LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-JRNL-827421

A Global Data Mining Approach to Surface Complexation Database Development

M. Zavarin, E. Chang, H. M. Wainwright, N. Parham, R. Kaukuntla, J. Zouabe, A. Rainier, V. Genetti, S. Shipman, F. Bok, V. Brendler

October 1, 2021

Environmental Science and Technology

Disclaimer

A Community Data Mining Approach for Surface Complexation Database Development

Mavrik Zavarin[1*], Elliot Chang[1*], Haruko Wainwright[2,3], Nicholas Parham[1], Rahul Kaukuntla[1], Jadallah Zouabe[1,4], Amanda Deinhart[1], Victoria Genetti[1], Sam Shipman[1], Frank Bok[5], and Vinzenz Brendler[5]

[1] Seaborg Institute, Lawrence Livermore National Laboratory (zavarin1@llnl.gov, chang58@llnl.gov)

[2] Lawrence Berkeley National Laboratory, Earth and Environmental Sciences Area (hmwainwright@lbl.gov)

[3] U.C. Berkeley, Department of Nuclear Engineering (harukom@berkeley.edu)

[4] U.C. Berkeley, Department of Chemical Engineering (jzouabe@berkeley.edu)

[5] Helmholtz-Zentrum Dresden-Rossendorf e.V., Institute of Resource Ecology (F.Bok@hzdr.de)

*co-first authors

ABSTRACT

This paper 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.

INTRODUCTION

Sorption processes – reactions at the solid-water interface – play a key role in nutrient cycling and contaminant migration in the environment [1]. Traditionally, sorption has been represented by an empirical linear distribution coefficient (K_d), which is the ratio between the solid-phase and aqueous-phase concentrations. However, the K_d approach is limited by the fact that fluid composition, surface site saturation, and competitive effects are not addressed. In the past thirty years, surface complexation models (SCMs) and ion exchange models have been developed to provide a more robust mechanistic description of sorption reactions. These models are directly coupled to thermodynamic models of aqueous speciation and solubility and can account for dynamic changes in solid and fluid composition, site saturation, and competitive effects. These models have become an integral part of reactive transport models (RTMs) that describe contaminant transport or biogeochemical cycling [2]. A recent review of manuscripts published in the ACS journal Environmental Science & Technology has revealed that sorption remains one of the most highly used words in the last twenty year history of this journal (second only to the term "water") [3]. A Web of Science search on the terms "sorption" and "adsorption" suggests that interest in sorption phenomena across all scientific disciplines continues to grow exponentially with nearly 100,000 manuscripts published in 2020 (Supplementary Information Figure SI.1). How might we capitalize on (i.e. mine) this rich source of data to improve RTM fidelity and effectively quantify model uncertainty?

A large number of SCM constructs have been proposed in the literature with divergent assumptions. While some studies implement multi-site SCMs that account for sorption-based accumulation of surface charge such as the CD-MUSIC approach [4-7], others implement

models without electrostatic characterization of the electrical double layer (the non-electrostatic approach) [8-10]. The presence of electrical double layers have been implemented in the form of constant capacitance [11, 12] and triple layer models [13, 14], whereby linear surface charge-to-surface potential relationships have been created based on fitted sorption data. Such SCM constructs, however, have often been an obstacle for developing RTMs, since modelers do not have an objective way to evaluate and integrate individual SCM constructs reported in the literature. Westall and Hohl [15] discovered this challenge first when fitting various SCM constructs to sorption data, concluding that constants determined from these fitting exercises are indeed model dependent. Most published studies focus on data collected by a particular researcher or institution for a particular combination of aqueous species and solid phases with little regard for previously published data or SCM constructs. Unified analysis of large sorption datasets has rarely been performed due to the lack of access to digitized literature data and the lack of common standards and approaches to archiving these data types (i.e. findable, accessible, interoperable, reusable "FAIR" data[16]).

Recently, the material and chemical sciences have had numerous successful demonstrations of community data collection across multiple institutions and the synthesis of data for scientific discoveries using machine learning (ML). For example, Chen et al. [17] developed a ML method to combine data from multiple sources of measurements and simulations to predict material properties. Natural language processing (NLP) has also been used to find related papers and synthesize insights. In another ongoing case, the US Department of Energy's Material Project (materialsproject.org) currently hosts over 130,000 distinct inorganic compounds, with over

125,000 users from around the world, in an effort to predict the functions of materials for engineering and experimentation purposes.

There have been some historical efforts to manually compile sorption data and develop SCM databases. For example, Dzombak and Morel [18] developed a diffuse layer SCM database for the sorption of 22 elements to the mineral hydrous ferric oxide (HFO). A follow-on effort by Karamalidis and Dzombak [19] as well as Mathur and Dzombak [20] developed similar databases for gibbsite and goethite, respectively. The HFO SCM database was developed by critically evaluating 169 literature references (109 references were rejected due to data quality issues). Manual digitization of the 60 high quality references yielded an estimated ~1910 batch sorption data which were fit to a 2-site diffuse layer SCM. The HFO surface area, reactive site densities, and surface acidity constants were held constant and based on a critical review of HFO characterization data. To this day, this database remains the most widely used SCM database for predicting metal interaction with HFO. Nevertheless, it is limited in that the database is static (no updates) and reactive transport models that use this database of reaction constants are constrained by the original choice of thermodynamic data and SCM construct employed by the authors (i.e. 2-site diffuse layer model, 0.2 mol-sites/mol-Fe and 0.005 mol-sites/mol-Fe, respectively, pK_{a1} = -7.29, pK_{a2} = -8.93).

The RES^{3}T database (www.hzdr.de/res3t) [21] is an alternative approach to SCM database development. RES^{3}T is an open-source database of SCM reaction constants and associated meta-data (i.e. experimental conditions, DOI references, etc) compiled from the literature and mainly related to radioactive waste management, nuclear decommissioning, environmental remediation

as well as NORM sites (SI.1). At present, this database contains information from 3172 literature references that include reactions between 147 minerals and 148 ligands and a total of 7062 surface complexation reaction constants that span across all known SCMs. However, the RES^{3}T database does not include the raw sorption data contained in these references and, as such, does not allow for refitting of sorption data to a particular SCM construct of interest. The compilation of reaction constants is the common approach taken in traditional thermodynamic database developments (e.g. aqueous speciation, solubility, etc.) for which data treatment and reaction parameterization is relatively well established in the scientific community [22-25]. However, SCMs have not yet reached that level of scientific consensus with many SCMs having divergent assumptions as mentioned above. Importantly, the reaction constants are not transferable from one SCM construct to another.

These past developments in SCM databases suggest the importance of compiling raw sorption data so that multiple SCM constructs can be evaluated across the whole community dataset. In addition, well-curated metadata are critical, since, for example, SCM reaction constants are impacted not only by the choice of SCM construct but also by the choice of aqueous speciation data and reactive site density. If a framework to evaluate multiple SCM constructs against the community data were available, we could more readily develop a self-consistent set of SCM reaction constants in a database form for use in reactive transport codes. Based on the data density in the Dzombak and Morel database (34 data per reference), we can estimate that the RES^{3}T database alone may provide on the order of 108,000 individual sorption data (if raw sorption data were included in the database). Indeed, we hypothesize that the narrow scope of the RES^{3}T database effort suggests that historical sorption data in the literature is likely to be on the

order of millions to tens of millions of data (Figure SI.1). Thus, SCM database development and future applications of ML and hybrid ML-SCM approaches is not limited by the paucity of data, but, rather, the paucity of FAIR data.

In this study, we developed a comprehensive data-to-model workflow for the development of a self-consistent SCM database, including a community database of digitized raw sorption data and a data fitting workflow to optimize the reaction constants based on multiple SCM constructs. A particular innovation is to automatically evaluate all available community ligand-mineral data in a comprehensive manner, yielding a set of reaction constants (and associated uncertainties). Such automated procedures not only permit the evaluation of new/extended SCM constructs but also allow for relatively routine re-fitting as new data sets become available. We demonstrate the power of this data-to-model workflow using U(VI) sorption to quartz as a case study. Based on these FAIR community sorption data, we test various U(VI)-quartz SCMs reported in the RES^{3}T database and compare various SCM constructs.

MATERIALS AND METHODS

The need to develop SCM databases, in concert with traditional thermodynamic models (for aqueous speciation and mineral solubility), for reactive transport modeling and nuclear waste repository performance assessment in particular, was identified many years ago [26]. This issue was expressly identified in Nuclear Energy Agency (NEA) sorption project reports [27, 28] and associated publications [29]. However, significant progress on this issue has been made only recently in various international nuclear waste repository programs (e.g. [30], [31], [32]). Hybrid approaches have also been attempted [33]. The best path forward for developing such databases

has remained an open question [32], particularly in cases where generic repositories are being investigated resulting in a need to model radionuclide behavior over a very broad range of solution and mineralogic conditions. The approach presented herein is based on principles of FAIR data that allow a SCM database to evolve along with updates in thermodynamic data, SCM constructs, and reactive transport modeling codes. Figure 1 presents an overview of our data mining/data driven workflow for SCM database development.

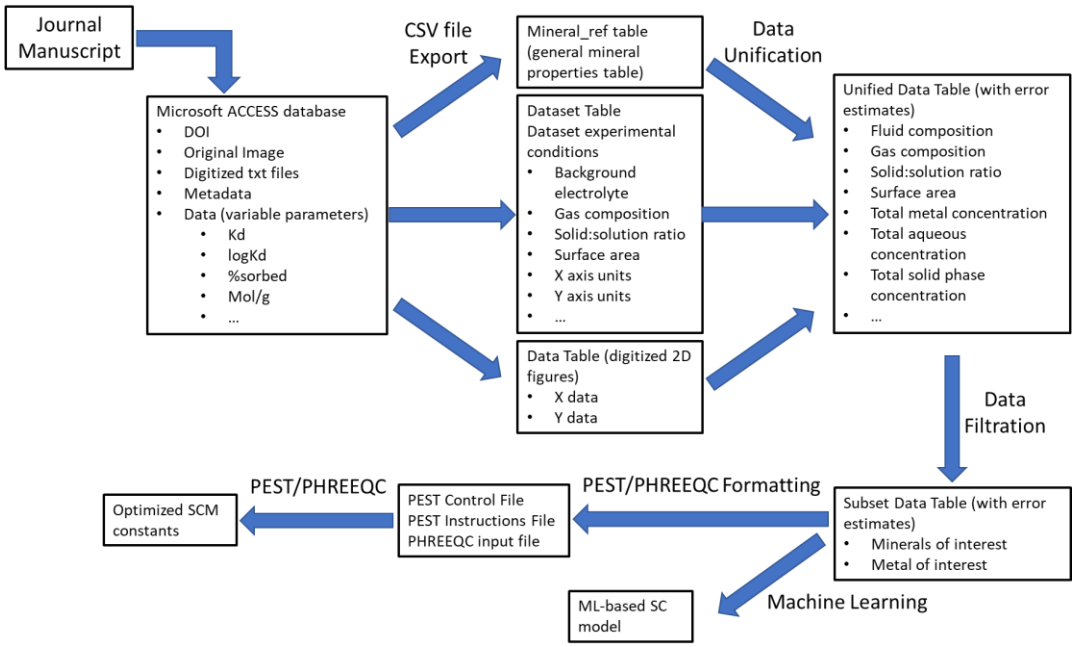

Figure 1. Workflow describing the process from literature data digitization, incorporation into the LLNL SCIE database, unit conversion and data unification, data filtration, data formatting, the SCM reaction constant fitting.

The L-SCIE Database

In coordination with our RES^{3}T partners, we have focused on developing a data digitization pipeline that would merge the approach pioneered by Dzombak and Morel (i.e. data mining, digitization, and comprehensive fitting of raw data) with the SCM reaction constant mining efforts in RES^{3}T (i.e. reaction constant mining and metadata digitization) to develop a robust approach to database development that is amenable to any SCM construct as well as ML and

hybrid ML approaches in the future. To date, the manual digitization of sorption data has yielded a LLNL Surface Complexation/Ion Exchange (L-SCIE) database that includes 214 references and a total of 23,106 individual digitized raw sorption data and associated metadata. Importantly, this approach to manual data mining is labor intensive. However, ML can be used to streamline the data mining operation in the future and the L-SCIE database could be used to train ML mining algorithms [34].

The L-SCIE database stores raw digitized sorption data from the literature, with a particular focus on references contained in RES^{3}T. At present, the L-SCIE database is stored as a Microsoft Access relational database with a series of linked tables. The tables include a Reference Table, Dataset Table, Atmospheric CO_2 Table, Mineral Site Density Table, and Data Table (the contents of each table are described in SI.2).

1. *Data mining process*

The first step in mining raw sorption data is to identify target references for digitization. Target references were identified primarily through the RES^{3}T database which also provides metadata (e.g. mineral surface area) and reported surface complexation reaction constants. The metadata are primarily recorded in the L-SCIE Dataset Table. The raw data contained in the reference (typically in the form of X-Y figures) are digitized using the DataThief software [35] which allows the user to capture data reported in plots, *including data uncertainties*. Most sorption data figures are reported as percent sorbed vs. pH. However, some data are reported in terms of K_d (mL/g), K_a (m^2/g), or simply equilibrium aqueous concentration. Regardless of the units included in the plotted data, the DataThief software captures the raw information contained in the plot and the raw data are imported into the L-SCIE database *as-is*. The capture of data uncertainty is limited by the presence or absence of error bars on the figures. If error bars are included, the data

uncertainties can be captured directly. If error bars are not included, an estimate of the data uncertainty can be made by the user based on expert judgement and/or information provided in the reference. Default data uncertainties and associated unit conversion and error propagation is reported in Table SI.3. We note that our data mining process does not explicitly address the critical evaluation of data quality; this critical evaluation must be performed prior to importing data into L-SCIE.

2. Data Unification

Once raw data are imported, a series of unit conversion and uncertainty estimates are necessary to make data amenable to SCM data fitting. Automating this process is essential. The data unification processes were scripted using the R environment [36]. The goal is to convert raw data contained in the L-SCIE database into a unified .CSV dataset with standard parameter units (mol/L is preferred unit). The most challenging aspect of this is the conversion of various types of X-Y plots into representative values of aqueous and sorbed concentrations.

An example of equations used to convert K_d, percent sorbed, and fraction sorbed parameters into aqueous and sorbed concentrations is presented in SI.3. While unit conversion and error propagation is straightforward for these parameters, certain error propagations are challenging (e.g. conversion of reported log K_d uncertainties to aqueous concentration uncertainties). The plots digitized from the literature include as many as 38 different parameters and units that represent values associated with X-Y plots. A complete set of calculations required to convert the L-SCIE database to a dataset with uniform parameter units and associated uncertainty estimates is reported in Table SI.3. Upon completion of the unit conversion and uncertainty calculations, a unified database of raw sorption data is produced as a .CSV file (L-SCIE database

and codes available through LLNL distribution: https://ipo.llnl.gov/technologies/software/llnl-surface-complexation-database-converter-scdc).

3. SCM development

The end goal of this effort is the development of a self-consistent database of surface complexation reaction constants for use in reactive transport codes. As such, the digitization of community sorption data is simply the first step. To achieve the end goal, community sorption data captured in the L-SCIE database must be fitted to SCM constructs to produce reaction constants. To this end, scripts were developed to extract and format the relevant data as input into the PEST parameter estimation code [37] that is linked to the SCM algorithms contained in the PHREEQC code [38]. The details of the input files will not be described here. However, it should be noted that all data processing was written in R to facilitate the data fitting workflow (all code is available through the LLNL distribution and the associated github site). These formatted data are passed over to PHREEQC/PEST data fitting routines (and ML applications not discussed here) and resuts visualized using scripts written in Python.

To perform the SCM reaction constant estimation, PHREEQC requires a thermodynamic database to simulate the aqueous speciation and precipitation of mineral phases at each solution condition for each sorption data point captured in the L-SCIE database. The PHREEQC package contains a thermodynamic database based on LLNL's SUPCRT[22] database that was supported by the US nuclear waste repository program at Yucca Mountain. More recently, the NEA-TDB effort [23, 24, 29] is providing updates to thermodynamic reaction constant data for radionuclides relevant to nuclear waste repository science. Our SCM reaction constant fitting workflow can be integrated with thermodynamic reaction constant database updates as they become available to provide a self-consistent set of reactions both from the standpoint of

speciation and solubility as well as reactions at the mineral-water interface, as we illustrate below. As new thermodynamic data become available, community sorption data captured in the L-SCIE database can be routinely re-fit to provide a consistent representation of aqueous speciation, solubility, and mineral-water interface reaction constants (Figure 1). While this work is focused on automating the development and updating of surface complexation databases, we note that a similar approach may be warranted for thermodynamic reaction constant database development as well.

All PEST optimizations were performed by minimizing the difference between the measured and modeled aqueous metal concentration in solution. Importantly, a weighting factor was assigned to each datapoint based on the reciprocal of the standard deviation of the aqueous concentration (e.g. 1/SD). The application of a weighting factor is essential to provide a balanced relevance to measured concentrations that range over many orders of magnitude. It is also essential to addressing uncertainties in measured values that arise from a combination of measurement error and data digitization error. Two parameters are examined from the standpoint of "goodness of fit". The first is the sum of squared weighted residuals (phi term) that is the objective function in PEST:

$$phi = \sum(c_i w_i - c_{0i} w_i)^2 \quad (1)$$

where c_i and c_{0i} are the measured and simulated aqueous concentration for the i'th observation and w_i is the weight associated with the i'th observation (1/SD here). When w_i is the uncertainty in the value c_i, the *phi* term is equivalent to weighted sum of squares (WSOS) parameter that is commonly reported in surface complexation data fitting [18].

The second parameter that provides a general assessment of "goodness of fit" is the correlation coefficient (R):

$$R = \frac{\sum(c_i w_i - m)(c_{0i} w_i - m_0)}{(\sum(c_i w_i - m)^2 \sum(c_{0i} w_i - m_0)^2)^{1/2}} \quad (2)$$

where m and m_0 are the mean value of weighted aqueous concentration and mean value of weighted simulated aqueous concentration, respectively. The correlation coefficient is valuable in comparing various models as it is independent of the number of observations in the model run. The correlation coefficient R is equivalent to a *weighted* Pearson correlation coefficient that accounts for the uncertainties associated with individual observations. A value of 0.9 is considered an acceptable fit to the data [37].

RESULTS AND DISCUSSION

The L-SCIE database

General statistics for data included in the L-SCIE database are reported in Table SI.4. Data (>23,000 datapoints) capture sorption to individual minerals (i.e. binary sorption experiments) and includes potentiometric titration data. The sorbing species include those that are common to nuclear waste performance assessment (i.e. Am, U, Np, Pu, Se, Sr, Cs, etc.) as well as others (i.e. Cr, Cu, P, Rb) (Table SI.5). Major cation/anion sorption data are included as they are an essential component of the ion exchange process (not discussed here) or may contribute to competitive surface sorption processes (e.g. Ca, Mg, K, Na, S).

Sorption experiments are typically performed in simple electrolytes (e.g. $NaClO_4$, NaCl) and across a range of ionic strength. The full list of electrolytes and the associated concentration range is given in Table SI.6. The HCO_3(-) electrolyte is a special case in that concentrations may be controlled either by direct addition to solution or by controlling the $CO_{2(g)}$ fugacity. Thus, either the solution or the gas phase concentrations were captured. When experiments are performed under atmospheric $CO_{2(g)}$ conditions, the L-SCIE database accounts for the general

trends in atmospheric $CO_{2(g)}$ concentrations over the past century. The minerals included in the L-SCIE database are listed in Table SI.7 which includes the source/type of sorbent material. The source information can be used to identify potential variations in data that result from characteristics of particular source materials (e.g. comparison of SWy-1 montmorillonite, SAz-1 montmorillonite, and synthetic montmorillonite).

U(VI) sorption to quartz – Digitized Data

To check the feasibility of using a FAIR sorption database (i.e. L-SCIE) to build a self-consistent SCM database, a test-case of U(VI) sorption to quartz was evaluated. The RES^{3}T database contains 12 publications [39-50] in total related to this binary system (Table SI.9) and reports 37 different surface complexation reactions constants based on a number of different SCM constructs (diffuse layer, non-electrostatic, and triple layer models; one-site, multisite sorption; variable site concentrations; etc.). The reaction constants from a single reference may include multiple fitting exercises performed on all or a subset of data included in that reference. For example, for reference DK01 [42], 11 reaction constants are reported. Upon detailed examination of this reference, the 11 reaction constants are associated with four separate data fitting exercises in which the authors chose particular subsets of data and SCM constructs. The four separate fitting exercises and associated reactions constants are identified in Table 1 (DK01_1, DK01_3, DK01_4, and DK01-c4).

Table 1. Fitting results using 17 RES^{3}T SCM constructs combined with the PHREEQC "llnl.dat" database updated with NEA-TDB U(VI) complexes from their online database https://www.oecd-nea.org/jcms/pl_37223/electronic-database-of-the-tdb-project. "Orig" and "Fit" keywords refer to published reaction constants from the RES^{3}T database and optimized values using the L-SCIE database, respectively; "w" and "s" keywords represent weak and strong affinity sites.

SCM Type	Surface Area m^2/g	Site Density sites/nm^2	Orig pK_1	Orig pK_2	Species	Orig logK	SD[a]	Fit logK	Fit SD	Orig R	Fit R	Orig WSOS	Fit WSOS	Model Reference[c]
DDL	0.2	4.81		5.62	>Si-O_2-UO_2	-5.72	0.026	-5.61	0.03	0.85	0.89	6629	6459	AZBN00a
DDL	0.2	4.81		5.62	>Si-O_2-UO_2	-5.51	0.027	-5.62	0.03	0.89	0.89	6631	6459	AZZBN01
DDL	0.2	2.3		7.2	$(>SiOH)_2$-UO_2^{2+} $(>SiOH)_2$-UO_2CO_3	5.12 16.94	0.07 0.06	3.50 16.70	0.13 0.04	0.76	0.89	15326	6024	CSB18

DDL	0.2	2.3		7.2	$(>SiOH)_2$-UO_2^{2+}	5.12	0.07	3.53	0.12	0.76	0.89	15309	6024	DW14
					$(>SiOH)_2$-UO_2CO_3	16.94	0.06	16.70	0.04					
DDL	10	10	-1.6	7.6	>Si-O-UO_2^+	-1.978	0.12	-1.28	0.04	0.88	0.90	8733	5206	JHLCH99
					>Si-O-UO_2(OH)	-7.259	0.16	-7.52	0.28					
					>Si-O-$UO_2CO_3^-$	9.529	0.07	9.85	0.05					
DDL	0.31	2.3	-1.24	7.06	>Si-O-UO_2^+	-0.3	0.04	-0.41	0.03	0.80	0.88	10475	6049	PJTP01
					>Si-O-$UO_2(OH)_3^{2-}$	-18.7	0.10	-17.25	0.07					
DDL	0.03	2.3		7.2	>Si-O-UO_2^+	0.3	0.06	-1.43	0.33	0.79	0.89	13060	5833	PTBP98
					>Si-O-UO_2(OH)	-5.65	0.19	-5.21	0.04					
					>Si-O-$(UO_2)_3(OH)_5$	-16.75	0.43	-16.31	0.51					
DDL	0.1	2.31		7.2	>Si-O-$UO_2(OH)_2^-$	-8.45	0.08	-10.12	0.03	0.83	0.90	21461	6961	VT98
NE	0.33	0.64w 0.0078s		7.2	>Si(w)-O-UO_2^+	-0.03	0.13	-2.18	7.18	0.84	0.89	8490	6352	DK01_3
					>Si(w)-O-UO_2(OH)	-4.95	0.06	-4.78	0.04					
					>Si(s)-O-UO_2^+	1.2	0.95	-2.18	386.2					
					>Si(s)-O-UO_2(OH)	-3.19	0.35	-2.48	0.13					
NE	0.33	0.64w 0.012s		7.2	>Si(w)-O-UO_2(OH)	-5.28	0.05	-5.46	0.06	0.94	0.94	3686	3615	DK01_4
					>Si(w)-O-$UO_2CO_3^-$	10.18	0.06	10.21	0.06					
					>Si(s)-O-UO_2(OH)	-2.56	0.06	-2.52	0.05					
NE	0.33	0.63w 0.0077s		7.2	>Si(w)-O-UO_2(OH)	-4.98	0.05	-4.77	0.04	0.87	0.89	7587	6328	DK01_c4
					>Si(s)-O-UO_2^+	1.06	0.48	-1.18	107.4					
					>Si(s)-O-UO_2(OH)	-3.28	0.28	-2.47	0.21					
NE	0.33	0.53		7.2	>Si(w)-O-UO_2(OH)	-4.64	0.03	-4.45	0.03	0.88	0.89	6900	6437	DK01_1
NE	0.33	0.53		7.2	>Si-O-UO_2(OH)	-4.73	0.03	-4.45	0.03	0.88	0.89	7437	6437	KCKD96_1
NE	0.33	0.64w 0.012s		7.2	>Si(w)-O-UO_2(OH)	-5.32	0.05	-4.80	0.04	0.87	0.89	7682	6315	KCKD96_2
					>Si(s)-O-UO_2(OH)	-2.65	0.07	-2.73	0.10					
NE	0.33	0.64w 0.012s		7.2	>Si(w)-$O_{(0.5)}$-UO_2(OH)	-5.28	0.05	-5.48	0.07	0.94	0.94	3685	3610	K02b_2
					>Si(w)-$O_{(0.5)}$-$UO_2CO_3^-$	10.17*	0.06	10.21	0.06					
					>Si(s)-$O_{(0.5)}$-UO_2(OH)	-2.56	0.05	-2.53	0.051					
NE	0.33	0.64w 0.012s		7.2	>Si(w)-$O_{(0.5)}$-UO_2(OH)	-6.56	0.41	-5.48	0.07	0.71	0.94	32701	3609	K02b_1
					>Si(w)-$O_{(0.5)}$-$UO_2CO_3^-$	8.89*	0.11	10.21	0.06					
					>Si(s)-$O_{(0.5)}$-UO_2(OH)	-5.57	2.09	-2.53	0.05					
TL	0.32	4.6051w 0.00184s		8.4	>Si(s)-O-UO_2^+	1.98	0.78	0	14.3	0.9	0.92	5338	4177	FDZ06[b]
					>Si(s)-O-UO_2(OH)	-1.88	0.29	-2.27	0.3					
					>Si(w)-O-UO_2^+	-2.08	0.58	-0.82	0.07					
					>Si(w)-O-UO_2(OH)	-5.76	0.04	-6.48	0.10					
					>Si(w)-O-UO_2(OH)CO_3^{2-}	2.47	0.33	4.45	0.09					

* Reaction equation rewritten using CO_3^{--} rather than CO_2(aq)

[a] SD calculated in PEST based on original reaction constant.

[b] Surface species incorrectly identified in the RES^{3}T database.

[c] AZBN00a = [39], AZZBN01 = [40], CSB18 = [41], DW14 = [43], JHLCH99 = [45], PJTP01 = [49], PTBP98 = [48], VT98 = [50], DK01_3 = [42], DK01_4 = [42], DK01_c4 = [42], DK01_1 = [42], KCKD96_1 = [46], KCKD96_2 = [46], K02b_2 = [47], K02b_1 = [47], FDZ06b = [44].

Importantly, it should be noted that the RES^{3}T database includes some references in which no original sorption data are presented (e.g. [47, 50]). It also overlooks literature references that report original sorption data but do not report surface complexation reactions constants. Consequently, we are able to supplement the data contained in RES^{3}T references with sorption data from several additional literature references (i.e. [51-53]). The original figures contained in all references, once digitized and unified, yield 526 individual data for the U(VI)-quartz system. These digitized and unified data (Figure 2) are captured in the L-SCIE database and are further subjected to the workflow described in Figure 1 for developing SCM reaction constants.

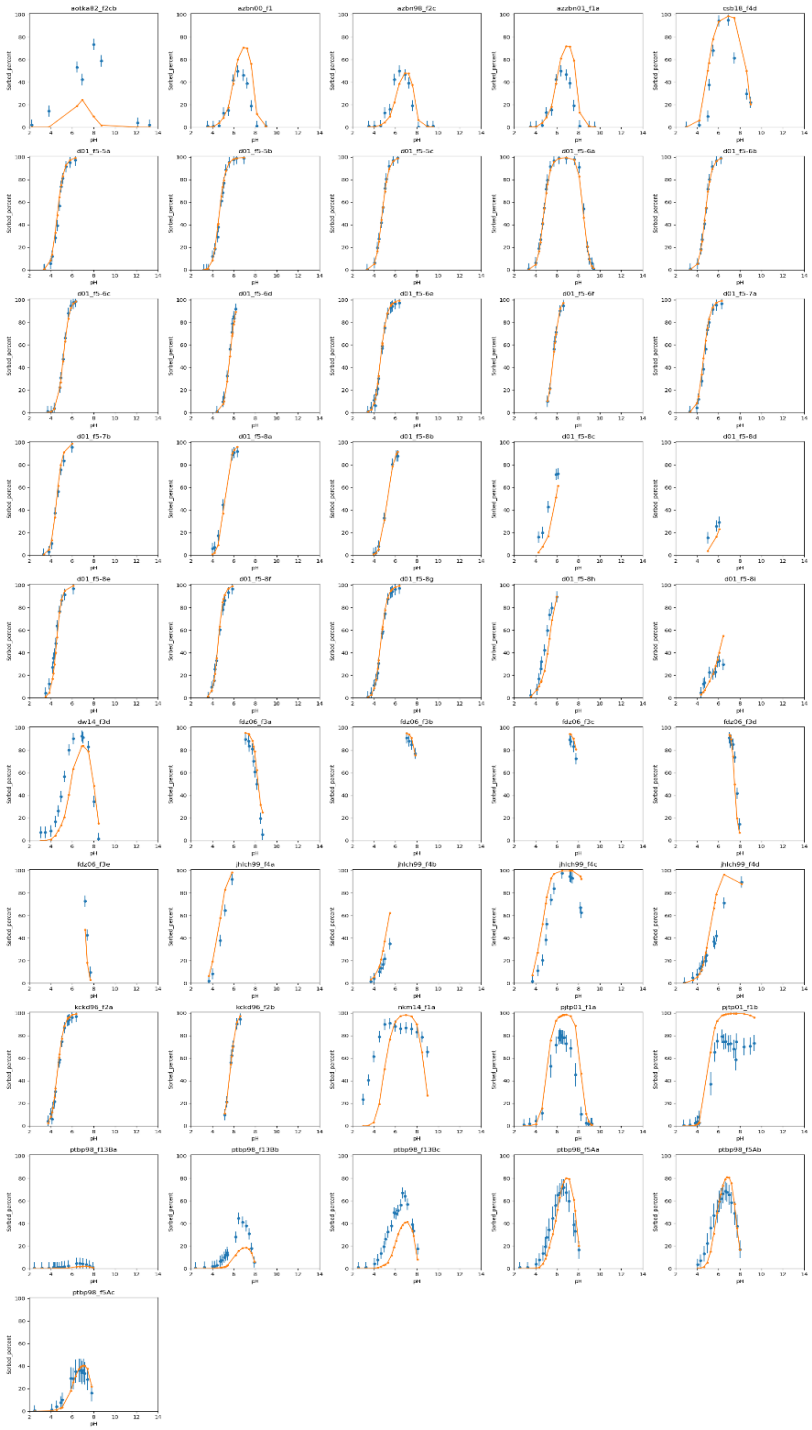

Figure 2. Complete set of sorption data for the U(VI)-quartz system in the L-SCIE database after passing through the data unification workflow. Error bars represent reported or estimated uncertainties at one standard deviation. Curves represent data fit across the multiple datasets using the optimized valued from the 2-site, 4 surface complex non-electrostatic model (Table 2).

U(VI) sorption to quartz – comparison of RES^{3}T database reaction constants

The 17 sets of published reaction constants (Table 1) included in the RES^{3}T database were tested against the full U(VI)-quartz dataset available in L-SCIE. Model performance was tested both for the original published constants and optimized values. Performance testing was also paired with two thermodynamic databases: 1) the "llnl.dat" database that is provided with PHREEQC and is derived from LLNL's SUPCRT[22] database (results reported in Table SI.10) and 2) the same database but updated with missing and revised U(VI) reaction constants taken from the latest NEA-TDB effort[23, 24] (results reported in Table 1). These models were used to evaluate performance of published reaction constants across a larger community dataset and evaluate the impact of thermodynamic databases on model performance. Goodness of fit parameters (R and WSOS) and reaction constants are reported both before and after optimization using PEST.

When coupled to the PHREEQC "llnl.dat" thermodynamic database, most SCM constructs reported in RES^{3}T fail to adequately represent the global dataset (R = 0.72-0.90); only 2 met the R=0.9 threshold (DK01_4 and K02b_2) to be considered an adequate representation of the complete dataset (Figure 3 and Table SI.10). Notably poor fits were characterized by datasets focused on relatively low initial U(IV) concentrations *and* low sorbent surface area estimates (DW14). When reaction constants are optimized, the low-quality SCM constructs are improved but none yield a correlation coefficient above 0.9 (R = 0.85-0.90). Importantly, the optimized reaction constants are shifted by as much as 2.9 logK units, suggesting that many of the published reaction constants were affected by the limited number of data (and range of conditions) used in the original fitting process. The relative non-uniqueness in the performance of the fitted reaction equations and surface complexes (data fit quality for DK01_4, K02b_2, K02b_1, FDZ06b was nearly equivalent) is a common challenge in implementing SCMs and is

apparent in this analysis. Nevertheless, the highest performing SCM constructs 1) include both strong and weak sorption sites and 2) include both hydroxide and carbonate surface complexes of U(VI). It is also true that model performance is not strongly correlated to the choice of SCM and the choice of bidentate or monodentate surface complexes does not appear to have a significant impact on performance.

To test the impact of thermodynamic reaction constant databases, the PHREEQC "llnl.dat" database was updated with most recent aqueous complexation constants from the NEA database ("llnl.dat+NEA" database contains updated constants for U(VI) phosphate, bicarbonate, fluoride, chloride, and silicate aqueous complexes as well as Mg and Ca uranyl carbonate aqueous complexes). The resulting fits to the global dataset improved substantially for most SCM constructs (Figure 3). Prior to optimization, all SCM construct R values increased by 0.01 to 0.04 relative to the "llnl.dat" simulations. Surprisingly, this indicates that, in some cases, an update to the thermodynamic database impacts model performance more so than the choice of reaction equation! Once optimized, all SCM constructs performed substantially better using the "llnl.dat+NEA" (R = 0.88-0.94) than the "llnl.dat" (R = 0.85-0.90) thermodynamic database. Importantly, the earlier conclusions regarding the importance of strong and weak sites and carbonate surface complexes are unchanged. When comparing optimized reaction constants using the "llnl.dat" vs. "llnl.dat+NEA" thermodynamic databases, we find that reaction constants shifted by as much as 3 logK units (0.25 logK units on average) indicating that the choice of thermodynamic database has a significant impact on the reported SCM reaction constants.

It is particularly interesting to illustrate the importance of a uranyl carbonate surface complex on model performance. Including a carbonate surface complex yielded the best overall SCM performance regardless of SCM construct (DK01_4, K02b_2, K02b_1 and FDZ06 have WSOS ≤

4200 and $R \geq 0.92$ and all contain a carbonate surface species). Reaction sets that contain carbonate surface species perform better than those that do not contain carbonate surface species (e.g. compare fit of KCKD96_2 and DK01_4 reactions in Figure SI.3). Thus, we can conclude that uranyl carbonate surface complexes likely play a significant role in U(VI) sorption to quartz.

While SCM constructs emphasize differences in the nature of the electric double-layer that forms at the mineral-water interface, our results demonstrate that non-electrostatic, diffuse layer, and even triple layer representations of the electrical double layer do not necessarily lead to a more robust fit to the sorption data presented in L-SCIE. This is consistent with original findings from Westall and Hohl [15] that particular SCM constructs do not necessarily produce better quality fits to sorption data.

The global fitting approach demonstrated herein identifies the highest performing representation of mineral-water interface and associated surface complexation reactions. However, it must be acknowledged that this fitting exercise does not yield a single preferred representation of reactions at the mineral-water interface. Indeed, this same observation led Dzombak, Morel, Mathur, and Karamalidis [18-20] to adopt the rationale that stoichiometry of surface complexes should be a secondary consideration to the primary focus of producing good fits to the sorption data in SCM database development. While this result may be disappointing from the standpoint of identifying true surface complex stoichiometries on the quartz surface, the intention here is the development of reaction constants representative of the overall partitioning of U(VI) between the surface and the aqueous phases. From that perspective, the stoichiometry of the surface complex is a secondary consideration. Indeed, the ability of the model to effectively capture the sorption of U(VI) for use in reactive transport models and the associated uncertainties takes precedent. Further refinement of the surface complexation reaction stoichiometries may be warranted in the

long term and can be readily introduced with the inclusion of spectroscopic evidence from the literature. However, from a practical standpoint, this may not impact our reactive transport predictive ability unless system conditions deviate significantly from the range of solution conditions captured in the database.

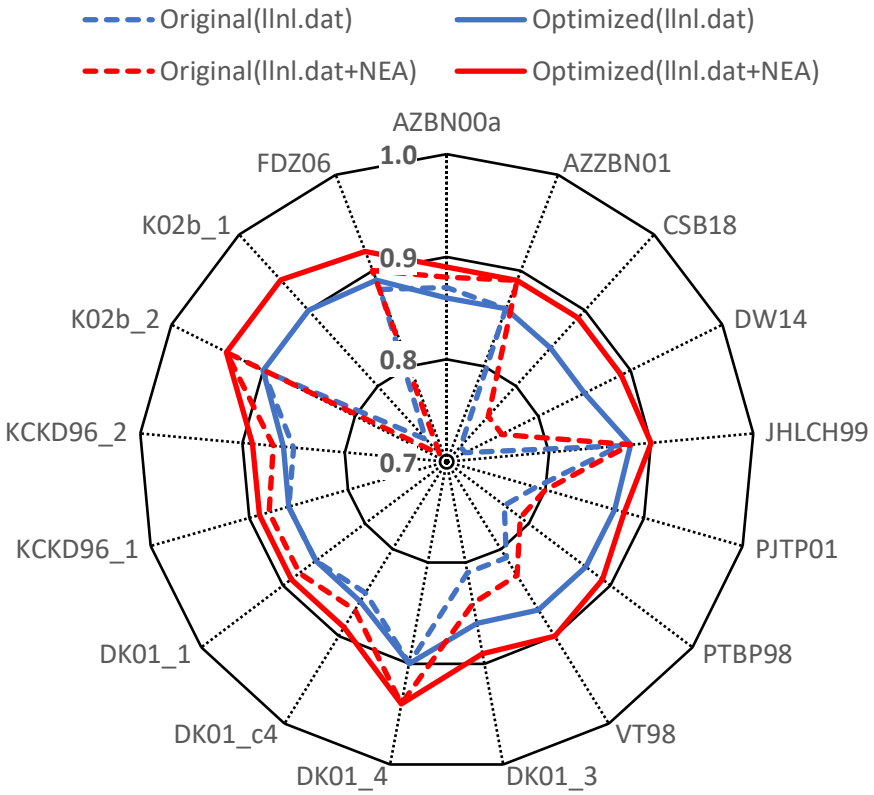

Figure 3. Comparison of R values for 17 sets of RES^{3}T reaction constants prior to and after global optimization using PEST/PHREEQC using the PHREEQC "llnl.dat" and "llnl.dat" updated with data from the NEA-TDB.

U(VI) sorption to quartz – systematic optimization of non-electrostatic and diffuse layer SCMs for database development

The 17 SCM constructs reported in Table 1 were based on a variety of SCMs (non-electrostatic, diffuse layer, and triple layer models), varying estimates of reactive site density (0.00184 to 10 sites/nm^2), the inclusion of one or two site types, surface acidity constants ranging from 5.62 to 8.4, and the absence or presence of a protonation constant that ranges from -1.24 to -1.6. As such, the reported and fitted reaction constants are only valid when paired with the site density and surface acidity constants used during data fitting. This represents a fundamental impasse for building a self-consistent set of reaction constants for multiple sorbing elements on a single mineral (e.g., [18-20]). Such a self-consistent model requires all reaction constants to be fit using common metadata with regards to the surface properties of a particular mineral.

In examining the results of Table 1, a 2-site non-electrostatic model yielded the best performance to the global dataset, but a 1-site diffuse layer model yielded a nearly equally good fit. Thus, the choice of single vs multisite sorption in the present example is somewhat ambiguous. The balance between model complexity and model uncertainty must be weighed by both the needs of the predictive model and the fit to the available data (i.e. the principle of parsimony).

For the purposes of developing a uniform comparison of one-site vs multi-site and non-electrostatic vs diffuse layer model constructs of the quartz surface, we chose the following pragmatic approach to testing model performance. A total site density commonly reported in SCMs (2.31 sites/nm^2) was used (a "strong site" density equivalent to 1% of the total site density was used in the 2-site models). We applied a pK_a that is most often reported in RES^{3}T for this mineral (7.20) and assumed that the protonation constant is sufficiently low to have little impact on the surface charge properties of quartz. Finally, all surface species were assumed to be mono-dentate inner-sphere reactions such that:

$$>SiOH + UO_2^{2+} + X\ H_2O \rightarrow >SiOUO_2(OH)_X^{+1-X} + (1+X)\ H^+ \qquad (3)$$

where X = 0-3. Similarly, for the carbonate species,

$$>SiOH + UO_2^{2+} + Y\ CO_3^{2-} \rightarrow >SiOUO_2(CO_3)_Y^{+1-2Y} + 1\ H^+ \quad (4)$$

where Y = 1-2. Based on the U(VI)-quartz reaction stoichiometries reported in RES^{3}T, the typical charge of surface complexes ranges from +1 to -2. For our analysis, we limited the number of potential monodentate inner sphere surface species to 6 ($>SiOUO_2^+$, $>SiOUO_2OH^0$, $>SiOUO_2(OH)_2^-$, $>SiOUO_2(OH)_3^{-2}$, $>SiOUO_2CO_3^-$, and $>SiOUO_2(CO_3)_2^{-3}$) for a single site model and 12 for a two-site model (with equivalent strong site and weak site stoichiometries). We then systematically tested global fits to all L-SCIE sorption data starting with 6 or 12 surface complexation reactions using non-electrostatic and diffuse layer SCMs. Reactions were removed from the simulation sequentially starting with the least sensitive reaction equations. The resulting plots of R-values are presented in Figure 4. The 1-site model achieved optimal performance with ≥3 reaction constants. a 2-site model was superior to a 1-site model even when the total number of reactions were the same (optimal performance achieved with ≥4 reaction constants), and the diffuse layer model led to a nominally better fit for the 1-site model but a nominally poorer fit in the two-site model. Thus, consistent with Davis et al.'s findings [54], it appears that the inclusion of a diffuse layer electrostatic term does not necessarily improve model performance. Table 2 lists the optimal results with the minimum number of surface reactions for the 4 SCM constructs examined here. As we concluded in the earlier RES^{3}T analysis, the presence of a carbonate surface species is universal for all models and the 2-site models consistently outperform the 1-site models.

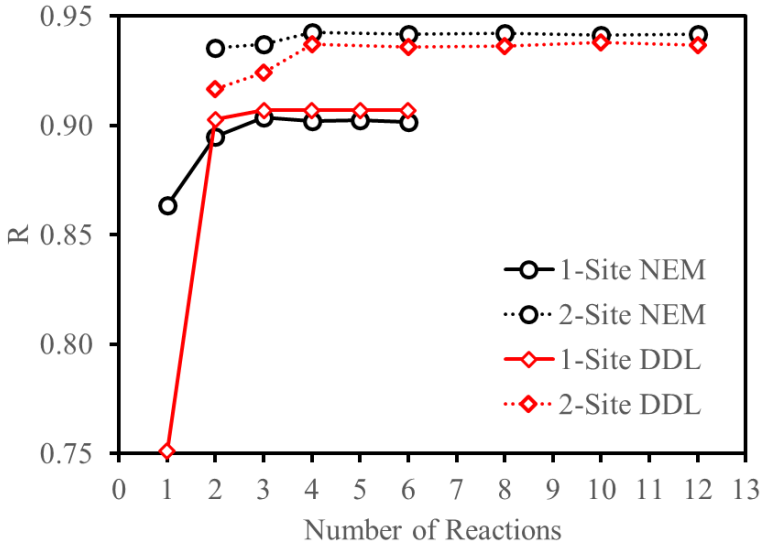

Figure 4. Comparison of quality of global fit as a function of the number of surface complexation reactions included in 4 models of U(VI) sorption to quartz.

Table 2. Optimized surface complexation reaction logK constants and associated error based on non-electrostatic and diffuse layer SCM constructs and 1-site and 2-site models.

1-Site NEM		2-Site NEM	
>SiOUO$_2$OH	-5.37(0.03)	>SiOwUO$_2$OH	-6.31(0.08)
>SiOUO$_2$(OH)$_2^-$	-12.43(0.14)	>SiOwUO$_2$(OH)$_2^-$	-12.13(0.06)
>SiOUO$_2$(CO3)$_2^{3-}$	15.23(0.14)	>SiOsUO$_2$OH	-2.96(0.04)
		>SiOsUO$_2$(CO3)$_2^{3-}$	17.70(0.17)
1-Site DDL		2-Site DDL	
>SiOUO$_2^+$	-0.56(0.03)	>SiOwUO$_2$OH	-6.28(0.07)
>SiOUO$_2$(OH)$_3^{2-}$	-20.22(0.53)	>SiOwUO$_2$CO$_3^-$	10.20(0.07)
>SiOUO$_2$CO$_3^-$	10.46(0.05)	>SiOsUO$_2^+$	1.34(0.07)
		>SiOsUO$_2$CO$_3^-$	13.43(0.10)

IMPLICATIONS

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.

ACKNOWLEDGEMENTS

This work was supported by the Spent Fuel and Waste Science and Technology campaign of the Department of Energy's Nuclear Energy Program, and the Office of Science, Biological and Environmental Research (SCW1053), and the University Engagement program of LLNL's University Relations office under Contract DE-AC52-07NA27344 to LLNL and DE-AC02-05CH11231 to LBNL.

References

1. Steefel, C.I., D.J. DePaolo, and P.C. Lichtner, *Reactive transport modeling: An essential tool and a new research approach for the Earth sciences.* Earth and Planetary Science Letters, 2005. **240**(3): p. 539-558.
2. Steefel, C.I., *Reactive Transport at the Crossroads.* Reviews in Mineralogy and Geochemistry, 2019. **85**(1): p. 1-26.

3. Zhu, J.-J., et al., *ES&T in the 21st Century: A Data-Driven Analysis of Research Topics, Interconnections, And Trends in the Past 20 Years.* Environmental Science & Technology, 2021. **55**(6): p. 3453-3464.
4. Antelo, J., et al., *Analysis of phosphate adsorption onto ferrihydrite using the CD-MUSIC model.* Journal of Colloid and Interface Science, 2010. **347**(1): p. 112-119.
5. Hiemstra, T. and W.H. Van Riemsdijk, *A Surface Structural Approach to Ion Adsorption: The Charge Distribution (CD) Model.* Journal of Colloid and Interface Science, 1996. **179**(2): p. 488-508.
6. Mangold, J.E., et al., *Surface complexation modeling of Hg(II) adsorption at the goethite/water interface using the Charge Distribution Multi-Site Complexation (CD-MUSIC) model.* Journal of Colloid and Interface Science, 2014. **418**: p. 147-161.
7. Venema, P., T. Hiemstra, and W.H. van Riemsdijk, *Multisite Adsorption of Cadmium on Goethite.* Journal of Colloid and Interface Science, 1996. **183**(2): p. 515-527.
8. Arora, B., et al., *Comparison of Electrostatic and Non-Electrostatic Models for U(VI) Sorption on Aquifer Sediments.* Groundwater, 2018. **56**(1): p. 73-86.
9. Newcombe, G. and M. Drikas, *Adsorption of NOM onto activated carbon: Electrostatic and non-electrostatic effects.* Carbon, 1997. **35**(9): p. 1239-1250.
10. Pivovarov, S., *Acid–Base Properties and Heavy and Alkaline Earth Metal Adsorption on the Oxide–Solution Interface: Non-Electrostatic Model.* Journal of Colloid and Interface Science, 1998. **206**(1): p. 122-130.
11. Lützenkirchen, J., *The Constant Capacitance Model and Variable Ionic Strength: An Evaluation of Possible Applications and Applicability.* Journal of Colloid and Interface Science, 1999. **217**(1): p. 8-18.
12. Schindler, P.W., et al., *Ligand properties of surface silanol groups. I. surface complex formation with Fe3+, Cu2+, Cd2+, and Pb2+.* Journal of Colloid and Interface Science, 1976. **55**(2): p. 469-475.
13. Leroy, P. and A. Revil, *A triple-layer model of the surface electrochemical properties of clay minerals.* Journal of Colloid and Interface Science, 2004. **270**(2): p. 371-380.
14. James, A.D. and O.L. James, *Surface ionization and complexation at the oxide/water interface. 3. Adsorption of anions.* Journal of Colloid and Interface Science, 1980. **74**(1): p. 32-43.
15. John, W. and H. Herbert, *A comparison of electrostatic models for the oxide/solution interface.* Advances in Colloid and Interface Science, 1980. **12**(4): p. 265-294.
16. Wilkinson, M.D., et al., *The FAIR Guiding Principles for scientific data management and stewardship.* Scientific Data, 2016. **3**(1): p. 160018.
17. Chen, C., et al., *Learning properties of ordered and disordered materials from multi-fidelity data.* Nature Computational Science, 2021. **1**(1): p. 46-53.
18. Dzombak, D.A. and F.M.M. Morel, *Surface complexation modeling : hydrous ferric oxide*. 1990, New York: Wiley. 393.
19. Karamalidis, A.K. and D.A. Dzombak, *Surface Coplexation Modeling: Gibbsite*. 2010, New Jersey: Wiley and Sons.
20. Mathur, S.S. and D.A. Dzombak, *Chapter 16 - Surface complexation modeling: Goethite*, in *Interface Science and Technology*, J. Lützenkirchen, Editor. 2006, Elsevier. p. 443-468.
21. Brendler, V., et al., *RES3T-Rossendorf expert system for surface and sorption thermodynamics.* J Contam Hydrol, 2003. **61**(1-4): p. 281-91.
22. Johnson, J.W. and S.R. Lundeen, *GEMBOCHS thermodynamic datafiles for use with the EQ3/6 modeling package*. 1997, Lawrence Livermore National Laboratory: Livermore.
23. Ragoussi, M.E. and S. Brassinnes, *The NEA Thermochemical Database Project: 30 years of accomplishments.* Radiochimica Acta, 2015. **103**(10): p. 679-685.

24. Ragoussi, M.E. and D. Costa, *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, 2019. **196**: p. 225-231.
25. Zimmer, K., et al., *SUPCRTBL: A revised and extended thermodynamic dataset and software package of SUPCRT92.* Computers & Geosciences, 2016. **90**: p. 97-111.
26. Bradbury, M.H. and B. Baeyens, *A general application of surface complexation modeling radionuclide sorption in natural systems.* J. Colloid Interface Sci., 1993. **158**: p. 364-371.
27. Davis, J., et al., *Interpretation and prediction of radionuclide sorption onto substrates relevant for radioactive waste disposal using thermodynamic sorption models*. 2005, OECD/Nuclear Energy Agency: Paris. p. 286.
28. Ochs, M., T.E. Payne, and V. Brendler, *Thermodynamic sorption modeling in support of radioactive waste disposal safety cases. A guideline document*. 2012, OECD/NEA: Paris. p. 153.
29. Payne, T.E., et al., *Guidelines for thermodynamic sorption modelling in the context of radioactive waste disposal.* Environmental Modelling & Software, 2013. **42**: p. 143-156.
30. Bradbury, M.H. and B. Baeyens, *Sorption modelling on illite. Part II: Actinide sorption and linear free energy relationships.* Geochim. Cosmochim. Acta, 2009. **73**(4): p. 1004-1013.
31. Dresden-Rossendorf, H.-Z. *RES³T - Rossendorf Expert System for Surface and Sorption Thermodynamics*. 2013 [cited 2013; Available from: https://www.hzdr.de/db/RES3T.disclaimer.
32. Geckeis, H., et al., *Mineral–Water Interface Reactions of Actinides.* Chemical Reviews, 2013. **113**(2): p. 1016-1062.
33. Bradbury, M.H., B. Baeyens, and T. Thoenen, *Sorption Data Bases for Generic Swiss Argillaceous Rock Systems*. 2010, Nagra: Wettingen, Switzerland.
34. Hiszpanski, A.M., et al., *Nanomaterial Synthesis Insights from Machine Learning of Scientific Articles by Extracting, Structuring, and Visualizing Knowledge.* Journal of Chemical Information and Modeling, 2020. **60**(6): p. 2876-2887.
35. Tummers, B., *DataThief III*. 2006.
36. Team, R.C., *R: A language and environment for statistical computing* 2018, R Foundation for Statistical Computing: Vienna, Austria.
37. Doherty, J., *PEST: Model-Independent Parameter Estimation*. 2003, Watermark Numerical Computing.
38. Parkhurst, D.L. and C.A.J. Appelo, *User's guide to PHREEQC (Version 2)--a computer program for speciation, batch-reaction, one-dimensional transport, and inverse geochemical calculations*, in *U.S. Geological Survey Water-Resources Investigations Report*. 1999. p. 312.
39. Arnold, T., et al., *Applying the DDLM to model the sorption of uranium onto quartz and muscovite*, in *Annual report 1999*. 2000, Forschungszentrum Rossendorf e.V., Institute of Radiochemistry: Dresden (Germany).
40. Arnold, T., et al., *Sorption behavior of U(VI) on phyllite: experiments and modeling.* Journal of Contaminant Hydrology, 2001. **47**(2): p. 219-231.
41. Coutelot, F.M., J.C. Seaman, and M. Baker, *Uranium(VI) adsorption and surface complexation modeling onto vadose sediments from the Savannah River Site.* Environmental Earth Sciences, 2018. **77**(4): p. 148.
42. Davis, J. and M. Kohler, *Surface complexation modeling of uranium(VI) adsorption on quartz.*, in *Surface Complexation Modeling of Uranium (VI) Adsorption on Natural Mineral Assemblages*. 2001, Office of Nuclear Regulatory Research, U.S. Nuclear Regulatory Commission: Washington, DC.
43. Dong, W. and J. Wan, *Additive Surface Complexation Modeling of Uranium(VI) Adsorption onto Quartz-Sand Dominated Sediments.* Environmental Science & Technology, 2014. **48**(12): p. 6569-6577.

44. Fox, P.M., J.A. Davis, and J.M. Zachara, *The effect of calcium on aqueous uranium(VI) speciation and adsorption to ferrihydrite and quartz.* Geochimica et Cosmochimica Acta, 2006. **70**(6): p. 1379-1387.
45. Jung, J., et al., *Adsorption of UO22+ on natural composite materials.* Journal of Radioanalytical and Nuclear Chemistry, 1999. **242**(2): p. 405-412.
46. Kohler, M., et al., *Experimental Investigation and Modeling of Uranium (VI) Transport Under Variable Chemical Conditions.* Water Resources Research, 1996. **32**(12): p. 3539-3551.
47. Kulik, D.A., *Sorption modelling by Gibbs energy minimisation: Towards a uniform thermodynamic database for surface complexes of radionuclides.* Radiochimica Acta, 2002. **90**(9-11): p. 815-832.
48. Pabalan, R.T., et al., *Chapter 3 - UraniumVI Sorption onto Selected Mineral Surfaces: Key Geochemical Parameters*, in *Adsorption of Metals by Geomedia*, E.A. Jenne, Editor. 1998, Academic Press: San Diego. p. 99-130.
49. Prikryl, J.D., et al., *UraniumVI sorption behavior on silicate mineral mixtures.* Journal of Contaminant Hydrology, 2001. **47**(2): p. 241-253.
50. Viani, B.E. and P.C. Torretto, *Sorption and transport of uranium on hematite*. 1998, Lawrence Livermore National Laboratory: Livermore. p. 1-12.
51. Allard, B., et al., *Sorption of Actinides in Well-Defined Oxidation States on Geologic Media.* MRS Online Proceedings Library, 2011. **11**(1): p. 775.
52. Arnold, T., et al., *Sorption of uranium(VI) onto phyllite.* Chemical Geology, 1998. **151**(1): p. 129-141.
53. Nair, S., L. Karimzadeh, and B.J. Merkel, *Sorption of uranyl and arsenate on SiO2, Al2O3, TiO2 and FeOOH.* Environmental Earth Sciences, 2014. **72**(9): p. 3507-3512.
54. Davis, J.A., et al., *Application of the Surface Complexation Concept to Complex Mineral Assemblages.* Environmental Science & Technology, 1998. **32**(19): p. 2820-2828.
55. Davis, J.A., et al., *Application of the surface complexation concept to complex mineral assemblages.* Environmental Science and Technology, 1998. **32**: p. 2820-2828.

SUPPORTING INFORMATION

A Community Data Mining Approach for Surface Complexation Database Development

Mavrik Zavarin[1*], Elliot Chang[1*], Haruko Wainwright[2,3], Nicholas Parham[1], Rahul Kaukuntla[1], Jadallah Zouabe[1,4], Amanda Deinhart[1], Victoria Genetti[1], Sam Shipman[1], Frank Bok[5], and Vinzenz Brendler[5]

[1] Seaborg Institute, Lawrence Livermore National Laboratory (zavarin1@llnl.gov, chang58@llnl.gov)

[2] Lawrence Berkeley National Laboratory, Earth and Environmental Sciences Area (hmwainwright@lbl.gov)

[3] U.C. Berkeley, Department of Nuclear Engineering (harukom@berkeley.edu)

[4] U.C. Berkeley, Department of Chemical Engineering (jzouabe@berkeley.edu)

[5] Helmholtz-Zentrum Dresden-Rossendorf e.V., Institute of Resource Ecology (F.Bok@hzdr.de)

*co-first authors

SI.1 Description of the RES^{3}T Database

The RES³T project is an ongoing effort by the Helmholtz-Zentrum Dresden-Rossendorf (HZDR) to develop a digital web-based open-source surface complexation model (SCM) database (https://www.hzdr.de/res3t). It includes mineral-specific surface complexation constants that can be used in component additivity models [55] of more complex solid phases such as rocks or soils. It offers an integrated graphical user interface to access selected mineral and sorption data and export data into formats suitable for other modeling software. Data records comprise mineral properties, specific surface areas, characteristics of surface binding sites and their protolysis constants, sorption ligand information, and surface complexation reactions (SCM include the Non-Electrostatic, Diffuse Double Layer, Constant Capacitance, Triple Layer, Basic Stern, and the 1-pK Model as extended to CD-MUSIC). The database also provides a comprehensive list of publications (3172) that are the primary sources of the surface complexation reaction constants. One limitation of this project is its lack of recommendations for specific reaction constants under relevant geochemical situations. It also does not capture the primary (raw) sorption data or provide information on the aqueous speciation constants used in determining those SCM reaction constants. As a result, the RES³T project provides a foundation for developing a self-consistent database of SCM reaction constants but does not go so far as to provide one. Three key issues prevent the direct application of the RES^{3}T database in reactive transport models. They are:

- An inability to integrate published SCM reaction constants and SCM constructs into single unified model and associated set of reaction constants
- An inability to produce self-consistent reaction constants based on a common set of aqueous speciation constants and surface properties.

- The absence of error propagation in the sorption data and/or database constants needed to assess model uncertainties.

To address the limitations of the RES^{3}T database, mining of the raw sorption data from publications contained in this database is necessary. These mined data could then be used to develop SCMs that are informed by the reported reaction constants in the RES^{3}T database. In a next step, it can be made self-consistent by using a specified thermodynamic database, applying mineral-specific metadata (e.g. site density, acidity constants, etc.) uniformly, capturing uncertainties associated with the data fitting process, and incorporating the reactions and associated parameter uncertainties into a database for use in reactive transport models.

SI.2 Organization and Details of the LLNL SCIE (L-SCIE) database structure

The **Reference Table** captures information regarding each unique reference, such as a unique identifier, document DOI, and data digitization status (Table SI.1).

The **Dataset Table** (Table SI.2) captures metadata for each "set" of data included in a reference used to describe the boundary conditions of batch sorption experiments. A "set" of data is generally defined as a set of batch sorption data that have a common sorbate concentration and/or other unifying features. A typical dataset would be one sorption envelope contained in a figure in which sorption was measured as a function of pH under otherwise similar solution conditions. However, a dataset could also comprise a single datapoint or an isotherm, i.e. varying sorbate concentration but fixed pH). The Dataset Table contains experimental information, such as name and chemical formula of the mineral sorbent, the source of the mineral, the electrolyte composition, the sorbing species (mol/L), and any competing adsorptive features such as cation exchange capacity or equilibrated gaseous species (Table SI.2).

The **Atmospheric CO_2 Table** contains the fugacity of CO_2 in the atmosphere as a function of time. A common problem in understanding sorption data from the literature is to quantify the equilibrium gas conditions in the headspace of the experimental vessel. This information is often not measured or simply not reported. However, experiments are often described as having been performed in an N_2 glovebox or under atmospheric conditions. In the case of atmospheric conditions, the headspace gas composition can be estimated from the composition of gases in the laboratory at the time of data collection. The CO_2 fugacity is critical to any modeling effort as it is well known that formation of carbonate complexes in solution can significantly affect the sorption behavior of many elements. Thus, for experiments performed under atmospheric conditions, the CO_2 fugacity in the experiments can be estimated by the atmospheric CO_2 fugacity at the date of publication (which is presumed to be close to the date of the actual experiments).

The **Mineral_Site_Density Table** addresses a second issue regarding datasets reported in the literature having varied mineral binding site densities. The site density on the mineral surface can be measured but is often assumed to be a value that has been previously reported for that mineral or simply is based on common values reported in the literature for a variety of minerals. Thus, most modeling efforts use surface site densities that are chosen somewhat arbitrarily. To develop consistency across multiple datasets, we include this table to regularize mineral site densities by overwriting the original site density values reported in the Dataset Table with a common site density for all data reported for a specific mineral. Of course, during any modeling effort, the user may choose to use a common site density for a mineral or an author-specified site density.

The **Data Table** simply captures the actual sorption data contained within each dataset. This table contains the raw X-Y data extracted from figures in the associated reference and associated

with a specific dataset in the Dataset Table. The Data Table contains information regarding the dataset associated with the data and the x- and y- values reported in the figures. Standard deviations for the datapoints are included when available.

SI.3 Example Unit Conversion Calculations

Data Reported as Percent Sorbed

To calculate the sorbate concentration in the aqueous phase, Aq(mol/L), total sorbate concentration in the sample, Sorbate_val(mol/L), and sorbate concentration on the solid phase, Sorbed(mol/L), the following calculations are necessary:

Sorbate_val(mol/L) = Reported in L-SCIE database

Sorbed(mol/L) = Sorbate_val(mol/L) × Sorbed(%) / 100

Aq(mol/L) = Sorbate_val(mol/L) - Sorbed(mol/L)

Data Reported as Fraction Sorbed

To calculate the sorbate concentration in the aqueous phase, Aq(mol/L), total sorbate concentration in the sample, Sorbate_val(mol/L), and sorbate concentration on the solid phase, Sorbed(mol/L), the following calculations are necessary:

Sorbate_val(mol/L) = Reported in L-SCIE database

Sorbed(mol/L) = Sorbate_val(mol/L) × Sorbed(fraction)

Aq(mol/L) = Sorbate_val(mol/L) - Sorbed(mol/L)

Data Reported as Kd or Rd

To calculate the sorbate concentration in the aqueous phase, Aq(mol/L), total sorbate concentration in the sample, Sorbate_val(mol/L), and sorbate concentration on the solid phase, Sorbed(mol/L), from the reporting of Kd (or Rd), the following calculations are necessary (similar calculations necessary if data reported as log(Kd) or log(Rd)):

Kd(mL/g) = Reported in L-SCIE database

Aq(mol/L) = Reported in L-SCIE database

Note: Kd or Rd assumed to be in units of mL/g.

Sorbate_val(mol/L) = Kd(mL/g) × Aq(mol/L) × (1/1000) × Mineral_val(g/L) + Aq(mol/L)

= Aq(mol/L) × (Kd(mL/g) × (1/1000) × Mineral_val(g/L) + 1)

Sorbed(mol/L) = Sorbate_val(mol/L) - Aq(mol/L)

If Aq(mol/L) is not reported in the database, alternative equations are needed but are not illustrated here for the sake of brevity.

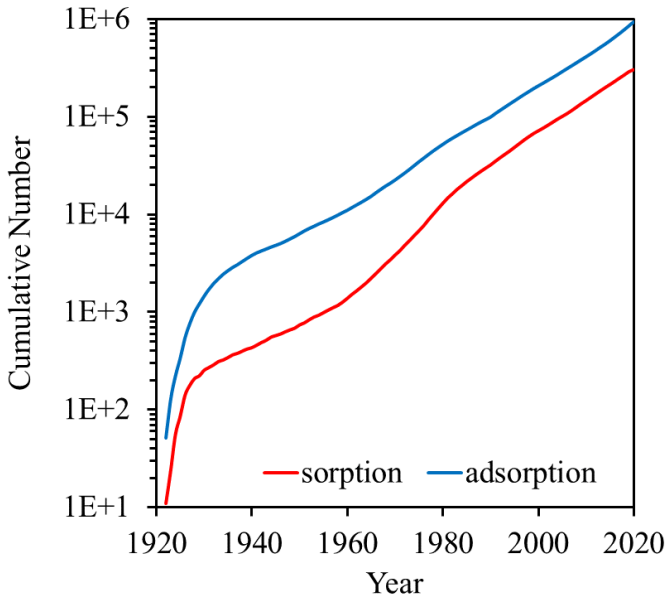

Figure SI.1. Cumulative number of publications referencing the topics of "sorption" and "adsorption" based on the compiled data available in the Web of Science.

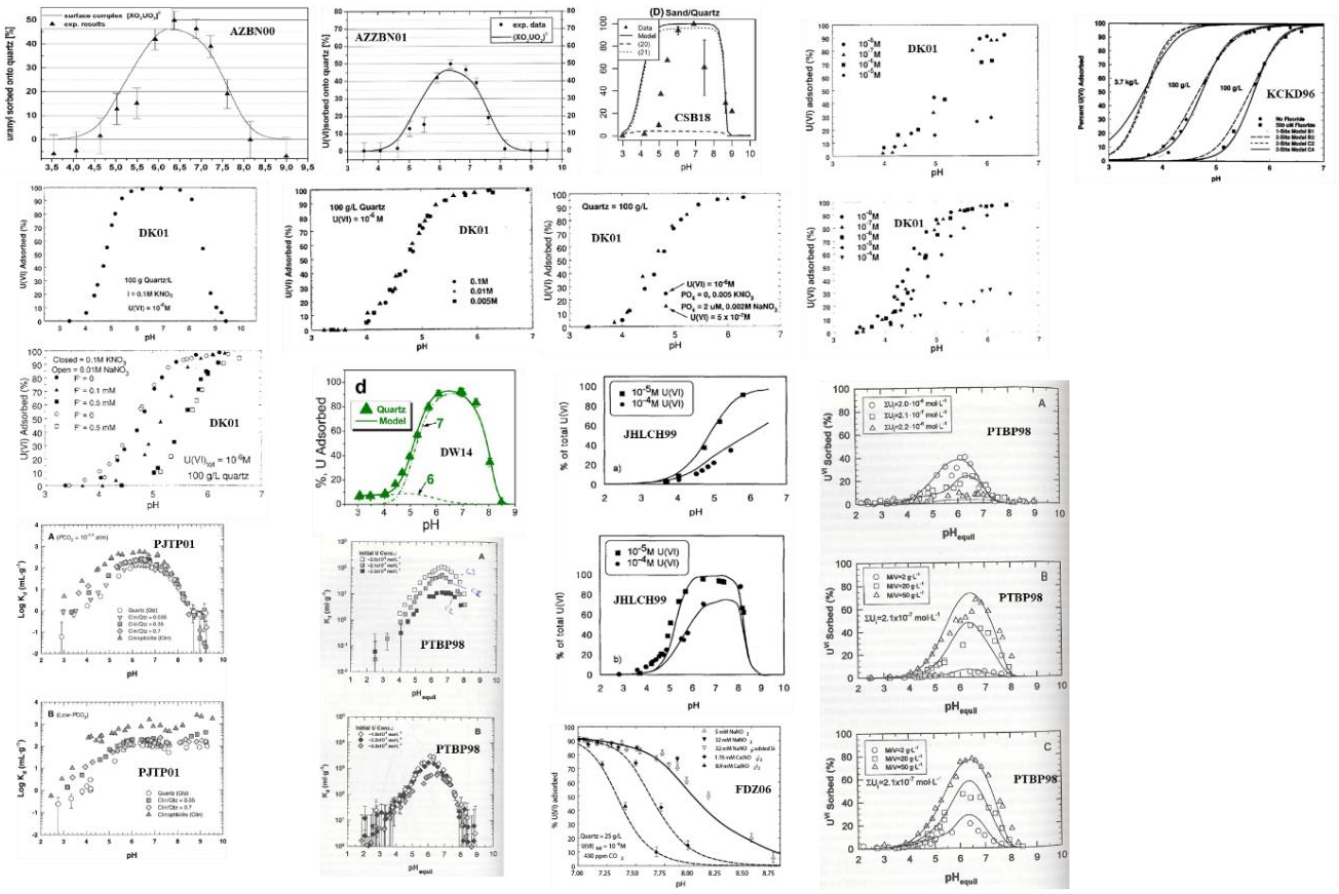

Figure SI.2. Sorption figures contained in the RES^{3}T references for U(VI) sorption to quartz.

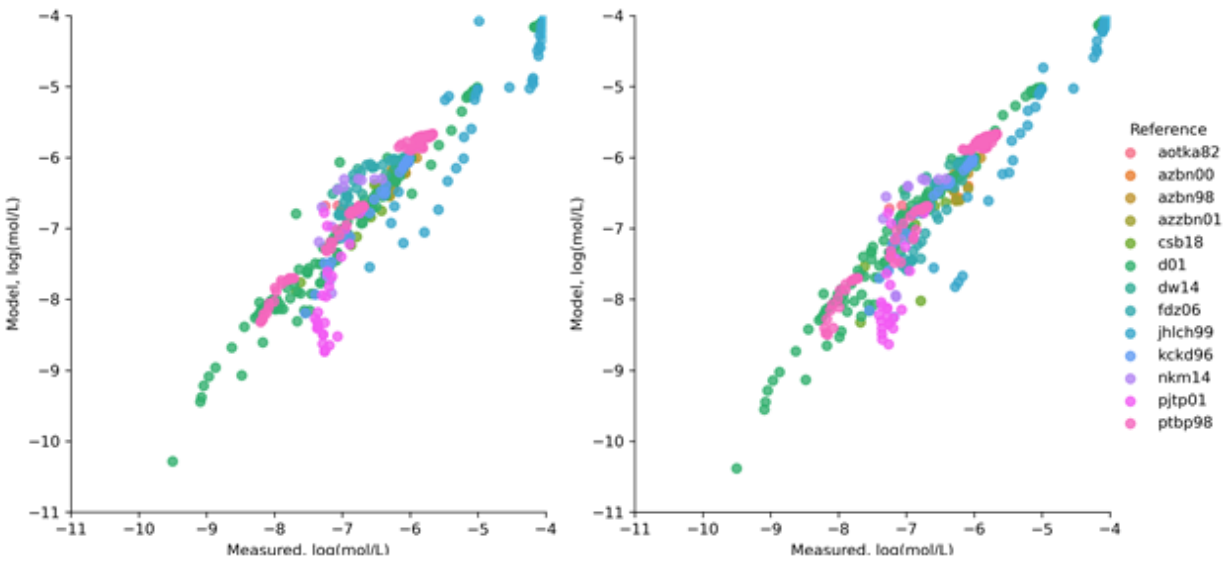

Figure SI.3. Comparison of U(VI)-quartz data fits of (a) KCKD96_2 (R=0.89) that includes >Si(w)-O-UO_2(OH) and >Si(s)-O-UO_2(OH) surface complexes to (b) DK01_4 (R=0.94) that includes >Si(w)-O-UO_2(OH), >Si(s)-O-UO_2(OH), and >Si(w)-O-$UO_2CO_3^-$ surface complexes.

Table SI.1 Information included in L-SCIE Reference Table.

Item	Description
Reference name[a]	Unique name for reference based on the nomenclature established in the RES^{3}T database.
PDF	Original pdf of manuscript (if DOI not available)
DOI	Document DOI
Importer	Name of person performing the data import
Original data	Y/N answer to whether the data appear to be original to the referenced manuscript
RES^{3}T reference	Y/N answer to whether the reference is contained in the RES^{3}T database
In database	Y/N answer to whether the raw data are captured in the database
Comments	Ancillary information regarding data digitization

[a]In the RES^{3}T database, the reference format is based on the author name initial and year of publication such that JHLCH99 = J. Jung, S. P. Hyun, J. K. Lee, Y. H. Cho, P. S. Hahn. 1999. Adsorption of $U0_2^{2+}$ on natural composite materials, Journal of Radioanalytical and Nuclear Chemistry, 242(2): 405-412. This same formalism is used in our digitized sorption database.

Table SI.2 Information included in the L-SCIE Dataset Table.

Item	Description
Reference name	Unique name for reference
Dataset name	Reference name followed by a figure and dataset number
Attachment	.jpg file of original figure and a .txt file with associated digitized data
Mineral	name of mineral
Mineral formula	chemical formula of mineral
Mineral_source	the source of the mineral which may be naturally sourced mineral, synthesized, or purchased through a specific brand
ElectrolyteX	Name of electrolyte ion X
ElectrolyteX_val*	Value for electrolyte X concentration
ElectrolyteX_units	Units of ElectrolyteX_val (e.g. mol/L)
Sorbate	Name of sorbate
Sorbate_val*	Value of sorbate concentration
Sorbate_unit	Units of Sorbate_val (e.g. mol/L)
Mineral_val*	Value of mineral concentration
Mineral_units	Units of mineral_val (e.g. g/L)
MineralSA*	surface area of mineral
MineralSA_units	Units of mineralSA (e.g. m^2/g)
Mineralsites*	mineral reactive sites value
Mineralsites_units	Units of Mineralsites (e.g. sites/nm^2)
CEC*	cation exchange capacity
CEC_units	units of CEC (e.g. meq/100g)
GasX	Name of gas in equilibrium with solution
GasX_val*	Value of gas concentration
GasX_units	units of GasX_val (e.g. bar)
X_axis	Name of X axis in figure
X_units	Units of X axis
Y_axis	Name of Y axis on figure
Y_units	Units of Y axis

* Standard deviations for these values are included when available.

Table SI.3 Unit conversions and uncertainty calculations contained in digitized X-Y plots.

Axis	Units	Conversion	SD Calculation
Alkalinity	meq/L	= axis / 1000 = mol/L HCO3-	= SD / 1000 (5% default)
Aqueous	mol/L	NA	NA (5% default)
Aqueous	log(mol/L)	= 10^{axis} = mol/L aqueous	As with electrolytes (10% default)
Aqueous	mmol/L	= axis / 1000 = mol/L aqueous	= SD / 1000 (5% default)
Aqueous	ppb	= axis / 1000000 / MM = mol/L aqueous	= SD / 1000000 / MM (5% default)
Aqueous	ppm	= axis / 1000 / MM = mol/L aqueous	= SD / 1000 / MM (5% default)
Charge	microC/cm^2	**NA**	NA (5% default)
Charge	meq/kg	= axis * 96485 (C/eq) / mineral_SA (m^2/g) / 10000 = microC/cm^2	= SD * 96485 (C/eq) / mineral_SA (m^2/g) / 10000 = microC/cm^2 (5% default)
Charge	micromol/m^2	= axis * 96485 / 10000 = microC/cm^2	= SD * 96485 / 10000 = microC/cm^2 (5% default)
Charge	C/g	= axis / mineral_SA (m^2/g) / 10000 * 1000000 = microC/cm^2	= SD / mineral_SA (m^2/g) / 10000 * 1000000 = microC/cm^2 (5% default)
Charge	C/m^2	= axis / 10000 * 1000000 = microC/cm^2	= SD / 10000 * 1000000 = microC/cm^2 (5% default)
Charge	microC/m^2	= axis /10000 = microC/cm^2	= SD /10000 = microC/cm^2 (5% default)
Charge	cmol/kg	= axis / 100 / 1000 * 96485 (C/eq) / mineral_SA (m^2/g) / 10000 * 1000000 = microC/cm^2	= SD / 100 / 1000 * 96485 (C/eq) / mineral_SA (m^2/g) / 10000 * 1000000 = microC/cm^2 (5% default)
Charge	eq/g	= axis * 96485 (C/eq) / mineral_SA (m^2/g) / 10000 * 1000000 = microC/cm^2	= SD * 96485 (C/eq) / mineral_SA (m^2/g) / 10000 * 1000000 = microC/cm^2 (5% default)
H(+)	-log(mol/L)	= pH	NA (0.1 default)
Kd	mL/g	= Aq(mol/L) * (Kd(mL/g) * (1/1000) * Mineral_val(g/L) + 1) = Sorbate_val(mol/L) = Sorbate_val(mol/L) / (Kd(mL/g) * (1/1000) * Mineral_val(g/L) + 1) = Aq(mol/L) = Sorbate_val(mol/L) - Aq(mol/L) = Sorbed(**mol/L**)	(50% default Kd) = (Sorbate_val(mol/L) / (Kd(mL/g) / 1000 * Mineral_val(g/L) + 1)) - (Sorbate_val(mol/L) / ([Kd(mL/g) * **1.5**)] / 1000 * Mineral_val(g/L) +1)) = Aq_SD (mol/L) = ((Sorbate_SD(mol/L)^2 + (Aq_SD (mol/L)) ^{2}) $^{0.5}$ = Sorbed_SD(mol/L)
Kd	log(mL/g)	Same as above, but, convert log first	(±0.2 Default logKd) = (Sorbate_val(mol/L) / ($10^{log(Kd(mL/g)}$ / 1000 * Mineral_val(g/L) + 1)) - (Sorbate_val(mol/L) / ([$10^{log(Kd(mL/g+0.2)}$] / 1000 * Sorbate_val(g/L) +1)) = Aq_SD (mol/L)

Axis	Units	Conversion	SD Calculation
			= ((Sorbate_SD(mol/L)^2 + (Aq_SD (mol/L)) ^{2}) $^{0.5}$ = Sorbed_SD(mol/L) = ((Sorbate_SD(mol/L)^2 + (Aq_SD (mol/L)) ^{2}) $^{0.5}$ / Mineral_val(g/L) = Sorbed_SD(mol/g)
Kd	log(L/kg)	Same as above, but, convert log first (L/kg = mL/g)	
Na(+)	mol/L	No conversion	NA (5% default)
pH	pH	No conversion	NA (0.1 default)
Rd	log(L/kg)	Same as above, but, convert log first (L/kg = mL/g)	
Rd	m3/kg	Same as above, but, first convert to mL/g (m3/kg *1000 = mL/g)	
Solid	g/L	NA	NA (5% default)
sorbed	mol/kg	= axis / 1000 = mol/g or = axis / 1000 * mineral_val(g/L) = mol/L	As with electrolytes (5% default)
sorbed	%	Sorbate_val(mol/L) = this value should be reported in the table. = Sorbate_val(mol/L) * Sorbed(%) / 100 = Sorbed(**mol/L**) = Sorbed(mol/L) / Mineral_val(g/L) = Sorbed(**mol/g**) = Sorbate_val(mol/L) - Sorbed(**mol/L**) = Aq(mol/L)	(5% default) = 5 / 100 * Sorbate_val (mol/L) = Sorbed_SD(mol/L) = 5 / 100 * Sorbate_val (mol/L) / Mineral_val(g/L) = Sorbed_SD(mol/g) = 5 / 100 * Sorbate_val (mol/L) = Aq_SD(mol/L)
sorbed	mol/g	No conversion of mol/g or = axis * mineral_val(g/L) = mol/L	As with electrolytes (5% default)
sorbed	log(mol/kg)	= 10^{axis} / 1000 = mol/g or = 10^{axis} / 1000 * mineral_val(g/L) = mol/L	As with electrolytes (10% default)
sorbed	fraction	Sorbate_val(mol/L) = this value should be reported in the table. = Sorbate_val(mol/L) * Sorbed(fraction) = Sorbed(**mol/L**) = Sorbed(mol/L) / Mineral_val(g/L) = Sorbed(**mol/g**) = Sorbate_val(mol/L) - Sorbed(**mol/L**) = Aq(mol/L)	(0.05 default) = 0.05 * Sorbate_val (mol/L) = Sorbed_SD(mol/L) = 0.05 * Sorbate_val (mol/L) / Mineral_val(g/L) = Sorbed_SD(mol/g) = 0.05 * Sorbate_val (mol/L) = Aq_SD (mol/L)
sorbed	mmol/kg	= axis / 1000000 = mol/g or	As with electrolytes (5% default)

Axis	Units	Conversion	SD Calculation
		= axis / 1000000 * mineral_val(g/L) = mol/L	
sorbed	log(mol/m2)	= 10^{axis} * mineral_SA(m2/g) = mol/g or = 10^{axis} * mineral_SA(m2/g) * mineral_val(g/L) = mol/L	As with electrolytes (10% default)
sorbed	log(mol/mol)	= 10^{axis} * Sorbate_val(mol/L) / mineral_val(g/L) = mol/g or = 10^{axis} * Sorbate_val(mol/L) = mol/L	As with electrolytes (10% default)
sorbed	mmol/g	= axis / 1000 = mol/g or = axis / 1000 * mineral_val(g/L) = mol/L	As with electrolytes (5% default)
sorbed	micromol/g	= axis / 1000000 = mol/g or = axis / 1000000 * mineral_val(g/L) = mol/L	As with electrolytes (5% default)
sorbed	log(mol/g)	= 10^{axis} = mol/g or = 10^{axis} * mineral_val(g/L) = mol/L	As with electrolytes (10% default)
total H(+)	mol/L	= axis * 96485 (C/eq) / mineral_val(g/L) / mineral_SA (m^2/g) / 10000 * 1000000 = microC/cm^2	As with electrolytes (5% default)
total H(+)	mmol/L	= axis / 1000 * 96485 (C/eq) / mineral_val(g/L) / mineral_SA (m^2/g) / 10000 * 1000000 = microC/cm^2	As with electrolytes (5% default)
total OH(-)	mol/L	= -axis * 96485 (C/eq) / mineral_val(g/L) / mineral_SA (m^2/g) / 10000 * 1000000 = microC/cm^2	As with electrolytes (5% default)

Table SI.4. General statistics of the L-SCIE database.

Data Table	Count
References	214
Datasets	2,006
Data	23,106

Table SI.5. Sorbing species included in the database and associated concentration range.

Species	**min**	**max**	**units**
Am(+3)	5.00E-10	2.90E-07	mol/L
Ba(+2)	3.12E-08	1.00E-07	mol/L
Ca(+2)	2.00E-06	2.52E-01	mol/L
Cd(+2)	1.00E-08	3.20E-05	mol/L
Cm(+3)	2.00E-07	2.00E-07	mol/L
Co(+2)	5.60E-09	1.00E-02	mol/L
Cr(+3)	9.62E-04	9.62E-04	mol/L
Cr(+6)	5.00E-06	5.00E-06	mol/L
Cs(+)	1.51E-09	7.95E-04	mol/L
Cu(+2)	3.10E-07	6.35E+02	mol/L
Eu(+3)	1.00E-09	1.11E-04	mol/L
H(+)			mol/L
HCO3(-)	5.50E-05	8.30E-04	mol/L
humic acid	9.00E+00	9.00E+00	mg/L
K(+)	1.24E-03	5.04E-01	mol/L
Mg(+2)	5.77E-05	3.63E-03	mol/L
Mn(+2)	1.00E-07	1.00E-02	mol/L
Ni(+2)	5.00E-09	1.00E-02	mol/L
NpO2(+)	1.00E-14	6.00E-06	mol/L
Pb(+2)	8.50E-08	2.07E+00	mol/L
PO3(-3)	1.00E-04	1.00E-04	mol/L
Pu	1.00E-14	6.00E-08	mol/L
Rb(+)	1.27E-09	7.95E-04	mol/L
SeO3(-2)	1.90E-05	1.90E-05	mol/L
SO4(-2)	1.72E-04	4.00E-03	mol/L
Sr(+2)	1.00E-08	5.00E-02	mol/L
Th(+4)	1.00E-11	3.00E-07	mol/L
UO2(+2)	1.00E-08	1.00E-03	mol/L
Yb(+3)	2.10E-05	2.10E-05	mol/L
Zn(+2)	1.00E-09	6.54E+02	mol/L

Table SI.6. Electrolytes included in the database and associated concentration range.

Electrolyte	min	max	Units
Acetate	1.0E-02	1.0E-02	mol/L
Ba(+2)	2.0E-04	5.0E-04	mol/L
Ca(+2)	1.4E-05	4.0E-01	mol/L
Citrate	1.0E-02	1.0E-02	mol/L
Cl(-)	4.0E-04	6.0E+00	mol/L
ClO4(-)	1.0E-03	5.0E+00	mol/L
F(-)	1.0E-04	5.0E-04	mol/L
humic acid	9.0E+00	9.0E+00	mg/L
HCO3(-)	7.8E-06	1.0E+00	mol/L
I(-)	4.0E-03	1.0E-01	mol/L
K(+)	1.4E-05	2.0E+00	mol/L
Mg(+2)	7.4E-05	5.4E-02	mol/L
NO3(-)	1.0E-03	2.5E+00	mol/L
NH4(+)	8.7E-03	8.7E-03	mol/L
Na(+)	2.5E-04	6.0E+00	mol/L
oxalate	5.0E-04	5.0E-04	mol/L
PO4(-3)	2.0E-06	1.0E-04	mol/L
pH	2.5E+00	1.0E+01	pH
pyromellitate	5.0E-04	1.0E-03	mol/L
SO4(-2)	3.8E-05	2.8E-02	mol/L
S2O4(-2)	1.0E-03	1.0E-03	mol/L
salicylate	5.0E-04	5.0E-04	mol/L
SiO2(aq)	1.0E-04	1.0E-04	mol/L
Sr(+2)	1.5E-04	5.0E-04	mol/L
UO2(+2)	1.0E-06	1.0E-03	mol/L

Table SI.7. Minerals included in the database.

Mineral	Mineral_Formula	Mineral_Source
Al(OH)3	gamma-Al(OH)3	synthetic
albite		natural
alumina	alpha-Al2O3	Ventron
alumina	alpha-Al2O3	Wako
alumina	gamma-Al2O3	Degussa
alumina	gamma-Al2O3	Fisher
alumina	gamma-Al2O3	synthetic
amorphous silica	SiO2	Merck
amorphous silica	SiO2	synthetic
bentonite		Fluka
bentonite		Kunippia F
bentonite		MX-80
bentonite		Jinchuan
biotite		Nihon
birnessite	delta-MnO2	synthetic
boehmite	gamma-AlOOH	BA Chemical
boehmite	gamma-AlOOH	Rare Mettalic
calcite	CaCO3	Spectrum
calcite	CaCO3	Wyandotte
calcite	CaCO3	Mexico
calcite	CaCO3	Fisher
chlorite		Grimsel
clinoptilolite		Death Valley Junction
ferrihydrite	Fe(OH)3	synthetic
gibbsite	gamma-Al(OH)3	commercial
gibbsite	gamma-Al(OH)3	Merck
gibbsite	gamma-Al(OH)3	Alcan
gibbsite	gamma-Al(OH)3	synthetic
gibbsite	gamma-Al(OH)3	Dalian Luming
gibbsite	gamma-Al(OH)3	Alcan SF-4
gibbsite	gamma-Al(OH)3	Alcoa
goethite	alpha-FeOOH	BASF
goethite	alpha-FeOOH	synthetic
goethite	alpha-FeOOH	Fe-52
goethite	alpha-FeOOH	Bayer
goethite	alpha-FeOOH	Nihon
hematite	alpha-Fe2O3	synthetic
hematite	alpha-Fe2O3	Nihon
hematite	alpha-Fe2O3	Wako
hematite	alpha-Fe2O3	natural
hematite	alpha-Fe2O3	Aldrich
hydrous aluminum oxide		synthetic
hydrous manganese oxide		synthetic
illite		Morris illite
illite		Socodis SA
illite		Marblehead
illite		du Puy
illite		Silver Hill
kaolinite		KGa-1
kaolinite		KGa-1B
kaolinite		Aldrich
kaolinite		KGa-2
lepidocrosite	gamma-FeOOH	Rare Mettalic
magnetite	Fe3O4	Puratronic
magnetite	Fe3O4	Nihon
magnetite	Fe3O4	Wako
montmorillonite		SWy-1

Mineral	Mineral Formula	Mineral Source
montmorillonite		Aldrich, K10
montmorillonite		SAz-1
montmorillonite		SWy-2
muscovite		natural
quartz	SiO2	Merck
quartz	SiO2	Min-U-Sil 30
quartz	SiO2	BDH
quartz	SiO2	Sigma
quartz	SiO2	Wedron #510
vernadite	delta-MnO2	synthetic

Table SI.8. Surface complexation constants reported in the RES^{3}T database for U(VI) sorption to quartz.

SCM Type	Mineral	Surface Area, m^2/g	SiteDensity, sites/nm^2	pK_1	pK_2	logK	Chemical Equation	Literature Reference
DDL	Quartz	0.2	4.81		5.62	-5.72	>Si-$(OH)_2$ + UO_2^{2+} = >Si-O_2-UO_2 + 2 H^+	AZBN00a
DDL	Quartz	0.2	4.81		5.62	-5.51	>Si-$(OH)_2$ + UO_2^{2+} = >Si-O_2-UO_2 + 2 H^+	AZZBN01
DDL	Quartz	0.2	2.3		7.2	5.12	2 >SiOH + UO_2^{2+} = $(>SiOH)_2$-UO_2^{2+}	CSB18
DDL	Quartz	0.2	2.3		7.2	16.943	2 >SiOH + UO_2^{2+} + CO_3^{2-} = $(>SiOH)_2$-UO_2CO_3	CSB18
DDL	Quartz	0.2	2.3		7.2	5.12	2 >SiOH + UO_2^{2+} = $(>SiOH)_2$-UO_2^{2+}	DW14
DDL	Quartz	0.2	2.3		7.2	16.94	2 >SiOH + UO_2^{2+} + CO_3^{2-} = $(>SiOH)_2$-UO_2CO_3	DW14
DDL	Quartz	10	10	-1.6	7.6	-7.259	>X-OH + UO_2^{2+} + H_2O = >X-O-UO_2(OH) + 2 H^+	JHLCH99
DDL	Quartz	10	10	-1.6	7.6	9.529	>X-OH + UO_2^{2+} + CO_3^{2-} = >X-O-$UO_2CO_3^-$ + H^+	JHLCH99
DDL	Quartz	10	10	-1.6	7.6	-1.978	>X-OH + UO_2^{2+} = >X-O-UO_2^+ + H^+	JHLCH99
DDL	Quartz	0.31	2.3	-1.24	7.06	-0.3	>Si-OH + UO_2^{2+} = >Si-O-UO_2^+ + H^+	PJTP01
DDL	Quartz	0.31	2.3	-1.24	7.06	-18.7	>Si-OH + UO_2^{2+} + 3 H_2O = >Si-O-$UO_2(OH)_3^{2-}$ + 4 H^+	PJTP01
DDL	Quartz	0.03	2.3		7.2	0.3	>Si-OH + UO_2^{2+} = >Si-O-UO_2^+ + H^+	PTBP98
DDL	Quartz	0.03	2.3		7.2	-16.75	>Si-OH + 3 UO_2^{2+} + 5 H_2O = >Si-O-$(UO_2)_3(OH)_5$ + 6 H^+	PTBP98
DDL	Quartz	0.03	2.3		7.2	-5.65	>Si-OH + UO_2^{2+} + H_2O = >Si-O-UO_2(OH) + 2 H^+	PTBP98
DDL	Quartz	0.1	2.31		7.2	-8.45	>Si-OH + UO_2^{2+} + 2 H_2O = >Si-O-$UO_2(OH)_2^-$ + 3 H^+	VT98
NE	Quartz	0.33	0			-4.95	>Si(w)-OH + UO_2^{2+} + H_2O = >Si(w)-O-UO_2(OH) + 2 H^+	DK01
NE	Quartz	0.33	0			1.06	>Si(s)-OH + UO_2^{2+} = >Si(s)-O-UO_2^+ + H^+	DK01
NE	Quartz	0.33	0			-3.19	>Si(s)-OH + UO_2^{2+} + H_2O = >Si(s)-O-UO_2(OH) + 2 H^+	DK01
NE	Quartz	0.33	0			-2.56	>Si(s)-OH + UO_2^{2+} + H_2O = >Si(s)-O-UO_2(OH) + 2 H^+	DK01
NE	Quartz	0.33	0			-4.98	>Si(w)-OH + UO_2^{2+} + H_2O = >Si(w)-O-UO_2(OH) + 2 H^+	DK01
NE	Quartz	0.33	0			1.2	>Si(s)-OH + UO_2^{2+} = >Si(s)-O-UO_2^+ + H^+	DK01
NE	Quartz	0.33	0			-4.64	>Si(w)-OH + UO_2^{2+} + H_2O = >Si(w)-O-UO_2(OH) + 2 H+	DK01
NE	Quartz	0.33	0			-0.03	>Si(w)-OH + UO_2^{2+} = >Si(w)-O-UO_2^+ + H^+	DK01
NE	Quartz	0.33	0			-5.28	>Si(w)-OH + UO_2^{2+} + H_2O = >Si(w)-O-UO_2(OH) + 2 H^+	DK01
NE	Quartz	0.33	0			10.183	>Si(w)-OH + UO_2^{2+} + CO3<2-> = >Si(w)-O-$UO_2CO_3^-$ + H^+	DK01
NE	Quartz	0.33	0			-3.28	>Si(s)-OH + UO_2^{2+} + H_2O = >Si(s)-O-UO_2(OH) + 2 H^+	DK01
NE	Quartz	0.33	0			-4.73	>Si-OH + UO_2^{2+} = >Si-O-UO_2(OH) + 2 H^+	KCKD96
NE	Quartz	0.33	0			-5.32	>Si(w)-OH + UO_2^{2+} + H2O = >Si(w)-O-UO_2(OH) + 2 H^+	KCKD96
NE	Quartz	0.33	0			-2.65	>Si(s)-OH + UO_2^{2+} + H2O = >Si(s)-O-UO_2(OH) + 2 H^+	KCKD96
NE	Quartz	0.33	0			-2.56	>Si(s)-$O_{(0.5)}$H + UO_2^{2+} + H2O = >Si(s)-$O_{(0.5)}$-UO_2(OH) + 2 H^+	K02b
NE	Quartz	0.33	0			-7.78	>Si(w)-$O_{(0.5)}$H + UO_2^{2+} + CO_2 + H_2O = >Si(w)-$O_{(0.5)}$-$UO_2CO_3^-$ + 3 H^+	K02b
NE	Quartz	0.33	0			-6.56	>Si(w)-$O_{(0.5)}$H + UO_2^{2+} + H2O = >Si(w)-$O_{(0.5)}$-UO_2(OH) + 2 H^+	K02b
NE	Quartz	0.33	0			-5.57	>Si(s)-$O_{(0.5)}$H + UO_2^{2+} + H2O = >Si(s)-$O_{(0.5)}$-UO_2(OH) + 2 H^+	K02b
NE	Quartz	0.33	0			-6.5	>Si(w)-$O_{(0.5)}$H + UO_2^{2+} + CO2 + H2O = >Si(w)-$O_{(0.5)}$-$UO_2CO_3^-$ + 3 H^+	K02b
NE	Quartz	0.33	0			-5.28	>Si(w)-$O_{(0.5)}$H + UO_2^{2+} + H_2O = >Si(w)-$O_{(0.5)}$-UO_2(OH) + 2 H^+	K02b
TL	Quartz	0.32	0.00184		8.4	1.98	>Si(s)-OH + UO_2^{2+} = >Si(s)-O-UO_2^+ + H^+	FDZ06
TL	Quartz	0.32	0.00184		8.4	-1.88	>Si(s)-OH + UO_2^{2+} + H_2O = >Si(s)-O-UO_2(OH) + H^+	FDZ06

Table SI.9. Fitting results using RES^{3}T surface complexation models coupled with the PHREEQC "llnl.dat" thermodynamic database.

SCM Type	Surface Area m^2/g	Site Density sites/nm^2	Orig pK_1	Orig pK_2	Species	Orig logK	SD[a]	Fit logK	Fit SD	Orig R	Fit R	Orig WSOS	Fit WSOS	Model Reference
DDL	0.2	4.81		5.62	>Si-O_2-UO_2	-5.72	0.03	-5.76	0.03	0.87	0.86	7686	7663	AZBN00a
DDL	0.2	4.81		5.62	>Si-O_2-UO_2	-5.51	0.03	-5.76	0.03	0.86	0.86	8563	7663	AZZBN01
DDL	0.2	2.3		7.2	(>SiOH)$_2$-UO_2^{2+} (>SiOH)$_2$-UO_2CO_3	5.12 16.943	0.09 0.07	3.65 16.51	0.11 0.04	0.72	0.85	18829	7654	CSB18
DDL	0.2	2.3		7.2	(>SiOH)$_2$-UO_2^{2+} (>SiOH)$_2$-UO_2CO_3	5.12 16.94	0.08 0.07	3.65 16.51	0.11 0.04	0.72	0.85	18806	7654	DW14
DDL	10	10	-1.6	7.6	>X-O-UO_2^+ >X-O-UO_2(OH) >X-O-$UO_2CO_3^-$	-1.98 -7.26 9.53	0.11 0.16 0.07	-1.33 -7.55 9.59	0.04 0.30 0.06	0.87	0.88	8117	6055	JHLCH99
DDL	0.31	2.3	-1.24	7.06	>Si-O-UO_2^+ >Si-O-$UO_2(OH)_3^{2-}$	-0.3 -18.7	0.04 0.10	-0.55 -17.29	0.03 0.07	0.79	0.87	10629	6645	PJTP01
DDL	0.03	2.3		7.2	>Si-O-UO_2^+ >Si-O-UO_2(OH) >Si-O-$(UO_2)_3(OH)_5$	0.3 -5.65 -16.75	0.06 0.10 0.47	-0.96 -5.55 -19.31	0.10 0.04 0.95	0.77	0.87	15759	7040	PTBP98
DDL	0.1	2.31		7.2	>Si-O-$UO_2(OH)_2^-$	-8.45	0.08	-10.18	0.04	0.81	0.87	22891	9469	VT98
NE	0.33	0.64w 0.0078s		7.2	>Si(w)-O-UO_2^+ >Si(w)-O-UO_2(OH) >Si(s)-O-UO_2^+ >Si(s)-O-UO_2(OH)	-0.03 -4.95 1.2 -3.19	0.15 0.06 1.02 0.44	-2.18 -4.91 -0.76 -2.58	7.55 0.05 2.17 0.21	0.81	0.86	9850	7869	DK01_3
NE	0.33	0.64w 0.012s		7.2	>Si(w)-O-UO_2(OH) >Si(w)-O-$UO_2CO_3^-$ >Si(s)-O-UO_2(OH)	-5.28 10.18 -2.56	0.07 0.08 0.07	-5.56 9.97 -2.62	0.08 0.07 0.07	0.90	0.90	5976	5495	DK01_4
NE	0.33	0.63w 0.0077s		7.2	>Si(w)-O-UO_2(OH) >Si(s)-O-UO_2^+ >Si(s)-O-UO_2(OH)	-4.98 1.06 -3.28	0.05 0.53 0.33	-4.90 -0.62 -2.47	0.04 49.34 0.19	0.85	0.86	8351	7843	DK01_c4
NE	0.33	0.53		7.2	>Si(w)-O-UO_2(OH)	-4.64	0.03	-4.56	0.03	0.86	0.86	8038	7987	DK01_1
NE	0.33	0.53		7.2	>Si-O-UO_2(OH)	-4.73	0.03	-4.58	0.03	0.86	0.86	8275	7987	KCKD96_1
NE	0.33	0.64w 0.012s		7.2	>Si(w)-O-UO_2(OH) >Si(s)-O-UO_2(OH)	-5.32 -2.65	0.05 0.08	-4.92 -2.82	0.05 0.11	0.85	0.86	8719	7894	KCKD96_2
NE	0.33	0.64w 0.012s		7.2	>Si(w)-$O_{(0.5)}$-UO_2(OH) >Si(w)-$O_{(0.5)}$-$UO_2CO_3^-$ >Si(s)-$O_{(0.5)}$-UO_2(OH)	-5.28 10.17* -2.56	0.07 0.08 0.07	-5.56 9.97 -2.64	0.07 0.07 0.07	0.90	0.90	5994	5504	K02b_2
NE	0.33	0.64w 0.012s		7.2	>Si(w)-$O_{(0.5)}$-UO_2(OH) >Si(w)-$O_{(0.5)}$-$UO_2CO_3^-$ >Si(s)-$O_{(0.5)}$-UO_2(OH)	-6.56 8.89* -5.57	0.38 0.10 1.90	-5.56 9.96 -2.65	0.08 0.07 0.07	0.73	0.90	29414	5503	K02b_1
TL	0.32	4.6051w 0.00184s		8.4	>Si(s)-O-UO_2^+ >Si(s)-O-UO_2(OH) >Si(w)-O-UO_2^+ >Si(w)-O-UO_2(OH) >Si(w)-O-UO_2(OH)CO_3^{2-}	1.98 -1.88 -2.08 -5.76 2.47	0.87 0.33 0.65 0.05 0.37	- -2.07 -1.00 -6.42 4.08	- 0.24 0.08 0.10 0.12	0.88	0.89	6815	5585	FDZ06[b]

* Reaction equation rewritten using CO_3^{--} rather than CO_2(aq)

[a] SD calculated in PEST based on original reaction constant.

[b] Surface species incorrectly identified in RES^{3}T database.