures for finding the important features, as the value of each coefficient depends on the scale of the input features. To resolve this problem, we can normalize the features to have a mean of 0 and a standard deviation of 1.

For the evaluation of the feature importance, we can combine “Game Theory” and “Machine Learning” to achieve a perfect measure called the SHAP value (Lundberg and Lee, 2017; Sundararajan and Najmi, 2020). This measure initiated from the game theory question: *How should we divide up the payoff among the players with different skills in a coalition?* For answering this question, the marginal contribution of each player can be calculated by adding them in a sequence to the coalition. Then, the marginal contributions should be averaged over every possible sequence in which the players could have been joined to the coalition. Now we can use the same solution for determining the importance (payoff) of different features (players) in a prediction model (coalition). The SHAP value is the average contribution of a feature value to a prediction model over every possible sequence. Hence, it is an indication of the importance of that feature. For calculating the contribution of a feature, we can say a feature has joined a model when we know its value and has not joined the model when we do not know its value.

It can be shown that, if we assume that features are independent, the SHAP value of a linear prediction model for a feature value, (x_i), can be written as:

$$\beta_i(x_i - E\{x_i\}) \quad (4)$$

where β_i is the coefficient of the linear model corresponding to the i th feature and $E\{x_i\}$ is the expected value of the i th feature [8]. As it can be seen in calculating the SHAP value, the scale of the feature is not important anymore as we have the multiplication of β_i by $x_i - E\{x_i\}$ instead of β_i .

2.2.2. Tree-based regression

Although small tree-based regression models can be interpretable (Rudin, 2019); it is hard to interpret large tree-based models. Exact calculation of the SHAP values for large tree-based models is complicated (NP-hard problem) (Matsui and Matsui, 2001). Hence, recently, some algorithms are proposed for tree-based SHAP value calculation that run in linear time (Lundberg et al., 2020, 2019).

2.3. Evaluation measure and model setup

In this paper, for the evaluation of the predictions, we use the mean R^2 -score, i.e., the coefficient of determination, over 1000 runs, to represent the estimated effect. The R^2 -score for the prediction of $Y = [y_1, y_2, \dots, y_n]^T$, when the predicted vector is $\tilde{Y} = [\tilde{y}_1, \tilde{y}_2, \dots, \tilde{y}_n]^T$, is defined as (Gujarati and Porter, 2009):

$$R^2\text{-score} = 1 - \frac{\sum_{i=1}^n (y_i - \tilde{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (5)$$

Where $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$. The train and test sets are randomly changed in 1000 runs, while the test size is 30 % of the size of the dataset. The coefficient of determination, R^2 -score, is then calculated for all the methods on the dataset. The R^2 -score value is less than or equal to 1, where higher values for the R^2 -score suggests higher accuracy of the model. Therefore, the best model results have an R^2 -score of 1.

For generating SHAP plots (See Section 4), as with the mean R^2 -score, we considered 1000 runs with random training and test sets. The SHAP values from each run were then collected and averaged, and the plots related to different runs were combined into a final SHAP value plot to see how each feature effects the prediction of the PCP concentration in a model.

3. Case Studies

The two experimental case studies included in this paper are former sawmill industries located in southern Sweden; one in Hjortsberga and

the other in Lidhult. They both have left a legacy of high concentrations of PCP and dioxins (Johansson, 2006) in the subsurface. The Swedish Geological Survey (SGU) is responsible for orphan contaminated sites in Sweden and is currently investigating and running test remediation trials at both these sites. PCP concentrations at both sites are up to 5 orders of magnitude higher than permissible according to regulations for drinking water use (of 1 µg per liter) (SGU, 2017). Currently, there is no drinking water use anticipated for these sites and site managers have set a remediation target of 100 µg per liter for the sites. Continuous monitoring is crucial for both sites as these and many more contaminated sites in Sweden are close to lakes. Such lakes are important not only for recreational purposes, but also for biodiversity and conservation in the nearby areas. This makes observing and understanding their dynamics more crucial. The long-lasting effects of recalcitrant contaminants like PCP makes long term monitoring practices even more important.

In addition to the noted similarities, these two sites have been selected for their different site investigation histories. The site investigation campaigns have been implemented by different consultant and contractor companies, in different time intervals, and some different sets of chemicals have been chosen to sample and measure. Missing values, which are common in site investigation datasets, are also more commonly found at one site than in the other. These differences might be due to many unknown site-specific considerations, e.g. financial or technical, that cannot be known in advance to plan modeling practices accordingly. Since such differences are common in site management practices, we took advantage of having two sites, one with missing values and one with complete datasets in order to evaluate the impact of such real-world situations with the ML methods implemented here. A model that works for only one site, for only one norm of sampling intervals, and one set of investigated chemicals may be insufficient to prove a model is successful in identifying the governing interrelationships and predicting contaminant concentrations. This is of more importance for ML based methods as these are largely based on data and their reliability might be questioned if they are too site-specific.

Both sites are in areas typically consisting of bedrock beneath a layer of till, which is characterized as “complex and heterogeneous” (Naseri-Rad et al., 2021), primarily composed of coarse-grained glaciofluvial sediments, varying tills, clay, and peat. The crystalline bedrock at the site contains water-bearing fractures. This geological formation is widespread in the southern and central regions of Sweden.

As shown in Fig. 1, PCP contamination in the Hjortsberga groundwater is threatening the nearby Lake Sjöatorpasjön.

Although, chlorophenol usage has been banned in Sweden since 1978 (Swedish EPA, 2009), these contaminants are persistent in the environment and still widely detected (Liu et al., 2019; Naseri-Rad et al.,

2022). According to the Stockholm convention, PCPs are recognized as persistent organic pollutants (POPs) in the environment (UNTC, 2001). Fortuitously, indigenous microbes can sometimes remove low levels of PCP (<1000 µg/L) to approach the regulatory standard of 1 µg/L with the addition of oxygen, with or without nutrient amendment (L. M. Schmidt et al., 1999), with a typical PCP half-life ranging between 0.1 and 3 years (USEPA, 1991). It should be noted that PCP is a fungicide that inhibits the activity of microorganisms above certain concentrations. For example, PCP concentrations as low as 20,000 µg/L in water can inhibit microbial activity (Davis et al., 1994).

Many studies have been conducted to investigate both sites with more than 20 different sampling campaigns at each site over the last two decades. Notably, not all wells were sampled in every campaign. The maximum number of samples taken from one well over the entire period was 10, while most wells were sampled only once or twice. This makes it challenging to discern trends in contamination concentration changes for individual wells. It is important to note that, while PCP in groundwater can also be in nonaqueous phase liquid form, the contamination assessed here is in the dissolved form, consistent with field observations.

The inherently persistent characteristics of PCP, together with the complex geology of the sites, i.e., heterogeneity and complexity of the till layer and fractured bedrock, the latter being filled with denser PCP solutions, which might be releasing PCP to the groundwater flowing above it, makes long-term monitoring and prediction of PCP distributions at such sites of great importance. Persistent contaminants in cases like the ones we consider here may need decades of continuous monitoring and plume controlling, which is expensive and requires much effort. Especially surrounding areas of such sites are affected by the contamination in the adjacent land, resulting in reduced land prices and social anxiety for neighboring residents. To improve the understanding of the contaminant plume dynamics in such cases, it is necessary to combine modeling with continuous observations in an iterative manner.

3. Results and discussion

3.1. PCP prediction

Table 2 shows the results for the three linear regressors (standard linear regressor, Ridge regressor, and LASSO) and the three tree-based regressors (random forest, extra trees, and XGB regressors) applied for predicting PCP concentrations using the three different datasets described in section 2.1.

As seen from Table 2, linear regressors for Dataset 1 and Dataset 2 predict the PCP value with the R^2 -score of greater than 0.977. For these datasets, which were without missing values, the LASSO regressor

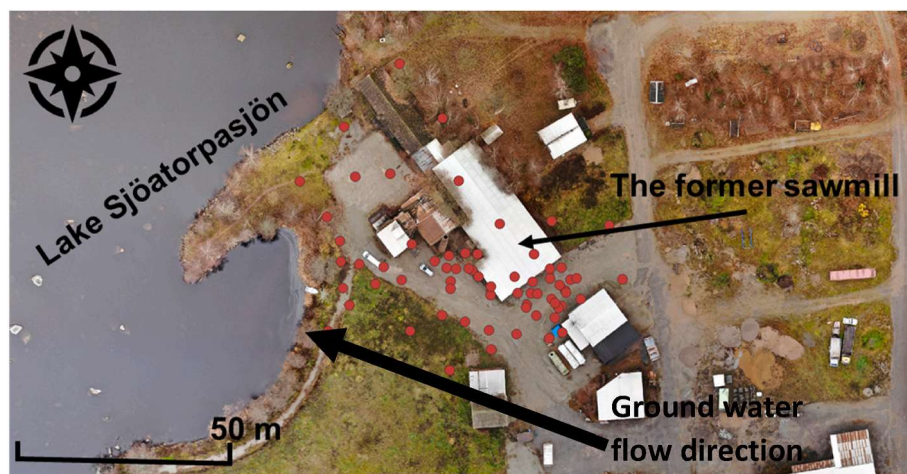


Fig. 1. Location of the former sawmill plant in Hjortsberga, and sampling wells (red dots) used as a case study. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 2
Mean R^2 -score for the PCP prediction.

Regressor methods	Dataset 1	Dataset 2	Dataset 3
Standard Linear Regressor	0.9774	0.9776	−14,232
Ridge Regressor	0.9873	0.9884	−2.1935
LASSO Regressor	0.9981	0.9996	−1.5079
Random Forest Regressor	0.9054	0.8255	0.6463
XGB Regressor	0.8887	0.7077	0.476
Extra Trees Regressor	0.9552	0.9295	0.7579

achieved the best results (with an R^2 -score of more than 0.998) while having the lowest complexity due to its automatic feature selection. However, linear regressors, including LASSO, were not useful for Dataset 3, which contained missing data. All tree-based regressors could predict PCP concentrations for all observations to some degree (R^2 -score value ranging between 0.48 and 0.96). In particular, extra trees consistently obtains the highest R^2 -scores among the tree-based models, including for Dataset 3 (R^2 -score = 0.76). Overall, the tree-based regressors significantly outperform the linear regressors, and the extra trees regressor was the best model for predicting PCP concentrations for all 3 datasets.

3.2. Interpretability

In Fig. 2a and b, we present the coefficients for the standard linear regressor averaged over 1000 runs for Dataset 1 and Dataset 2, respectively. In these figures, the features were scaled and centered to have a mean of 0 and a standard deviation of 1. As depicted, “Chlorophenols. Sum” had the highest coefficients for both datasets. This is not unexpected as chlorophenols are daughter products of the PCP transformation in the subsurface media and contribute the most in PCP concentration prediction.

Fig. 3a and b shows the SHAP value bar plots for the standard linear regressor when Dataset 1 and Dataset 2 were used. In this plot, the importance of the features decreases from top to bottom. This importance was determined by calculating the mean absolute value of the SHAP values over all feature values for all runs. According to Fig. 3a and b “Chlorophenols. Sum” is the most important feature in the standard linear model, as it is the first feature in the list of the SHAP value bar plots.

The SHAP value bar plots for the LASSO regressor, which had the best R^2 -score among the exploited linear regressors, are illustrated in Fig. 3c and d for Dataset 1 and Dataset 2, respectively. Once again, as for the standard linear regressor, “Chlorophenols.Sum” is the most

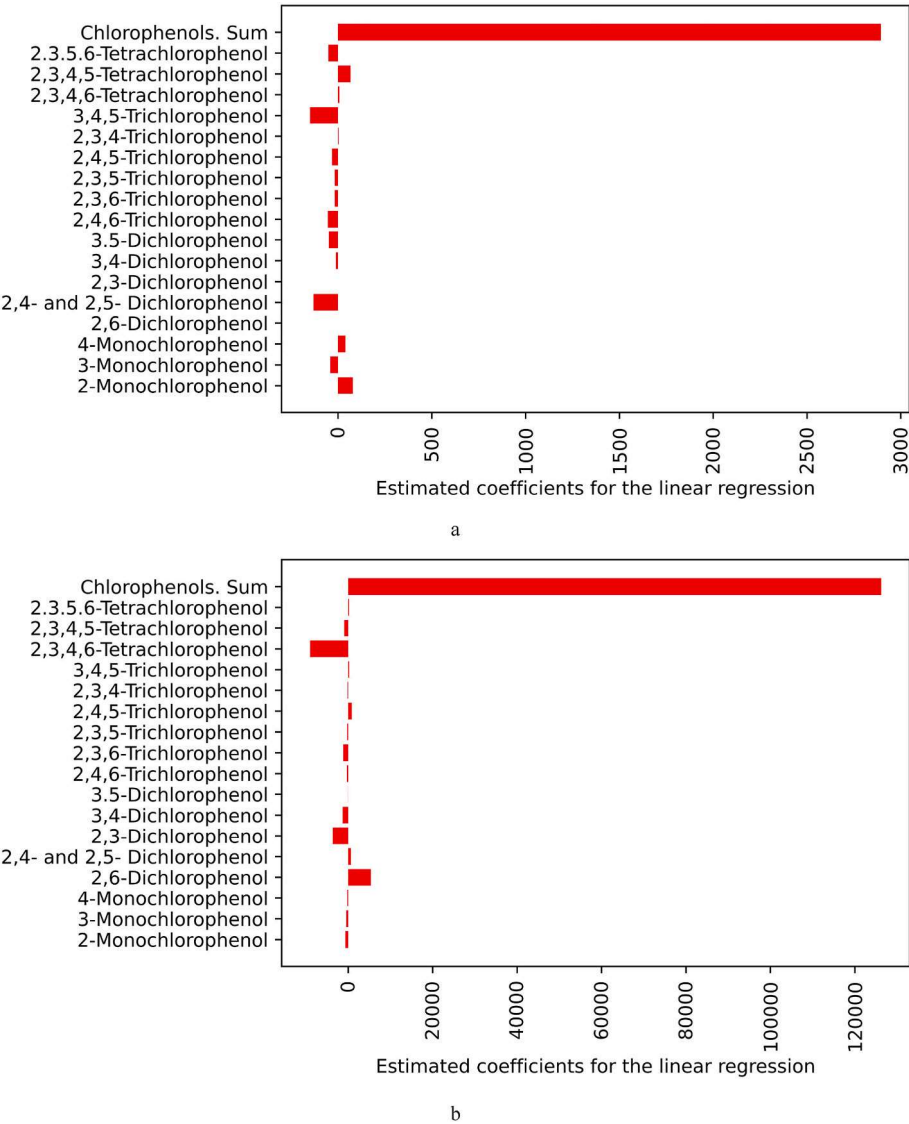


Fig. 2. Coefficients for standard linear regressor based on a) Dataset 1 and b) Dataset 2.

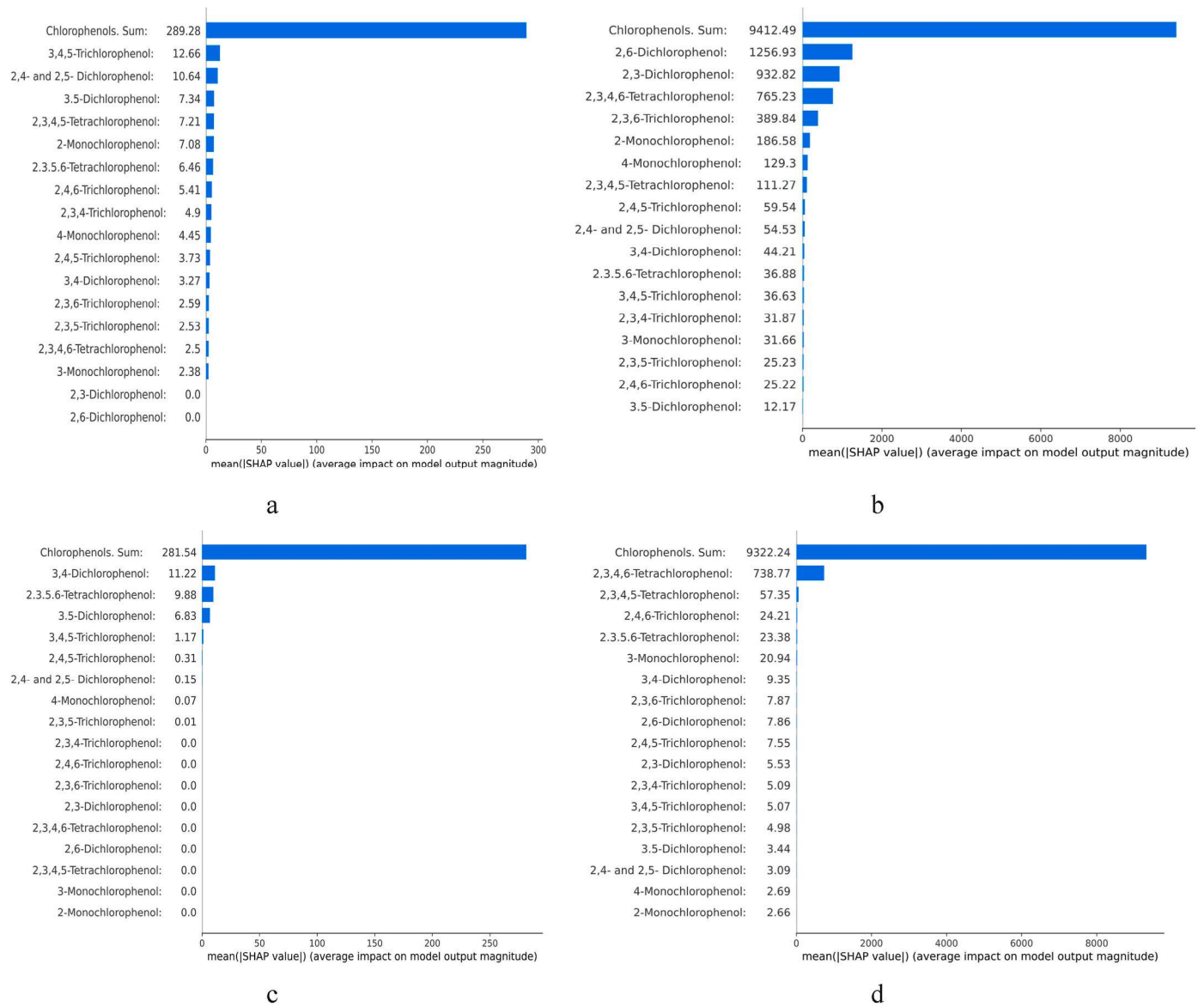


Fig. 3. SHAP value bar plot for a) the standard linear regressor and Dataset 1, b) the standard linear regressor and Dataset 2, c) LASSO regressor and Dataset 1, and d) LASSO regressor and Dataset 2.

important feature in the LASSO model. Notably, in Fig. 3c using Dataset 1, 9 features are classified as not important in the LASSO regressor and eliminated from the model. This is understandable as the LASSO regressor automatically does feature selection, which lowers model complexity.

SHAP value beeswarm summary plots for SHAP and the extra trees regressor are shown in Fig. 4a, b, and c for Datasets 1, 2, and 3, respectively. In a SHAP value beeswarm summary plot, the interpretation for each feature is represented by a single dot on each feature row. In the plot, the importance of the features decreases from top to bottom and is determined by calculating the mean absolute value of the SHAP values over all feature values for all runs. In this plot, the positive SHAP values are displayed on the left side of the horizontal axis and negative values are displayed on the right side. As the SHAP value is associated with impact of a feature on the output of the model, the horizontal axis shows whether the effect of a feature is associated with higher or lower prediction value (PCP concentration) and the color shows the observation value for the feature; for high values it is red and for low values it is blue. As seen, “Chlorophenols.Sum” is also the most important feature for the extra trees regressor. Moreover, a high “Chlorophenols. Sum”

value has a high (shown by the horizontal location of the red color points) and positive (look at the sign of the horizontal location) impact on the predicted PCP concentration by the standard linear regressor. In other words, “Chlorophenols. Sum” greatly affected the predicted PCP concentration value, where higher values of “Chlorophenols. Sum” resulted in higher predicted PCP concentrations, and its lower value resulted in lower predicted PCP concentrations.

3.3. Ablation study

Next, we removed “Chlorophenols. Sum” from the datasets to better understand the contribution of this component to the prediction models. This is because “Chlorophenols. Sum” was the most important feature based on the SHAP value plots. Furthermore, we evaluated the model performance by considering only this feature as input. Here we relied on extra trees regressors as it had the overall best R^2 -score (performance). The results are shown in Table 3. According to the results, when using only the “Chlorophenols. Sum”, we can predict the PCP value accurately.

Fig. 5a, b, and c illustrate summary plots of SHAP when we remove “Chlorophenols. Sum” from Datasets 1, 2, and 3, respectively. These

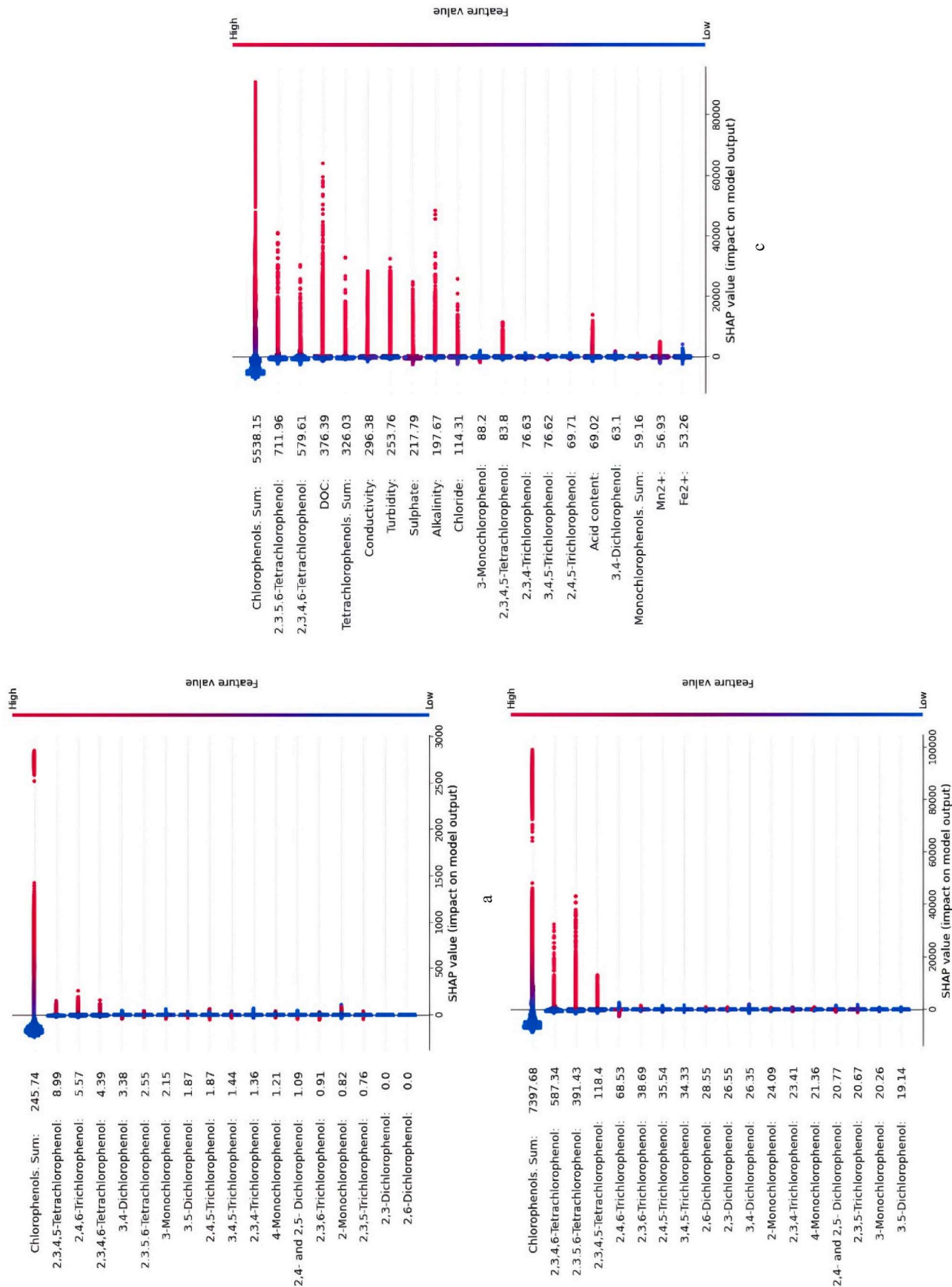


Fig. 4. SHAP value beeswarm summary plot for extra trees regressor and a) Dataset 1, b) Dataset 2, and c) Dataset 3.

Table 3
Mean R^2 –score for the PCP prediction by exploiting the Extra Trees Regressor.

	Dataset 1	Dataset 2	Dataset 3
Without <i>Chlorophenols.Sum</i>	0.6160	0.4504	0.4967
Only <i>Chlorophenols.Sum</i> as input	0.9586	0.9500	0.7525

figures illustrate that, for Datasets 1 and 2 and for without “Chlorophenols. Sum”, two most important chemicals are from tetra-chlorophenol family. However, for Dataset 3, which again contains missing values of features, DOC and conductivity appear among the top-5 most important constituents when “Chlorophenols. Sum” is removed from the dataset.

3.4. Interpretability per sample

We show observations versus predictions for the extra trees regressor and the three datasets in Fig. 6a, b, and c. The figure was generated considering 10 runs with random training and test sets. In general, predictions are close to observations for these datasets when the extra trees regressor was exploited. For Dataset 3, the existence of missing values resulted in decreasing accuracy compared to the other datasets. However, the accuracy of the prediction was still high. This is supported by heat maps, shown in Appendix A, which further illustrate the

correlation between the features for both Dataset 1 and Dataset 2.

3.5. Cross dataset evaluation

Finally, we trained the regressor based on one dataset and tested the trained model exploiting the other dataset in a cross-dataset evaluation. This evaluation lets us know how well our model can adapt to the changes in the datasets. When we trained the model by using Dataset 2 and test it on Dataset 1, the mean R^2 –score was 0.947. Fig. 6d shows observations versus predictions for the Extra Trees Regressor in this case. This indicates that our model could adapt well to the change of dataset. Furthermore, it gives an important insight into the quality of the training dataset.

4. Conclusions

Reliable prediction of contaminant plume behavior is of paramount importance for risk assessment and sustainable handling of the contaminants present in soil and groundwater. In this paper, we focused on PCP concentration prediction and found the most important hydro-chemical constituents by combining ML regressor algorithms with SHAP values. Specifically, we implemented six regressor methods, three linear (standard linear regressor, ridge regressor, and LASSO) and three tree-based (random forest, extra trees, and XGB regressor) to predict and

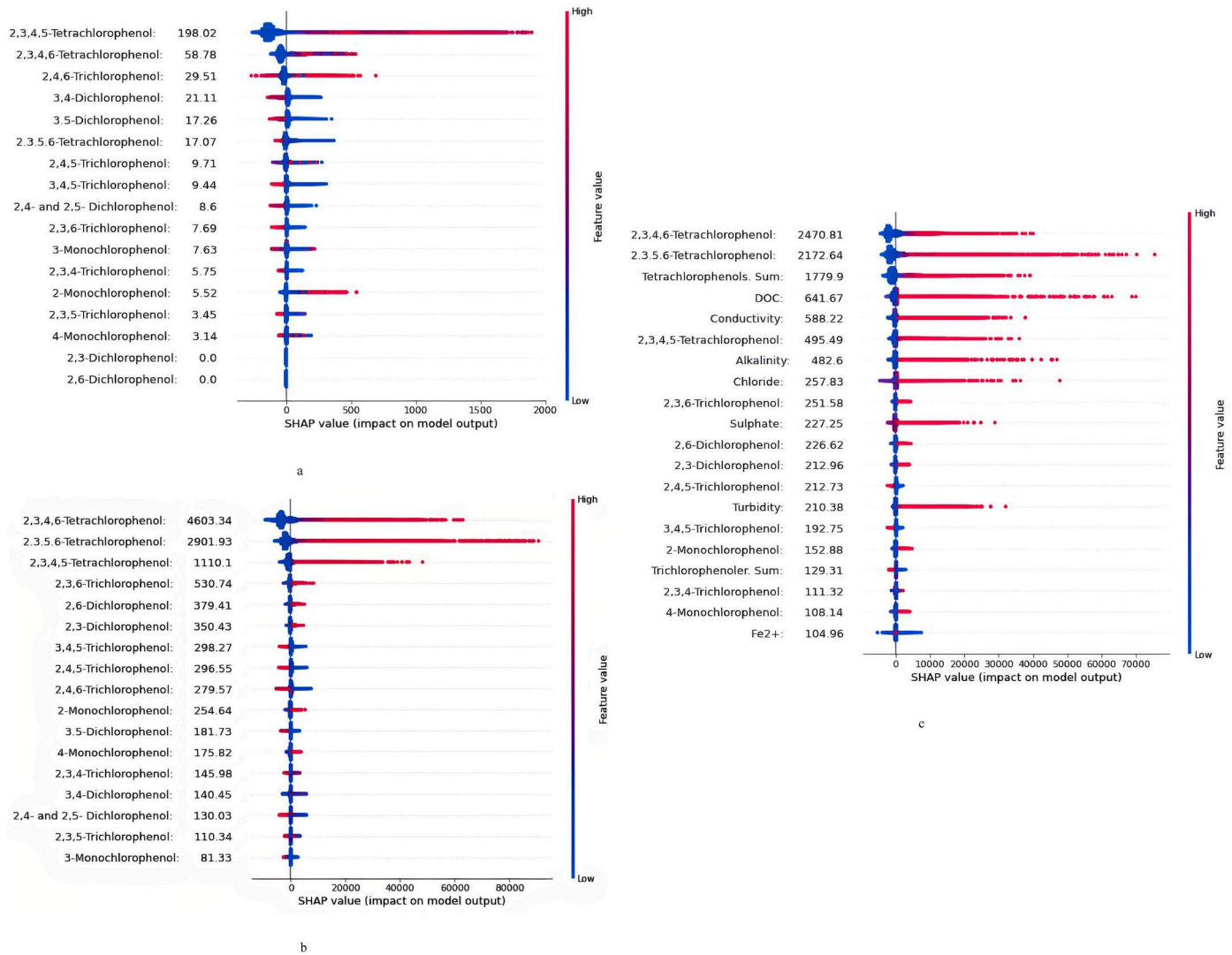


Fig. 5. SHAP value beeswarm summary plot for the Extra Trees Regressor and a) Dataset 1, b) Dataset 2, and c) Dataset 3 when we eliminate “Chlorophenols. Sum”.

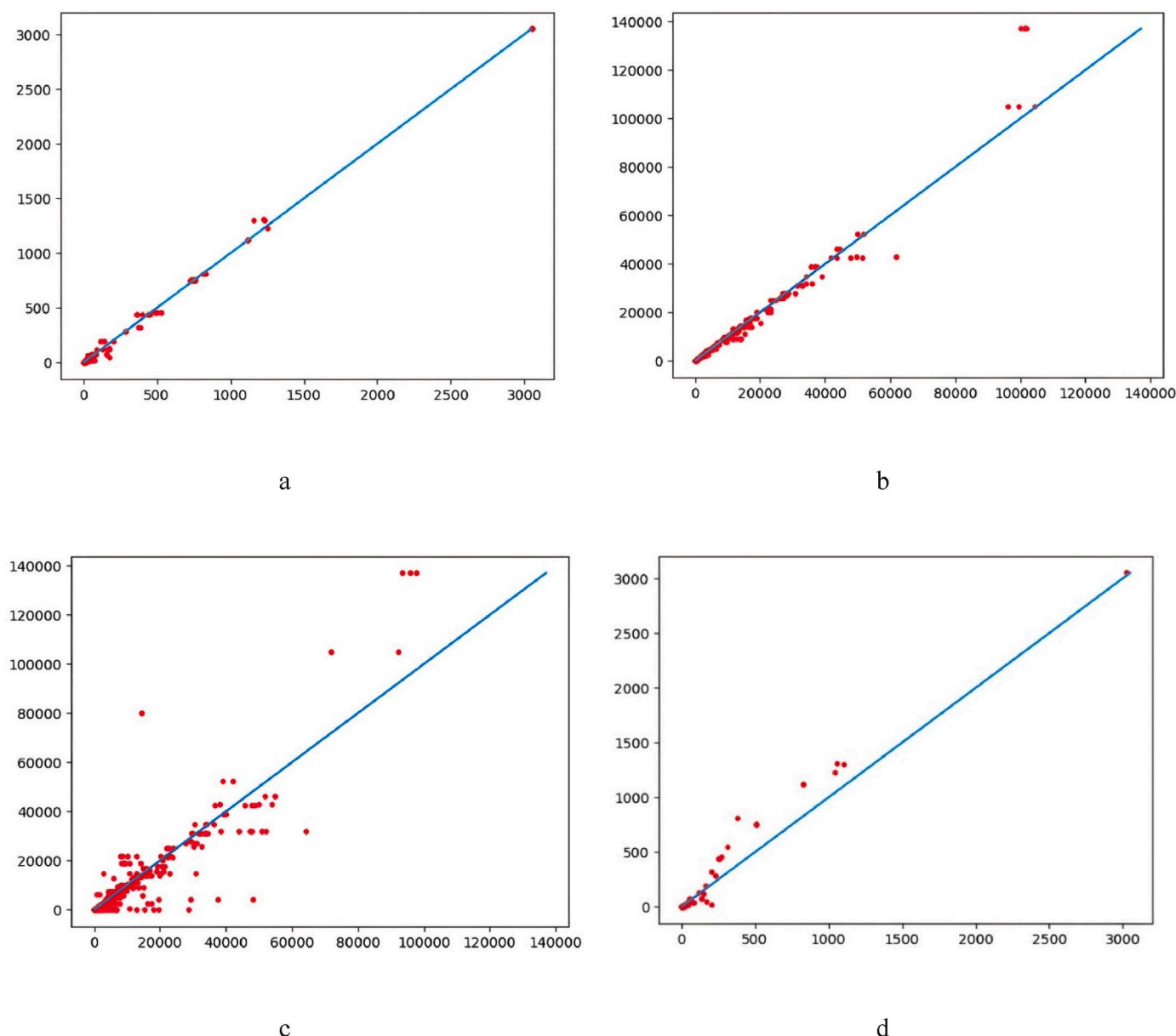


Fig. 6. Observed vs. predicted values for the extra trees regressor and a) Dataset 1, b) Dataset 2, c) Dataset 3, and d) when we train the model by using Dataset 2 and then test it by Dataset 1.

improve the understanding of PCP concentration dynamics in groundwater, unveiling the more significantly contributing hydrochemical constituents in the dataset influencing PCP concentrations downstream of the source.

As shown by the results of the models, even simple linear ML can achieve notable success in predicting PCP concentrations for datasets devoid of missing values. For these datasets (Dataset 1 and Dataset 2), the LASSO Regressor can achieve the best results (with an R^2 -score of more than 0.998) while having the lowest complexity due to its automatic feature selection. On the other hand, when dealing with datasets containing missing values, linear regressors, including LASSO, prove ineffective. In contrast, tree-based regressors, which are more complex compared to linear ones, emerge as a preferable choice for accurately predicting PCP concentrations in such datasets. All tree-based regressors demonstrate the ability to predict PCP concentrations to some degree, with extra trees consistently yielding the highest R^2 -score values among the tree-based models across all datasets (ranging between 0.48 and 0.96). Extra Trees achieves the R^2 -score of 0.76 for Dataset 3, where

linear models were not helpful. Overall, the extra trees regressor was the best model for predicting PCP concentrations with missing observations.

In a next step, we interpreted the proposed regressor models and the importance of different constituents for PCP concentration prediction by exploiting the SHAP value approach. The regressor models were further explored to improve the understanding of how the predictions may be backed by knowledge gained for how individual features, included in the models, may affect the prediction outputs. Subsequently, and based on these interpretations, an ablation analysis was conducted where the constituent “Chlorophenols. Sum” was removed from the datasets (as the most important feature based on the SHAP value plots) to determine the contribution of this component for the predictions. Results indicate that this was a significantly influential component that could meaningfully determine the relevance of the model (with R^2 -score generally above 0.7).

Notably, the exact calculation of the SHAP values for large tree-based models is intractable (NP-hard problem) and some efficient algorithms were applied for enabling the tree-based SHAP value calculation that

run in linear time, instead of exact calculation of SHAP values. Hence, for a larger number of features, we may have less accurate SHAP-values. Also, from a hydrological point of view it is important to note that we rely on existing data here that may not be sufficiently comprehensive for a long-time span of investigations. This, regarding climate variability-induced changes in the subsurface media, might affect the model capabilities significantly in certain applications. This suggests a way forward for such research on long lasting contaminants in the subsurface.

Having a dataset with missing values is quite common in contaminated site investigations, and improved understanding resulting from the approach developed and implemented in this paper is expected to be of great help in driving future field investigations. This is especially important for persistent contaminants such as PCP requiring longer periods of site monitoring and plume control which in turn needs a reliable understanding of the plume dynamics. Complex hydrogeological media, as found in the two contaminated sites dealt with in this paper, makes planning monitoring and remediation actions even more time and resource demanding. Moreover, the need for long term monitoring and remediation actions at a site may significantly affect the livelihood of the surrounding areas. Thus, having reliable models to predict the plume's dynamics and deepen the understanding of significant chemicals involved in these dynamics, provided by such models as developed and demonstrated here, may significantly enhance the cost-efficiency of investigative actions and reduce any life cycle consequences from chosen remedial actions. This approach could thus pave the way to improve our understanding of other constituents typically found in groundwater that also have complex subsurface behavior dynamics and of which elucidation could be important for improving existing, as well as designing future monitoring and remediation projects more efficiently.

CRedit authorship contribution statement

Mehran Rad: Writing – review & editing, Writing – original draft, Funding acquisition, Data curation, Conceptualization. **Azra Abtahi:** Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Ronny Berndtsson:** Writing – review & editing, Supervision, Resources. **Ursula S. McKnight:** Writing – review & editing, Supervision, Methodology, Formal analysis, Conceptualization. **Amir Aminifar:** Supervision, Software, Methodology, Funding acquisition, Formal analysis, Data curation, 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.

Data availability

Data will be made available on request.

Acknowledgement

This research was financially supported by the Åke och Greta Lissheds Foundation, and by the Swedish Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.envpol.2024.123449>.

References

- Breiman, L., 2001. Random forests. *Mach. Learn.* 45, 5–32. <https://doi.org/10.1023/A:1010933404324>.
- Chen, T., Guestrin, C., 2016. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794. <https://doi.org/10.1145/2939672.2939785>.
- Chowdhury, A., Rahnuma, M., 2023. Groundwater contaminant transport modeling using MODFLOW and MT3DMS: a case study in Rajshahi City. *Water Pract. Technol.* 18, 1255–1272. <https://doi.org/10.2166/wpt.2023.076>.
- Davis, A.K., Campbell, J., Gilbert, C., Ruby, M., Bennett, M., Tobin, S., 1994. Attenuation and biodegradation of chlorophenols in ground water at a former wood treating facility. *Ground Water* 32, 248–257.
- Elzain, H.E., Chung, S.Y., Senapathi, V., Sekar, S., Lee, S.Y., Roy, P.D., Hassan, A., Sabarathinam, C., 2022. Comparative study of machine learning models for evaluating groundwater vulnerability to nitrate contamination. *Ecotoxicol. Environ. Saf.* 229, 113061 <https://doi.org/10.1016/j.ecoenv.2021.113061>.
- Elzain, H.E., Chung, S.Y., Venkatramanan, S., Selvam, S., Ahemd, H.A., Seo, Y.K., Bhuyan, MdS., Yassin, M.A., 2023. Novel machine learning algorithms to predict the groundwater vulnerability index to nitrate pollution at two levels of modeling. *Chemosphere* 314, 137671. <https://doi.org/10.1016/j.chemosphere.2022.137671>.
- Freedman, D., 2009. *Statistical Models: Theory and Practice*.
- Geurts, P., Ernst, D., Wehenkel, L., 2006. Extremely randomized trees. *Mach. Learn.* 63, 3–42. <https://doi.org/10.1007/s10994-006-6226-1>.
- Gruber, M., 1998. *Improving Efficiency by Shrinkage: the James–Stein and Ridge Regression Estimators*.
- Gujarati, D.N., Porter, D.C., 2009. *Basic Econometrics*, fifth ed. McGraw-Hill Irwin, Boston.
- Hou, D., 2020. *Sustainable Remediation of Contaminated Soil and Groundwater: Materials, Processes, and Assessment*. Butterworth-Heinemann, An imprint of Elsevier.
- Jobson, J.D., 2012. *Applied Multivariate Data Analysis: Regression and Experimental Design*. Springer Science & Business Media.
- Johansson, A., 2020. *Groundwater Flow Modelling to Address Hydrogeological Response of a Contaminated Site to Remediation Measures at Hjortsberga southern Sweden*.
- Johansson, L., 2006. Södra skogsägarna ekonomisk förening. Utökad undersökning vid Hjortsberga sågverk (southern forest owners cooperative society. Extended investigation at Hjortsberga sawmill) [WWW Document]. URL: https://www.lansstyrelsen.se/download/18.4e0415ee166afb59324449/154055835286/Sågverk_i_Hjortsberga-Utökad_undersökning.-SWEKO_VIAK_2006.pdf.
- Jordan, M.I., Mitchell, T.M., 2015. Machine learning: trends, perspectives, and prospects. *Science* 349, 255–260. <https://doi.org/10.1126/science.aaa8415>.
- Lemaire, G.G., Jessen Rasmussen, J., Höss, S., Figari Kramer, S., Schittich, A.-R., Zhou, Y., Köppl, C.J., Traunsperger, W., Bjerg, P.L., McKnight, U.S., 2022. Land use contribution to spatiotemporal stream water and ecological quality: implications for water resources management in peri-urban catchments. *Ecol. Indic.* 143, 109360 <https://doi.org/10.1016/j.ecolind.2022.109360>.
- Libera, A., de Barros, F.P.J., Faybishenko, B., Eddy-Dilek, C., Denham, M., Lipnikov, K., Moulton, D., Maco, B., Wainwright, H., 2019. Climate change impact on residual contaminants under sustainable remediation. *J. Contam. Hydrol.* 226, 103518 <https://doi.org/10.1016/j.jconhyd.2019.103518>.
- Liu, F., Xu, B., He, Y., Brookes, P.C., Xu, J., 2019. Co-transport of phenanthrene and pentachlorophenol by natural soil nanoparticles through saturated sand columns. *Environ. Pollut.* 249, 406–413. <https://doi.org/10.1016/j.envpol.2019.03.052>.
- Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.-I., 2020. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* 2, 56–67. <https://doi.org/10.1038/s42256-019-0138-9>.
- Lundberg, S.M., Erion, G.G., Lee, S.-I., 2019. Consistent Individualized Feature Attribution for Tree Ensembles. <https://doi.org/10.48550/arXiv.1802.03888>.
- Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. In: *Advances in Neural Information Processing Systems*. Curran Associates, Inc.
- Matsui, Y., Matsui, T., 2001. NP-completeness for calculating power indices of weighted majority games. *Theor. Comput. Sci., Combinatorics and Computer Science* 263, 305–310. [https://doi.org/10.1016/S0304-3975\(00\)00251-6](https://doi.org/10.1016/S0304-3975(00)00251-6).
- McLean, M.I., Evers, L., Bowman, A.W., Bonte, M., Jones, W.R., 2019. Statistical modelling of groundwater contamination monitoring data: a comparison of spatial and spatiotemporal methods. *Sci. Total Environ.* 652, 1339–1346. <https://doi.org/10.1016/j.scitotenv.2018.10.231>.
- Meray, A.O., Sturla, S., Siddiquee, M.R., Serata, R., Uhlemann, S., Gonzalez-Raymat, H., Denham, M., Upadhyay, H., Lagos, L.E., Eddy-Dilek, C., Wainwright, H.M., 2022. PyLEnM: a machine learning framework for long-term groundwater contamination monitoring strategies. *Environ. Sci. Technol.* 56, 5973–5983. <https://doi.org/10.1021/acs.est.1c07440>.
- Milasinović, M., Randelović, A., Jaćimović, N., Prodanović, D., 2019. Coupled groundwater hydrodynamic and pollution transport modelling using Cellular Automata approach. *J. Hydrol.* 576, 652–666. <https://doi.org/10.1016/j.jhydrol.2019.06.062>.
- Motavalli, A., Naghibi, S.A., Hashemi, H., Berndtsson, R., Pradhan, B., Gholami, V., 2019. Inverse method using boosted regression tree and k-nearest neighbor to quantify effects of point and non-point source nitrate pollution in groundwater. *J. Clean. Prod.* 228, 1248–1263. <https://doi.org/10.1016/j.jclepro.2019.04.293>.
- Naseri-Rad, M., Berndtsson, R., Aminifar, A., McKnight, U.S., O'Connor, D., Persson, K. M., 2022. DynSus: dynamic sustainability assessment in groundwater remediation practice. *Sci. Total Environ.* 832, 154992 <https://doi.org/10.1016/j.scitotenv.2022.154992>.

- Naseri-Rad, M., Berndtsson, R., McKnight, U.S., Persson, M., Persson, K.M., 2021. INSIDE-T: a groundwater contamination transport model for sustainability assessment in remediation practice. *Sustainability* 13, 7596. <https://doi.org/10.3390/su13147596>.
- Naseri-Rad, M., Berndtsson, R., Persson, K.M., Nakagawa, K., 2020. INSIDE: an efficient guide for sustainable remediation practice in addressing contaminated soil and groundwater. *Sci. Total Environ.* 740, 139879 <https://doi.org/10.1016/j.scitotenv.2020.139879>.
- O'Connor, D., Hou, D., Ok, Y.S., Song, Y., Sarmah, A.K., Li, X., Tack, F.M.G., 2018. Sustainable in situ remediation of recalcitrant organic pollutants in groundwater with controlled release materials: a review. *J. Contr. Release* 283, 200–213. <https://doi.org/10.1016/j.jconrel.2018.06.007>.
- Quinlan, J.R., 1986. Induction of decision trees. *Mach. Learn.* 1, 81–106. <https://doi.org/10.1007/BF00116251>.
- Radelyuk, I., Naseri-Rad, M., Hashemi, H., Persson, M., Berndtsson, R., Yelubay, M., Tussupova, K., 2021. Assessing data-scarce contaminated groundwater sites surrounding petrochemical industries. *Environ. Earth Sci.* 80, 351. <https://doi.org/10.1007/s12665-021-09653-z>.
- Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* 1, 206–215. <https://doi.org/10.1038/s42256-019-0048-x>.
- Schmidt, F., Wainwright, H.M., Faybishenko, B., Denham, M., Eddy-Dilek, C., 2018. In situ monitoring of groundwater contamination using the kalman filter. *Environ. Sci. Technol.* 52, 7418–7425. <https://doi.org/10.1021/acs.est.8b00017>.
- Schmidt, L.M., Delfino, J.J., Preston 3rd, J.F., St Laurent 3rd, G., 1999. Biodegradation of low aqueous concentration pentachlorophenol (PCP) contaminated groundwater. *Chemosphere* 38, 2897–2912. [https://doi.org/10.1016/s0045-6535\(98\)00480-9](https://doi.org/10.1016/s0045-6535(98)00480-9).
- Schmidt, L.M., Delfino, J.J., Preston, J.F., St Laurent, G., 1999. Biodegradation of low aqueous concentration pentachlorophenol (PCP) contaminated groundwater. *Chemosphere* 38, 2897–2912. [https://doi.org/10.1016/s0045-6535\(98\)00480-9](https://doi.org/10.1016/s0045-6535(98)00480-9).
- SGU, 2017. Hjortsberga F.D Sägverk (In Swedish) [Hjortsberga F.D Sawmill]. Uppsala.
- Sonne, A.Th, McKnight, U.S., Rønde, V., Bjerg, P.L., 2017. Assessing the chemical contamination dynamics in a mixed land use stream system. *Water Res.* 125, 141–151. <https://doi.org/10.1016/j.watres.2017.08.031>.
- Sundararajan, M., Najmi, A., 2020. The many shapley values for model explanation. In: *Proceedings of the 37th International Conference on Machine Learning*. Presented at the International Conference on Machine Learning, pp. 9269–9278. PMLR.
- Swedish EPA, 2009. The role of pentachlorophenol treated wood for emissions of dioxins into the environment [WWW Document]. URL. <https://www.naturvardsverket.se/Documents/publikationer/978-91-620-5935-4.pdf?pid=3534>.
- Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Methodol.* 58, 267–288.
- UNTC, 2001. Stockholm Convention on Persistent Organic Pollutants. Stockholm.
- USEPA, E.P.A.U., 1991. On-Site Treatment of Creosote and Pentachlorophenol Sludges and Contaminated Soil. Washington, DC.
- Yu, X., Cui, T., Sreekanth, J., Mangeon, S., Doble, R., Xin, P., Rassam, D., Gilfedder, M., 2020. Deep learning emulators for groundwater contaminant transport modelling. *J. Hydrol.* 590, 125351 <https://doi.org/10.1016/j.jhydrol.2020.125351>.
- Zhang, H., Yang, R., Wang, Y., Ye, R., 2019. The evaluation and prediction of agriculture-related nitrate contamination in groundwater in Chengdu Plain, southwestern China. *Hydrogeol. J.* 27, 785–799. <https://doi.org/10.1007/s10040-018-1886-z>.