Title

A chemistry-informed hybrid machine learning approach to predict metal adsorption onto mineral surfaces

Authors

Chang, Elliot
Zavarin, Mavrik
Beverly, Linda
et al.

Publication Date

2023-08-01

DOI

10.1016/j.apgeochem.2023.105731

Peer reviewed

Applied Geochemistry 155 (2023) 105731

Contents lists available at ScienceDirect

Applied Geochemistry

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

A chemistry-informed hybrid machine learning approach to predict metal adsorption onto mineral surfaces

Elliot Chang [a,*], Mavrik Zavarin [a], Linda Beverly [b], Haruko Wainwright [b,c]

[a] Seaborg Institute, Lawrence Livermore National Laboratory, 7000 East Avenue, Livermore, CA, 94550, USA
[b] Lawrence Berkeley National Laboratory, Earth and Environmental Sciences Area, 1 Cyclotron Road, Berkeley, CA, 94720, USA
[c] U.C. Berkeley, Department of Nuclear Engineering, 4153 Etcheverry Hall, Berkeley, CA, 94720, USA

ARTICLE INFO

Editorial Handling By: Dr. Zimeng Wang

Keywords:
Adsorption
Metal
Mineral
Surface complexation modeling
Random forest
Hybrid machine learning

ABSTRACT

Historically, surface complexation model (SCM) constants and distribution coefficients (K_d) have been employed to quantify mineral-based retardation effects controlling the fate of metals in subsurface geologic systems. Our recent SCM development workflow, based on the Lawrence Livermore National Laboratory Surface Complexation/Ion Exchange (L-SCIE) database, illustrated a community FAIR data approach to SCM development by predicting uranium(VI)-quartz adsorption for a large number of literature-mined data. Here, we present an alternative hybrid machine learning (ML) approach that shows promise in achieving equivalent high-quality predictions compared to traditional surface complexation models. At its core, the hybrid random forest (RF) ML approach is motivated by the proliferation of incongruent SCMs in the literature that limit their applicability in reactive transport models. Our hybrid ML approach implements PHREEQC-based aqueous speciation calculations; values from these simulations are automatically used as input features for a random forest (RF) algorithm to quantify adsorption and avoid SCM modeling constraints entirely. Named the LLNL Speciation Updated Random Forest (L-SURF) model, this hybrid approach is shown to have applicability to U(VI) sorption cases driven by both ion-exchange and surface complexation, as is shown for quartz and montmorillonite cases. The approach can be applied to reactive transport modeling and may provide an alternative to the costly development of self-consistent SCM reaction databases.

1. Introduction

The high reactivity of mineral surfaces (Dong and Wan, 2014; Durrant et al., 2018) enables metals to adsorb to soils and sediments, limiting their bioavailability and influencing their overall mobility. Scientists have traditionally used surface complexation models (SCMs) or linear distribution coefficients (K_d) to quantify this adsorption phenomenon and to predict metal partitioning in immobile solid versus mobile aqueous phases (Appelo et al., 2002; Goldberg, 1992; Nair et al., 2014b). SCMs not only account for sorbate complexation to mineral surfaces, but, when paired with thermochemical databases (e.g. Ragoussi and Costa (2019)), they account for aqueous speciation, mineral solubility, and liquid-gas exchange (Parkhurst and Appelo, 2013; Romero-González et al., 2007). Over the course of last several decades, SCMs have become well-established thermodynamic components to larger-scale reactive transport models that also include hydrological and fluid-dynamics transport processes, such as advection, diffusion, and dispersion (Nitzsche and Merkel, 1999; Steefel, 2019; Zhu et al., 2001). Furthermore, recent years have seen deeper mechanistic information incorporated into SCMs: Estes and Powell (2020) established the temperature dependence of U(VI) adsorption onto hematite through calibrating SCMs to multi-temperature batch adsorption data while Tournassat et al. (2018) addressed basal plane electrostatic surface potential spillover effects on montmorillonite edge-surface SCMs (Estes and Powell, 2020; Tournassat et al., 2018). Additionally, Ren et al. (2020) used attenuated total reflectance-Fourier transform infrared spectroscopy and extended X-ray absorption fine structure spectroscopy to discern that U(VI) adsorption onto manganese oxide is driven by a bidentate, binuclear surface structure under acidic to neutral pH conditions. The use of various spectroscopic techniques in combination with batch adsorption methods demonstrated a path forward for enhancing the mechanistic accuracy during the construction of SCMs.

While calibrated SCMs yield valuable aqueous- and surface-speciation predictions under the investigated geochemical conditions,

* Corresponding author.
E-mail addresses: elliotc@berkeley.edu (E. Chang), zavarin1@llnl.gov (M. Zavarin), lcbeverly@lbl.gov (L. Beverly), hmwainwright@lbl.gov (H. Wainwright).

https://doi.org/10.1016/j.apgeochem.2023.105731
Received 1 February 2023; Received in revised form 20 June 2023; Accepted 23 June 2023
Available online 27 June 2023

the implementation of SCMs also poses some key limitations. The most notable of these challenges that has not been solved to-date is the non-uniqueness of SCMs that are implemented with various divergent assumptions regarding the nature of the surface electrostatic potential (Westall and Hohl, 1980). Because adsorption is largely driven by the favorable interactions between oppositely charged sorbent surfaces and aqueous sorbate molecules, surface complexation is impacted by surface charge, which in turn may influence the stability of various surface-bound complexes (Davis and Kent, 2018). Historically, there have been numerous conceptualizations of the electrical double layer, ranging from constant capacitance (Schindler et al., 1976) and triple layer models (Leroy and Revil, 2004) that assume linear surface charge-potential relationships to non-electrostatic approaches that do not account for charge buildup altogether (Newcombe and Drikas, 1997; Pivovarov, 1998). These various SCMs capture the electrostatic effects on adsorption through fundamentally different assumptions—yet, they are all able to sufficiently match data from batch adsorption experiments with similar accuracy (Westall and Hohl, 1980). Because the sorbate-sorbent stability constants extracted from these SCMs are model-dependent, a significant present-day challenge exists in comparing and co-utilizing various historic SCMs and associated reaction constants that have very different underlying assumptions.

The various electrostatic model descriptions of the mineral electrical double-layer are simplifications of the reality that inherently pose limitations in their usage. For instance, constant capacitance SCMs assume that all sorbate metals bound to the mineral sorbent are strong, inner-sphere complexes, eliminating any nuance to outer-sphere or diffuse-swarm based surface complexation. Because the theoretical basis for such mineral-fluid interface descriptions may be limited, implementing a data-driven model to represent the mineral-fluid interface provides a new path forward in directly exploiting the continual growth of adsorption data available in the literature. Among ML techniques, the random forest (RF) algorithm has received significant attention for providing a flexible learning framework that can effectively capture nonlinear behavior commonly found in adsorption dynamics, such as impacts associated with pH and ionic strength solution changes (Breiman, 2001). Merzlikine et al. (2011) illustrated the applicability of RF models to predict solubility changes induced by the complexation of cyclodextrins to various drug molecules (Merzlikine et al., 2011). In a similar fashion, Chaube et al. (2020) implemented RF modeling to probe lanthanide binding affinities to many different ligand compounds (Chaube et al., 2020). Both Merzlikine et al. (2011) and Chaube et al. (2020) demonstrated the effectiveness of RF modeling in predicting aqueous complexation processes. In recent years, authors have also discussed the quantification of solid-phase adsorption reactions through RF methods. Zhu et al. (2019) used RF learning to predict adsorption of six heavy metals onto biochar surfaces and Hafsa et al. (2020) demonstrated the broad generalizability of RF learning to predict adsorption of heavy metals onto various biosorbent interfaces. At a larger environmental scale, Dalla Libera et al. (2020) also implemented unsupervised ML algorithms in the form of self-organizing maps (SOMs) to study redox-controlled dissolution-precipitation processes of hydrous ferric oxides as a key parameter controlling arsenic mobility in alluvial aquifers near Venice, Italy. Latest research further deploys AI-supported surrogate model development to rapidly predict SCM parameters in a more robust manner than historically evaluated (Li and Zarzycki, 2022), pushing the boundary of predictive modeling of sorption processes. Many of these studies have implemented various forms of uncertainty quantification. Root mean square error and mean absolute error are among the most common metrics to evaluate overall ML model performance and errors between observed and predicted data points (Karunasingha, 2021). These rigorous error quantification methods provide a major added benefit for the implementation of ML approaches for environmental risk and performance assessments as uncertainties can be propagated through downstream analyses.

The goal of this study is to develop a hybrid-ML approach that exploits thermodynamic aqueous speciation calculations while also including RF-ML regression modeling of the mineral-fluid interface. While the traditional RF studies utilize ML to model sorption, these works wholly use ML, eliminating the mechanistic underpinnings of SCMs altogether. This highlights a dramatic shift from quasi-mechanistic SCM constructs to data-driven, black-box predictions of the adsorption process. The Lawrence Livermore National Laboratory-Speciation Updated Random Forest (L-SURF) model operates as an alternative hybrid approach. Because reactive transport codes can effectively simulate aqueous speciation but the relevant SCM data suffer from diverging descriptions of surface reactions (e.g. electrostatics, reaction stoichiometries, etc.), we exploit the solution chemistry description found in traditional thermochemical databases while replacing the SCM interfacial chemical modeling with a data-driven, RF-ML approach. By doing so, we develop a new model that is not hindered by limitations of explicit surface descriptions: we eliminate challenges associated with assumptions on electrostatic surface effects and complicated permutations of relevant reaction stoichiometries that potentially convolute overall mechanism. Here, we demonstrate how L-SURF can be trained to predict metal adsorption onto mineral surfaces. This L-SURF approach can be integrated with reactive transport modeling codes that account for aqueous speciation and solubility using traditional thermodynamics.

2. Methods

2.1. Data acquisition and pre-processing

Extensive raw adsorption data in addition to an aqueous speciation database are needed for the application of L-SURF. The Lawrence Livermore National Laboratory-Surface Complexation/Ion Exchange (L-SCIE) database is a recent effort to unify community adsorption experiments and metadata in a findable, accessible, interoperable, and reusable (FAIR) format (Zavarin et al., 2022). It has already mined over 23,000 raw adsorption data from the literature. Briefly, L-SCIE mines sorption data (K_d, % sorbed, surface excess) and dataset experimental conditions (background electrolyte, mineral surface area, gas composition, etc.) from journal manuscripts and loads them into a database. The sorption data undergo a series of unit conversions to yield a unified database which includes propagated conversion errors from the original extracted data (Zavarin et al., 2022) The database can then be filtered for a mineral-metal pair of interest in order to display a corresponding experimental dataset.

2.2. L-SURF algorithm part 1: Aqueous speciation modeling of raw sorption data

The first step of L-SURF requires an aqueous speciation database and compilation of raw experimental adsorption data for a given metal-mineral pair (Fig. 1) to perform aqueous speciation calculations. Here, aqueous speciation calculations were performed using the PHREEQC software (Parkhurst and Appelo, 2013). Notably, other speciation codes may also be implemented if desired. We used the thermodynamic database that is provided with PHREEQC and is derived from LLNL's SUPCRT (Johnson and Lundeen, 1997) database but updated with missing and revised U(VI) reaction constants taken from the latest NEA-TDB effort (Ragoussi and Brassinnes, 2015; Ragoussi and Costa, 2019) as implemented in our previous work (Zavarin et al., 2022). The L-SCIE database was used as the source of raw experimental adsorption data, for data unification, and data filtering. Raw adsorption data consist of total sorbate concentration, aqueous equilibrium sorbate concentration, and associated metadata consisting of gas composition, sorbent properties (concentration, surface area, reactive site density), background electrolyte concentrations, and pH conditions compiled in a comma-separated values format (Table 1). The metadata and the total aqueous sorbate concentration are used for each sorption datapoint to create PHREEQC simulations of solution chemistry conditions. Upon the

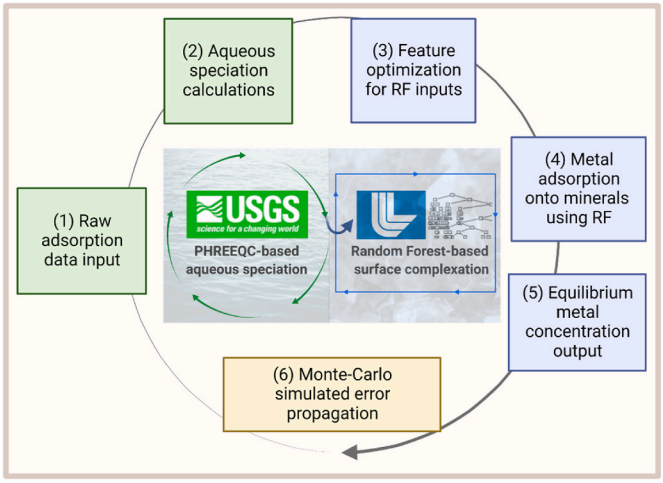

Fig. 1. L-SURF workflow chart with chronological steps: (1) Adsorption data and selected thermodynamic database are imported into L-SURF module, (2) Aqueous speciation calculations are conducted and all geochemical features are output, (3) Choice of most impactful geochemical features and hyperparameters are optimized, (4) Optimal features are used to train and test a random forest adsorption model, (5) Equilibrium aqueous metal sorbate concentrations are output, and (6) Steps 1–5 are repeated using Monte-Carlo simulations with randomly sampled input data ± experimentally-determined measurement uncertainty.

Table 1
Generic data frame structure used as feature inputs into RF adsorption model.

Input feature	Method for obtaining feature values	Input feature units
Mineral source	Extracted from L-SCIE metadata	Number associated with unique mineral source
Log_{10}(Total Sorbate Concentration)	Extracted from L-SCIE metadata	Log_{10}(Molar)
Log_{10}(Total Site Concentration)	Calculated using L-SCIE metadata[a]	Log_{10}(sites/L)
Ionic Strength	Calculated using PHREEQC speciation	Log_{10}(Molar)
Log_{10}(Aqueous sorbate species concentrations)[b]	Calculated using PHREEQC speciation	Log_{10}(Molar)
$Log_{10}(HCO_3^-$ concentration)	Calculated using PHREEQC speciation	Log_{10}(Molar)

[a] log_{10}(Total Site Concentration) = Mineral density (g/L) × Mineral surface area (m^2/g) × Mineral site density (sites/m^2). All parameters to compute Total Site Concentration are available in the L-SCIE metadata.

[b] Here, aqueous U(VI) sorbate species features include the testing of all of the following under variable permutations: UO_2OH^+, $(UO_2)_2CO_3(OH)_3^-$, $UO_2(OH)_2$, UO_2CO_3, UO_2^{2+}, $(UO_2)_3(OH)_5^+$, $(UO_2)_2(OH)_2^{2+}$, $UO_2(CO_3)_2^{2-}$, UO_2Cl^+, $(UO_2)_4(OH)_7^+$, $UO_2NO_3^+$, $(UO_2)_3(OH)_4^{2+}$, $(UO_2)_3(OH)_7^-$, $(UO2)_2OH^{3+}$, $UO_2(CO_3)_3^{4-}$, UO_2Cl_2, $(UO_2)_3O(OH)_2(HCO3)^+$, $UO_2(OH)_4^{2-}$, $(UO_2)_3(CO_3)_6^{6-}$, $(UO_2)_{11}(CO_3)_6(OH)_{12}^{2-}$, and $UO_2(OH)_3^-$.

completion of aqueous speciation calculations, relevant geochemical variables are assessed as input features for the subsequent RF adsorption model development. An example data frame for raw data of U (VI)-quartz adsorption as extracted directly from L-SCIE is available in the Supplemental Information (Table SI1).

2.3. L-SURF algorithm part 2: RF regression of mineral-based adsorption

The second, RF-based step of L-SURF is executed after aqueous speciation modeling determines the equilibrium solution conditions for each individual adsorption data point. RF is an ensemble machine learning algorithm that uses a combination of decision trees whereby individual trees are built upon a randomly and independently sampled set of training data (Breiman, 2001). The RF regression method was chosen as the ensemble ML algorithm because of its effectiveness in capturing non-linear relationships between various dependent variables and the target valuable. This poses a particular advantage in characterizing adsorption isotherm and edge data, where ionic strength, adsorbate and adsorbent concentrations can non-linearly impact the overall adsorption phenomena and the resulting equilibrium sorbate concentration (Pereira et al., 2019). The RF algorithm uses bootstrapped samples obtained from the provided training dataset to construct individual decision trees. Decision trees are a common form of supervised learning whereby nodes are created to subset the data to meet certain criteria, such as evaluation of mean square error. In this RF model, all tree predictions are defined as the equilibrium aqueous sorbate concentration. These tree predictions are bagged, and bootstrap aggregated to output the best-predicting output value. At each splitting node within a decision tree, a random subset of dependent variables is selected as input features. Training data not selected via this method are called out-of-bag (OOB) samples and are used to estimate error of the built model. OOB predicted values for all decision trees are averaged and compared to the experimental values based on RMSE calculation:

$$\mathrm{RMSE}_{\mathrm{OOB}} = \sqrt{\frac{\sum_{1}^{n}\left[y_i - \bar{y}_i^{OOB}\right]^2}{n}} \quad (1)$$

where y_i is the experimental equilibrium aqueous adsorbate concentration extracted directly from the L-SCIE database of raw adsorption data, $\bar{y}_i^{OOB}$ is the average predicted equilibrium aqueous adsorbate concentration based on OOB-sampled data, and n is the number of samples employed in the RF model.

As applied to the L-SURF workflow, previously described metadata variables and PHREEQC output variables (Table 1) are pushed through the RF regression model. The root-mean-square-error (RMSE) of the regression outputs are then evaluated against a training subset of the experimental raw adsorption data. A constraint on the number of input features is tested at this step to ascertain the most impactful geochemical features to incorporate into the RF regression. Specifically, the number of input features to obtain the highest RMSE is determined by adding an individual variable up to the specified input feature count. The selection of the individual variables is explored using 10-fold cross-validation splitting. A backward sequential feature selection algorithm was also explored (one variable removed from 6 to 1 feature counts) with minimal alteration in the result. Upon the optimal choice of input features that yielded the lowest RMSE, the RF regression model is trained using a 10-fold cross-validation (Marcot and Hanea, 2021) repeated 20 times to randomly search for hyperparameters that yielded the best training RMSE. The hyperparameters tuned in this study's implementation of RF modeling were the number of estimators, maximum tree depth, minimum samples split, and minimum samples in a leaf. The number of estimates determines the number of decision trees in a given forest and the maximum tree depth constrains the maximum number of levels to implement in each decision tree. The minimum sample split determines the minimum number of datapoints to place in a given node before splitting occurs and the minimum samples in a leaf determines how many data points are allowed in each leaf node. These hyperparameters are adjusted after the choice of the most impactful geochemical input features in order to optimize the overall RF adsorption model performance. Upon completion of training, the RF regression model is validated against another subset of experimental raw adsorption data, and test predictions are finally made. The model is trained using 80% of the experimental data while 20% of the remaining data are equally split (10%/10%) and used as validation and test datasets. The training, validation, and test errors are determined using RMSE. The test error is additionally calculated using a weighted Pearson correlation coefficient, R, to allow for direct comparison with a traditional SCM constructed for U(VI)-quartz adsorption (Zavarin et al., 2022):

$$R = \frac{\sum(c_i w_i - m)(c_{oi} w_i - m_o)}{\left(\sum(c_i w_i - m)^2 \sum(c_{io} w_i - m_o)^2\right)^{1/2}} \quad (2)$$

where c_i and c_{oi} are the measured and simulated aqueous concentration for the ith observation, respectively, w_i is the weight (1/standard deviation) of the ith observation, and m and m_o are the mean measured and simulated values of weighted aqueous concentration, respectively. The Pearson correlation coefficient was chosen because it is independent of the number of observations and additionally accounts for uncertainties associated with each individual observation.

For a more intuitive understanding of the model outputs, the equilibrium aqueous sorbate concentrations that were output from the RF regression model were converted to K_d values using the following expression:

$$K_d(L/g) = \frac{n\left(\frac{mol}{g}\right)}{c\left(\frac{mol}{L}\right)}, \quad (3)$$

where c = model output equilibrium aqueous sorbate concentration and n = surface excess as computed using the following expression:

$$n\ (\mathrm{mol}/\mathrm{g}) = \frac{Sorbed\ concentration\ \left(\frac{mol}{L}\right)}{Mineral\ density\ \left(\frac{g}{L}\right)} = \frac{Total\ sorbate\ concentration\ \left(\frac{mol}{L}\right) - c}{Mineral\ density\ \left(\frac{g}{L}\right)} \quad (4)$$

2.4. Partial dependency plot for assessing effects of important geochemical features

A well-trained RF model can provide useful predictive capabilities and also elucidate dependent variables that are particularly important in the prediction process (Nguyen et al., 2022). This approach has yielded descriptions of the most important geochemical features that impact contaminant presence in aquifers (Lopez et al., 2021; Ransom et al., 2017; Wheeler et al., 2015). The marginal effects contributed by a given feature on a predicted outcome can be visualized using a partial dependence plot (PDP) (Friedman, 2001). The partial dependence function for a regression is defined as

$$\hat{f}_s(x_s) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}_s\left(x_s, x_c^{(i)}\right) \quad (5)$$

where x_s are the plotted features and $x_c^{(i)}$ are the features in the ML model $\hat{f}$ we are not interested in. An average over the n instances of the data is taken, where a Monte-Carlo method is used n times and an average of x_s partial dependencies while marginalizing effects of $x_c^{(i)}$ is used to calculate the global relationship of a feature x_s with its predicted value.

2.5. Error propagation from experimental uncertainty

Each adsorption datapoint and its associated metadata possess experimental uncertainties that are extracted directly from L-SCIE. For each datapoint that is selected, a normal distribution is commonly used, and a random sample is chosen within ± *1* standard deviation of the average value. After a random sample for each variable is selected, a Monte-Carlo simulation encompassing the full L-SURF workflow (aqueous speciation calculations + RF mineral sorption modeling) is run Y times, where Y iterations may be specified. For the U(VI)-quartz test case, Y = 200 iterations were run to demonstrate the Monte-Carlo iterative process. Upon completing the L-SURF iterations, a mean value and standard deviation is computed from the L-SURF output values to quantify error propagated throughout the full modeling process (Anderson, 1976). The model output values (equilibrium sorbate concentrations) were then converted to associated K_d values following equations (3) and (4).

2.6. U(VI)-quartz and -montmorillonite test cases

For the test case of U(VI)-quartz adsorption, the features tested included mineral site concentration, ionic strength, HCO_3^- aqueous species concentration, and aqueous UO_2OH^+, $(UO_2)_2CO_3(OH)_3^-$, $UO_2(OH)_2$, UO_2CO_3, UO_2^{2+}, $(UO_2)_3(OH)_5^+$, $(UO_2)_2(OH)_2^{2+}$, $UO_2(CO_3)_2^{2-}$, UO_2Cl^+, $(UO_2)_4(OH)_7^+$, $UO_2NO_3^+$, $(UO_2)_3(OH)_4^{2+}$, $(UO_2)_3(OH)_7^-$, $(UO2)_2OH^{3+}$, $UO_2(CO_3)_3^{4-}$, UO_2Cl_2, $(UO_2)_3O(OH)_2(HCO3)^+$, $UO_2(OH)_4^{2-}$, $(UO_2)_3(CO_3)_6^{6-}$, $(UO_2)_{11}(CO_3)_6(OH)_{12}^{2-}$, and $UO_2(OH)_3^-$ species concentrations. Each aqueous species concentration was treated as an individual feature within the RF model. Ionic strength was chosen as a feature representing chemical effects associated with aqueous species activity corrections and surface electrostatic potential. The HCO_3^- aqueous species concentration was used as an input feature to account for CO_2 liquid-gas exchange and speciation as a function of pH. The U(VI) aqueous species were chosen as features in an attempt to capture surface complexation of U(VI) onto quartz, such as through a monodentate, inner-sphere reaction as described in Zavarin et al. (2022):

$$>SiOH + UO_2^{2+} + X\ H_2O \rightarrow >SiOUO_2(OH)_X^{+1-X} + (1+X)\ H^+, \text{ where } X = 0\text{–}3. \quad (6)$$

Among the U(VI) species, carbonate complexes such as the $UO_2CO_3{}_{(aq)}$ aqueous species were also tested as features. For instance, $UO_2CO_3{}_{(aq)}$ is an important species that may participate in U(VI)-carbonate surface complexation with quartz (Zavarin et al., 2022):

$$>SiOH + UO_2^{2+} + Y\ CO_3^{2-} \rightarrow >SiOUO_2(CO_3)_Y^{+1-2Y} + 1H^+, \text{ where } Y = 1\text{–}2. \quad (7)$$

Ultimately, these variables were tested specifically to account for the most relevant liquid-gas exchange processes, aqueous complexes, and activities of aqueous species.

A second test case of U(VI)-montmorillonite adsorption was investigated to assess the flexibility of L-SURF to evaluate adsorption processes driven by a combination of surface complexation and ion-exchange as described in Troyer et al. (2016) in which the ion-exchange is formulated as a competition between cationic species:

$$2\ X\text{–}Na + UO_2^{2+} \rightarrow X_2\text{-}UO_2 + 2\ Na^+ \quad (8)$$

where X is the permanent structural charge associated with the montmorillonite basal plane.

The same geochemical features were tested under the second test case in order to compare any mechanistic differences as determined by the L-SURF model. At ambient CO_2 conditions, it is expected that the U(VI)-quartz test case will be a representative example of the surface complexation of UO_2^{2+}, U(VI)-hydroxide and U(VI)-carbonate species whereas the U(VI)-montmorillonite test case will be representative of a combination of UO_2^{2+} ion-exchange and surface complexation of the various U(VI) species in solution. The most impactful RF input features for each test case is discussed in the Results and Discussion section.

3. Results and Discussion

3.1. Model performance of U(VI) adsorption onto quartz

Among the aqueous speciation calculations and metadata information provided as input features, the most impactful geochemical variables selected for U(VI)-quartz adsorption were $UO_2(CO_3)_3^{4-}$ aqueous concentration, total uranium concentration, and mineral site concentration (Table 2). The optimized hyperparameters chosen for the RF adsorption model using these three features involved an unconstrained maximum depth, a 200-estimator count, 1 minimum sample per leaf, and 2 minimum samples per split (Table 2). The model performance as evaluated using the Pearson correlation coefficient showed no significant improvement after the incorporation of the aforementioned three geochemical features (Fig. 2). The choice of 3–6 input features results in

Table 2
Most impactful geochemical input features and hyperparameters for RF models.

	U(VI)-Quartz	U(VI)-Montmorillonite
Features:		
Feature 1	Total Uranium Concentration Log_{10}(mol/L)	Total U(VI) Concentration Log_{10}(mol/L)
Feature 2	Total Site Concentration Log_{10}(mol/L)	Total Site Concentration Log_{10}(mol/L)
Feature 3	$UO_2(CO_3)_3^{4-}$ Concentration Log_{10}(mol/L)	Ionic Strength Log_{10}(mol/L)
Feature 4		UO_2^{2+} Concentration Log_{10}(mol/L)
Hyperparameters:		
Max depth	None [no maximum constraint]	None [no maximum constraint]
Number of estimators	200	200
Min samples in a leaf	1	1
Min samples in a split	2	2

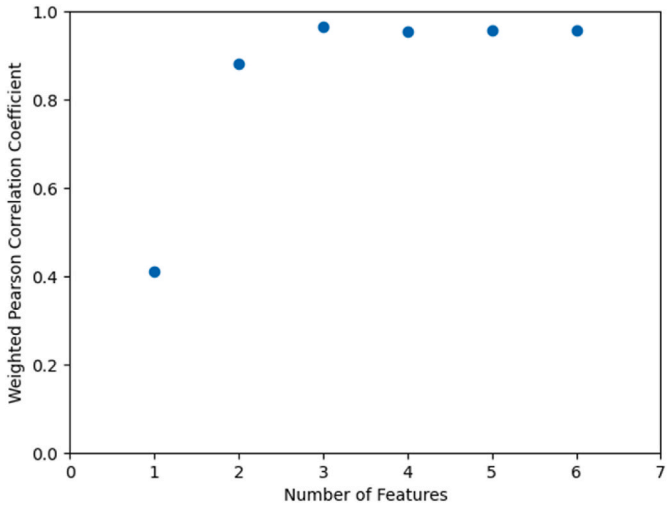

Fig. 2. U(VI)-quartz model performance against the training dataset as the number of RF input features is varied. The weighted Pearson correlation coefficient is not significantly improved with the inclusion of more than 3 features.

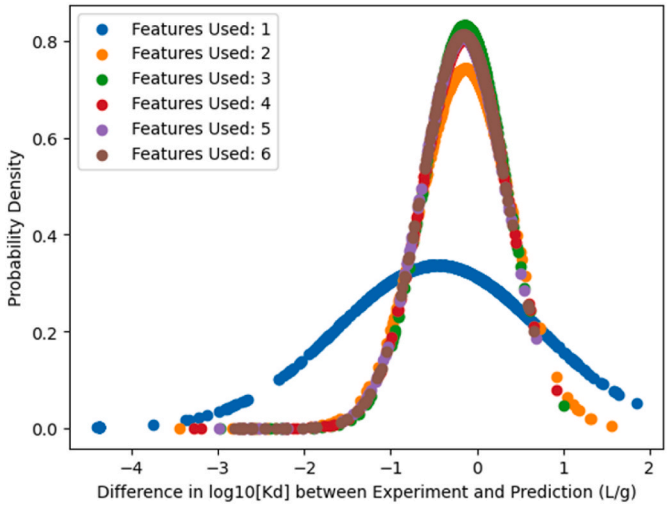

Fig. 3. Probability Density Function of Error between Experimentally Derived and Predicted K_d values for U(VI)-quartz adsorption. The number of feature inputs were varied until no further improvement was observed.

comparable error between the experimental and predicted K_d values (Fig. 3). Notably, while the choice of one input feature results in large error (±1 standard deviation falls within 2 $log_{10}(K_d)$ units of experimental values), the choice of 2+ geochemical features results in predictions falling within 1 $log_{10}(K_d)$ unit of the experimental values. Among the input feature counts between 2 and 6, the choice of 3 input features yielded the best RMSE against the training data. Total uranium concentration and mineral site concentrations are relevant for predicting adsorption dynamics as these features establish the solid: solution ratio. The automatic feature selection of $UO_2(CO_3)_3^{4-}$ aqueous concentration suggests that a surface complexation mechanism that includes U(VI)-carbonate species is important for predicting U(VI)-quartz sorption.

Model performance of the optimal RF adsorption model was reached using 3 input features. A training RMSE score = 0.065, validation RMSE score = 0.18, and test RMSE score = 0.13 are reported, demonstrating high-quality predictions across each data subset. When accounting for experimentally derived uncertainty, a weighted Pearson correlation coefficient R score = 0.96 was determined. As the RF training and validation scores yield low RMSE against their respective subsets of data and the R score is greater than 0.90 (Zavarin et al., 2022), this study presents an ML method that successfully accepts aqueous chemistry based features to accurately predict U(VI)-quartz interactions (Fig. 2). Additionally, Monte-Carlo iterations of L-SURF applied to U(VI)-quartz adsorption were implemented to propagate measurement derived uncertainties associated with electrolyte concentrations, pH, total site concentration, and CO_2 gas fugacity, yielding standard deviation values for each L-SURF test prediction of equilibrium aqueous sorbate concentrations. These standard deviations were then propagated to compute errors associated with predicted K_d values, which may have particular relevance in downstream modeling efforts, such as performance assessment approaches.

The adsorption predictions generated from L-SURF can be compared with recent efforts that use the same U(VI)-quartz dataset to generate optimized non-electrostatic, diffuse layer, and triple layer SCMs. Briefly, Zavarin et al. (2022) used PHREEQC/PEST to optimize a number of SCMs and associated reactions constants reported in the literature. The global fits yielded R values that ranged from 0.88 to 0.94 depending on the choice of SCM. In comparison, the L-SURF model yielded an R of 0.96, demonstrating the ability of hybrid ML modeling to predict pH-dependent radionuclide-mineral adsorption with comparable accuracy to the best traditional 2-site, 4 surface complex non-electrostatic surface complexation model (Fig. 4).

3.2. Model performance of U(VI) adsorption onto montmorillonite

In contrast to U(VI)-quartz adsorption, U(VI) adsorption onto montmorillonite required the presence of 4 geochemical input features to fit the training dataset (Fig. 5). The important features were total U(VI) concentration, total site concentration, ionic strength, and UO_2^{2+} aqueous species concentration. The optimized hyperparameters for this RF model with 4 input features were the same as that optimized for the U(VI)-quartz RF model (Table 2). Notably, the lack of U(VI)-carbonate species and the presence of ionic strength and a cationic UO_2^{2+} aqueous species suggests the RF algorithm's distinction between two different adsorption mechanisms. Rather than a U(VI)-carbonate surface complexation driven regression seen in the U(VI)-quartz RF model, the importance of ionic strength and aqueous UO_2^{2+} features suggests that the U(VI)-montmorillonite RF model is most impacted by ion-exchange. The resultant RF model that utilized 4 input features yielded K_d prediction values that were within ±1 $log_{10}(K_d)$ units of the experimental values (Fig. 6).

The best-performing RF adsorption model was evaluated to have training, validation, and test RMSE of 0.087, 0.30, and 0.29, respectively. A larger RMSE for validation and test errors suggest a potential for overfitting of the training data. Modifications to the execution of hyperparameter tuning or choice of training-validation-test sub-datasets

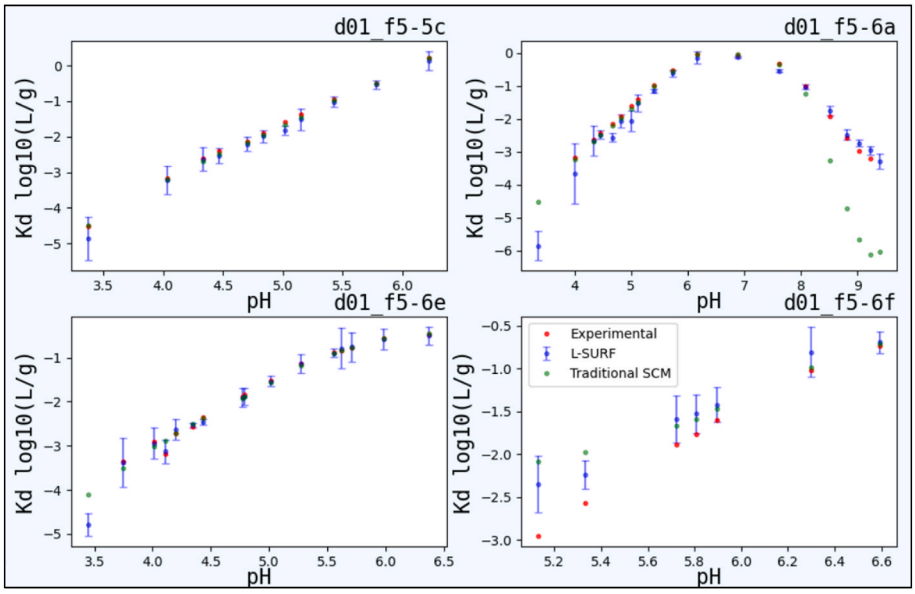

Fig. 4. Example pH-dependent model predictions of K_d values for U(VI)-quartz adsorption. The community-data optimized 2-site, 4 surface complex non-electrostatic SCM reported in Table 2 of Zavarin et al. (2022)is represented by green circles; L-SURF modeling is represented by blue stars ± 1 standard deviation error bars; experimentally measured values are indicated by red dots. Adsorption edge data are sub-divided into individual datasets from the D01 (Davis, 2001) original publication as extracted from L-SCIE (Zavarin et al., 2022). Datapoint references, such as f5 through 5c, are included as originates from the L-SCIE database nomenclature to track individual datapoints from reference D01. The full pH-dependent model predictions may be found in the Supplemental Information (SI.1) and utilize the following references: AOTKA82 = Allard et al. (2011); AZBN00a = (Arnold et al., 1999); AZBN98 = (Arnold et al., 1998); AZZBN01 = (Arnold et al., 2001); CSB18 = (Coutelot et al., 2018); D01 = (Davis, 2001); DW14 = (Dong and Wan, 2014); FDZ06b = (Fox et al., 2006).; JHLCH99 = (Jung et al., 1999); KCKD96 = (Kohler et al., 1996); NKM14 = (Nair et al., 2014a); PJTP01 = (Prikryl et al., 2001); PTBP98 = (Pabalan et al., 1998a).

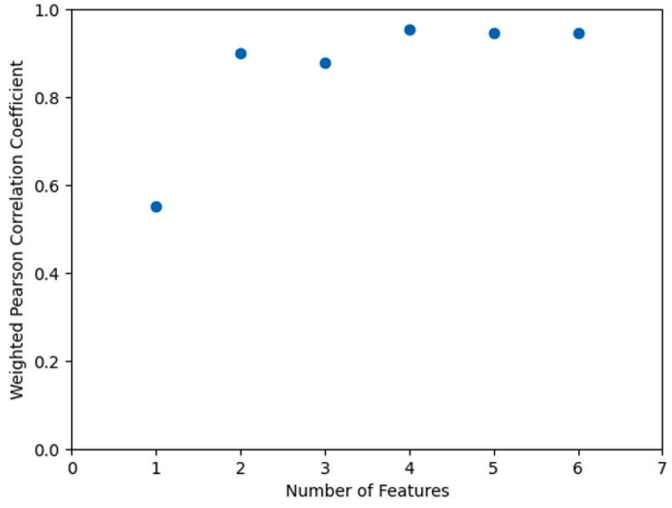

Fig. 5. U(VI)-montmorillonite model performance against the training dataset as the number of RF input features is varied. The weighted Pearson correlation coefficient is not significantly impacted after the incorporation of 4 input features.

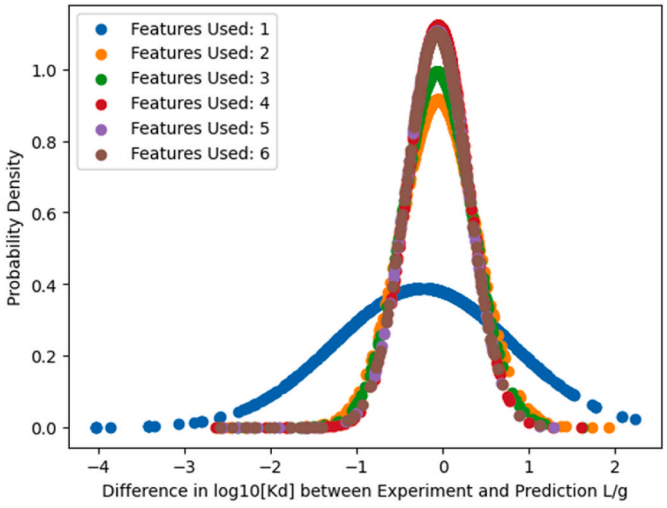

Fig. 6. Probability Density Function of Error between Experimentally Derived and Predicted K_d values for U(VI)-montmorillonite adsorption. The number of feature inputs were varied until no further improvement was observed.

may mitigate such effects in future applications. When accounting for experimental uncertainty, a Pearson correlation coefficient of the test prediction was reported as 0.96. Uncertainty associated with the raw adsorption data (e.g. electrolyte and sorbate concentration, pH, mineral concentration, and gas fugacity) were propagated into the model K_d predictions using a Monte-Carlo based method. This resulted in test K_d predictions with standard deviation errors that fell within a range of ±1 $\log_{10}(K_d)$ units of the experimental values (Fig. 7).

3.3. RF partial dependence contour plots for U(VI)-Quartz and -montmorillonite models

Low RMSE training and validation scores and a high >0.90 test weighted prediction R score allow for the RF regression modeling to be well-equipped for further model analytics that elucidate feature relationships and partial dependencies of prediction values. As part of the L-SURF work package, PDPs are generated to illustrate how the aqueous chemistry features of the RF adsorption model contribute to equilibrium aqueous U(VI) concentration predictions (Fig. 8). For the U(VI)-quartz test case, the impacts on the predicted equilibrium U(VI) aqueous concentration associated with changes in total U(VI) sorbate concentration and $UO_2(CO_3)_3^{4-}$ aqueous concentration were explored (Fig. 8a). At the high U(VI) total sorbate concentration feature region (greater than −4.5 $\log_{10}$(Molar) units), the RF regressor is dominated by impacts associated with the total U(VI) concentration. However, at lower U(VI) total sorbate concentrations, $UO_2(CO_3)_3^{4-}$ aqueous concentration becomes more influential on the model predictions. This is evidenced by the high total U(VI) sorbate concentration region experiencing minimal change as $UO_2(CO_3)_3^{4-}$ aqueous concentration varies while the low total U(VI) sorbate concentration region experiences large variability as $UO_2(CO_3)_3^{4-}$ aqueous concentration changes.

For the U(VI)-montmorillonite test case, impacts on the predicted equilibrium U(VI) aqueous concentration were observed with variable ionic strength and UO_2^{2+} aqueous concentration (Fig. 8b). The most sensitive region where both features impact the model prediction most

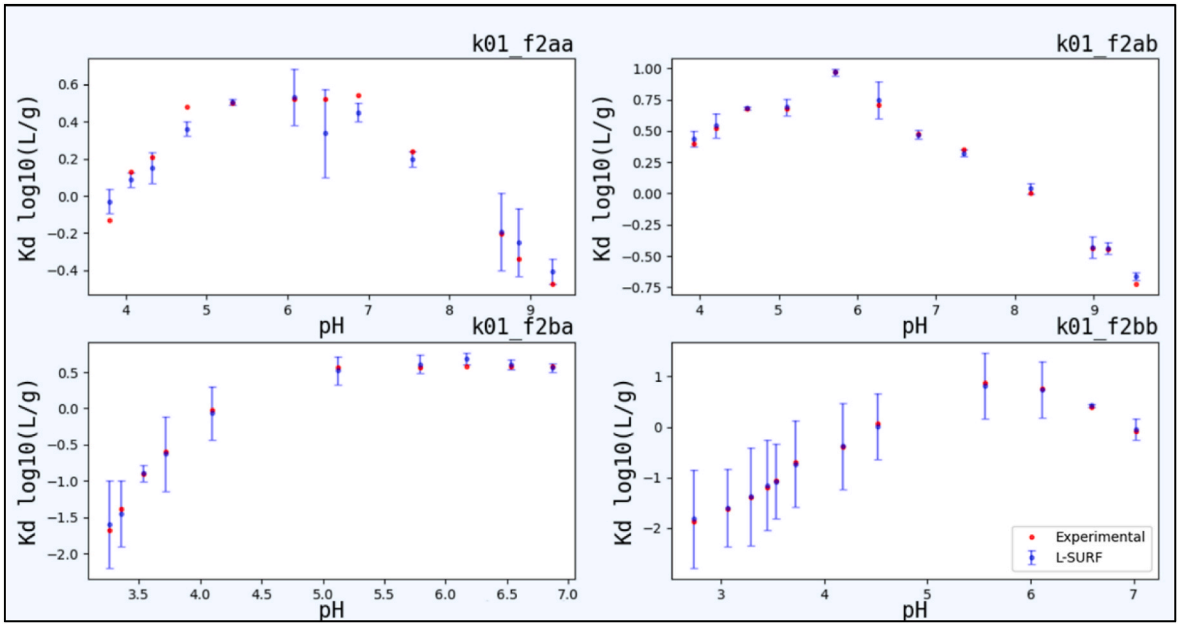

Fig. 7. Example pH-dependent model predictions of K_d values for U(VI)-montmorillonite adsorption. L-SURF modeling is represented by blue stars ±1 standard deviation error bars; experimentally measured values are indicated by red dots. Adsorption edge data are sub-divided into individual datasets from the K01 (Kim, 2001) original publication as extracted from L-SCIE (Zavarin et al., 2022). The full pH-dependent model predictions may be found in the Supplemental Information (SI.2) and utilize the following references: BB05 = (Bradbury and Baeyens, 2005); BCHSSY98 = (Boult et al., 1998); BM11 = (Bachmaf and Merkel, 2011); K01 = (Kim, 2001); KDSE04 = (Kowal-Fouchard et al., 2004); MBDSB12 = (Marques Fernandes et al., 2012); MZST95 = (McKinley et al., 1995); PTBP98 = (Pabalan et al., 1998b).

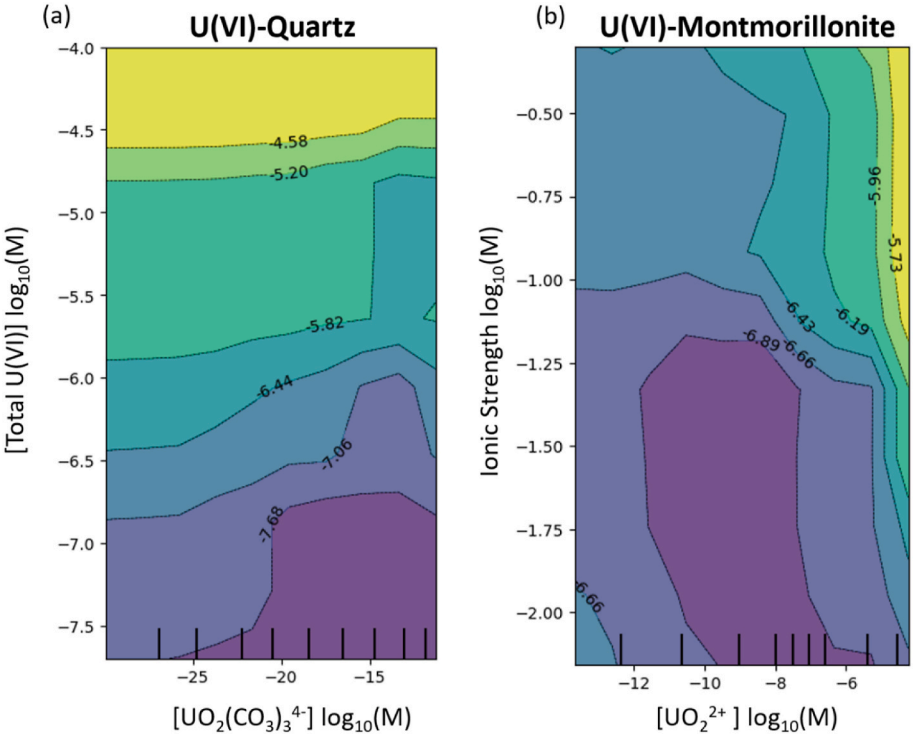

Fig. 8. Partial dependence contour plots of (a) $UO_2(CO_3)_3^{4-}$ aqueous concentration and U(VI) total sorbate concentration impacts on modelled U(VI)-quartz adsorption and (b) UO_2^{2+} aqueous concentration and ionic strength impacts on modelled U(VI)-montmorillonite adsorption. Solid black dashes on x-axis indicate deciles of the feature values (every 10th value among the full list of feature values used in the contour plot). Various colored regions indicate discretized zones of partial dependency as quantified by the $\log_{10}$ values expressed along the boundary lines.

heavily is under high ionic strength, high UO_2^{2+} aqueous concentration conditions. This is consistent with reaction (8), where UO_2^{2+} and electrolyte (Na^+) aqueous concentrations directly impact the thermodynamics of the montmorillonite ion-exchange process. Conversely, lower ionic strengths and UO_2^{2+} aqueous concentrations yield conditions with the least sensitive feature impacts on the predicted equilibrium U(VI) aqueous concentration.

While PDPs inform us of how important input features impact the model output, the true mechanisms explaining these trends may still be difficult to elucidate. For instance, in discussing U(VI)-montmorillonite ion-exchange, we cannot de-convolute whether the aforementioned trends are a result of changes in (1) aqueous species activities, (2)

electrical double layer thickness, or (3) electrolyte competition for sorption sites on the mineral surface. The PDPs presented in this study are thus more illustrative of the impacts from specific features on the RF regressor rather than an elucidation of exact sorption mechanisms occurring at the mineral-fluid interface.

3.4. L-SURF use cases and implications to other modeling tools

The hybrid framework posed by the L-SURF method enables the extraction of chemically relevant variables through the use of aqueous speciation programs such as PHREEQC. A major benefit of this approach is to efficiently translate experimental variables (e.g., pH, elemental concentrations) acquired from datasets into meaningful geochemical features (e.g., ionic strength, aqueous species concentrations). While this study focuses on the use of RF regressors as coupled with a geochemical feature engineering process, the hybrid machine learning approach is more broadly applicable to the incorporation of different regression-based models. Gaussian process models, for instance, may serve as a regression model alternative by solving for equilibrium aqueous metal concentrations as a function of the pre-determined input features. Specifically, a correlated multivariate Gaussian distribution in the multi-dimensional input feature parameter space may be used to interpolate output values (e.g., equilibrium aqueous concentrations). The covariance functions used in these models may explore several different kernels, such as the commonly used radial-basis function. An advantage of using the Gaussian process approach may include the computation of uncertainty estimates without ensemble simulations. Furthermore, the availability of multiple kernels allows for flexibility of fitting diverse sets of adsorption data as impacted by variable geochemical conditions.

The use of non-linear multivariable regressions may also serve as a valuable use case once the hybrid framework is exploited to generate chemistry-informed features. Similar to the Gaussian process method that possesses flexibility in the choice of kernel, the non-linear multivariable regression approach may be deployed by choosing from a variety of different equations, such as exponential, logarithmic, and logistic function shapes. As the RF regressor in this study is optimized by using a training dataset to conduct hyper-parameter tuning, a similar method may be executed for the quantification of non-dependent and -independent variable parameters for non-linear functions. Such explorations and optimizations of kernel or function shape are well-suited in the context of Python-based scripting and may serve as an efficient data-processing tool to understand preliminary trends among various geochemical features as determined from the aqueous speciation results of the hybrid data-driven approach.

3.5. Potential trade-offs in the use of hybrid-machine learning approaches

This study explores the use of a hybrid algorithm to quantify U(VI)-mineral sorption processes. As discussed previously, the L-SURF approach enables the flexible application of various other regression-based modeling tools to estimate fluid-mineral interactions, posing a benefit for systems that possess large amounts of sorption data under diverse geochemical conditions. Recent work further suggests the "imperative need" to quantitatively evaluate large and varied datasets to build more robust models that quantify surface reactions (Satpathy et al., 2021). The L-SURF approach explores this topic by exploiting a data unification tool (L-SCIE) and poses an alternative, data-driven means to quantify metal-mineral sorption.

An advantage of the L-SURF method that relies on high data density can be considered a limitation under conditions of data paucity. Consistent with other data-driven methods, the lack of sufficient representative data may hinder the applicability of the L-SURF approach and result in overfitting or bias towards geochemical conditions with greater data density. Additionally, because the currently described algorithm trains a RF regressor through computing RMSE values, large ranges of data input conditions may result in unintended biases as reported by the RMSE metric if proper scaling methods are not implemented. Such effects may be mitigated through, for example, the normalization of RMSE to either the mean or the range of the model output (equilibrium aqueous concentration values). While large datasets enable data-driven methods to effectively characterize trends, the RF algorithm does not lend itself to extrapolation. Furthermore, the replacement of a traditional surface complexation model with an RF regressor currently prevents the direct mechanistic interpretation of sorption modes and mechanisms (e.g., polydentate or redox transformation behaviors). Thus, the work presented herein should be considered an alternative approach for quantifying adsorption processes; one that is neither more accurate nor more precise compared to traditional surface complexation models yet enables the use of regression-based modeling techniques to estimate fluid-mineral interaction effects and may lead to more rapid application of large community datasets to reactive transport model parameterization.

3.6. Concluding remarks

The L-SURF work package has yielded successful predictions of U(VI) adsorption onto quartz and montmorillonite mineral surfaces, while suggesting distinct sorption processes of surface complexation and ion-exchange. This work is also a significant contribution in hybrid physics-integrated ML applications in the field of Earth Sciences (Reichstein et al., 2019); physical models that are known to be reliable (i.e., aqueous speciation) are combined with data-driven models for processes that have greater uncertainty (i.e., surface complexation).

This study explores L-SURF using single element-mineral interface datasets. There is a need to study real-world systems that incorporate complex mixtures of aqueous species and mineral phases. Future work will thus include incorporation of more complex features, such as the ability to distinguish between different mineral structures or adsorbate oxidation states. In addition, work will be done to integrate L-SURF into higher-order reactive transport codes as a substitute to more complex SCM approaches to adsorption and retardation. As the L-SURF method is novel in the space of adsorption modeling, the authors emphasize the need to test L-SURF rigorously across numerous different elements and minerals under varied environmental conditions. Additionally, comparisons with other SCM constructs may yield new insights into which alternative methods may serve as a more robust tool for computing non-linear K_d-based retardation factors. The work presented herein captures a first-of-its-kind method to bridge the gap between mechanistic surface complexation modeling and ML-based regression modeling, successfully achieving prediction accuracy equivalent to traditional SCMs for U(VI)-quartz surface complexation and demonstrating applicability to U(VI)-montmorillonite ion-exchange/surface complexation processes as well.

Computer code availability

L-SURF work package is made available for review under a free LLNL license: https://ipo.llnl.gov/technologies/software/l-surf. The code is authored by Elliot Chang, Linda Beverly, Sol-Chan Han, and Jadallah Zouabe. The primary developer is Elliot Chang (email: elliotc@berkeley.edu). The work package is written in the Python programming language and is accessible as a Jupyter notebook. Additional software required include the latest packages of sklearn and an executable version of PHREEQC.

Authorship contribution statement

Elliot Chang: Conceptualization, Investigation, Coding, Writing-draft, Review. Mavrik Zavarin: Conceptualization, Supervision, Funding acquisition, Writing-methodology, results/discussion, Review. Linda Beverly: Coding optimization, Writing-methods. Haruko Wainwright: Conceptualization, Supervision, Writing-editing, Review.

Code distribution: Code is available through LLNL through free licensing: (https://ipo.llnl.gov/technologies/software/lsurf

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.

Acknowledgements

Thank you to Jadallah Zouabe for early code ideas and implementations, which assisted in the inception of the L-SURF work package. Financial support was provided in part by the U.S. Department of Energy's Earth and Environmental Systems Sciences Division of the Office of Science Biological and Environmental Research program under contract SCW1053 for Lawrence Livermore National Laboratory and contract DE-AC02-05CH11231 for Lawrence Berkeley National Laboratory. This work was also supported by the Spent Fuel and Waste Science and Technology campaign of the Department of Energy's Nuclear Energy Program. This work was performed under the auspices of the U. S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. Funding was provided by the University Engagement program of LLNL's University Relations office.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.apgeochem.2023.105731.

References

Allard, B., Olofsson, U., Torstenfelt, B., Kipatsi, H., Andersson, K., 2011. Sorption of actinides in well-defined oxidation states on geologic media. MRS Online Proc. Libr. 11, 775.
Anderson, G.M., 1976. Error propagation by the Monte Carlo method in geochemical calculations. Geochem. Cosmochim. Acta 40, 1533–1538.
Appelo, C.A.J., Van Der Weiden, M.J.J., Tournassat, C., Charlet, L., 2002. Surface complexation of ferrous iron and carbonate on ferrihydrite and the mobilization of arsenic. Environ. Sci. Technol. 36, 3096–3103.
Arnold, T., Zorn, T., Bernhard, G., Nitsche, H., 1998. Sorption of uranium(VI) onto phyllite. Chem. Geol. 151, 129–141.
Arnold, T., Zorn, T., Zanker, H., Bernhard, G., Nitsche, H., 1999. Sorption behavior of U (VI) on phyllite: experiments and modeling. J. Contam. Hydrol. 47 (Medium: X; Size: vp).
Arnold, T., Zorn, T., Zänker, H., Bernhard, G., Nitsche, H., 2001. Sorption behavior of U (VI) on phyllite: experiments and modeling. J. Contam. Hydrol. 47, 219–231.
Bachmaf, S., Merkel, B.J., 2011. Sorption of uranium(VI) at the clay mineral–water interface. Environ. Earth Sci. 63, 925–934.
Boult, K.A., Cowper, M.M., Heath, T.G., Sato, H., Shibutani, T., Yui, M., 1998. Towards an understanding of the sorption of U(VI) and Se(IV) on sodium bentonite. J. Contam. Hydrol. 35, 141–150.
Bradbury, M.H., Baeyens, B., 2005. Modelling the sorption of Mn(II), Co(II), Ni(II), Zn (II), Cd(II), Eu(III), Am(III), Sn(IV), Th(IV), Np(V) and U(VI) on montmorillonite: linear free energy relationships and estimates of surface binding constants for some selected heavy metals and actinides. Geochem. Cosmochim. Acta 69, 875–892.
Breiman, L., 2001. Random Forests, Machine Learning. Kluwer Academic Publishers, Netherlands, pp. 5–32.
Chaube, S., Goverapet Srinivasan, S., Rai, B., 2020. Applied machine learning for predicting the lanthanide-ligand binding affinities. Sci. Rep. 10, 14322-14322.
Coutelot, F.M., Seaman, J.C., Baker, M., 2018. Uranium(VI) adsorption and surface complexation modeling onto vadose sediments from the Savannah River Site. Environ. Earth Sci. 77, 148.
Dalla Libera, N., Pedretti, D., Tateo, F., Mason, L., Piccinini, L., Fabbri, P., 2020. Conceptual model of arsenic mobility in the shallow alluvial aquifers near Venice (Italy) elucidated through machine learning and geochemical modeling. Water Resour. Res. 56, e2019WR026234.
Davis, J., 2001. Surface Complexation Modeling of Uranium(VI) Adsorption on Natural Mineral Assemblages.
Davis, J.A., Kent, D.B., 2018. Chapter 5. Surface complexation modeling. In: Hochella, M. F., White, A.F. (Eds.), AQUEOUS ; GEOCHEMISTRY Mineral-Water Interface Geochemistry. De Gruyter, pp. 177–260.
Dong, W., Wan, J., 2014. Additive surface complexation modeling of uranium(VI) adsorption onto quartz-sand dominated sediments. Environ. Sci. Technol. 48, 6569–6577.
Durrant, C.B., Begg, J.D., Kersting, A.B., Zavarin, M., 2018. Cesium sorption reversibility and kinetics on illite, montmorillonite, and kaolinite. Sci. Total Environ. 610–611, 511–520.
Estes, S.L., Powell, B.A., 2020. Enthalpy of uranium adsorption onto hematite. Environ. Sci. Technol. 54, 15004–15012.
Fox, P.M., Davis, J.A., Zachara, J.M., 2006. The effect of calcium on aqueous uranium (VI) speciation and adsorption to ferrihydrite and quartz. Geochem. Cosmochim. Acta 70, 1379–1387.
Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 29, 1189–1232.
Goldberg, S., 1992. Use of surface complexation models in soil chemical systems. In: Sparks, D.L. (Ed.), Advances in Agronomy. Academic Press, pp. 233–329.
Hafsa, N., Rushd, S., Al-Yaari, M., Rahman, M., 2020. A Generalized Method for Modeling the Adsorption of Heavy Metals with Machine Learning Algorithms. Water 12.
Johnson, J.W., Lundeen, S.R., 1997. GEMBOCHS Thermodynamic Datafiles for Use with the EQ3/6 Modeling Package. Lawrence Livermore National Laboratory, Livermore.
Jung, J., Hyun, S.P., Lee, J.K., Cho, Y.H., Hahn, P.S., 1999. Adsorption of UO22+ on natural composite materials. J. Radioanal. Nucl. Chem. 242, 405–412.
Karunasingha, D.S.K., 2021. Root Mean Square Error or Mean Absolute Error? Use Their Ratio as Well. Information Sciences.
Kim, S.J., 2001. Sorption mechanism of U(VI) on a reference montmorillonite: binding to the internal and external surfaces. J. Radioanal. Nucl. Chem. 250, 55–62.
Kohler, M., Curtis, G.P., Kent, D.B., Davis, J.A., 1996. Experimental investigation and modeling of uranium (VI) transport under variable chemical conditions. Water Resour. Res. 32, 3539–3551.
Kowal-Fouchard, A., Drot, R., Simoni, E., Ehrhardt, J.J., 2004. Use of spectroscopic techniques for uranium(VI)/Montmorillonite interaction modeling. Environ. Sci. Technol. 38, 1399–1407.
Leroy, P., Revil, A., 2004. A triple-layer model of the surface electrochemical properties of clay minerals. J. Colloid Interface Sci. 270 2, 371–380.
Li, C., Zarzycki, P., 2022. A computational pipeline to generate a synthetic dataset of metal ion sorption to oxides for AI/ML exploration. Front. Nuclear Eng. 1.
Lopez, A.M., Wells, A., Fendorf, S., 2021. Soil and aquifer properties combine as predictors of groundwater uranium concentrations within the central valley, California. Environ. Sci. Technol. 55, 352–361.
Marcot, B.G., Hanea, A.M., 2021. What is an optimal value of k in k-fold cross-validation in discrete Bayesian network analysis? Comput. Stat. 36, 2009–2031.
Marques Fernandes, M., Baeyens, B., Dähn, R., Scheinost, A.C., Bradbury, M.H., 2012. U (VI) sorption on montmorillonite in the absence and presence of carbonate: a macroscopic and microscopic study. Geochem. Cosmochim. Acta 93, 262–277.
McKinley, J.P., Zachara, J.M., Smith, S.C., Turner, G.D., 1995. The influence of uranyl hydrolysis and multiple site-binding reactions on adsorption of U(VI) to montmorillonite. Clay Clay Miner. 43, 586–598.
Merzlikine, A., Abramov, Y.A., Kowsz, S.J., Thomas, V.H., Mano, T., 2011. Development of machine learning models of β-cyclodextrin and sulfobutylether-β-cyclodextrin complexation free energies. Int. J. Pharm. 418, 207–216.
Nair, S., Karimzadeh, L., Merkel, B.J., 2014a. Sorption of uranyl and arsenate on SiO2, Al2O3, TiO2 and FeOOH. Environ. Earth Sci. 72, 3507–3512.
Nair, S., Karimzadeh, L., Merkel, B.J., 2014b. Surface complexation modeling of Uranium (VI) sorption on quartz in the presence and absence of alkaline earth metals. Environ. Earth Sci. 71, 1737–1745.
Newcombe, G., Drikas, M., 1997. Adsorption of NOM onto activated carbon: electrostatic and non-electrostatic effects. Carbon 35, 1239–1250.
Nguyen, X.C., Ly, Q.V., Nguyen, T.T.H., Ngo, H.T.T., Hu, Y., Zhang, Z., 2022. Potential application of machine learning for exploring adsorption mechanisms of pharmaceuticals onto biochars. Chemosphere 287, 132203.
Nitzsche, O., Merkel, B., 1999. Reactive transport modeling of uranium 238 and radium 226 in groundwater of the Königstein uranium mine, Germany. Hydrogeol. J. 7, 423–430.
Pabalan, R.T., Turner, D.R., Bertetti, F.P., Prikryl, J.D., 1998a. Chapter 3 UraniumVI Sorption onto Selected Mineral Surfaces: Key Geochemical Parameters.
Pabalan, R.T., Turner, D.R., Paul Bertetti, F., Prikryl, J.D., 1998b. Chapter 3 - UraniumVI sorption onto selected mineral surfaces: key geochemical parameters. In: Jenne, E.A. (Ed.), Adsorption of Metals by Geomedia. Academic Press, San Diego, pp. 99–130.
Parkhurst, D.L., Appelo, C.A.J., 2013. Description of Input and Examples for PHREEQC Version 3: a Computer Program for Speciation, Batch-Reaction, One-Dimensional Transport, and Inverse Geochemical Calculations. Techniques and Methods, Reston, VA, p. 519.
Pereira, R.C., Anizelli, P.R., Di Mauro, E., Valezi, D.F., da Costa, A.C.S., Zaia, C.T.B.V., Zaia, D.A.M., 2019. The effect of pH and ionic strength on the adsorption of glyphosate onto ferrihydrite. Geochem. Trans. 20, 3.
Pivovarov, S., 1998. Acid–base properties and heavy and alkaline earth metal adsorption on the oxide–solution interface: non-electrostatic model. J. Colloid Interface Sci. 206, 122–130.
Prikryl, J.D., Jain, A., Turner, D.R., Pabalan, R.T., 2001. UraniumVI sorption behavior on silicate mineral mixtures. J. Contam. Hydrol. 47, 241–253.
Ragoussi, M.E., Brassinnes, S., 2015. The NEA thermochemical database project: 30 years of accomplishments. Radiochim. Acta 103, 679–685.

Ragoussi, M.E., Costa, D., 2019. Fundamentals of the NEA Thermochemical Database and its influence over national nuclear programs on the performance assessment of deep geological repositories. J. Environ. Radioact. 196, 225–231.

Ransom, K.M., Nolan, B.T., J, A.T., Faunt, C.C., Bell, A.M., Gronberg, J.A.M., Wheeler, D. C., C, Z.R., Jurgens, B., Schwarz, G.E., Belitz, K., S, M.E., Kourakos, G., Harter, T., 2017. A hybrid machine learning model to predict and visualize nitrate concentration throughout the Central Valley aquifer, California, USA. Sci. Total Environ. 601–602, 1160–1172.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., Prabhat, 2019. Deep learning and process understanding for data-driven Earth system science. Nature 566, 195–204.

Ren, Y., Bao, H., Wu, Q., Wang, H., Gai, T., Shao, L., Wang, S., Tang, H., Li, Y., Wang, X., 2020. The physical chemistry of uranium (VI) immobilization on manganese oxides. J. Hazard Mater. 391, 122207.

Romero-González, M.R., Cheng, T., Barnett, M.O., Roden, E.E., 2007. Surface complexation modeling of the effects of phosphate on uranium(VI) adsorption. Radiochim. Acta 95, 251–259.

Satpathy, A., Wang, Q., Giammar, D.E., Wang, Z., 2021. Intercomparison and refinement of surface complexation models for U(VI) adsorption onto goethite based on a metadata analysis. Environ. Sci. Technol. 55 (13), 9352–9361. https://doi.org/10.1021/acs.est.0c07491.

Schindler, P.W., Fürst, B., Dick, R., Wolf, P.U., 1976. Ligand properties of surface silanol groups. I. surface complex formation with Fe3+, Cu2+, Cd2+, and Pb2+. J. Colloid Interface Sci. 55, 469–475.

Steefel, C.I., 2019. Reactive transport at the crossroads. Rev. Mineral. Geochem. 85, 1–26.

Tournassat, C., Tinnacher, R.M., Grangeon, S., Davis, J.A., 2018. Modeling uranium(VI) adsorption onto montmorillonite under varying carbonate concentrations: a surface complexation model accounting for the spillover effect on surface potential. Geochem. Cosmochim. Acta 220, 291–308.

Troyer, L.D., Maillot, F., Wang, Z., Wang, Z., Mehta, V.S., Giammar, D.E., Catalano, J.G., 2016. Effect of phosphate on U(VI) sorption to montmorillonite: ternary complexation and precipitation barriers. Geochem. Cosmochim. Acta 175, 86–99.

Westall, J.C., Hohl, H., 1980. A comparison of electrostatic models for the oxide/solution interface. Adv. Colloid Interface Sci. 12, 265–294.

Wheeler, D.C., Nolan, B.T., Flory, A.R., DellaValle, C.T., Ward, M.H., 2015. Modeling groundwater nitrate concentrations in private wells in Iowa. Sci. Total Environ. 536, 481–488.

Zavarin, M., Chang, E., Wainwright, H., Parham, N., Kaukuntla, R., Zouabe, J., Deinhart, A., Genetti, V., Shipman, S., Bok, F., Brendler, V., 2022. Community data mining approach for surface complexation database development. Environ. Sci. Technol. 56, 2827–2838.

Zhu, C., Hu, F.Q., Burden, D.S., 2001. Multi-component reactive transport modeling of natural attenuation of an acid groundwater plume at a uranium mill tailings site. J. Contam. Hydrol. 52, 85–108.

Zhu, X., Li, Y., Wang, X., 2019. Machine learning prediction of biochar yield and carbon contents in biochar based on biomass characteristics and pyrolysis conditions. Bioresour. Technol. 288, 121527.