Integrating process-based reactive transport modeling and machine learning for electrokinetic remediation of contaminated groundwater

Sprocati, R.; Rolle, M.

Published in:
Water Resources Research

Link to article, DOI:
10.1029/2021WR029959

Publication date:
2021

Document Version
Peer reviewed version

Integrating process-based reactive transport modeling and machine learning for electrokinetic remediation of contaminated groundwater

R. Sprocati[1], M. Rolle[1]

[1]Department of Environmental Engineering, Technical University of Denmark, Bygningstorvet, Building 115, 2800 Kgs. Lyngby, Denmark

Corresponding author: Massimo Rolle (masro@env.dtu.dk)

Key Points:

- Development of a surrogate modeling framework for electrokinetic bioremediation from a multidimensional multiphysics process-based model
- The surrogate model predicts well (R^2>0.90) the outputs of the reactive transport model including EK transport and biogeochemical reactions
- The surrogate modeling framework allows computationally efficient model exploration, sensitivity analysis, and uncertainty quantification

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1029/2021WR029959.

Abstract

Advanced reactive transport models of fluid flow and solute transport in subsurface porous media are instrumental for the assessment of contaminant environmental fate and for the design of in situ remediation interventions. However, the increasing complexity of process-based reactive transport simulators often leads to long runtimes, which poses severe restrictions for tasks that require numerous model evaluations. To overcome this limitation, we demonstrate how machine learning surrogate models, trained on the outputs of a limited number of process-based reactive transport simulations, can predict the evolution of complex subsurface systems. We focus on electrokinetic enhanced bioremediation of chlorinated solvents in low-permeability porous media, which is an in situ remediation technology entailing a suite of complex and coupled physical, chemical and biological processes. A process-based, multicomponent reactive transport model, capable of describing the key mechanisms of electrokinetic flow and transport, is setup in a two-dimensional domain. The model accounts for electromigration and electroosmosis, the electrostatic interactions between charged species, the chemistry of the pore water solution, the microbially-mediated degradation of the organic compounds, and the dynamics of different degraders. We develop a response surface surrogate framework using an artificial neural network as approximation function and we show that the surrogate model has the capability and the flexibility to capture the complex dynamics of electrokinetic remediation in subsurface porous media and allows computationally efficient model exploration, sensitivity analysis and uncertainty quantification.

1. Introduction

Multiphysics reactive transport models are essential tools in many fields of science and engineering (Prommer et al., 2019; Steefel et al., 2015). Process-based reactive transport simulators coupling flow and transport processes to complex networks of equilibrium and kinetically controlled biogeochemical reactions (e.g., Appelo & Rolle, 2010; Mayer et al., 2002; Nardi et al., 2014; Rolle et al., 2018) are instrumental to describe multispecies and multicomponent transport in natural subsurface systems as well as in engineering applications. Despite advances in computational power and the increasing accessibility of large high-performance-computing clusters that allow for model parallelization and scalability, there are still significant limitations regarding the use of complex process-based models which frequently require either a significant amount of computational time or a large number of processors for each model evaluation (Damiani et al., 2020; Sohrabi et al., 2019; Su et al., 2017). For instance, tasks such as parameter calibration, sensitivity analysis, optimization and uncertainty analysis can require hundreds or thousands model evaluations, thus leading to prohibitive resources requirement in terms of time and infrastructure for advanced reactive transport codes.

To overcome this issue, one solution is to develop surrogate models (also called metamodels) from the process-based models, which are able to predict the relations between inputs and outputs, with the advantage of providing results very quickly, almost in real-time (Forrester et al., 2008). The largest computational time required to develop surrogate models is indeed in the evaluation of a finite number of process-based simulations that are used for the training of the surrogate, which however are only a fraction of the number of simulations required for any of the tasks described above. Surrogate models have been successfully developed for several applications in environmental sciences and water resources (Asher et al., 2015; Razavi et al., 2012; Tahmasebi et al., 2020). Applications on flow and transport problems have been recently investigated to predict contaminant transport and source location (Mo et al., 2019; Wang et al., 2020; X. Yu et al., 2020) as well as optimal well location and pumping schedule for pump-and-treat operations (Yan & Minsker, 2006; Yin & Tsai, 2020). Other implementations also included models for the multiphase transport of dense non-aqueous phase liquids (Jiang & Na, 2020; Luo et al., 2020; Ouyang et al., 2017). However, to the best of our knowledge, applications of surrogate models to complex reactive transport problems involving both physical flow and transport processes as well as comprehensive networks of biogeochemical reactions for natural transport and engineered in situ remediation of contaminated groundwater are still lacking.

In this study we focus on electrokinetic bioremediation (EK-Bio), which is an emerging in situ remediation technology promoting biological degradation of organic contaminants in low-permeability porous media through enhanced amendment delivery by the application of a DC current electric field in the subsurface. Modelling of such system is particularly challenging as it involves the multidimensional solution of fluid flow, electrokinetic transport considering the Coulombic interactions between the different charged species, and a wide range of geochemical and biological reactions. We consider a two-dimensional process-based model entailing a cathode-anode doublet for electrokinetically enhanced distribution of amendments stimulating in situ biodegradation of chlorinated ethenes. We develop a response surface surrogate framework and we show how an approximation function, trained on a limited set of simulations performed with the complex process-based reactive transport simulator, allows preforming model exploration, sensitivity analysis and uncertainty analysis almost in real time, while accounting for all the input-output relations of the original process-based model.

2. Modeling approach

2.1. Conceptual model of EK-Bio

Electrokinetic remediation (EK) consists in the application of an electric potential field in the subsurface to enhance transport processes and represents one of the few promising approaches to cleanup groundwater contamination in impervious, low-permeability formations. During EK, the following mechanisms drive the transport processes as a result of the induced electric potential gradient (Chowdhury et al., 2017; Lima et al., 2017; Reddy & Cameselle, 2009; Reynolds et al., 2008): (i) electromigration, resulting in the transport of charged species towards the electrode of opposite polarity and (ii) electroosmosis, consisting in the movement of pore water as a result of the movement of the excess of charges in the diffuse double layer of soil particles.

In this study, we consider an application of electrokinetically-enhanced bioremediation (EK-Bio) of chlorinated ethenes in a low-permeability groundwater flow system. Lactate ($C_3O_5H_3^-$), a negatively charged substrate, is electrokinetically delivered from a cathodic electrode to promote the biodegradation of the chlorinated contaminants initially present in the subsurface. The delivery of lactate stimulates the microbial activity of indigenous organohalide-respiring bacteria (OHRB), which are able to sequentially degrade tetrachloroethylene (PCE), the chlorinated compound initially present in the domain, to trichloroethylene (TCE) and dichloroethylene (DCE) (Buttet et al., 2018; Murray et al., 2019; S. Yu et al., 2005). Specialized OHRB (KB-1), able to perform complete dehalogenation of chlorinated solvents including the conversion of DCE to vinyl chloride (VC) and subsequently to the non-toxic ethene, are distributed with the electroosmotic flow from the anode to the cathode. Figure 1a,b schematically illustrates an EK-Bio system, including the main EK process, the recirculation system to buffer electrolysis reactions preventing extreme pH conditions at the electrodes, and the biodegradation of the chlorinated contaminants in which the parent compound (PCE) is initially present in the domain both as dissolved species and as segregated NAPL (Non-Aqueous Phase Liquid) phase. This type of EK technique has been previously tested in laboratory experiments (Mao et al., 2012) and in a pilot-scale field application in Skuldelev, Denmark (NIRAS, 2011) for which a process-based model of EK-Bio has been recently proposed to elucidate the system dynamics (Sprocati et al., 2020).

In this study, we simulate the complex processes occurring during EK-Bio in a simplified geometry consisting of a 2D horizontal domain (5 m × 3 m) with two electrodes (one cathode and one anode), placed at three meters distance from each other.

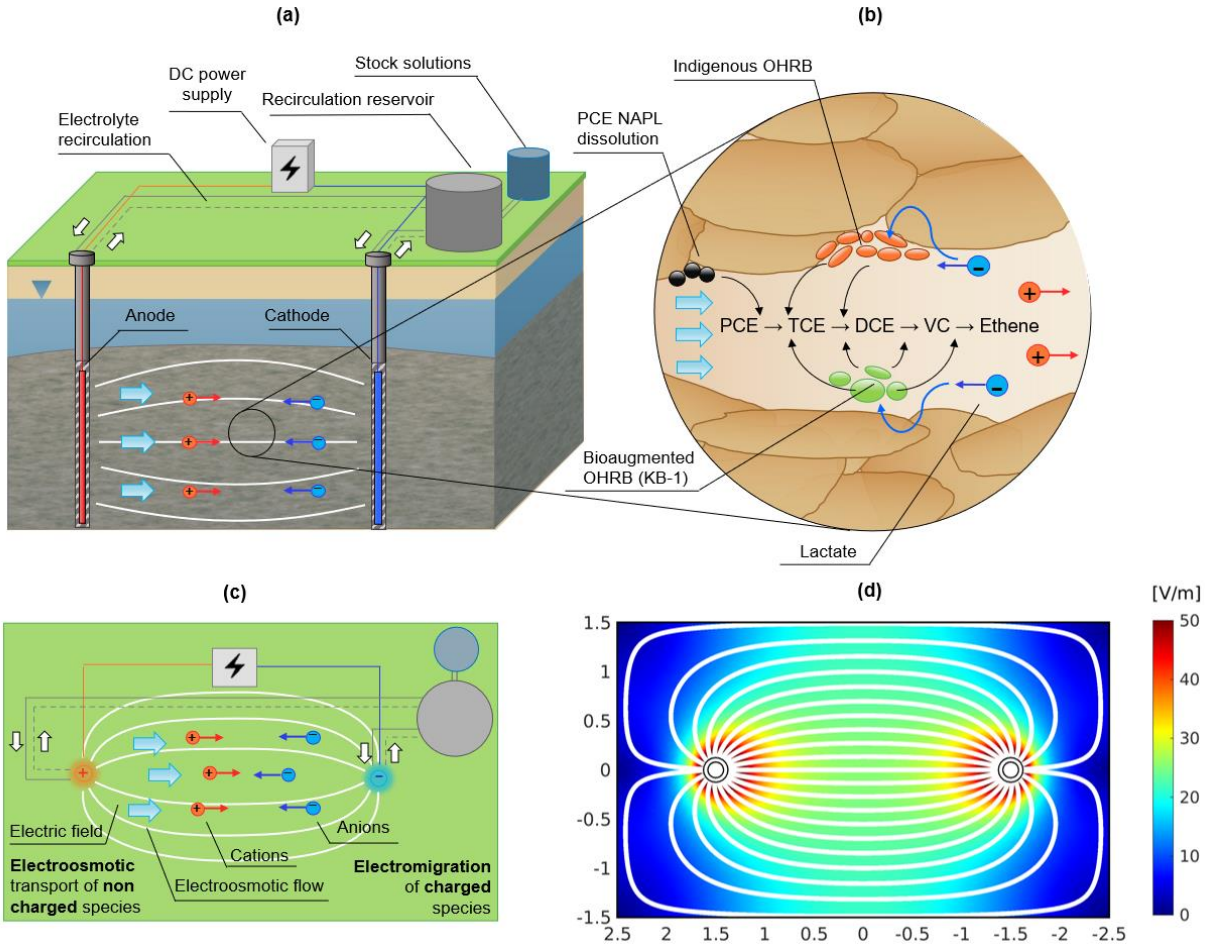

Figure 1. Schematic of EK-Bio for a system consisting of two electrodes. **(a)** illustration of the key elements of the EK setup, **(b)** details of EK transport and biodegradation processes, **(c)** plan view showing the 2D model domain with indication of the governing transport mechanisms and **(d)** simulated electric potential gradients in the domain with indication of the electric potential streamlines.

During EK treatment, the electric field is applied for a certain time to ensure delivery of both the electron donor and the bioaugmented degraders, KB-1 (Figure 1). After mixing the delivered electron donor (lactate), the electron acceptors (chlorinated compounds) and the degrading microorganisms, the electric field is turned off and biodegradation reactions can proceed.

2.1.1. Electrokinetic transport processes

The flux of charged species in a porous medium can be described with the Nernst-Planck equation (Alt-Epping et al., 2015; Rasouli et al., 2015; Rolle et al., 2018; Steefel & Tournassat, 2020; Tournassat & Steefel, 2019; Wu et al., 2020), which accounts for the contribution of diffusion, migration, and electroosmosis:

$$\boldsymbol{J}_i^{Tot} = \underbrace{-nD_i\nabla c_i}_{\boldsymbol{J}_i^{Dif}} \underbrace{-nD_i\frac{z_iF}{RT}c_i\nabla\Phi}_{\boldsymbol{J}_i^{Mig}} \underbrace{+n\,\boldsymbol{v}_{\boldsymbol{eo}}c_i}_{\boldsymbol{J}_i^{Adv}} \tag{1}$$

where n is the accessible porosity, $D_i = D_i^{aq}\tau$ is the pore diffusion/dispersion coefficient in which D_i^{aq} is the aqueous diffusion coefficient of the species i, τ the tortuosity, c_i the molar concentration, ∇c_i the concentration gradient, z_i the charge, F the Faraday constant, R the gas constant, T the temperature, $\nabla\Phi$ the electric potential gradient and $\boldsymbol{v}_{\boldsymbol{eo}}$ the average velocity resulting from electroosmotic flow, calculated as $\boldsymbol{v}_{eo} = -k_{eo}\nabla\Phi/n$ where k_{eo} represents the electroosmotic permeability (Alizadeh et al., 2021; Yustres et al., 2020). The terms $\boldsymbol{J}_i^{Tot}, \boldsymbol{J}_i^{Dif}, \boldsymbol{J}_i^{Mig}$ and $\boldsymbol{J}_i^{Eo}$ represent the total, diffusive, migration and electroosmotic fluxes, respectively.

The mass balance allows deriving the governing equations of multicomponent ionic transport (e.g., Muniruzzaman & Rolle, 2015, 2017; Rolle et al., 2013):

$$\frac{\partial(nc_i)}{\partial t} + \nabla\cdot\boldsymbol{J}_i^{Tot} = r_i \tag{2}$$

in which r_i is the source/sink term.

The Poisson's equation regulates the charge interactions in the system (Newman & Thomas-Alyea, 2004):

$$\nabla^2\Phi = -\frac{F}{\varepsilon}\sum_{i=1}^{N} z_ic_i = -\frac{\rho_e}{\varepsilon} \tag{3}$$

where ε is the dielectric constant of the porous medium, N is the number of charged species in solution and ρ_e is the charge density of the solution. At the continuum scale the overall charge density of the solution is zero and Eq. (3) is equivalent to the electroneutrality condition.

As a flux of electrons corresponds to a flux of current, the current density is defined as:

$$\boldsymbol{I} = F\sum_{i=1}^{N} z_i\boldsymbol{J}_i^{Tot} = -F\sum_{i=1}^{N} z_inD_i\nabla c_i - \left(F^2\sum_{i=1}^{N} z_i^2\frac{nD_i}{RT}c_i\right)\nabla\Phi + Fn\boldsymbol{v}_{\boldsymbol{eo}}\sum_{i=1}^{N} z_ic_i \tag{4}$$

therefore, the current balance in the system reads as:

$$F\sum_{i=1}^{N} z_i\frac{\partial(nc_i)}{\partial t} + \nabla\cdot\left(F\sum_{i=1}^{N} z_i\boldsymbol{J}_i^{Tot}\right) = F\sum_{i=1}^{N} z_ir_i \tag{5}$$

Equations (1)-(5) represent the Poisson-Nernst-Planck equations that account for solute transport, electric field distribution, charge balance, Coulombic interactions and current conservation.

2.1.2. Biogeochemical reactions network

The process-based model used in this study accounts for fast, equilibrium reactions for aqueous speciation and kinetically-controlled reactions for contaminant biodegradation and microbial population dynamics. In the proposed conceptual model, lactate is used as electron donor by both

indigenous OHRB and bioaugmented OHRB (KB-1) (Duhamel & Edwards, 2007; Sprocati et al., 2020) to degrade PCE to TCE, DCE, VC and ethene. The general microbially-mediated redox reaction for the first step of PCE degradation to TCE using lactate as electron donor reads as:

$$\frac{1}{6}C_3O_5H_3^- + C_2Cl_4 + H_2O \rightarrow C_2HCl_3 + Cl^- + \frac{4}{3}H^+ + \frac{1}{2}HCO_3^- \tag{6}$$

Similar reactions describe the other steps of reductive dehalogenation, coupling the oxidation of lactate to the reduction of TCE, DCE, and VC.

The rate of all reductive dehalogenation reactions is defined according to a double Monod kinetics including competitive inhibition of the chlorinated ethenes (Murray et al., 2020; Zhou et al., 2006):

$$\frac{dc_{EA,i}}{dt} = -\eta\, \kappa_{max,EA_i,X_k}\left(\frac{c_{ED}}{c_{ED} + K_{s,ED}}\right)\left(\frac{c_{EA_i}}{c_{EA_i} + K_{s,EA_i}\left(1 + \sum_{j=1}^{n}\frac{c_{EA_j}}{K_{inh_j}}\right)}\right)X_k \tag{7}$$

In Eq. (7), c_{ED} is the molar concentration of lactate, $c_{EA,i}$ is the molar concentration of the i-th electron acceptor (EA), K_{s,EA_i} is the half-saturation of the i-th EA whereas K_{inh_j} is the inhibition constant of the j-th EA which accounts for all the other chlorinated compounds except for the i-th EA and $K_{s,ED}$ is the half-saturation constant of lactate. The term η is a generic coefficient that accounts for variability of the biodegradation rate (e.g., under different conditions and/or between field and laboratory settings). κ_{max,EA_i,X_k} is the maximum specific degradation rate of the i-th electron acceptor caused by the k-th bacterial consortium present with X_k molar concentration.

The dynamics of both indigenous (X_{ind}) and bioaugmented biomass (X_{KB-1}) is modeled considering a growth and a linear decay term:

$$\frac{dX_k}{dt} = -Y\frac{dc_{EA,i}}{dt} - b_kX_k \tag{8}$$

where Y is the bacterial yield. The indigenous biomass is considered attached to the solid matrix, whereas the bioaugmented organisms are transported by the electroosmotic flow.

We considered that the parent compound PCE is also present as non-aqueous phase liquid (NAPL) at the beginning of the simulation. The dissolution of PCE from the NAPL phase to the aqueous phase, in which the biological reactions occurs, is described with a linear mass transfer expression (e.g., Ramsburg et al., 2011)

$$\frac{dc_{PCE}}{dt} = \omega_{PCE}(S_{PCE} - c_{PCE}) \tag{9}$$

where ω_{PCE} is the mass-transfer rate coefficient of PCE and S_{PCE} is the aqueous solubility.

In addition to the kinetic degradation reactions, Eq. (6)-(9), we also considered equilibrium aqueous speciation reactions in the pore water with equilibrium constants according to the database phreeqc.dat.

2.1.3. *Process-based numerical model*

The EK-Bio system was simulated with the code NP-Phreeqc-EK (Sprocati et al., 2019), which is a MATLAB implementation specifically developed for electrokinetic applications that couples the flow and transport software COMSOL Multiphysics with the geochemical code PhreeqcRM (Muniruzzaman & Rolle, 2019; Parkhurst & Wissmeier, 2015). NP-Phreeqc-EK is based on a sequential non-iterative approach: flow and transport equations are solved with COMSOL Multiphysics for short time steps, after which the concentrations, evaluated for each mesh point, are passed to PhreeqcRM, which performs the equilibrium and kinetic reactions calculations, considering each mesh point as an independent cell. At the end of the reaction step, the resulting concentrations from PhreeqcRM are set as initial values in COMSOL Multiphysics for the subsequent flow and transport time step. Such communication is repeated until the end of the defined simulation time. NP-Phreeqc-EK is optimized for multi-processing, using both the OpenMP features of PhreeqcRM and multithreading options available in COMSOL Multiphysics.

In the considered system we included 17 species (9 of which are charged) with the diffusion coefficients at 15 °C listed in Table S1. Overall, the biogeochemical model accounted for 7 kinetically-controlled reactions and 6 equilibrium reactions (Table S3). The model domain was discretized into 7290 elements resulting in 67770 degrees of freedom. The total simulation time was 360 days, subdivided in 180 coupling time steps between the flow and transport simulator and the geochemical code. The simulations were performed in a High-Performance Computing (HCP) cluster, running up to five process-based models in parallel in four nodes. Overall, using max 7.5 Gb of RAM and 4 cores, each model simulation took approximately 11 hours. For each process-based simulation the reactive steps performed by PhreeqcRM took from 2 to 5 seconds whereas the transport steps performed by Comsol had a variable duration ranging from 30 to 300 seconds. In this study the solution of multicomponent ionic transport was the bottleneck of the process-based simulations due to the nonlinearity of the coupled transport equations and the number of mesh elements. In cases in which the reaction step is the bottleneck, different approaches have been proposed to integrate machine learning surrogates in the reaction step to speed up the process-based models (Hennig & Kühn, 2021; Jatnieks et al., 2016; Leal et al., 2020; De Lucia & Kühn, 2021).

To compare the outputs of the process-based model, we evaluated different metrics, which provide information on both the distribution of the species in the domain and the extent of the degradation.

The first metric used to evaluate the performance of the EK-Bio system is the relative area (RA) of reactant delivery. *RA* is used to assess the amendments' distribution in the domain and its values range from 0 (no distribution) to 1 (complete distribution in the whole domain):

$$RA_i(t) = \frac{1}{A_{Tot}} \int_{\Omega} H(c_i(\boldsymbol{x}, t) - c_{i,lim}) d\Omega \tag{10}$$

where A_{Tot} is the total area of the considered domain Ω, H is the Heaviside step function and $c_{i,lim}$ indicates a threshold concentration above which the species i is considered to be distributed in a significant amount. The relative area was evaluated for both lactate and KB-1, as the effective distribution of these two amendments is necessary for the degradation of the chlorinated ethenes in the system. In this work, $c_{Lac,lim}$ was set to 2.16 mM to account for the amount of lactate necessary to convert 1 mM of PCE to ethene and $c_{KB-1,lim}$ was set to 1×10^8 cells/L.

The second metric is the relative mass (RM) of the contaminants in the domain and quantifies the effectiveness of EK-enhanced biodegradation by accounting for the relative mass of a species i with respect to the total mass of a reference species j (Sprocati et al., 2020):

$$RM_i(t) = \frac{1}{M_{Tot,j}} \int_{\Omega} c_i(\boldsymbol{x}, t)\, d\Omega \tag{11}$$

In Eq. (11), $M_{Tot,j}$ is the total initial mass of a reference species j, which in this study is the total initial mass of PCE in the domain, both as NAPL and as dissolved species. We evaluated the relative mass for all chlorinated parent compounds and metabolic products, including PCE NAPL, PCE, TCE, DCE, VC and ethene.

2.2. Surrogate modeling approach

The proposed surrogate modeling approach was implemented considering a data-driven model, also referred as an approximation function, that used as inputs for the training the outputs of multiple runs of the process-based model (Razavi et al., 2012). With this method, also defined as a response surface approach, the modeling framework begins with the definition of the explanatory (input) variables of the approximation function (Figure 2).

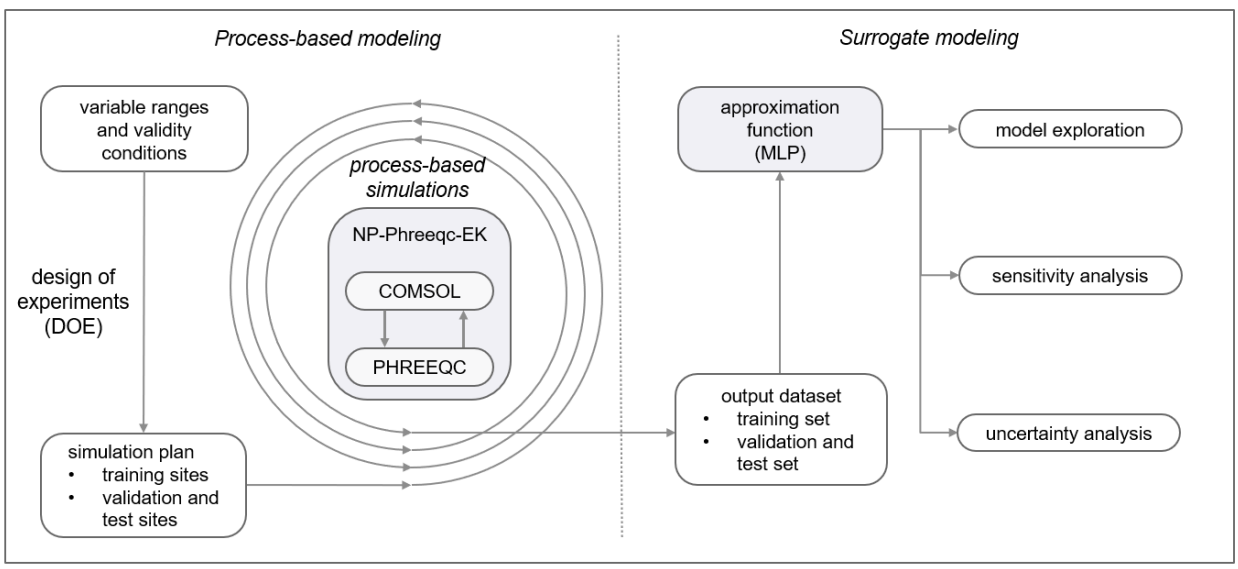

Figure 2. Schematic overview of the proposed modeling framework including the process-based reactive transport model and the data-driven surrogate model.

The explanatory variables are selected based on their relevance for the problem and their number should be as low as necessary to limit the model runs of the process-based simulations. To define realistic values of the variables for the problem, variable ranges and validity conditions are then defined for each variable. Subsequently, combinations of explanatory variables are sampled with a design of experiments (DOE) procedure to obtain a simulation plan, consisting of several design sites. Each design site consists of a unique combination of the explanatory variables and the

number of sites determines the number of process-based simulations executed. During the DOE, the design sites are also divided into training, validation, and test sites, which are then used during the development of the approximation function. The multiphysics simulations are executed with NP-Phreeqc-EK, which performs process-based simulations for each design site. Once all the process-based simulations are evaluated and the results collected in an output dataset, the approximation function is trained using the training set and cross-validation is implemented using the validation set to avoid overfitting and increasing generalization performances. After training and testing the model performances with the test set, the approximation function is used as a surrogate of the process-based model to explore relations between inputs and outputs, to perform global sensitivity analysis and to assess output uncertainty based on uncertainties on input variables.

The selection of explanatory variables considered physical parameters dependent on the actual field condition such as tortuosity (τ), electroosmotic coefficient (k_{eo}) and mass-transfer coefficient (ω_{PCE}), as well as operational parameters which can be controlled during the implementation of the in situ remediation technology. The latter include the electric potential applied at the anode (V_{an}) and the time of application of electric potential (t_{EK}). In addition, we also accounted for the differences in reaction rates of dehalogenation reactions with respect to the values obtained from laboratory tests (η).

Explanatory variables ranges have been selected based on literature values. The six explanatory variables and the ranges used for the surrogate model are listed in Table 1. The same table also includes values of the variables which are constant for all simulations and the values of the variables which have been used to perform a base case scenario. The values of all the other kinetic parameters are provided in Table S2.

Table 1. Explanatory variables and parameters used as input for the surrogate model.

Ranges of explanatory variables					
Variable name	**Symbol**	**Units**	**Min**	**Max**	**Base case**
Tortuosity	τ	-	0.20	0.75	0.60
Electroosmotic coeff.	k_{eo}	$m^2/(V\ s)$	5.00×10^{-10}	1.00×10^{-8}	2.00×10^{-9}
Reaction rate factor	η	-	0.30	2	1.0
Log of NAPL mass transfer coeff.	$\log_{10}(\omega_{PCE})$	1/s	-7	-5.3	-6.3
Electric potential at the anode	V_{an}	V	60	360	110
Time of EK active phase	t_{EK}	days	60	180	120

Fixed parameters for all simulations		
Variable name	**Units**	**Values**
Temperature	°C	15
Porosity	-	0.5
Distance between electrodes	m	3
Electric potential at the cathode	V	0
Initial PCE in the domain (uniform)	mM	1.24
PCE NAPL in the domain (uniform)	mM	2
Conc. Lactate injected	mM	18
Conc. Bacteria injected	mM	2.60×10^{-3}
Initial concentration (HCO_3^-)	mM	10
Initial concentration (Ca^{2+})	mM	5

Regarding the validity conditions for the surrogate model, we considered electromigration of lactate from the cathode to the anode as the main electrokinetic transport process; thus, the electromigration velocity for lactate is always larger than the electroosmotic velocity:

$$v_{Mig} = \tau D_i^{aq} \frac{z_i F}{RT} \nabla\Phi > v_{Eo} = k_{eo} \nabla\Phi / n \tag{12}$$

After definition of ranges and validity conditions for the explanatory variables, a design of experiments is performed with the optimal Latin hypercube sampling algorithm (Jin et al., 2005; Morris & Mitchell, 1995; Park, 1994). Such method has been selected as it provides uniform spread of design points across the design space (Forrester et al., 2008) and for the property of generating points that are never projected into each other when projected from a space of dimension k to a $k-1$ space (Crombecq et al., 2011; Van Dam et al., 2007; Morris & Mitchell, 1995). In this study, the approximation function is trained on a simulation plan $\boldsymbol{X_{TR}} = \left[\boldsymbol{x}_{TR,1}\ \boldsymbol{x}_{TR,2} \dots \boldsymbol{x}_{TR,n}\right]^T$ of n training points (design sites), in which $\boldsymbol{x}_{TR,i} = \left[x_i^{(1)}\ x_i^{(2)} \dots x_i^{(k)}\right]$ represents the vector of explanatory variables used for the ith process-based simulation on a k-dimensional input space. To calculate the number of n training points for the surrogate model, some authors suggest to use at least $n = 10k$ points (Alwosheel et al., 2018; Loeppky et al., 2009). Given k = 6 in our case, to include a margin in the design of training points we decided to select n = 100. After considering the physical validity criterion about electrokinetic velocities in Eq. (12), only n = 87 design sites were retained for the training procedure. Indeed, the training set for the approximation function contained a number of observations equal to $n_{TR} = n \times N_{timesteps}$, where $N_{timesteps}$ is the number of recorded time steps generated in every design site by the process-based model (93 in this study). Therefore, the final training dataset had 8091 observations and 7 input variables when including also the time.

In addition to the training set, we also evaluated additional sets for validation ($\boldsymbol{X_{VA}}$), included to avoid overfitting during training of the approximation function (Eason & Cremaschi, 2014), and a test set ($\boldsymbol{X_{TE}}$), used to compare the performance of the surrogate model on a set of process-based simulations that were not used for training or validation. Both the validation and test sets have been sampled using a random combination of explanatory variables inside the validity ranges and each accounted for approximately the 20% of the total number of training points. In this study, the validation set included 14 design sites and the test set 15 design sites.

Data-driven surrogate models have been typically made using different approximation functions for regression problems including polynomial, Kriging, radial basis function, support vector machine, Gaussian process, random trees and random forests (Asher et al., 2015; Razavi et al., 2012; Tahmasebi et al., 2020). In recent years, also Artificial Neural Networks (ANN) of different sizes and complexities have been increasingly used in the environmental and water resources field (Mo et al., 2019; Prasianakis et al., 2020; Taormina et al., 2012). Their widespread use has been facilitated by the development of advanced computational algorithms (Haghighat & Juanes, 2021; Pedregosa et al., 2011) allowing improvements in the design and the control during training ANN are also referred as deep neural networks (DNN) with architectures ranging from the simpler multi-layer perceptron (MLP), to convolutional neural networks (CNN) and recurrent neural networks (RNN).

Here, we chose as approximation function a neural network using a stack of MLP, which are commonly applied architectures for regression analysis in machine learning problems (Géron, 2019; Maier et al., 2010) and have already been used to develop surrogate models for transport problems in porous media (Behzadian et al., 2009; Hou et al., 2017; Johnson & Rogers, 2000; Razavi et al., 2012; Yan & Minsker, 2006, 2011). The modeling of integral quantities can also be performed with other approximation functions (Razavi et al., 2012) but here we restricted the scope of the investigation to artificial neural networks. To obtain a simple network structure, we used MLPs with dense layers, in which all neurons in a layer are connected to every neuron in the previous layer. For every dense layer, the outputs are calculated as:

$$h_{\boldsymbol{W},\boldsymbol{b}}(\boldsymbol{X}) = \phi(\boldsymbol{X}\boldsymbol{W} + \boldsymbol{b}) \tag{13}$$

where $h_{\boldsymbol{W},\boldsymbol{b}}$ is the output value, $\boldsymbol{W}$ is the weight matrix containing all the connection weights except the ones from the bias neuron, $\boldsymbol{b}$ is the bias vector and contains all the connection weights between the bias neuron and the artificial neurons and ϕ is the activation function, which in this study is the Rectified Linear Unit function, $ReLU(z) = max(0, z)$.

Prior to training, all inputs and outputs have been normalized with respect to their minimum and maximum values in the training set, so that their values range from 0 to 1. To simplify the network structure, all hidden layers have been set to have the same number of neurons. The MLP is then trained with a stochastic gradient descent optimizer (Chollet, 2015) with momentum 0.9, learning rate 0.46 and mean squared error (MSE) as loss function. Early stopping with 10 steps is performed to prevent overfitting (Caruana et al., 2001). For the development of the ANN we used TensorFlow with a Keras interface (Chollet, 2015), using a GPU with CUDA to improve model performance (Sanders & Kandrot, 2010).

A randomized search algorithm using cross-validation (Bergstra & Bengio, 2012; Pedregosa et al., 2011) has been used to select optimal hyperparameters such as (i) the number of hidden layers, (ii) the number of neurons per hidden layer and (iii) the learning rate. The randomized approach consisted in generating random combinations of the hyperparameters, training the model with backpropagation and evaluating performances against the validation set. The configuration of parameters which provide the lowest MSE on the validation set is then used as architecture of the neural network. For this work, the final artificial neural network was composed of 7 hidden layers of 485 neurons each, with an input dimension of 7 (explanatory variables and time t) and an output with size of 8: 1) Relative Area of Lactate, 2) Relative Area of KB-1, 3) Relative Mass of PCE NAPL, 4) Relative Mass of PCE, 5) Relative Mass of TCE, 6) Relative Mass of DCE, 7) Relative Mass of VC and 8) Relative Mass of ethene. The randomized cross-validation required the training of 50 different models (10 different parameter combinations and 5-fold stratified cross validation) with a maximum number of epochs for each model set to 300. The selected model had a final MSE of the averages of all the scaled outputs of 4.95×10^{-5} for training, 1.686×10^{-3} for the validation and 1.516×10^{-3} for the test sets with. The total training time was 64 minutes to train all 50 models.

2.3. Surrogate model analysis

The proposed surrogate framework allowed us to perform detailed analysis of the dynamics and performances of the EK-Bio remediation systems. Specifically, we performed extensive model exploration, sensitivity analysis and uncertainty quantification.

We started exploring the surrogate model by using a single design site and showing the comparison with a process-based simulation which can be considered a first base case scenario. Subsequently, we used the surrogate model to provide partial dependence plots in the form of surface plots, as they allow to visualize values of the output variable on a continuous surface representing a two-dimensional input space. For such task, a grid of 100×100 points was generated considering two inputs quantities, τ and k_{eo}, assuming all the other input variables as in the base case scenario. These two input quantities were selected for analysis as they represent the main unknowns regarding the transport during in situ electrokinetic remediation and should be experimentally determined for every field site. In this work we also performed Monte Carlo simulations considering random distributions of inputs to evaluate a correlation matrix between inputs and outputs. The aim was to obtain a simple overview on the relations between the variables involved in the defined problem, across all parameter's ranges. The correlation matrix was evaluated with the Pearson r correlation, which indicates the degree of linearity between two variables. For this task, we performed 1×10^5 evaluations of the surrogate model with random combinations of inputs at two specific times.

Global sensitivity analysis methods allow the evaluation of changes in the output of a model from different sources of uncertainty from model inputs (Iooss & Lemaître, 2015; Saltelli et al., 2000). In this work we used the Sobol' indices method (Formaggia et al., 2013; Saltelli et al., 2010; Sobol', 2001), sampling the combination of the input variables using the method described by Saltelli et al. (2010). The ranges for the sampling was the same range used during the design of experiment whereas to comply with the validity condition we assigned a value of zero to all output values which did not satisfy the validity conditions. For every output variable, we evaluated separate indices (Iooss & Lemaître, 2015; Saltelli et al., 2010): (i) the first order Sobol' index S_i, accounting for the contribution of a single parameter i to the output variance, (ii) the second-order Sobol index S_{ij} to quantify the output variance explained by the interactions between the input variables i and j, and (iii) the total Sobol index S_{T_i} accounting for the effect on the output variance of the first, second and higher-order effects. The evaluation of S_i, S_{ij} and S_{T_i} required $N(2k+2)$ model evaluations, where N is the initial size of Monte Carlo sampling (set to 10^4) and k is the number of input variables (6, as the time was fixed).

Finally, we used the surrogate model to quantify uncertainty, considered here as the combined effect of probabilistic distributions of independent input values on the output metrics over the time of the simulation. To evaluate the uncertainties, we performed Monte Carlo simulations randomly selecting combinations of input parameters sampled from normal distributions with mean and standard deviation reported in Table S4. Here, we considered 100 time steps from 0 to 360 days and for each instant we calculated 1×10^4 input combinations. Subsequently, for each time step we selected the 10th, 50th and 90th percentiles of each output variable. With the evaluation of the percentiles, it was possible to define the lower and upper confidence intervals (respectively the 10th and 90th percentiles) and the median value (50th percentile) for each output variable.

3. Results and discussion

3.1. Process-based simulations and surrogate model performance

In the considered EK-Bio setup, upon application of an electric potential between the electrodes at $t = 0$, lactate is delivered in the domain from the cathode. The transport of such electron donor is the result of the larger electromigration velocity of lactate that is higher than the electroosmotic velocity, which occurs in the opposite direction. In fact, electroosmosis results in a flow of water from the anode to the cathode and it is the main transport mechanism of the non-charged species, such as the chlorinated compounds and the KB-1 degraders. When lactate is delivered in the system it is consumed by the immobile indigenous OHRB and the mobile, bioaugmented OHRB (KB-1). These microorganisms use lactate as electron donor to transform the parent compound PCE into the degradation products TCE and DCE. The subsequent degradation steps of DCE to VC and ethene are performed only by the KB-1, provided their effective distribution and contact with lactate and the chlorinated ethenes.

Figure 3a-l shows the distribution of the electrokinetically delivered amendments, as well as the main contaminant (PCE) present in the system and the final metabolic product (ethene) of reductive dehalogenation. The plots show results at different times considering the parameters of the base case scenario reported in Table 1 with an active EK phase (i.e., voltage applied at the electrodes) during the first 120 days. Specifically, Figure 3a,e,i illustrates the spatial distribution of lactate after 60, 120 and 360 days. The concentration of lactate is lower in the right part of the domain due to its consumption as electron donor during reductive dehalogenation carried out by both indigenous and bioaugmented microorganisms. Moreover, the concentration of lactate in the domain (around 10 mM) is lower than the injection concentration (18 mM) as a result of charge interactions, which limit the maximum concentration of charged reactants that can be transported in the domain by electromigration (Sprocati et al., 2021; Sprocati & Rolle, 2020).

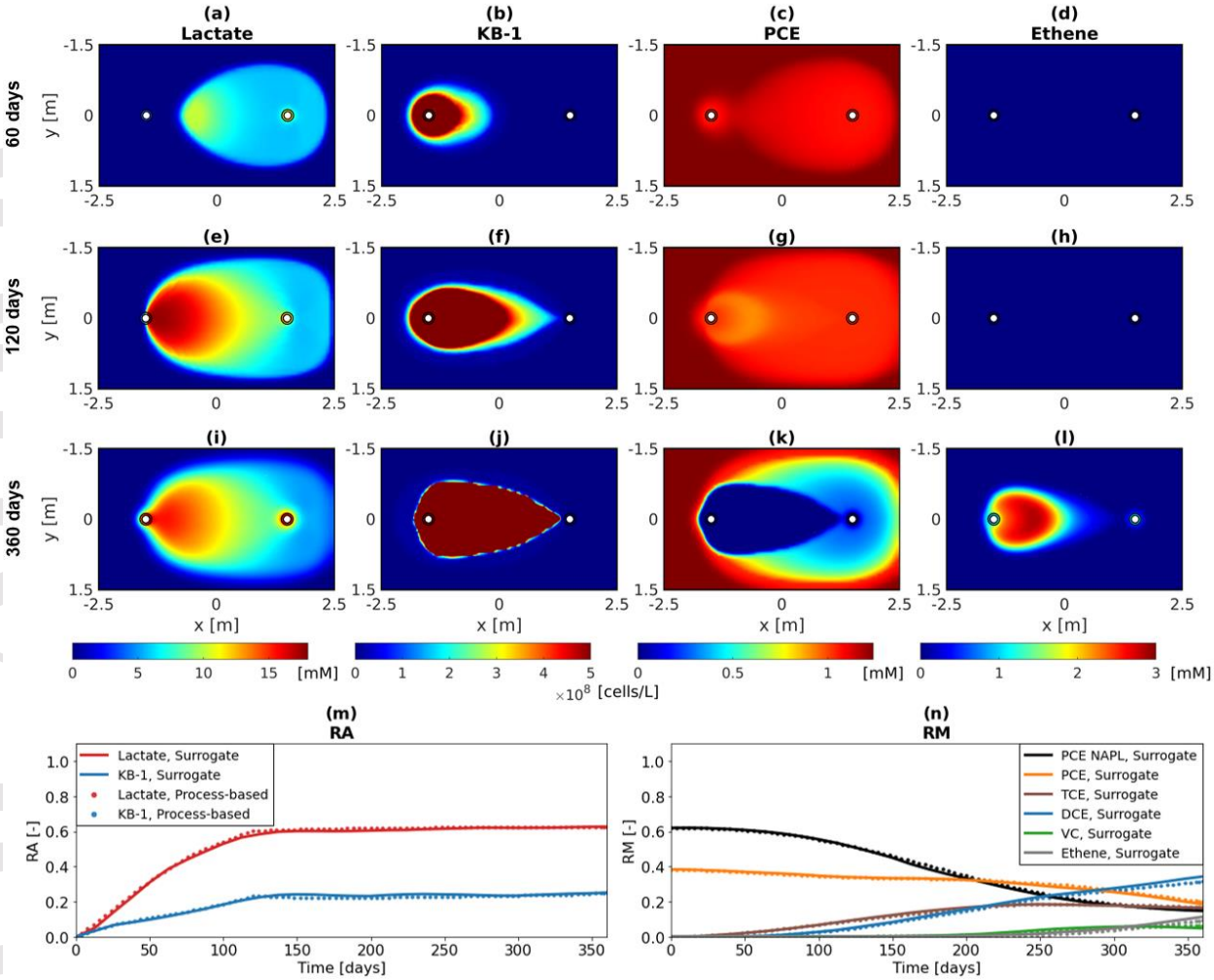

Figure 3. Process-based simulation of EK-Bio for the base case scenario: **(a-i)** spatial distribution of selected species at 60, 120 and 360 days. Comparison of process-based and surrogate models: **(m)** relative area and **(n)** relative mass.

Figure 3b,f,j shows the spatial distribution of the KB-1 degraders in the domain after 60, 120 and 360 days, resulting from their transport by electroosmosis from the anode. The distribution of KB-1 is less effective than the distribution of lactate due to the higher electromigration velocity of the charged electron donor with respect to the electroosmotic transport of the bioaugmented OHRB. PCE, initially uniformly present in the domain (Figure 3c), is consumed over time in the area between the electrodes, where both lactate and the degraders are present (Figure 3g,k). However, close to the edges of the domain, PCE is still present at the initial concentration due to the absence of mixing between electron donor and acceptors. Figure 3d,h indicate that ethene was not produced in the domain until late times due to the limited distribution of the OHRB, which are critical for complete degradation, and to the slower kinetics of the last dehalogenation steps.

The temporal evolution of the system in terms of amendment distribution (RA) and degradation efficiency (RM) is displayed in Figure 3m,n. Considering the entire domain, Figure 3m shows the relative area in which the amendments have been delivered and indicates that both lactate and KB-1 are increasingly distributed in the domain over the initial EK phase of 120 days. Subsequently,

their distribution stabilizes due to the much smaller transport velocity by natural diffusion compared to the induced processes of electromigration and electroosmosis. The more effective distribution of lactate is indeed dependent on the higher transport velocity of this species due to electromigration with respect to electroosmosis, resulting in almost three times RA_{lac} compared to RA_{KB-1}. The relative mass of the species in the system is shown in Figure 3n for all ethenes, showing that PCE initially present as NAPL decreases over time, due to its dissolution to PCE in the pore water and its subsequent biodegradation. At the same time, byproducts such as TCE and DCE increase as a consequence of biodegradation. Conversely, due to the limited distribution of the KB-1, conversion to ethene is less effective and can be appreciated only at late times. After the process-based model was run with the defined input dataset from the DOE procedure, the approximation function (MLP) and its hyperparameters were trained with a cross-validation procedure using both the training and the validation datasets. Surrogate model predictions were then compared with the results from a single run of the process-based model, using a combination of input parameters (Table 1) which were not part of the training or validation datasets. Figure 3m-n shows that the predictions of the surrogate model are in very good agreement with the outcomes of the process-based model.

Figure 4 shows the scatter plot comparing process-based and surrogate model outputs for all output variables differentiating between training, validation, and test data points.

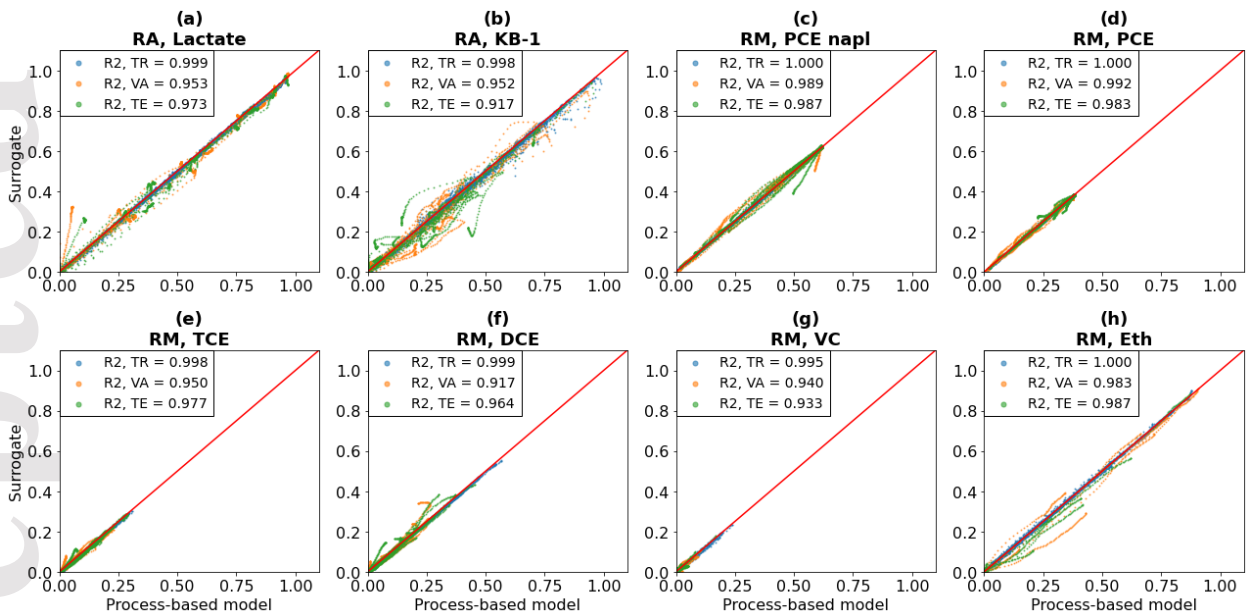

Figure 4. Scatter plots comparing outputs of the process-based model (x-axis) and predictions of the response surface surrogate (y-axes) for training (TR), validation (VA) and test (TE) sets. The red solid line represents the 1:1 line, indicating perfect prediction performances of the surrogate model.

It is important to note that MLP are not exact predictors, thus model outputs do not pass exactly through the training points (Razavi et al., 2012). Nevertheless, the coefficient of determination R^2 of the scatter plots for the training set for all cases is larger than 0.99. The lower scores of R^2 in the test sets are above 0.96 except for RA_{KB-1} (R^2 =0.92) and RM_{VC} (R^2 =0.93), indeed indicating

excellent prediction performances, limited overfitting and ability to generalize well to new combinations of input variables. Higher accuracy can be achieved either by improving the approximation function (changing different combinations of hyperparameters or changing the type of approximation function) or by providing more design sites to be used for training. The surrogate model is also able to maintain the underlying relations between the output variables such as the sum of the relative masses of the different species, which by definition in Eq. 11 must be equal to one. Such physical constraint, although not enforced in the surrogate framework, is predicted with sufficient consistency as shown in Figure S1 in Supporting Information.

3.2. Model exploration

Surrogate models can be used to quickly explore the behavior of the system over a wide range of input combinations. Here we performed model exploration evaluating partial dependence plots of the output variables in the $\tau - k_{eo}$ space. These two input variables have been considered since the tortuosity (τ) directly controls the electromigration velocity of charged species (lactate), and the electroosmotic coefficient (k_{eo}) determines the electroosmotic velocity, which is the main transport mechanism of the KB-1 degraders and also impacts the transport of all the other species including lactate and the non-charged chlorinated ethenes. The partial dependence plots evaluated with the surrogate model are illustrated in Figure 5 at t = 360 days.

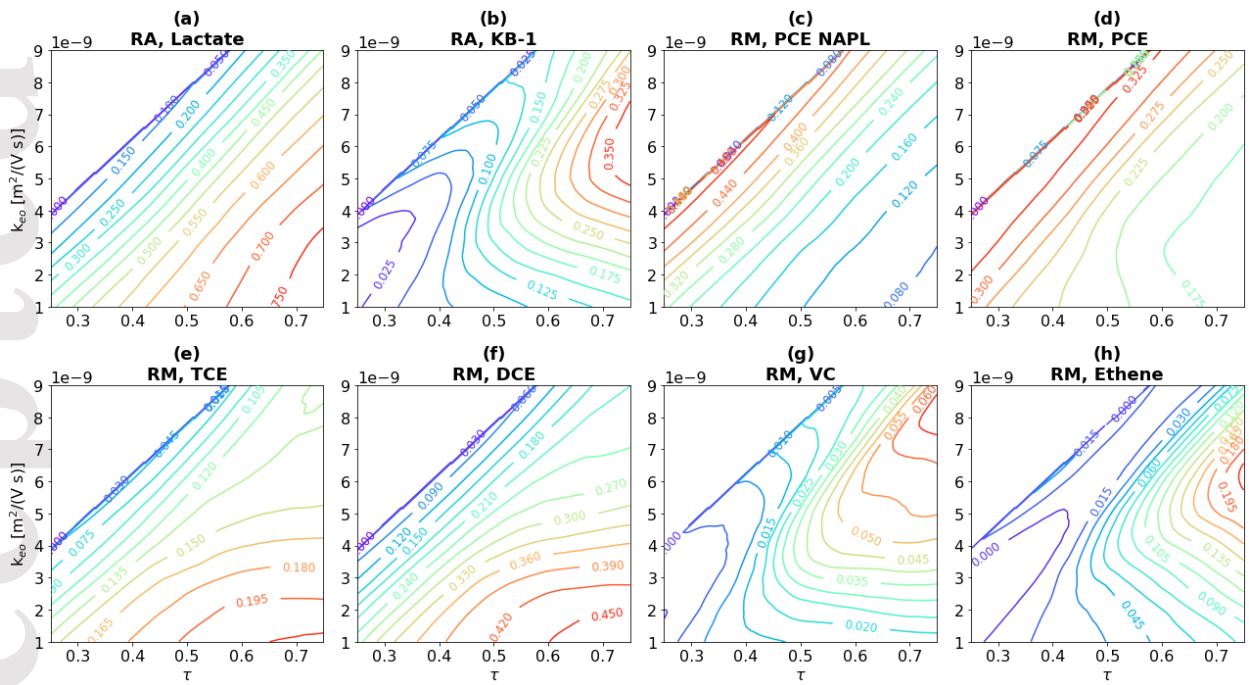

Figure 5. Partial dependence surface plots used for model exploration of the output variables depending on tortuosity τ (x-axes) and electroosmotic coefficient k_{eo} (y-axes). Cold colors represent low values of the computed efficiency metric (i.e., relative area and relative mass), warm colors indicate higher values. The model is evaluated after 360 days from the beginning of EK.

Figure 5a shows that RA_{lac} increases along a pattern perpendicular to the isolines in the plane $\tau - k_{eo}$ to reach the highest value in the bottom-right corner (high τ and low k_{eo}), as a result of the

net migration velocity of lactate which is directly dependent on the tortuosity of the porous medium and inversely dependent on the electroosmotic velocity. Therefore, it is possible to estimate the average slope of the lines by considering the ratio k_{eo}/τ in Eq. (12). In practical terms, moving along the isolines corresponds to consider $\tau - k_{eo}$ points resulting in the same net velocity of lactate. Figure 5b illustrates that for KB-1, values of RA_{KB-1} tend to increase differently with respect to the increase of RA_{lac}. The relative area of distribution of the degraders (RA_{KB-1}) increases with the k_{eo} since the bioaugmented biomass is transported with the electroosmotic flow. However, as a certain amount of lactate must be present in the domain to sustain the growth of KB-1 that otherwise would decay according to Eq. (8), the trend of RA_{KB-1} becomes oblique due to the dependency by RA_{lac}.

With the insights from the surface plots of RA_{lac} and RA_{KB-1}, it is possible to gain insight on the distribution and efficiency of in situ biodegradation of the chlorinated ethenes. In particular, Figure 5c-f shows that the shape of $RM_{PCE,NAPL}$, RM_{PCE}, RM_{TCE} and RM_{DCE} follows a trend similar to RA_{lac}, as lactate controls the biodegradation for such contaminants. However, the trend for RM_{VC} and RM_{Eth} is more similar to the one of RA_{KB-1}, since the delivery of KB-1 controls the last steps of reductive dehalogenation.

The trends displayed in Figure 5 indicate that poor performances are expected for low values of τ and k_{eo} and higher performances are anticipated in the opposite situation. Note that such considerations are also dependent on the time at which the plots are evaluated and on the other input quantities. The sum of the relative masses of the different species is shown on the $\mathrm{k_{eo}} - \tau$ space in Figure S2 in Supporting Information. Values of the sum of the relative masses very close to 1 indicate that the surrogate model is able to account for the conditions indicated in Eq. 11 despite the absence of an explicit constraint and support good prediction performances of the surrogate.

Further model exploration was performed with Monte Carlo simulations that allow the evaluation of the correlation matrix between input and output variables. The results of a set of 10^5 simulations are shown Figure 6 for two correlation matrices at different times (120 and 360 days), generated by sampling randomly the distribution of the input parameters.

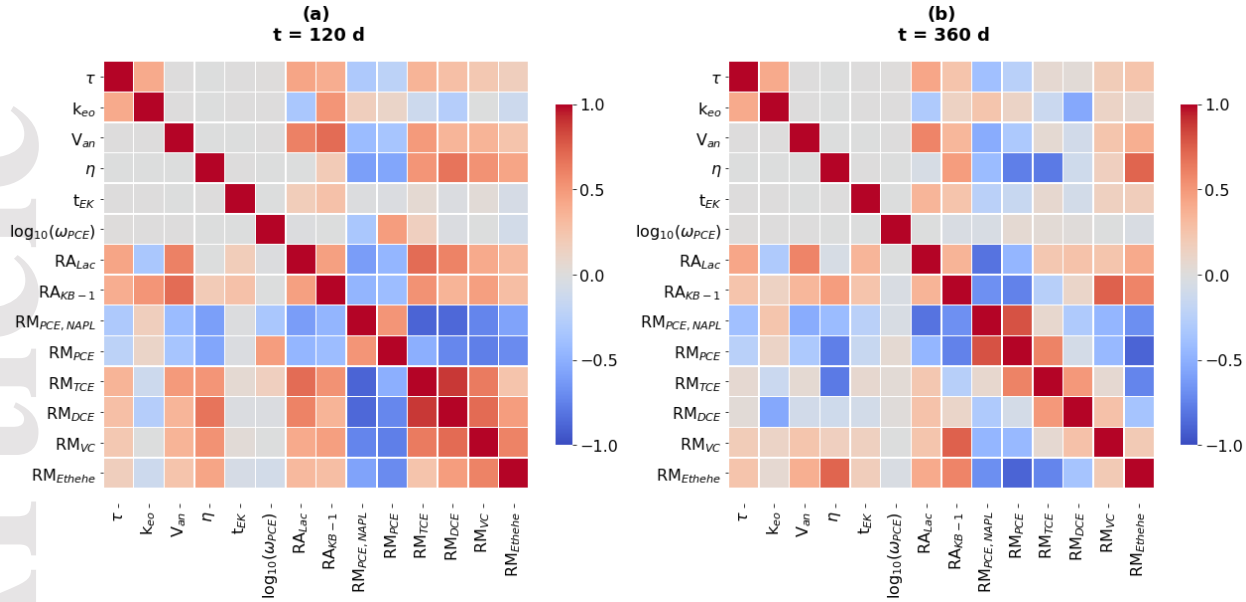

Figure 6. Correlation matrices between model input and output variables at (**a**) 120 days and (**b**) 360 days from the start of the simulation.

Figure 6 indicates that there is no correlation between the generated random input variables except for $\tau - k_{eo}$ as a result of the condition imposed by Eq. (12). For both t = 120 d and t = 360 d, an increase in τ results in an increase in RA_{lac} and RA_{KB-1}, with the simultaneous decrease of PCE, and increase of degradation products. Considering the effects of k_{eo}, an increase of such variable causes the RA_{lac} to decrease, whereas it contributes to increase transport of the KB-1 degraders. The implications of the electroosmotic flow on the chlorinated ethenes are not trivial, as such flow appears to support the release of PCE, with a consequent delay in the effects of the degradation reactions. An increase in the electric potential gradient accounted for by V_{an} appears to provide a more effective distribution of the amendments and consequently a higher rate of contaminant biodegradation, with similar effects also observed for the duration of the active electrokinetic phase, t_{EK}. Faster reaction rates, represented by higher η values, result in enhanced degradation of the chlorinated ethenes with positive correlations also for the bacteria population in the domain. Finally, an increase in the mass transfer coefficient of PCE, $log_{10}(\omega_{PCE})$, results in a decrease of PCE NAPL and a consequent increase of PCE in the aqueous phase. Comparing such effects at 120 days and 360 days, the effects of mass-transfer limitations appear to affect the system only at early times when dissolution from the NAPL to the aqueous phase is active. Overall, the correlation matrices provide an overview of the relations between inputs and outputs and can assist in identifying and understanding simple dependencies of complex systems with a great effectiveness for linear trends.

3.3. Sensitivity analysis

The global sensitivity analysis using the Sobol method was performed for all output metrics after 360 days from the start of the simulation. Figure 7 shows the results of the sensitivity analysis by displaying the Sobol first and second-order and total indices for each output variable.

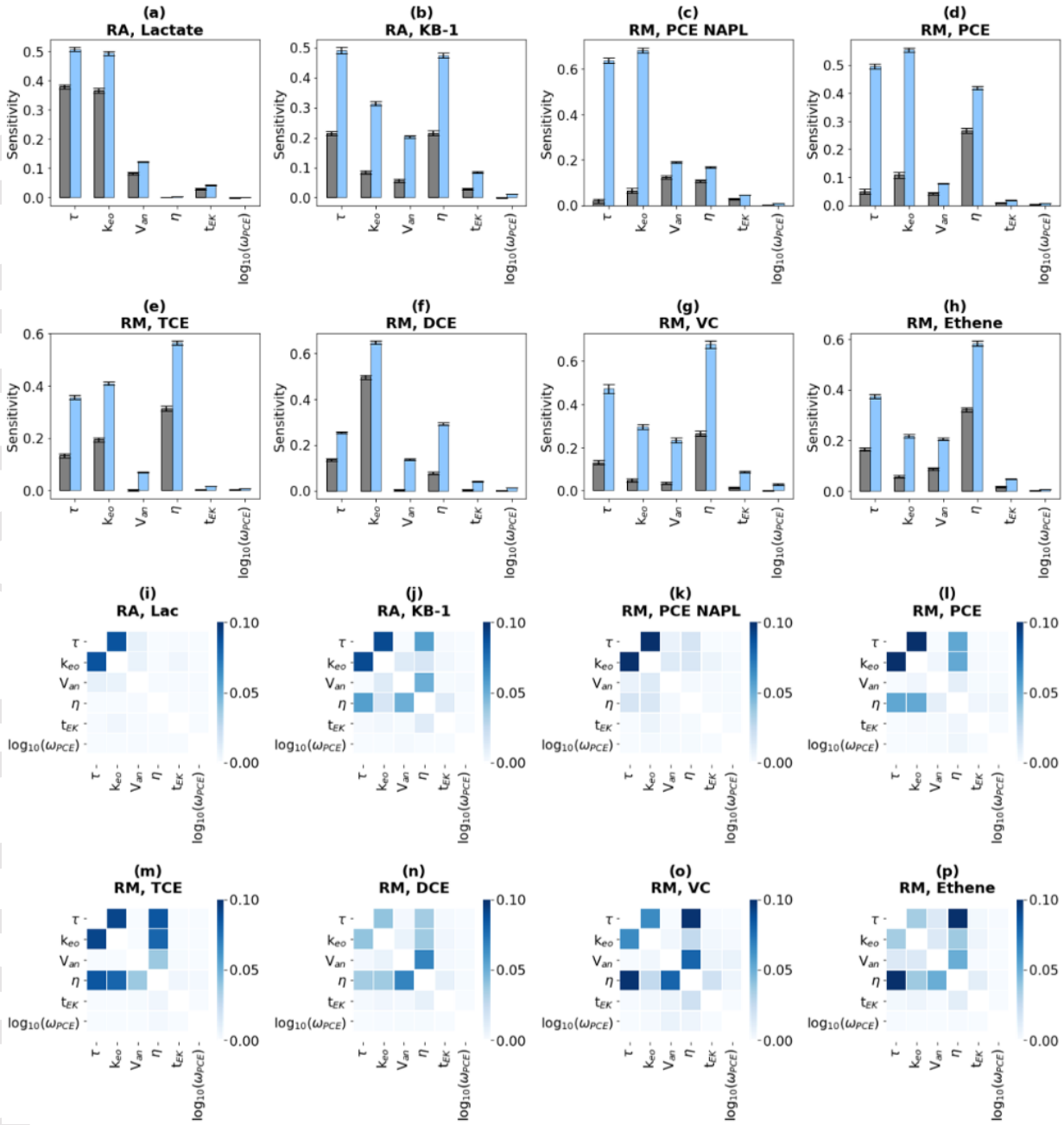

Figure 7. Results of the sensitivity analysis for each individual output variable. Sobol first-order and total sensitivity indices indicating parameter importance **(a-h)**. Correlation matrix of the Sobol second-order indices of indicating mutual interactions **(j-p)**.

Figure 7a shows that both first order and total indices provide similar results and indicates that the most sensitive variables for lactate distribution, RA_{Lac}, are the tortuosity (τ) and the electroosmotic coefficient (k_{eo}), followed by the applied voltage (V_{an}) and the time of active application of the electric field, t_{EK}. Such results indicate that the transport of lactate in the system is more influenced by the specific properties of the porous medium rather than by the reactions occurring in the

system. Indeed, this consideration applies in the present case as lactate is delivered in excess with respect to the amount necessary to perform reductive dehalogenation.

Regarding the distribution of KB-1, Figure 7b highlights that there are significant differences between the first and total indices and therefore strong nonlinearities control the behavior of RA_{KB-1}. Here, τ and the biodegradation rate (η) appear to have the largest importance, followed by k_{eo} and V_{an}. In this case, the higher importance of η suggests that the rate at which reactions occur influences the extent and concentration of KB-1 in the system.

The relative mass of PCE as NAPL after 360 days (Figure 7c) shows high total sensitivity indices for τ and k_{eo} which indicates that multiple phenomena control the mass of PCE as NAPL in the system. Indeed, after 360 days the mass of PCE NAPL does not seem to strongly depend on the mass-transfer coefficient $log_{10}(\omega_{PCE})$ which has the lowest sensitivity since, as indicated also by the correlation matrices discussed above, interphase mass transfer does not impact significantly the system at late times.

The dissolved chlorinated ethenes PCE (Figure 7d) and TCE (Figure 7e) have similar trends with the mass in the system mainly dependent on τ, k_{eo} and η. The sensitivity of DCE (Figure 7f) shows a different tendency, with a higher dependence on k_{eo} and to a lesser extent on τ and η. VC (Figure 7g) and ethene (Figure 7h) show similar trends, with high dependence on η and slightly lower on τ, k_{eo} and V_{an}.

Effects from high-order variance can be identified considering the second-order indices in Figure 7i-p, which highlight the relations between different input variables to the outputs. In particular, for all output variables both τ and k_{eo} appear to be strongly related, and this is particularly noticeable for RA_{Lac} (Figure 7i). However, considering RA_{KB-1}, it appears that η contributes with non-linear interactions with τ and V_{an}. Similar considerations can also be derived for the chlorinated ethenes and products in Figure 7k-p, which highlight a stronger second-order dependence of η with τ, k_{eo} and V_{an}.

The relations presented in Figure 7 provide an overview of the global sensitivity of the considered explanatory variables at 360 days. In the perspective of full-scale EK remediation applications, sensitivity analysis can help highlighting which operating parameters control the variables of interest and which properties of the system should be more accurately analyzed to improve model predictions. In this case, the highest sensitivity was observed for τ and k_{eo}, suggesting that such parameters should be better characterized than the mass-transfer coefficient $log_{10}(\omega_{PCE})$, which demonstrated weak sensitivities for all the output variables.

3.4. Uncertainty quantification

The uncertainty analysis was performed considering the frequency distributions of the input values and analyzing the impact of such distributions on the output metrics quantifying the capability of the EK-Bio system to effectively distribute the amendments in the low-permeability formation and to promote in situ biodegradation of the chlorinated compounds. Figure 8a-f shows the frequency distributions of 1×10^4 parameter combinations of the six input variables which have been sampled from Gaussian distributions with ranges reflecting the variability of physical values and operational conditions.

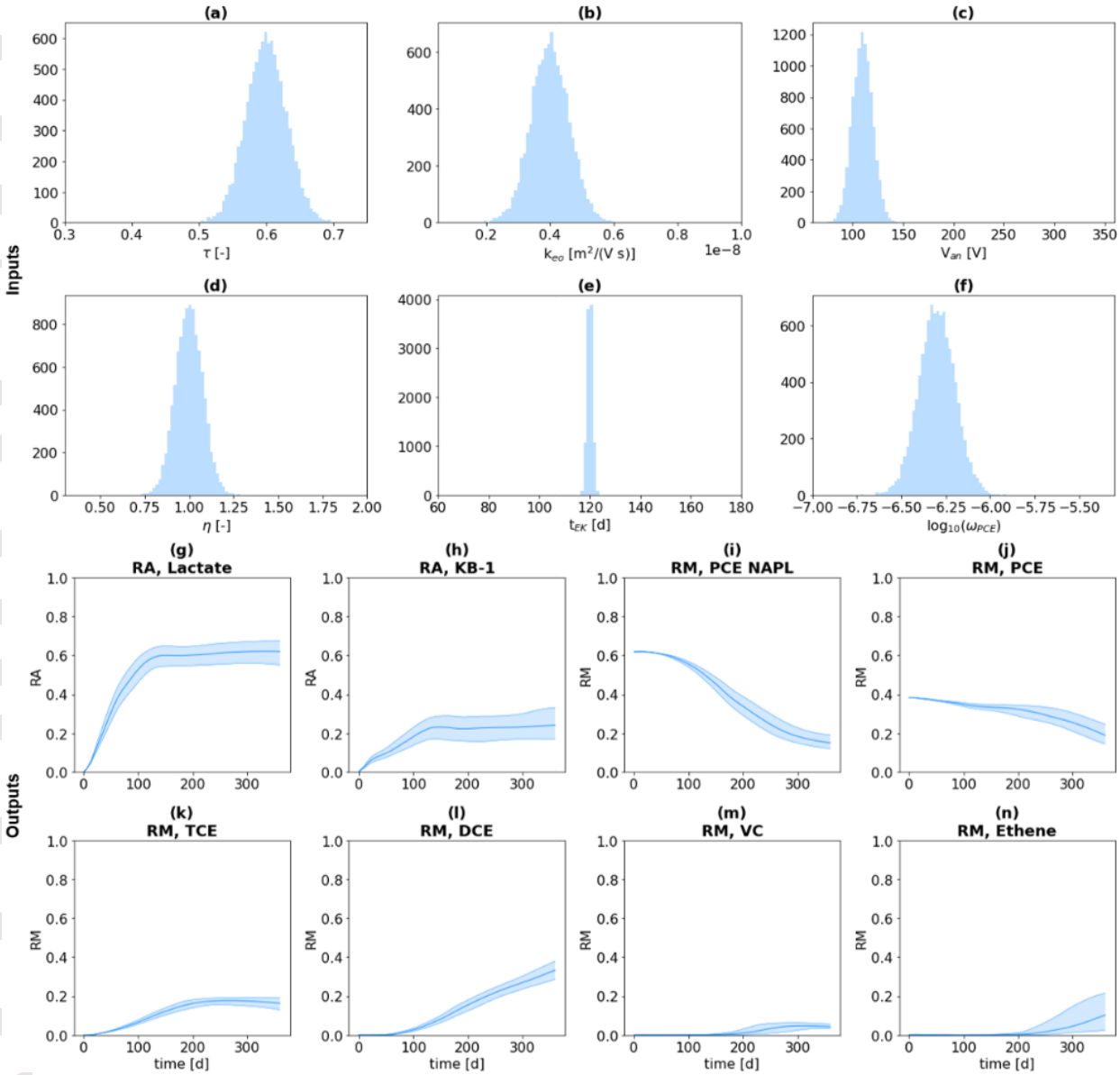

Figure 8. Results of the uncertainty analysis. Frequency distribution of the input variables subdivided in 100 equally spaced bins defined for each variable range **(a-f)**. Temporal evolution

of the output variables with median (solid line) and uncertainty intervals (10th and 90th percentiles) **(g-n)**.

Figure 8g-l shows the uncertainty ranges of the output variables over time, obtained with 10^6 surrogate model evaluations (10^4 parameter combinations × 100 time steps). The solid line indicates the median of all output simulations at each considered time step, whereas the shaded area represents the zone between the 10th and 90th percentile. The plots indicate that uncertainties in the predictions at the beginning of the simulation are small and increase with time, due to the increasing effect that different input variables have on the system over time. With the selected frequency distributions, Figure 8g shows that there is the potential of an adequate delivery of lactate by electromigration, with a relative area ranging from 0.55 to 0.70 from the time in which the electric field has been turned off. As discussed above, the delivery of KB-1 by electroosmosis is less efficient and the probability distribution of the relative area ranges from 0.15 to 0.30.

Considering the contaminant mass removal, PCE in NAPL phase decreased significantly with a relative mass that went from 0.60 at t = 0 to less than 0.20 at t = 360 d (median values in Figure 8c) and its dissolution to the aqueous phase (Figure 8d) resulted in high levels of PCE in the system until approximately 200 days from the start of the simulation, after which the relative mass of dissolved PCE decreased from approximately 0.40 to 0.20. During the same time, the degradation product TCE increased until 250 days reaching a relative mass of 0.20 (Figure 8e), before its trend shifted toward the dominance of its consumption and production of DCE. Such degradation product had the highest relative mass among all chlorinated ethenes, with values exceeding 0.35 at the end of the simulation (Figure 8l). Figure 8m-n indicate that the relative mass of VC and ethene in the domain is not likely to be large in the considered timeframe, with values respectively in the range 0-0.05 and 0.03-0.20. The final product ethene in this case starts to be present in the system after approximately 200 days and its concentration increases until 360 days.

With the considered input parameters, the biodegradation of chlorinated ethenes does not appear to be complete after 360 days and it is expected to continue in time, provided availability of substrate and specialized OHRB. The simulation uncertainties indicate that the confidence intervals tend to increase for all output variables with time since the combined uncertainty of input parameters propagates and becomes more significant. Note that the uncertainty analysis illustrated in Figure 8 presents only one of the possible scenarios of EK-Bio systems. The simulations performed with the surrogate model could be indeed quickly updated once new information becomes available. For instance, if additional investigations reveal changes in the probability distributions of the inputs, the surrogate model will allow to rapidly re-evaluate the uncertainties of the output metrics.

The computational gains provided by the surrogate is significant and, depending on the number of simulations required, it is possible to evaluate uncertainties also in real-time. For instance, it took 222 seconds to perform the 10^6 surrogate model evaluations (10^4 parameter combinations), whereas using a lower number of parameter combinations (e.g., 100), would make the surrogate model run in less than 3 seconds. Table S5 presents a summary of the surrogate model performances based on the number of model evaluations. As a comparison, considering running twenty process-based models at the same time (four nodes each running five simulations), it would have taken 230 days just for the uncertainty analysis (10^4 model runs).

4. Conclusions

In this work we presented a response surface surrogate framework to explore, assess and predict the performances of electrokinetic bioremediation, which was modeled based on complex process-based simulations accounting for Nernst-Planck-based electrokinetic transport, Coulombic interactions, interphase mass-transfer, as well as equilibrium and kinetically controlled biogeochemical reactions.

The surrogate model was set up using an artificial neural network with MLP and was trained with randomized cross-validation of hyperparameters. Such approximation function showed great flexibility with excellent prediction performances on training, validation, and test sets for most of the output variables. The use of the developed response surface surrogate for model exploration allowed us to analyze a wide range of input conditions. Computed correlation matrices provided information on the linearity and general trends of the variables, whereas global sensitivity analysis allowed identifying the effects and relative importance of selected input variables on the model's output. Finally, the fast runtime of each simulation allowed us to perform uncertainty analysis based on Monte Carlo simulations to derive confidence intervals of the output variables based on frequency distributions of the key input parameters.

Besides the greatly reduced computational time and the possibility to perform probabilistic assessment, the developed surrogate model also favored process understanding. In fact, the quick evaluation of a large number of different scenarios unveiled dependencies and relations among physical, chemical and biological properties as well as operational input parameters. Such relations would have been challenging to explore and would have probably remained hidden due to the computational costs of a fully coupled reactive transport model.

The results of this study indicate that surrogate models can provide several benefits to the modeling of many environmental processes as they could become essential in (i) supporting quick model exploration, (ii) indicating the optimal set of operating conditions, (iii) identifying and excluding unfeasible and inefficient configurations, (iv) providing an overview of the importance and effect of different input variables on different output metrics, and (v) quantifying uncertainties on the outputs based on uncertainties in input quantities. Besides such benefits, there are also disadvantages associated with the use of surrogate models. For instance, a drawback of the surrogate modeling approach presented in this study is that it still requires process-based simulations to generate the data on which response surface models are then trained, and the generation of this data requires process understanding, reactive transport modeling skills, and may be time consuming. Moreover, as data-driven models are dependent on the data provided, if an assumption in the process-based model changes and if such variation was not accounted for in the input parameters of the model, there might be the need to repeat all the process-based simulations for training, validation and testing as the trained model cannot extrapolate well in case a new input is radically different from the training data. Other disadvantages common to many machine learning models include the difficult interpretation/explanation of data-driven models that are considered as "black boxes" and the absence of physics, which result in a model predicting well the quantities of interest but not explicitly including all the physical principles such as conservation of mass and energy considered in process-based models.

An increasing use of surrogate models in the field of groundwater contamination and remediation would help lowering the technical barriers in the simulation and increase the trust in technologies that are currently considered very difficult to predict due to the complex interplay between several

physical and biogeochemical processes. In fact, complexity is a key feature of subsurface systems where inherently coupled flow, mass transfer processes, chemical and biological reactions control the fate of contaminants (Battistel et al., 2021; Fakhreddine et al., 2016; Guo et al., 2020; Li, 2019; Prommer et al., 2019; Rathi et al., 2017; Steefel et al., 2015; Stolze et al., 2019a, 2019b) and the efficiency of in situ remediation technologies (Ni et al., 2015; Piscopo et al., 2013; Sookhak Lari et al., 2019; Sprocati et al., 2020).

Due to the large flexibility and relatively low resources needed to execute surrogate models, we envision an increasing use of such tools, appropriately tested and validated against comprehensive and fully coupled process-based models. Response surface surrogates can become important as a screening tool for decision making as well as during the design phase. Other uses could also involve the fast calibration of model parameters based on measured quantities in pilot and full-scale implementations. In this context, surrogate models capable of accounting for complex non-linear interactions can be used during calibration procedures involving for instance genetic algorithms, which are known to require several model evaluations. In addition, we envision the integration of such models with dynamic real-time data acquisition for continuous monitoring of system performances. Finally, surrogate models could become precious as digital twins during operational phases where they could be used for real-time simulation of the system dynamics and for optimizing the performances and duration of the selected remediation technology.

Besides remediation applications, the proposed approach combining process-based, multi-physics and multicomponent reactive transport modeling with response surface surrogate modeling, could be adopted in other fields of subsurface research including radioactive waste confinement (Montes-H et al., 2005; Tournassat et al., 2015), CO_2 storage (Celia et al., 2015; Gislason et al., 2010; Saaltink et al., 2013), multiphase and unsaturated flow systems (Ahmadi et al., 2020; Molins & Mayer, 2007), as well as mining operations and risk assessments (Martens et al., 2021; Muniruzzaman et al., 2020; Sinclair & Thompson, 2015).

Acknowledgments and Data

The datasets obtained from the process-based simulations performed are available in the public repository https://doi.org/10.11583/DTU.14805741. The Supporting Information document includes the parameters and constants of the biogeochemical model, the equilibrium speciation reactions and constants, the probability distribution of input parameters, and an overview of the surrogate model performances.

This study was supported by the Capital Region of Denmark and by the Minerals Research Institute of Western Australia (grant 5044). The authors would like to thank Dr. John Flyvbjerg and Dr. Nina Tuxen for the discussions during the conceptualization of this work. Constructive comments of the associate editor and three anonymous reviewers helped improving the quality of the manuscript.

References

Ahmadi, N., Mosthaf, K., Scheutz, C., Kjeldsen, P., & Rolle, M. (2020). Model-based interpretation of methane oxidation and respiration processes in landfill biocovers: 3-D simulation of laboratory and pilot experiments. *Waste Management*, *108*, 160–171. https://doi.org/10.1016/j.wasman.2020.04.025

Alizadeh, A., Hsu, W.-L., Wang, M., & Daiguji, H. (2021). Electroosmotic flow: From microfluidics to nanofluidics. *Electrophoresis*, *0*, 1–35. https://doi.org/https://doi.org/10.1002/elps.202000313

Alt-Epping, P., Tournassat, C., Rasouli, P., Steefel, C. I., Mayer, K. U., Jenni, A., et al. (2015). Benchmark reactive transport simulations of a column experiment in compacted bentonite with multispecies diffusion and explicit treatment of electrostatic effects. *Computational Geosciences*, *19*(3), 535–550. https://doi.org/10.1007/s10596-014-9451-x

Alwosheel, A., van Cranenburgh, S., & Chorus, C. G. (2018). Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis. *Journal of Choice Modelling*, *28*. https://doi.org/10.1016/j.jocm.2018.07.002

Appelo, C. A. J., & Rolle, M. (2010). PHT3D: A reactive multicomponent transport model for saturated porous media. *Ground Water*, *48*(5), 627–632. https://doi.org/10.1111/j.1745-6584.2010.00732.x

Asher, M. J., Croke, B. F. W., Jakeman, A. J., & Peeters, L. J. M. (2015). A review of surrogate models and their application to groundwater modeling. *Water Resources Research*, *51*(8), 5957–5973. https://doi.org/10.1002/2015WR016967

Battistel, M., Stolze, L., Muniruzzaman, M., & Rolle, M. (2021). Arsenic release and transport during oxidative dissolution of spatially-distributed sulfide minerals. *Journal of Hazardous Materials*, *409*, 124651. https://doi.org/10.1016/j.jhazmat.2020.124651

Behzadian, K., Kapelan, Z., Savic, D., & Ardeshir, A. (2009). Stochastic sampling design using a multi-objective genetic algorithm and adaptive neural networks. *Environmental Modelling and Software*, *24*(4), 530–541. https://doi.org/10.1016/j.envsoft.2008.09.013

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, *13*(1), 281–305.

Buttet, G. F., Murray, A. M., Goris, T., Burion, M., Jin, B., Rolle, M., et al. (2018). Coexistence of two distinct Sulfurospirillum populations respiring tetrachloroethene-genomic and kinetic considerations. *FEMS Microbiology Ecology*, *94*(5). https://doi.org/10.1093/femsec/fiy018

Caruana, R., Lawrence, S., & Giles, L. (2001). Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. In *Advances in Neural Information Processing Systems* (pp. 402–408).

Celia, M. A., Bachu, S., Nordbotten, J. M., & Bandilla, K. W. (2015). Status of CO2 storage in deep saline aquifers with emphasis on modeling approaches and practical simulations. *Water Resources Research*, *51*(9), 6846–6892. https://doi.org/10.1002/2015WR017609

Chollet, F. (2015). Keras. *GitHub Repository*.

Chowdhury, A. I. A., Gerhard, J. I., Reynolds, D., Sleep, B. E., & O'Carroll, D. M. (2017).

Electrokinetic-enhanced permanganate delivery and remediation of contaminated low permeability porous media. *Water Research*, *113*, 215–222. https://doi.org/10.1016/j.watres.2017.02.005

Crombecq, K., Laermans, E., & Dhaene, T. (2011). Efficient space-filling and non-collapsing sequential design strategies for simulation-based modeling. *European Journal of Operational Research*, *214*(3), 683–696. https://doi.org/10.1016/j.ejor.2011.05.032

Van Dam, E. R., Husslage, B., Den Hertog, D., & Melissen, H. (2007). Maximin Latin hypercube designs in two dimensions. *Operations Research*, *55*(1), 158–169. https://doi.org/10.1287/opre.1060.0317

Damiani, L. H., Kosakowski, G., Glaus, M. A., & Churakov, S. V. (2020). A framework for reactive transport modeling using FEniCS–Reaktoro: governing equations and benchmarking results. *Computational Geosciences*, *24*, 1071–1085. https://doi.org/10.1007/s10596-019-09919-3

Duhamel, M., & Edwards, E. A. (2007). Growth and yields of dechlorinators, acetogens, and methanogens during reductive dechlorination of chlorinated ethenes and dihaloelimination of 1,2-dichloroethane. *Environmental Science and Technology*, *41*(7), 2303–2310. https://doi.org/10.1021/es062010r

Eason, J., & Cremaschi, S. (2014). Adaptive sequential sampling for surrogate model generation with artificial neural networks. *Computers and Chemical Engineering*, *68*, 220–232. https://doi.org/10.1016/j.compchemeng.2014.05.021

Fakhreddine, S., Lee, J., Kitanidis, P. K., Fendorf, S., & Rolle, M. (2016). Imaging geochemical heterogeneities using inverse reactive transport modeling: An example relevant for characterizing arsenic mobilization and distribution. *Advances in Water Resources*, *88*, 186–197. https://doi.org/10.1016/j.advwatres.2015.12.005

Formaggia, L., Guadagnini, A., Imperiali, I., Lever, V., Porta, G., Riva, M., et al. (2013). Global sensitivity analysis through polynomial chaos expansion of a basin-scale geochemical compaction model. *Computational Geosciences*, *7*(1), 25–42. https://doi.org/10.1007/s10596-012-9311-5

Forrester, A. I. J., Sbester, A., & Keane, A. J. (2008). *Engineering Design via Surrogate Modelling*. John Wiley & Sons. https://doi.org/10.1002/9780470770801

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media.

Gislason, S. R., Wolff-Boenisch, D., Stefansson, A., Oelkers, E. H., Gunnlaugsson, E., Sigurdardottir, H., et al. (2010). Mineral sequestration of carbon dioxide in basalt: A pre-injection overview of the CarbFix project. *International Journal of Greenhouse Gas Control*, *4*(3), 537–545. https://doi.org/10.1016/j.ijggc.2009.11.013

Guo, B., Zeng, J., & Brusseau, M. L. (2020). A Mathematical Model for the Release, Transport, and Retention of Per- and Polyfluoroalkyl Substances (PFAS) in the Vadose Zone. *Water Resources Research*, *56*(2). https://doi.org/10.1029/2019WR026667

Haghighat, E., & Juanes, R. (2021). SciANN: A Keras/TensorFlow wrapper for scientific computations and physics-informed deep learning using artificial neural networks.

Computer Methods in Applied Mechanics and Engineering, *373*, 113552. https://doi.org/10.1016/j.cma.2020.113552

Hennig, T., & Kühn, M. (2021). Surrogate model for multi-component diffusion of uranium through opalinus clay on the host rock scale. *Applied Sciences (Switzerland)*, *11*(2). https://doi.org/10.3390/app11020786

Hou, Z., Lu, W., Xue, H., & Lin, J. (2017). A comparative research of different ensemble surrogate models based on set pair analysis for the DNAPL-contaminated aquifer remediation strategy optimization. *Journal of Contaminant Hydrology*, *203*(June), 28–37. https://doi.org/10.1016/j.jconhyd.2017.06.003

Iooss, B., & Lemaître, P. (2015). A review on global sensitivity analysis methods. *Operations Research/ Computer Science Interfaces Series*, *59*, 101–122. https://doi.org/10.1007/978-1-4899-7547-8_5

Jatnieks, J., De Lucia, M., Dransch, D., & Sips, M. (2016). Data-driven Surrogate Model Approach for Improving the Performance of Reactive Transport Simulations. *Energy Procedia*, *97*, 447–453. https://doi.org/10.1016/j.egypro.2016.10.047

Jiang, X., & Na, J. (2020). Online surrogate multiobjective optimization algorithm for contaminated groundwater remediation designs. *Applied Mathematical Modelling*, *78*, 519–538. https://doi.org/10.1016/j.apm.2019.09.053

Jin, R., Chen, W., & Sudjianto, A. (2005). An efficient algorithm for constructing optimal design of computer experiments. *Journal of Statistical Planning and Inference*, *134*(1), 268–287. https://doi.org/10.1016/j.jspi.2004.02.014

Johnson, V. M., & Rogers, L. L. (2000). Accuracy of neural network approximators in simulation-optimization. *Journal of Water Resources Planning and Management*, *126*(2), 48–56. https://doi.org/10.1061/(ASCE)0733-9496(2000)126:2(48)

Leal, A. M. M., Kyas, S., Kulik, D. A., & Saar, M. O. (2020). Accelerating Reactive Transport Modeling: On-Demand Machine Learning Algorithm for Chemical Equilibrium Calculations. *Transport in Porous Media*, *133*(2), 161–204. https://doi.org/10.1007/s11242-020-01412-1

Li, L. (2019). Watershed Reactive Transport. *Reviews in Mineralogy and Geochemistry*, *85*(1), 381–418. https://doi.org/10.2138/rmg.2018.85.13

Lima, A. T., Hofmann, A., Reynolds, D., Ptacek, C. J., Van Cappellen, P., Ottosen, L. M., et al. (2017). Environmental Electrokinetics for a sustainable subsurface. *Chemosphere*, *181*, 122–133. https://doi.org/10.1016/j.chemosphere.2017.03.143

Loeppky, J. L., Sacks, J., & Welch, W. J. (2009). Choosing the sample size of a computer experiment: A practical guide. *Technometrics*, *51*(4), 366–376. https://doi.org/10.1198/TECH.2009.08040

De Lucia, M., & Kühn, M. (2021). DecTree v1.0 -- Chemistry speedup in reactive transport simulations: purely data-driven and physics-based surrogates. *Geoscientific Model Development Discussions*, *2021*, 1–26. https://doi.org/10.5194/gmd-2020-445

Luo, J., Lu, W., Yang, Q., Ji, Y., & Xin, X. (2020). An adaptive dynamic surrogate model using a constrained trust region algorithm: application to DNAPL-contaminated-groundwater-

remediation design. *Hydrogeology Journal*, *28*, 1285–1298. https://doi.org/10.1007/s10040-020-02130-0

Maier, H. R., Jain, A., Dandy, G. C., & Sudheer, K. P. (2010). Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions. *Environmental Modelling and Software*, *25*(8), 891–909. https://doi.org/10.1016/j.envsoft.2010.02.003

Mao, X., Wang, J., Ciblak, A., Cox, E. E., Riis, C., Terkelsen, M., et al. (2012). Electrokinetic-enhanced bioaugmentation for remediation of chlorinated solvents contaminated clay. *Journal of Hazardous Materials*, *213–214*, 311–317. https://doi.org/10.1016/j.jhazmat.2012.02.001

Martens, E., Prommer, H., Sprocati, R., Sun, J., Dai, X., Crane, R., et al. (2021). Toward a more sustainable mining future with electrokinetic in situ leaching. *Science Advances*, *7*(18). https://doi.org/10.1126/sciadv.abf9971

Mayer, K. U., Frind, E. O., & Blowes, D. W. (2002). Multicomponent reactive transport modeling in variably saturated porous media using a generalized formulation for kinetically controlled reactions. *Water Resources Research*, *38*(9), 1–13. https://doi.org/10.1029/2001wr000862

Mo, S., Zabaras, N., Shi, X., & Wu, J. (2019). Deep Autoregressive Neural Networks for High-Dimensional Inverse Problems in Groundwater Contaminant Source Identification. *Water Resources Research*, *55*(5), 3856–3881. https://doi.org/10.1029/2018WR024638

Molins, S., & Mayer, K. U. (2007). Coupling between geochemical reactions and multicomponent gas and solute transport in unsaturated media: A reactive transport modeling study. *Water Resources Research*, *43*(5). https://doi.org/10.1029/2006WR005206

Montes-H, G., Marty, N., Fritz, B., Clement, A., & Michau, N. (2005). Modelling of long-term diffusion-reaction in a bentonite barrier for radioactive waste confinement. *Applied Clay Science*, *30*(3–4), 181–198. https://doi.org/10.1016/j.clay.2005.07.006

Morris, M. D., & Mitchell, T. J. (1995). Exploratory designs for computational experiments. *Journal of Statistical Planning and Inference*, *43*(3), 381–402. https://doi.org/10.1016/0378-3758(94)00035-T

Muniruzzaman, M., & Rolle, M. (2015). Impact of multicomponent ionic transport on pH fronts propagation in saturated porous media. *Water Resources Research*, *51*(8), 6739–6755. https://doi.org/10.1002/2015WR017134

Muniruzzaman, M., & Rolle, M. (2017). Experimental investigation of the impact of compound-specific dispersion and electrostatic interactions on transient transport and solute breakthrough. *Water Resources Research*, *53*(2), 1189–1209. https://doi.org/10.1002/2016WR019727

Muniruzzaman, M., & Rolle, M. (2019). Multicomponent ionic transport modeling in physically and electrostatically heterogeneous porous media with PhreeqcRM coupling for geochemical reactions. *Water Resources Research*, *55*(12), 11121–11143. https://doi.org/10.1029/2019WR026373

Muniruzzaman, M., Karlsson, T., Ahmadi, N., & Rolle, M. (2020). Multiphase and

multicomponent simulation of acid mine drainage in unsaturated mine waste: Modeling approach, benchmarks and application examples. *Applied Geochemistry*, *120*(15), 104677. https://doi.org/10.1016/j.apgeochem.2020.104677

Murray, A. M., Maillard, J., Jin, B., Broholm, M., Holliger, C., & Rolle, M. (2019). A modeling approach integrating microbial activity, mass transfer, and geochemical processes to interpret biological assays: An example for PCE degradation in a multi-phase batch setup. *Water Research*, *160*, 1–13. https://doi.org/https://doi.org/10.1016/j.watres.2019.05.087

Murray, A. M., Maillard, J., Rolle, M., Broholm, M., & Holliger, C. (2020). Impact of iron- and/or sulfate-reduction on a cis-1, 2-dichloroethene and vinyl chloride respiring bacterial consortium: experiments and model-based interpretation. *Environmental Science: Processes & Impacts*, *22*(3), 740–750. https://doi.org/https://doi.org/10.1039/C9EM00544G

Nardi, A., Idiart, A., Trinchero, P., De Vries, L. M., & Molinero, J. (2014). Interface COMSOL-PHREEQC (iCP), an efficient numerical framework for the solution of coupled multiphysics and geochemistry. *Computers and Geosciences*, *69*, 10–21. https://doi.org/10.1016/j.cageo.2014.04.011

Newman, J., & Thomas-Alyea, K. E. (2004). *Electrochemical systems*. John Wiley & Sons.

Ni, Z., Van Gaans, P., Smit, M., Rijnaarts, H., & Grotenhuis, T. (2015). Biodegradation of cis-1,2-Dichloroethene in Simulated Underground Thermal Energy Storage Systems. *Environmental Science and Technology*, *49*(22), 13519–13527. https://doi.org/10.1021/acs.est.5b03068

NIRAS. (2011). *Område IV, Skuldelev Elektrokinetisk stimuleret biologisk nedbrydning, EK-BIO*.

Ouyang, Q., Lu, W., Miao, T., Deng, W., Jiang, C., & Luo, J. (2017). Application of ensemble surrogates and adaptive sequential sampling to optimal groundwater remediation design at DNAPLs-contaminated sites. *Journal of Contaminant Hydrology*, *207*, 31–38. https://doi.org/10.1016/j.jconhyd.2017.10.007

Park, J. S. (1994). Optimal Latin-hypercube designs for computer experiments. *Journal of Statistical Planning and Inference*, *39*(1), 95–111. https://doi.org/10.1016/0378-3758(94)90115-5

Parkhurst, D. L., & Wissmeier, L. (2015). PhreeqcRM: A reaction module for transport simulators based on the geochemical model PHREEQC. *Advances in Water Resources*, *83*, 176–189. https://doi.org/10.1016/j.advwatres.2015.06.001

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Piscopo, A. N., Neupauer, R. M., & Mays, D. C. (2013). Engineered injection and extraction to enhance reaction for improved in situ remediation. *Water Resources Research*, *49*(63618–3625). https://doi.org/10.1002/wrcr.20209

Prasianakis, N. I., Haller, R., Mahrous, M., Poonoosamy, J., Pfingsten, W., & Churakov, S. V. (2020). Neural network based process coupling and parameter upscaling in reactive transport simulations. *Geochimica et Cosmochimica Acta*, *291*, 126–143.

https://doi.org/10.1016/j.gca.2020.07.019

Prommer, H., Sun, J., & Kocar, B. D. (2019). Using reactive transport models to quantify and predict groundwater quality. *Elements*, *15*(2), 87–92. https://doi.org/10.2138/gselements.15.2.87

Ramsburg, C. A., Christ, J. A., Douglas, S. R., & Boroumand, A. (2011). Analytical modeling of degradation product partitioning kinetics in source zones containing entrapped DNAPL. *Water Resources Research*, *47*(3). https://doi.org/10.1029/2010WR009958

Rasouli, P., Steefel, C. I., Mayer, K. U., & Rolle, M. (2015). Benchmarks for multicomponent diffusion and electrochemical migration. *Computational Geosciences*, *19*(3), 523–533. https://doi.org/10.1007/s10596-015-9481-z

Rathi, B., Neidhardt, H., Berg, M., Siade, A., & Prommer, H. (2017). Processes governing arsenic retardation on Pleistocene sediments: Adsorption experiments and model-based analysis. *Water Resources Research*, *53*(5). https://doi.org/10.1002/2017WR020551

Razavi, S., Tolson, B. A., & Burn, D. H. (2012). Review of surrogate modeling in water resources. *Water Resources Research*, *48*(7). https://doi.org/10.1029/2011WR011527

Reddy, K. R., & Cameselle, C. (2009). Overview of Electrochemical Remediation Technologies. In *Electrochemical Remediation Technologies for Polluted Soils, Sediments and Groundwater* (pp. 1–28). Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/9780470523650.ch1

Reynolds, D. A., Jones, E. H., Gillen, M., Yusoff, I., & Thomas, D. G. (2008). Electrokinetic migration of permanganate through low-permeability media. *Ground Water*, *46*(4), 629–637. https://doi.org/10.1111/j.1745-6584.2008.00415.x

Rolle, M., Muniruzzaman, M., Haberer, C. M., & Grathwohl, P. (2013). Coulombic effects in advection-dominated transport of electrolytes in porous media: Multicomponent ionic dispersion. *Geochimica et Cosmochimica Acta*, *120*, 195--205. https://doi.org/10.1016/j.gca.2013.06.031

Rolle, M., Sprocati, R., Masi, M., Jin, B., & Muniruzzaman, M. (2018). Nernst-Planck-based Description of Transport, Coulombic Interactions, and Geochemical Reactions in Porous Media: Modeling Approach and Benchmark Experiments. *Water Resources Research*, *54*(4), 3176–3195. https://doi.org/10.1002/2017WR022344

Saaltink, M. W., Vilarrasa, V., De Gaspari, F., Silva, O., Carrera, J., & Rötting, T. S. (2013). A method for incorporating equilibrium chemical reactions into multiphase flow models for CO2 storage. *Advances in Water Resources*, *62*, 431–441. https://doi.org/10.1016/j.advwatres.2013.09.013

Saltelli, A., Chan, K., Scott, M., & others. (2000). *Sensitivity analysis. Probability and statistics series. John and Wiley & Sons, New York*.

Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., & Tarantola, S. (2010). Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. *Computer Physics Communications*, *181*(2), 259–270. https://doi.org/10.1016/j.cpc.2009.09.018

Sanders, J., & Kandrot, E. (2010). *CUDA by Example: An Introduction to General-Purpose*

GPU Programming. *Concurrency Computation Practice and Experience*.

Sinclair, L., & Thompson, J. (2015). In situ leaching of copper: Challenges and future prospects. *Hydrometallurgy*, *157*, 306–324. https://doi.org/10.1016/j.hydromet.2015.08.022

Sobol', I. M. (2001). Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and Computers in Simulation*, *55*(1–3), 271–280. https://doi.org/10.1016/S0378-4754(00)00270-6

Sohrabi, R., Omlin, S., & Miller, S. A. (2019). GEYSER: 3D thermo-hydrodynamic reactive transport numerical simulator including porosity and permeability evolution using GPU clusters. *Computational Geosciences*, *23*, 1317–1330. https://doi.org/10.1007/s10596-019-09885-w

Sookhak Lari, K., Rayner, J. L., & Davis, G. B. (2019). Toward Optimizing LNAPL Remediation. *Water Resources Research*, *55*(2), 923–936. https://doi.org/10.1029/2018WR023380

Sprocati, R., & Rolle, M. (2020). Charge interactions, reaction kinetics and dimensionality effects on electrokinetic remediation: A model-based analysis. *Journal of Contaminant Hydrology*, *229*, 103567. https://doi.org/10.1016/j.jconhyd.2019.103567

Sprocati, R., Masi, M., Muniruzzaman, M., & Rolle, M. (2019). Modeling electrokinetic transport and biogeochemical reactions in porous media: a multidimensional Nernst-Planck-Poisson approach with PHREEQC coupling. *Advances in Water Resources*, *127*, 134–147. https://doi.org/10.1016/j.advwatres.2019.03.011

Sprocati, R., Flyvbjerg, J., Tuxen, N., & Rolle, M. (2020). Process-based modeling of electrokinetic-enhanced bioremediation of chlorinated ethenes. *Journal of Hazardous Materials*, *397*, 122787. https://doi.org/10.1016/j.jhazmat.2020.122787

Sprocati, R., Gallo, A., Sethi, R., & Rolle, M. (2021). Electrokinetic Delivery of Reactants: Pore Water Chemistry Controls Transport, Mixing, and Degradation. *Environmental Science & Technology*, *55*(1), 719–729. https://doi.org/https://doi.org/10.1021/acs.est.0c06054

Steefel, C. I., & Tournassat, C. (2020). A model for discrete fracture-clay rock interaction incorporating electrostatic effects on transport. *Computational Geosciences*, *25*, 395–410. https://doi.org/10.1007/s10596-020-10012-3

Steefel, C. I., Appelo, C. A. J., Arora, B., Jacques, D., Kalbacher, T., Kolditz, O., et al. (2015). Reactive transport codes for subsurface environmental simulation. *Computational Geosciences*, *19*(3), 445–478. https://doi.org/10.1007/s10596-014-9443-x

Stolze, L., Zhang, D., Guo, H., & Rolle, M. (2019a). Model-Based Interpretation of Groundwater Arsenic Mobility during in Situ Reductive Transformation of Ferrihydrite. *Environmental Science and Technology*, *248*, 274–288. https://doi.org/10.1021/acs.est.9b00527

Stolze, L., Zhang, D., Guo, H., & Rolle, M. (2019b). Surface complexation modeling of arsenic mobilization from goethite: Interpretation of an in-situ experiment. *Geochimica et Cosmochimica Acta*. https://doi.org/10.1016/j.gca.2019.01.008

Su, D., Ulrich Mayer, K., & MacQuarrie, K. T. B. (2017). Parallelization of MIN3P-THCm: A high performance computational framework for subsurface flow and reactive transport simulation. *Environmental Modelling and Software*, *95*, 271–289.

https://doi.org/10.1016/j.envsoft.2017.06.008

Tahmasebi, P., Kamrava, S., Bai, T., & Sahimi, M. (2020). Machine learning in geo- and environmental sciences: From small to large scale. *Advances in Water Resources*, *142*, 103619. https://doi.org/10.1016/j.advwatres.2020.103619

Taormina, R., Chau, K. W., & Sethi, R. (2012). Artificial neural network simulation of hourly groundwater levels in a coastal aquifer system of the Venice lagoon. *Engineering Applications of Artificial Intelligence*, *25*(8), 1670–1676. https://doi.org/10.1016/j.engappai.2012.02.009

Tournassat, C., & Steefel, C. I. (2019). Modeling diffusion processes in the presence of a diffuse layer at charged mineral surfaces: a benchmark exercise. *Computational Geosciences*, 1–18. https://doi.org/10.1007/s10596-019-09845-4

Tournassat, C., Steefel, C. I., Bourg, I. C., & Bergaya, F. (2015). *Natural and engineered clay barriers*. Elsevier.

Wang, H., Lu, W., Chang, Z., & Li, J. (2020). Heuristic search strategy based on probabilistic and geostatistical simulation approach for simultaneous identification of groundwater contaminant source and simulation model parameters. *Stochastic Environmental Research and Risk Assessment*, *34*, 891–907. https://doi.org/10.1007/s00477-020-01804-1

Wu, T., Yang, Y., Wang, Z., Shen, Q., Tong, Y., & Wang, M. (2020). Anion Diffusion in Compacted Clays by Pore-Scale Simulation and Experiments. *Water Resources Research*, *56*(11), e2019WR027037. https://doi.org/10.1029/2019wr027037

Yan, S., & Minsker, B. (2006). Optimal groundwater remediation design using an Adaptive Neural Network Genetic Algorithm. *Water Resources Research*, *42*(5). https://doi.org/10.1029/2005WR004303

Yan, S., & Minsker, B. (2011). Applying dynamic surrogate models in noisy genetic algorithms to optimize groundwater remediation designs. *Journal of Water Resources Planning and Management*, *137*(3), 284–292. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000106

Yin, J., & Tsai, F. T. C. (2020). Bayesian set pair analysis and machine learning based ensemble surrogates for optimal multi-aquifer system remediation design. *Journal of Hydrology*, *580*, 124280. https://doi.org/10.1016/j.jhydrol.2019.124280

Yu, S., Dolan, M. E., & Semprini, L. (2005). Kinetics and inhibition of reductive dechlorination of chlorinated ethylenes by two different mixed cultures. *Environmental Science and Technology*, *39*(1), 195–205. https://doi.org/10.1021/es0496773

Yu, X., Cui, T., Sreekanth, J., Mangeon, S., Doble, R., Xin, P., et al. (2020). Deep learning emulators for groundwater contaminant transport modelling. *Journal of Hydrology*, *590*, 125351. https://doi.org/10.1016/j.jhydrol.2020.125351

Yustres, Á., López-Vizcaíno, R., Cabrera, V., Rodrigo, M. A., & Navarro, V. (2020). Donnan-ion hydration model to estimate the electroosmotic permeability of clays. *Electrochimica Acta*, *355*, 136758. https://doi.org/10.1016/j.electacta.2020.136758

Zhou, D.-M., Cang, L., Alshawabkeh, A. N., Wang, Y.-J., & Hao, X.-Z. (2006). Pilot-scale electrokinetic treatment of a Cu contaminated red soil. *Chemosphere*, *63*(6), 964–971. https://doi.org/10.1016/j.chemosphere.2005.08.059