LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-TR-838686

M4SF-22LL010301062-Surface Complexation Database and Hybrid ML Model Development for GDSA

M. Zavarin, E. Chang, S. Han, H. Wainwright

August 11, 2022

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

July 26, 2022

M4SF-22LL010301062-Surface Complexation Database and Hybrid ML Model Development for GDSA

M. Zavarin[1], Elliot Chang[1], Sol-Chan Han, Haruko Wainwright[2]

[1] Glenn T. Seaborg Institute, Physical & Life Sciences, Lawrence Livermore National Laboratory, 7000 East Avenue, Livermore, CA 94550, USA.
[2] Lawrence Berkeley National Laboratory, Berkeley, CA 94620, USA

DISCLAIMER

This information was prepared as an account of work sponsored by an agency of the U.S. Government. Neither the U.S. Government nor any agency thereof, nor any of their employees, makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness, of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. References herein to any specific commercial product, process, or service by trade name, trade mark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the U.S. Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the U.S. Government or any agency thereof.

Contents

1. Introduction

This progress report (Level 4 Milestone Number M4SF-22LL010301062) summarizes research conducted at Lawrence Livermore National Laboratory (LLNL) within the Argillite International Collaborations Activity Number Activity SF-22LL01030106. The activity is focused on our long-term commitment to engaging our partners in international nuclear waste repository research. The focus of this milestone is the establishment of international collaborations for surface complexation modeling and the associated impacts of unlocking larger, community-based datasets. More specifically, we are developing a database framework for Spent Fuel and Waste and Science Technology (SFWST) that is aligned with the Helmholtz Zentrum Dresden Rossendorf (HZDR) sorption database development group in support of the database needs of the SFWST program.

The FY21 effort focused primarily on building out the Access database of raw sorption data and developing a framework for surface complexation/ion exchange data fitting methods and surface complexation/ion exchange database development. Effort was coordinated with international partners involved in similar database development efforts (e.g. HZRD RES^{3}T). In FY22, significant effort has been placed in the buildout of a large set of surface complexation and ion exchange reactions constants for radionuclides and minerals of interest to nuclear waste repository performance assessment. A second effort included the expansion of our data interrogation effort by applying modern data science methods (e.g. machine learning (ML)). The combination of ML approaches with more traditional aqueous speciation calculations from PHREEQC modeling has yielded a new hybrid code that enhances GDSA efforts. By taking advantage of historically-well established aqueous speciation models of radionuclide solution chemistry and building ML models that can accept this information in a fully automated manner, we present a new hybrid ML method that successfully incorporates geochemistry knowledge. This effort has yielded a new capability to process large sets of sorption data. It is in this context that we suggest the continued support in the expansion of the surface complexation/ion exchange database work; with the growth of ‘big data’ in geochemistry, our hybrid ML model will demonstrate increased power in quantifying radionuclide-mineral partitioning (Kd values) under a variety of different geochemical conditions (e.g. pH, ionic strength, temperature).

2. Status of LLNL SCIE Sorption Database

To develop a comprehensive surface complexation database in support of the SFWST program, we continued to build a digital sorption database to support surface complexation database development. The LLNL SCIE (L-SCIE) digital sorption database and workflow was developed in Microsoft Access with a series of linked tables as reported previously. The structure of the database was recently reported in Zavarin et al. (2022) and will not be reported here. However, some enhancements to the database and workflow were pursued in FY22 and are highlighted below.

Our LLNL SCIE digital sorption database continues to expand. **At present, it includes 243 references, 2318 datasets, and 27,000 individual data points.** The database was recently linked to a large fraction of Kd data available from the JAEA Kd database. A python code was written to automate the import of JAEA Kd data into the L-SCIE workflow which increased our total data holdings to 44,000 data points. Details regarding international collaboration efforts in the development of L-SCIE are reported in the Argillite International milestone report M4SF-22LL010302062.

The L-SCIE workflow, written in R and associated with the database went through a large revision, particularly in terms of uncertainty quantification. The code no longer relies on a web hosted interface and can be run directly and locally from a web browser window. Table 1 lists the data density by element and/or species. Most data are focused on radionuclides relevant to the nuclear waste program. For example, over 8000 data are associated with uranium and over 3000 data are associated with various oxidation states of selenium. Nevertheless, the data density for some elements is quite low (e.g. technetium, iodine) for which additional data are needed. As our workflows begin to work toward a holistic analysis of radionuclide-mineral surface complexation and hybrid models, we will need to identify areas of low data density. We are presently working on python code to visualize data density across the range of radionuclides, minerals, and geochemical conditions to allow us to strategically fill in data gaps in low data density regions.

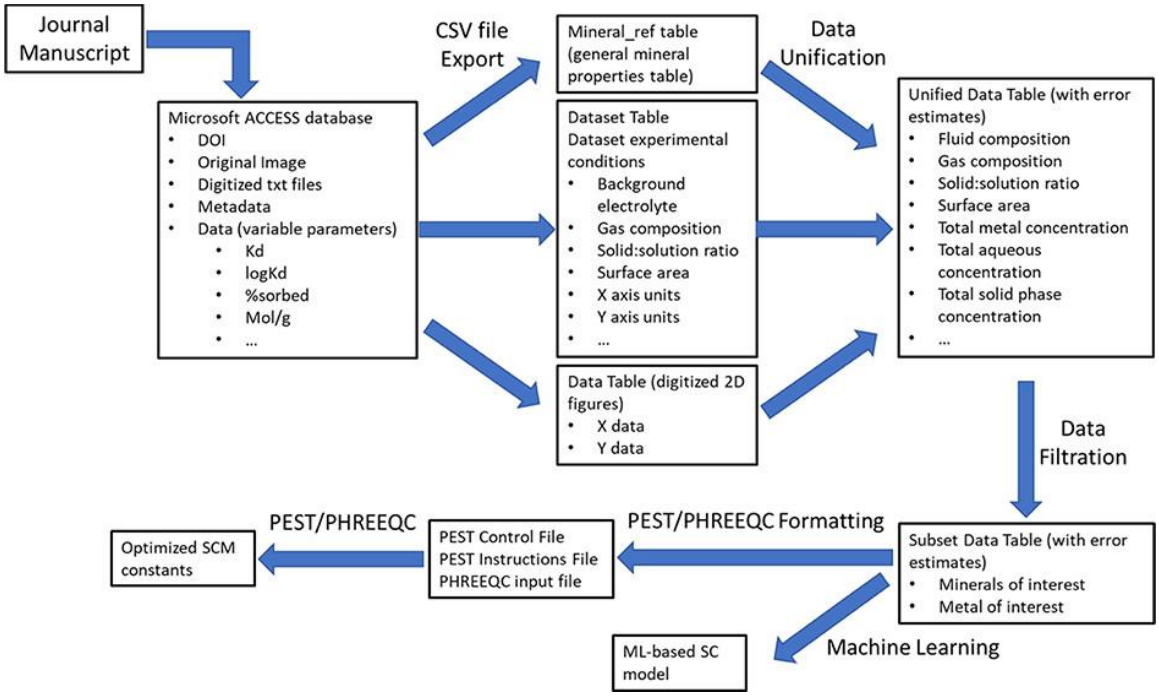
Table 1. Number of datapoints by sorbate species

Species	#	Species	#
Am(+3)	1000	P(+5)	211
As(+5)	52	Pa(+5)	117
Ba(+2)	89	Pa(n.r.)[a]	6
Bi(+3)	20	Pb(+2)	1506
C(+4)	227	Pd(+2)	16
Ca(+2)	403	Po(+4)	6
Cd(+2)	357	Pu	24
Ce(+3)	5	Pu(+4)	369
Cm(+3)	51	Pu(+5)	282
Co(+2)	510	Pu(n.r.)	93
Cr(+3)	26	Ra(+2)	152
Cr(+6)	74	Ra(+4)	3
Cs(+1)	4357	Rb(+)	87
Cu(+2)	317	Rb(+1)	7
Eu(+3)	1948	S(+6)	96
Fe(+2)	90	Sb(+5)	1
Fe(+3)	3	Se(+4)	2770
Fulvic_acid	9	Se(+6)	831
H(+1)	8150	Se(-2)	220
Humic_acid	11	Se(n.r.)	131
I(-1)	90	Sm(+3)	80
K(+1)	68	Sn(+4)	97
Mg(+2)	126	Sn(n.r.)	3
Mn(+2)	92	Sr(+2)	925
Na(+1)	6	Tc(+4)	5
Nb(+5)	8	Tc(+7)	131
Ni(+2)	3566	Tc(n.r.)	27
Np(+4)	230	Th(+4)	836
Np(+4,5)	35	U(+4)	34
Np(+5)	3189	U(+6)	8335
Np(+5,4)	2	U(n.r.)	110
Np(+6)	1	Yb(+3)	59
Np(n.r.)	92	Zn(+2)	958
		Zn(n.r.)	15
		Zr(+4)	25

[a] n.r. = not reported

3. A Community Data Mining Approach for Surface Complexation Database Development

Zavarin, M., Chang, E., Wainwright, H., Parham, N., Kaukuntla, R., Zouabe, J., Deinhart, A., Genetti, V., Shipman, S., Bok, F., and Brendler, V. 2022. A Community Data Mining Approach for Surface Complexation Database Development, Env. Sci. Technol. *56* (4), 2827-2838. https://doi.org/10.1021/acs.est.1c07109.



The information presented below is a summary from a manuscript published in January, 2022 in Environmental Science and Technology (Zavarin et al., 2022). The publication describes our L-SCIE sorption database, the workflow used to fit surface complexation constants using these data, and an overall approach to surface complexation model database development.

This manuscript presents a comprehensive data-to-model workflow, including a findable, accessible, interoperable, reusable (FAIR) community sorption database (newly developed L-SCIE database) along with a data fitting workflow to efficiently optimize surface complexation reaction constants with multiple surface complexation model (SCM) constructs. This workflow serves as a universal framework to mine, compile, and analyze large numbers of published sorption data as well as to estimate reaction constants for parameterizing reactive transport models. The framework includes (1) data digitization from published papers, (2) data unification including unit conversions, (3) data-model integration and reaction constant estimation using geochemical software

PHREEQC coupled with the universal parameter estimation code PEST. We demonstrate our approach using an analysis of U(VI) sorption to quartz based on a first L-SCIE implementation, concluding that a multi-site SCM construct with carbonate surface species yielded the best fit to community data. It yielded surface complexation reaction constants that capture all available sorption data available in the literature and provide insight into previously published reaction constants and surface complexation model constructs. The L-SCIE sorption database presented herein allows for automating this approach across a wide range of metals and minerals and implementing novel machine learning approaches to reactive transport in the future.

This work presents a comprehensive data analytics workflow for the mining of community sorption data from the literature, evaluation of published SCM constructs using a global fitting approach, and comparison of the performance metrics of different SCMs. Based on our U(VI)-quartz test case, we conclude that a uranyl carbonate species likely plays a role in uranium sorption to quartz and that multisite sorption significantly improves global data fits. However, a limitation that exists for all SCMs still remains: the non-uniqueness of SCM constructs and associated reaction constants. Nevertheless, our FAIR data approach, combined with automated workflows, provides a guide for developing surface complexation reaction databases that are flexible and easily updated as SCM constructs, thermodynamic databases, and reactive transport modeling codes evolve.

Ultimately, the work presented here provides the necessary tools and advancements in data mining and data processing to conduct novel sorption data analyses that are the ultimate goal of the FAIR data approach. The application of data-driven approaches to sorption and retardation processes could significantly increase model robustness, reduce the computational costs in modeling Earth systems, allow for uncertainty quantification and progressive model improvement when new data become available, and could help to bring consensus to the application of SCMs in reactive transport modeling.

4. A Chemistry-Informed Hybrid Machine Learning Approach to Quantify Mineral-Based Radionuclide Retardation

4.1 Introduction

The high reactivity of mineral surfaces (Dong and Wan, 2014; Durrant et al., 2018) enable radionuclides to adsorb to soils and sediments, limiting their bioavailability and influencing their overall mobility. Scientists have traditionally used surface complexation models (SCMs) 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., 2014). While SCMs calibrated to batch adsorption experiments yield valuable aqueous- and surface-speciation predictions under the investigated geochemical conditions, the implementation of SCMs also poses some key limitations. The most notable of these challenges is the non-uniqueness of SCMs that implement various divergent assumptions regarding the nature of the surface electrostatic potential. 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.

In contrast to the increasing amount of SCMs calibrated to different datasets and underpinned by divergent assumptions, machine learning (ML) methods may provide a new path forward in directly exploiting the continual growth of adsorption data available in the literature while avoiding the need to assume a specific numerical representation of electrostatics and reaction stoichiometries. 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.

The RF algorithm, acting as a black-box, data-driven approach, suffers from the inability to deduce mechanistic underpinnings of sorption processes altogether. This highlights a major distinction from quasi-mechanistic SCM constructs. The Lawrence Livermore National Laboratory-Speciation Updated Random Forest (L-SURF) model operates as a new, alternative hybrid approach. Our new method is described as a hybrid-ML approach because the initial steps involve performing thermodynamic aqueous speciation calculations while the later steps include RF-ML regression modeling of the mineral-fluid interface. Because SCMs effectively simulate aqueous speciation but 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.

4.2 Methods

4.2.1 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, a recent effort to unify community adsorption experiments and metadata in a findable, accessible, interoperable, and reusable (FAIR) format (Zavarin et al., 2022) is discussed in Section 3 of this report.

4.2.2 L-SURF Part 1: Aqueous Speciation Modeling of Raw Sorption Data

The first step of L-SURF requires an aqueous speciation database and compilation raw adsorption data for a given metal-mineral pair (Figure 1) to perform aqueous speciation calculations. Here, aqueous speciation calculations were performed using the PHREEQC software. Notably, other speciation codes may also be implemented if desired. We used the "llnl.dat" 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 and associated codes are used as the source of raw adsorption data, for data unification, and data filtering. Raw adsorption data consists of total sorbate concentration and aqueous equilibrium sorbate concentration and

associated metadata consists of gas composition, sorbent properties (concentration, surface area, reactive site density), background electrolyte concentrations, and pH conditions in a comma-separated values format (Table 2). The metadata and the total aqueous sorbate concentration are used for each sorption datapoint to create PHREEQC simulations of solution chemistry conditions. Upon the completion of aqueous speciation calculations, relevant geochemical variables are output as input features for the subsequent RF adsorption model development. For the current test case (U(VI)-quartz), the variables exported from PHREEQC output files include HCO_3^- aqueous species concentration, selected aqueous U(VI) species concentrations (e.g. UO_2^{2+} and $UO_2CO_{3(aq)}$), and ionic strength. 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 UO_2^{2+} aqueous species was chosen as a feature in an attempt to capture surface complexation of U(VI) onto quartz *via* a monodentate, inner-sphere reaction as described in Zavarin et al. (2022)

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

The $UO_2CO_{3(aq)}$ aqueous species was also tested as an input feature, replacing HCO_3^- in order to minimize covariance of carbonate-based features. $UO_2CO_{3(aq)}$ is an important species that may participate in uranyl-carbonate surface complexation with quartz:

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

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

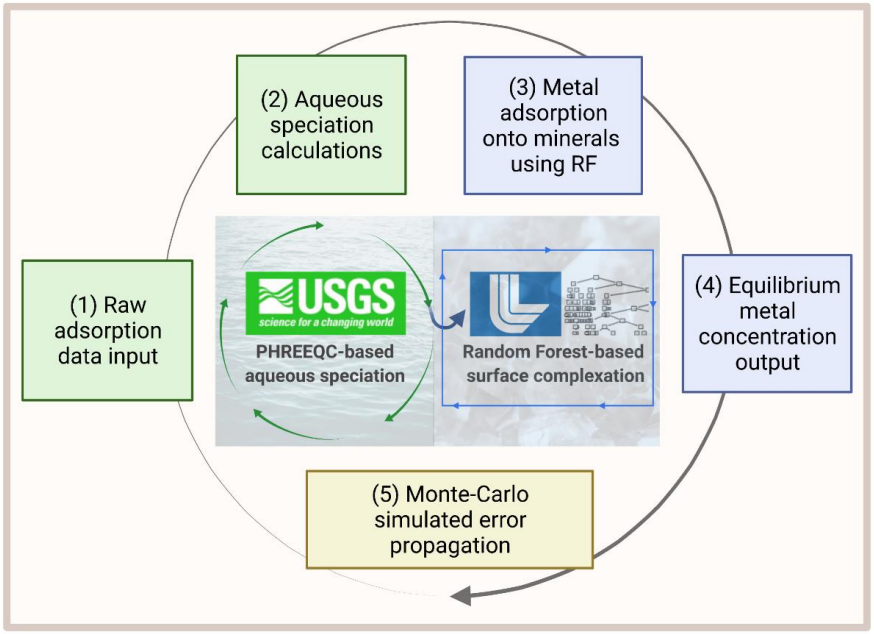


Figure 1. L-SURF workflow chart with chronological steps: (1) Adsorption data and selected thermodynamic database are imported into L-SURF module, (2) Aqueous speciation calculations are conducted and important geochemical features are output, (3) Calculated geochemical features and sorbate-sorbent metadata features are used to train and test a random forest adsorption model, (4) Equilibrium aqueous metal sorbate concentrations are output, and (5) Steps 1-4 are repeated using Monte-Carlo simulations with randomly sampled input data +/- experimentally-determined measurement uncertainty.

4.2.3 L-SURF Part 2: RF Regression of Mineral-Based Adsorption

The second step of L-SURF is executed after aqueous speciation modeling determines the equilibrium solution conditions for each individual adsorption data point. These PHREEQC speciation output variables are automatically input into the RF-ML algorithm alongside additional important metadata variables as RF features, which include total mineral site density, and mineral source (Table 2). RF is defined as an ensemble machine learning algorithm that uses a combination of tree predictors whereby individual trees are built upon a randomly and independently sampled set of training data (Breiman, 2001). The previously described metadata variables and PHREEQC output variables are pushed through the RF regression model to train, validate, and test predictions of metal-mineral surface interactions. This method was chosen as the ensemble ML algorithm because of its effectiveness in capturing non-linear relationships between various dependent variables. 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).

Table 2. Data frame used as feature inputs into RF model.

Feature #	Input feature names	Method for obtaining feature values	Input feature units
1	Mineral source	Extracted from L-SCIE metadata	Number associated with unique mineral source
2	Log_{10}(Total Sorbate Concentration)	Extracted from L-SCIE metadata	Log_{10}(Molar)
3	Log_{10}(Total Site Density)	Calculated using L-SCIE metadata: Mineral density (g/L) x Mineral surface area (m^2/g) x Mineral site density (sites/m^2)	Log_{10}(sites/L)
4	Ionic Strength	Calculated using PHREEQC speciation	Molar
5	Log_{10}(Basis species, UO_2^{2+})	Calculated using PHREEQC speciation	Log_{10}(Molar)
6	Log_{10}(HCO_3^- or UO_2CO_3 concentration)	Calculated using PHREEQC speciation	Log_{10}(Molar)

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:

$$(3)\ \hat{f}_s(x_s) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}_s(x_s, x_c^{(i)})$$

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.

4.2.4 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 assumed, and a random sample is chosen within ±*X* standard deviation of the average value, where *X* can be a specified constraint contingent on how narrow or broad the user wants to characterize the data. For the U(VI)-quartz test case, an *X* value of 1 (±1 standard deviation range) was used. 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).

4.3 Results and Discussion

4.3.1 Prediction of U(VI) Adsorption onto Quartz

Aqueous speciation calculations and metadata information (Table 2) were used as input features for the cross-validated RF model, generating a training RMSE score = 0.086 and a validation RMSE score = 0.222. A set of predictions for equilibrium aqueous U(VI) concentration was calculated based on a final test dataset, yielding a test RMSE score = 0.128. When accounting for model uncertainty, a weighted Pearson correlation coefficient R score = 0.943 was determined (Figure 2). 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), the authors present an ML method that successfully accepts aqueous chemistry based features to accurately predict U(VI)-quartz interactions. 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 density, and CO_2 gas fugacity, yielding standard deviation values for each L-SURF test prediction of equilibrium aqueous sorbate concentrations. These standard deviations are particularly useful for downstream error propagation in reactive transport modeling.

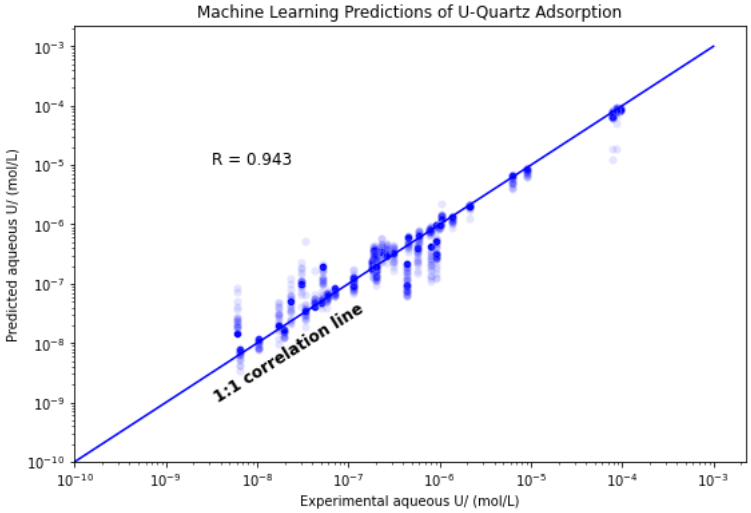


Figure 2. Correlation plot of measured equilibrium aqueous U(VI) concentration versus RF model equilibrium aqueous U(VI) concentration for the U(VI)-quartz sorption dataset. Solid blue dots indicate average prediction values; translucent blue dots are Monte-Carlo generated error propagated values.

4.3.2 RF Feature Partial Dependencies for U(VI)-Quartz Model

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 model contribute to prediction results (Figure 3). For one-way PDPs (Figures 3 a,b), the horizontal axis represents the feature of interest (x_s in equation 3) and the vertical axis represents target response of the RF-ML model, $\hat{f}$, which is the equilibrium total U(VI) concentration in solution. For two-way PDPs (Figure 3c), the vertical and horizontal axes express the relevant features of interest and the labeled contour colors represent the target response of the RF-ML model, $\hat{f}$, with changes to the two input features.

Increasing HCO_3^- aqueous species concentrations (10^{-8} to 10^{-6} M) generally result in a decreased impact on the total equilibrium aqueous U(VI) concentration output by the RF-ML model (Figure 3a). In the regime of low equilibrium UO_2^{2+} aqueous species concentration (10^{-12} to 10^{-8} M), an increase in UO_2^{2+} aqueous species concentration results in a minimal impact on the equilibrium total aqueous U(VI) in solution (Figure 3b). At higher concentrations of UO_2^{2+} aqueous species concentrations (>10^{-8} M), though, there is an increased effect on the output equilibrium total aqueous U(VI) concentration, evidenced by the dramatic increase in the partial dependence values. The contour PDP (Figure 3c) allows for visualization of the parameter space where there is greatest model response from input features, HCO_3^- and UO_2^{2+} aqueous species concentrations. Notably, the model is most sensitive to the feature interactions at high UO_2^{2+} and low HCO_3^- species concentrations. More specifically, under these concentrations, the model response is influenced significantly by the high UO_2^{2+} aqueous species. This is consistent with surface complexation modeling results, where equilibrium total aqueous U(VI) concentrations are impacted by the presence of a high UO_2^{2+} species concentration (reaction described in equation 1). Additionally, the model output is generally less sensitive to changes in UO_2^{2+} aqueous species concentrations at higher HCO_3^- aqueous species concentrations. This can be explained by the abundance of HCO_3^- aqueous species (10^{-5} to 10^{-6} M) with only a limited amount of UO_2^{2+} species (10^{-7} to 10^{-10} M) to form uranyl carbonate surface species (reaction described in equation 2). PDPs discussed herein demonstrate how a hybrid ML approach that integrates aqueous speciation calculations allows for the elucidation of feature impacts on model output.

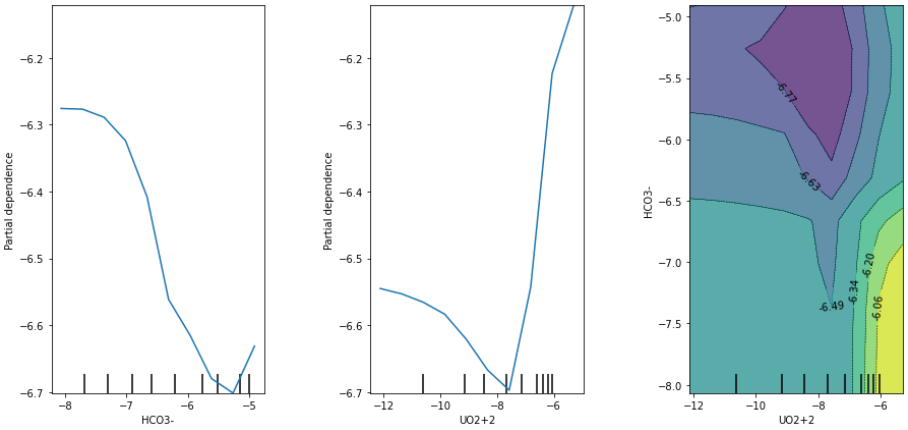


Figure 3. Partial dependence plots of (a) HCO_3^- aqueous species concentration and (b) aqueous UO_2^{2+} species concentration on equilibrium total aqueous U(VI) concentration, and (c) associated contour plot representation of partial dependencies. HCO_3^-, UO_2^{2+}, and contour legends expressed as $\log_{10}$(mol/L).

4.3.3 Improving L-SURF Predictions with Optimized Aqueous Species Features

When incorporating HCO_3^- aqueous species concentration as an RF input feature, the L-SURF prediction standard deviations noticeably increased at neutral to alkaline pH (Figure 4). CO_2(g) liquid-gas exchange and aqueous speciation will favor the presence of CO_3^{2-} aqueous species in solution and lead to the formation of uranyl carbonate complexes, particularly in alkaline conditions. To test the impact of our choice of RF input feature, $UO_2CO_{3(aq)}$ aqueous species concentration was used in lieu of HCO_3^- aqueous species concentration. Using $UO_2CO_{3(aq)}$ as a feature, the L-SURF predictions were significantly more stable with much lower model uncertainties at neutral to alkaline pH (Figure 4). This is a significant result as it aligns with traditional SCM approach modeling interpretation (Zavarin et al., 2022): Uranyl-carbonate species likely play a major role in U(VI) adsorption onto quartz surfaces. In the context of hybrid ML modeling, this result also signifies that incorporation of deeper aqueous speciation knowledge into the RF model can result in significantly improved propagated error under specific geochemical conditions and may provide additional insight into the solution conditions and aqueous species that influence adsorption.

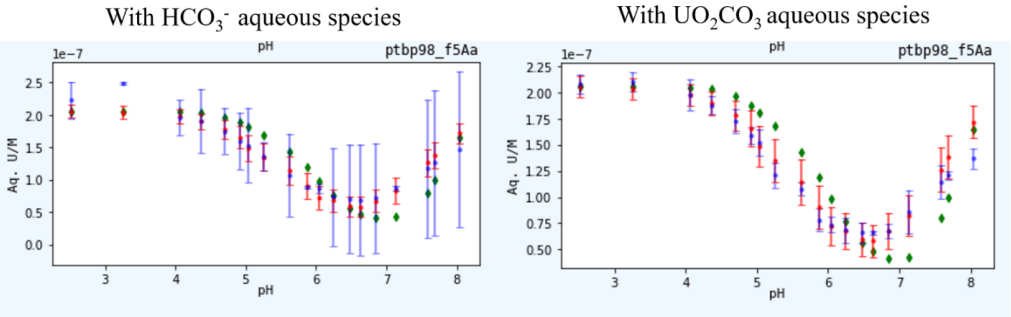


Figure 4. pH-dependent model predictions of equilibrium aqueous U(VI) concentrations for U(VI)-quartz adsorption for one literature source extracted dataset (Pabalan et al., 1998). PEST-optimized PHREEQC modeling is represented by green diamonds; L-SURF modeling is represented by blue stars ± 1 standard deviation error bars; experimentally measured values are indicated by red dots ± 1 standard deviation error bars. Improved propagated error bounds are visible under mid- to high- pH regimes as a result of replacing HCO_3^- species with a $UO_2CO_{3(aq)}$ species as an RF input feature.

4.3.4 L-SURF Work Package and Concluding Remarks

The L-SURF work package has yielded successful predictions of U(VI) adsorption onto the quartz mineral surface. Despite its strong predictive quality as a modeling tool, L-SURF currently focuses on 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 into the work package, 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. Importantly, an analysis is needed to evaluate

the computational burden of traditional and hybrid ML approaches and whether L-SURF can streamline the incorporation of adsorption data into reactive transport modeling, including uncertainty quantification. As the L-SURF hybrid ML approach 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. Nonetheless, 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 high prediction accuracy for the U(VI)-quartz system.

5. Planned FY23 Efforts

Here, we described a detailed analysis of U(VI) sorption to quartz through both traditional surface complexation modeling and through a hybrid ML framework. In FY23, effort will be placed on the continued growth of the surface complexation and ion exchange database (L-SCIE) in order to assess mineral-based radionuclide retardation under a wider variety of geochemical conditions (e.g., ionic strength, varying electrolyte compositions). This effort will also include the testing of L-SURF to quantify K_d values for numerous different radionuclide-mineral pairs under varying geochemical conditions. Additionally, we will conduct direct comparisons of the L-SURF approach with various surface complexation (Non-electrostatic, diffuse layer, etc.) and ion exchange (Vanselow, Gapon, etc.) models. These important model inter-comparisons will provide a clearer path forward in incorporating traditional SCM or/and modern hybrid ML approaches into repository performance assessment.

Key considerations for future modeling development will include (1) the ability to reduce computational burden on determining retardation coefficients for PA and (2) a new capability to quantify and predict radionuclide-mineral partitioning at a more efficient, rapid pace due to automated workflows. Upon the careful consideration of the most effective modeling approach to implement for PA, we will discuss efforts for the actual implementation of our surface complexation model or/and hybrid ML approaches into PA models. FY23 will see both the fine-tuning of our model developments and also defined requirements discussed to apply these model learnings to higher level GDSA work.

6. Acknowledgments

This work was supported by the Spent Fuel and Waste Science and Technology campaign of the Department of Energy's Nuclear Energy Program. Prepared by LLNL under Contract DE-AC52-07NA27344.

7. References

Anderson G. M. (1976) Error propagation by the Monte Carlo method in geochemical calculations. *Geochimica et Cosmochimica Acta* **40**, 1533-1538.

Appelo C. A. J., Van Der Weiden M. J. J., Tournassat C. and Charlet L. (2002) Surface Complexation of Ferrous Iron and Carbonate on Ferrihydrite and the Mobilization of Arsenic. *Environmental Science & Technology* **36**, 3096-3103.

Breiman L. (2001) Random Forests. In *Machine Learning*. Kluwer Academic Publishers, Netherlands. pp. 5-32.

Dong W. and Wan J. (2014) Additive Surface Complexation Modeling of Uranium(VI) Adsorption onto Quartz-Sand Dominated Sediments. *Environmental Science & Technology* **48**, 6569-6577.

Durrant C. B., Begg J. D., Kersting A. B. and Zavarin M. (2018) Cesium sorption reversibility and kinetics on illite, montmorillonite, and kaolinite. *Science of The Total Environment* **610-611**, 511-520.

Friedman J. H. (2001) Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics* **29**, 1189-1232.

Goldberg S. (1992) Use of Surface Complexation Models in Soil Chemical Systems. In *Advances in Agronomy* (ed. D. L. Sparks). Academic Press. pp. 233-329.

Johnson J. W. and Lundeen S. R. (1997) GEMBOCHS thermodynamic datafiles for use with the EQ3/6 modeling package. Lawrence Livermore National Laboratory, Livermore.

Lopez A. M., Wells A. and 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.

Nair S., Karimzadeh L. and Merkel B. J. (2014) Surface complexation modeling of Uranium(VI) sorption on quartz in the presence and absence of alkaline earth metals. *Environmental Earth Sciences* **71**, 1737-1745.

Nguyen X. C., Ly Q. V., Nguyen T. T. H., Ngo H. T. T., Hu Y. and Zhang Z. (2022) Potential application of machine learning for exploring adsorption mechanisms of pharmaceuticals onto biochars. *Chemosphere* **287**, 132203.

Pabalan R. T., Turner D. R., Paul Bertetti F. and Prikryl J. D. (1998) Chapter 3 - UraniumVI Sorption onto Selected Mineral Surfaces: Key Geochemical Parameters. In *Adsorption of Metals by Geomedia* (ed. E. A. Jenne). Academic Press, San Diego. pp. 99-130.

Pereira R. C., Anizelli P. R., Di Mauro E., Valezi D. F., da Costa A. C. S., Zaia C. T. B. V. and Zaia D. A. M. (2019) The effect of pH and ionic strength on the adsorption of glyphosate onto ferrihydrite. *Geochemical Transactions* **20**, 3.

Ragoussi M. E. and Brassinnes S. (2015) The NEA Thermochemical Database Project: 30 years of accomplishments. *Radiochim Acta* **103**, 679-685.

Ragoussi M. E. and Costa D. (2019) Fundamentals of the NEA Thermochemical Database and its influence over national nuclear programs on the performance assessment of deep geological repositories. *Journal of Environmental Radioactivity* **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. and 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.

Wheeler D. C., Nolan B. T., Flory A. R., DellaValle C. T. and 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. and Brendler V. (2022) Community Data Mining Approach for Surface Complexation Database Development. *Environmental Science & Technology* **56**, 2827-2838.