580369

SAND2023-01165R

FINAL EC 03/27/23

Sandia National Laboratories

Operated for the United States Department of Energy
National Nuclear Security Administration
by National Technology and Engineering
Solutions of Sandia, LLC.
Albuquerque, New Mexico 87185-0143
Livermore, California 94551-0969

date: January 30, 2023

to: WIPP Records Center

from: Martin B. Nemer and Paul Domski

subject: Bayesian Inference applied to the Solubility of Phase-3 and Phase-5

WIPP: 4.2.1 :PUB:QA-L:SAND 2023-01165R

Sandia National Laboratories
ENERGY NNSA

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Information Only

1 Introduction

This Section covers an introduction to the objectives and techniques used in this analysis. The objectives of the report are given in Subsection 1.1. An introduction to aqueous thermodynamics and how variance might propagate through the relevant thermodynamic equations is given in Subsection 1.2. An introduction to Bayesian inference and its application to thermodynamic modeling is given in Subsection 1.3.

1.1 Objective

The overall objective of this analysis is to estimate uncertainties in the log K values of phase-3 ($Mg_2(OH)_3Cl{\cdot}4H_2O$) and phase-5 ($Mg_3(OH)_5Cl{\cdot}4H_2O$) to answer EPA's questions regarding the use of phase-5 in future performance assessment (PA) analyses. A broader goal of this memo is to demonstrate the Bayesian inference technique (also known as Bayesian inversion) as applied to the problem of parameter estimation and parameter uncertainty estimation in thermodynamic geochemical modeling. In this report we focus on estimating both the log K values their uncertainty for phase-3 and phase-5 in aqueous systems using the Pitzer equation of state. It is the authors' opinion that the evidence presented herein supports the use of Bayesian inference for estimating uncertainty in thermodynamic geochemical model parameters and uncertainty in thermodynamic geochemical predictions.

1.2 Aqueous Thermodynamics

In this memo we present an initial and simplified Bayesian-inference analysis to estimate the $\log_{10}K$ and associated uncertainty for the dissolution of $Mg_2(OH)_3Cl{\cdot}4H_2O$(cr) and $Mg_3(OH)_5Cl{\cdot}4H_2O$(cr) in high-ionic-strength aqueous solutions. The dissolution reaction considered can be written as,

$$Mg_2(OH)_3Cl{\cdot}4H_2O + 3\ H^+ \leftrightarrow 2\ Mg^{2+} + Cl^- + 7\ H_2O, \quad (1)$$

or alternatively

$$Mg_2(OH)_3Cl{\cdot}4H_2O \leftrightarrow 2\ Mg^{2+} + Cl^- + 4\ H_2O + 3OH^-, \quad (2)$$

for the phase colloquially known as phase-3 in the literature. Similarly,

$$Mg_3(OH)_5Cl{\cdot}4H_2O + 5\ H^+ \leftrightarrow 3\ Mg^{2+} + Cl^- + 9\ H_2O, \quad (3)$$

or alternatively

$$Mg_3(OH)_5Cl{\cdot}4H_2O \leftrightarrow 3\ Mg^{2+} + Cl^- + 3\ H_2O + 5\ OH^-, \quad (4)$$

for phase-5. The final equilibrium concentrations of ions in solution in equilibrium with phase-3 can be calculated using standard aqueous equilibrium thermodynamics. At equilibrium

$$2\log_{10}(a_{Mg^{2+}}) + \log_{10}(a_{Cl^-}) + 7\log_{10}(a_{H_2O}) - 3\log_{10}(a_{H^+}) = \log_{10}K, \quad (5)$$

where a_X is the activity of species X in solution, and K is the equilibrium constant, which depends on temperature but is independent of the concentration of species in solution. A similar equation can be written for phase-5. Because equation (5) represents a linear relationship between $\log_{10}(a_X)$ and $\log_{10}K$, it follows [1] that any variance on the right-hand-side of equation (5) will linearly propagate to variance and covariance in the terms of the left-hand side of equation (5). That is for a sum of n random variables Z_i,

$$F = \sum_{i=1}^{n} c_i Z_i \quad (6)$$

Information Only

$$V(F)=\sum_{i=1}^{n}c_i^2V(Z_i)+2\sum_{i=1}^{n}\sum_{j=i+1}^{n}c_ic_jCov(Z_i,Z_j) \tag{7}$$

where c_i are fixed deterministic constants, $V(\cdot)$ is the variance and $Cov(\cdot,\cdot)$ is the covariance. The variance is defined as

$$V(Z_i)=E\left[\left(Z_i-\overline{Z}_i\right)\left(Z_i-\overline{Z}_i\right)\right], \tag{8}$$

where $E[\cdot]$ is the expected value of a random variable, and $\overline{Z}_i=E[Z_i]$. The covariance is then defined as

$$Cov(Z_i,Z_j)=E\left[\left(Z_i-\overline{Z}_i\right)\left(Z_j-\overline{Z}_j\right)\right]. \tag{9}$$

For the purpose of determining expected values in this report, we are always using a finite set of samples (realizations or observations) from a distribution, hence

$$E\left[Z_i\right]=\frac{1}{m}\sum_{j=1}^{m}z_{i,m} \tag{10}$$

Where $z_{i,m}$ are the m samples (realizations or observations) of random variable Z_i.

The activities and concentrations of ions in solution are related to each other through a thermodynamic equation of state (EOS), which maps activity to molality through an activity coefficient defined as,

$$a_X=\gamma_X m_X \tag{11}$$

where γ_X is the activity coefficient and m_X is the molality of species X in solution defined as moles of solute per kg water in solution. The activity coefficient is, in general, a function of the concentration of all the species in solution. For high ionic strength solutions ($I>2$), the Pitzer model is often used. The Pitzer EOS [2] models activity coefficients as a virial expansion in ionic strength,

$$\log_{10}\gamma_X=\frac{z_X^2}{2}f'(I)+\sum_{i=1}^{k}\lambda_{iX}(I)m_i+\sum_{i=1}^{k}\sum_{j=1}^{k}\lambda_{ij}'(I)m_im_j+\ldots \tag{12}$$

where the first term represents the impact of ionic strength alone, the second term represents pair interactions between the ion X and other species of interest, and additional terms represent triplet interactions followed by quartet interactions and so on. Along with equations (5)-(12), is a set of mass balance equations and charge balance equations. Every mole of each atom that was initially in solution must at equilibrium either remain in solution or be transferred to the solid phase (mass balance). Charge balance requires zero net charge in solution at equilibrium. Equations (5)-(12) plus mass balance and charge balance are solved simultaneously to determine the equilibrium activities and concentrations in solution, and the amount of solid phase that is formed or dissolved. In this work the EQ3/6 code [3] was used to determine the equilibrium concentrations of ions in solution, and the amount of solid that remained in equilibrium with the solution. This is discussed in Section 3.2, with a description of the EQ3/6 scripting used left to an appendix (Section 9).

1.3 Bayesian Inference

Bayesian inference or Bayesian inversion is a statistical technique that has gained popularity in recent years due to the availability of open-source code packages and the availability of computing power [4]. Bayesian inference has been applied to various inverse problems including those in aqueous thermodynamics [5], using the Pitzer EOS [6, 7], although the idea is still somewhat new in the field. A notable quote from Sakar (2012) [7] confirming this assertion reads "*A considerable amount of research has been reported on the forward propagation of assumed uncertainty in the chemical equilibrium*

Information Only

models in geochemistry and geophysics; but the inverse problem of uncertainty quantification in the chemical equilibrium modeling has not received much attention." Despite Sakar's dearth of found literature on this subject, the authors utilized a frequentist approach [8, 9] to determining the uncertainty in fitted parameters [10] in previous work. Other authors have subsequently applied a traditional frequentist approach to estimating parameter uncertainty in aqueous thermodynamic model fitting [11]. However, as will be shown, the Bayesian framework is flexible and doesn't require the assumption that the data is normally distributed. As an aside, INTERA Inc. is utilizing the linearized Bayesian method for parameter estimation within the PEST framework [12] for modeling hydrology; an explanation of the linearized Bayesian method is given in Section 1.3.1.1.

Probably theory [13, 14] describes the relationship between how experimental data is distributed for a given model + parameters and how the model's parameters are distributed for given experimental data using conditional probability,

$$p(\theta \mid y)p(y) = p(y \mid \theta)p(\theta) \tag{13}$$

where y is the set of experimental data that is being fit and analyzed, $p(y|\theta)$ is the probability density of the experimental data for a given vector of model parameters θ, $p(\theta|y)$ is the probability density of the model parameters for a given set of experimental data y, and $p(y)$ and $p(\theta)$ are the marginalized probability densities of the experimental data and model parameters, respectively.

Bayesian statistics [13, 14] takes equation (13) and assigns a meaning to $p(\theta)$ as the prior distribution, which represents our prior knowledge or lack thereof on the model parameters θ [13, 14]. Then $p(\theta|y)$ is the posterior distribution, which represents how our beliefs in the model parameters should be updated given the experimental data set y. The probability distribution $p(y|\theta)$ is then the data model (a.k.a the likelihood model) for the experimental data y for a given set of model parameters θ. The marginal distribution on y, $p(y)$ is then integrated out to obtain the posterior distribution

$$p(\theta \mid y) = \frac{p(y \mid \theta)p(\theta)}{\int p(y \mid \theta)p(\theta)d\theta} \tag{14}$$

Equation (14) provides information as to the set of possible model parameters and their probability, but depends heavily if not entirely on the data model.

1.3.1 Data Model

The data model is the key assumption in this and any future Bayesian inference model for determining uncertainty in the model parameters. The data model is often determined subjectively by visually examining how the data is distributed versus the model, although there exist quantitative tests for judging how closely data follows a given distribution [15]. Ultimately however, the choice of data model is a subjective choice guided by the data. In Section 4.2, the experimental data is plotted as a Q-Q plot to visually examine the adequacy of the current data-model assumption. A Q-Q plot plots a quantile or metric on a quantile of the experimental data assuming it follows some known distribution (here that's log normal for molality) versus the actual observed metric. Herein, the metric used for the Q-Q plot was the Z-score, which is explained in Section 4.2.1.1.

In aqueous thermodynamics, it is often found [16-23] that the difference between the experimental data and an adequate equilibrium thermodynamic model for the experiment is normally distributed on a log scale,

$$\log_{10}\left(m_{meas}\right) - \log_{10}\left(m_{\mathrm{mod}}(\xi,\theta)\right) \approx \varepsilon(\xi), \tag{15}$$

$$\varepsilon(\xi) \sim N\left(\mu(\xi), \sigma^2(\xi)\right) \tag{16}$$

Information Only

where m_{meas} is the experimentally measured concentration (herein molality) of some ion in solution and $m_{\mathrm{mod}}(\xi,\theta)$ is the modeled concentration of the same ion in solution obtained from an equilibrium thermodynamics code with model parameters θ and independent variables ξ.

Equation (15)-(16) follow from (7) when the covariances are small, the activity coefficients are relatively fixed as a function of the Pitzer parameters, and for a uniquely-valued $\log_{10}$K parameter that carries with it a normally distributed source of error. Alternatively, and perhaps preferably, one could consider the model parameters as being irreducibly distributed, e.g. $\log_{10}$K is (in the real world) a distributed parameter rather than having a single true value. We have not found any statistical-mechanics justification for a normally distributed $\log_{10}$K, but the authors speculate that the distribution of imperfections in the solid-phase crystal structure undergoing dissolution or a differently crystalized stochastic surface layer could account for normality in $\log_{10}$K. In future work it would be interesting to compare some TBD (to be determined) metric of crystallinity from XRD analysis to any subsequent solubility measurements and their uncertainty. A simple statistical mechanics analysis of the impact of crystal dislocations on the resulting distribution of $\log_{10}$K would also be useful but has not been found in the scientific literature by the authors. However, to be clear we did not assume $\log_{10}$K is a distributed parameter in this report; this is left to future work.

Readers with experience in electrode-based measurements may recognize equation (15). These measurements include pH = -$\log_{10}(a_{H^+})$, pmH, pcH = -$\log_{10}([\mathrm{H}^+])$, and other types of electro-potential based measurements, since

$$pmH = -\log_{10}(m_{H+}), \tag{17}$$

and normally distributed errors in the measured potential are often observed [24]. For matrix ions, whose concentrations vary little during the course of a solubility experiment, e.g. Mg^{2+}, Cl^-, Na^+, errors may not follow equation (15)-(16), and may be normally distributed on a linear scale due to Inductively Coupled Plasma (ICP) measurement error being the primary source of uncertainty. Hence, one size does not fit all, and uncertainty must be looked at on a case-by-case basis.

The mean $\mu(\xi)$ in equation (16) is the bias between the experiment and model, which could be due to either an imperfect model or a measurement bias from the "true" value. Given that both m_{meas} and m_{mod} are functions of the input chemical conditions ξ, we postulate that the stochastic error ε and the mean μ to be functions of those conditions

$$\log_{10}\left(m_{meas}\right) - \log_{10}\left(m_{\mathrm{mod}}(\xi,\theta)\right) \approx \mu(\xi,\theta) + \varepsilon(\xi), \tag{18}$$

$$\varepsilon(\xi) \sim N(0,\sigma^2(\xi)), \tag{19}$$

where $\xi \equiv$ set of input chemical conditions to the model (i.e. the independent variables, of which there may be none), and $\mu(\xi,\theta)$ is the bias between the data and the model of potentially unknown origin; here we allow the bias term to be a function of the model parameters to make it clear that the bias will vary with θ.

If the model and experiment are in some sense close, i.e. the model is "good", then the bias term can be neglected; herein, we neglected the bias term. Using equation (18)-(19), neglecting the bias term, and assuming that all the data are independently and identically distributed leads to the following data model,

$$p(y_{meas} \mid \theta,\sigma) = \left(2\pi\sigma^2\right)^{-\frac{nk}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{k}\left(\log_{10}\left[m_{meas,ij}\right] - \log_{10}\left[m_{\mathrm{mod},j}(\xi_i,\theta)\right]\right)^2\right], \tag{20}$$

Information Only

where the sum is taken over all the n independent experiments performed, and over k measurable ions (or in some cases perhaps species, if speciation is somehow measured) measured in solution in each experiment. Here $m_{\mathrm{mod},j}(\xi_i,\theta)$ is a model of the j^{th} ion in solution as a function of the input chemical conditions in the i^{th} independent experiment ξ_i, and model parameters θ. In equation (20) we define y_{meas} as the full set of experimental data

$$y_{meas} = \{m_{meas,ij}; 1 \le i \le n, 1 \le j \le k\}. \tag{21}$$

There is only one value of σ in equation (20), corresponding to the admittedly doubtful assumption that each measurement is identically distributed and that all the different ions in solution all have the same value of σ. This need not be assumed, and different groups of data could have their own value of σ; this issue will come up again in Section 2.1 where two datasets from two different researchers are combined and compared to the data model.

Another point is that the data model, in general, can accommodate both log-normally and normally distributed data (and other types of distributed data). E.g., suppose an experiment has both *pm*H and $m\mathrm{Mg}^{2+}$ measurements and both are believed to be critical to determining the $\log_{10}$K value. Further suppose that the $m\mathrm{Mg}^{2+}$ measurements are dominated by normally distributed ICP measurement error. Then we might re-write the objective function as,

$$p(y_{meas} \mid \theta,\sigma) \sim \exp\left[-\frac{1}{2\sigma_{pmH}^{\ 2}}\sum_{i=1}^{n}\left(pmH_{meas,i} - pmH_{\mathrm{mod}}(\xi_i,\theta)\right)^2 - \frac{1}{2\sigma_{ICP}^{\ 2}}\sum_{i=1}^{n}\left(mMg^{2+}_{meas,i} - mMg^{2+}_{\mathrm{mod}}(\xi_i,\theta)\right)^2\right], \tag{22}$$

Where now each measurement type has its own distribution (log normal for $m\mathrm{H}^{+}$ and normal for $m\mathrm{Mg}^{2+}$) and its own value of sigma.

We did not use equation (22) in the present analysis, but it is planned in a follow up analysis for this and future cases, as discussed in Section 5.2. The only real limitation that the authors' place on the data model herein is the requirement of linearity, i.e.

$$Y_{meas} - Y_{\mathrm{mod}}(\xi,\theta) \approx \varepsilon(\xi), \tag{23}$$

where Y_{meas} is some measurement or mapped measurement (e.g. $\log_{10}(m_{meas})$), $Y_{mod}(\xi,\theta)$ is a model for that measurement, and $\varepsilon(\xi)$ is an error term that is stochastic and distributed in some way that can be calculated and sampled. Equation (23) is not a fundamental statistical requirement; bootstrapping can be used (numerically) to circumvent the apparent difficulty of non-linearity in the error term associated with the experimental data if the uncertainty in the experimental data is well characterized.

Equation (20) can be re-written in terms of the sum of square error between the model and the experiment

$$p(y_{meas} \mid \theta,\sigma) = \left(2\pi\sigma^2\right)^{-\frac{nk}{2}} \exp\left[-\frac{1}{2\sigma^2}S(\theta)\right] \tag{24}$$

with $S(\theta)$ as the sum of square error as usually defined,

$$S(\theta) = \sum_{i=1}^{n}\sum_{j=1}^{k}\left(\log_{10}\left[m_{meas,ij}\right] - \log_{10}\left[m_{\mathrm{mod},i}(\xi_j,\theta)\right]\right)^2 \tag{25}$$

Equation (20) ignores any potential covariances between the errors of different ions in a given experiment (i.e. equation (7)), and any covariances between different experiments. These type of covariances could arise for example if Cl^- was determined from charge balance, given ICP measurements on Mg^{2+} and Na^+ in solution. Then clearly the Cl^- concentration measurement would

Information Only

have a covariance with Mg^{2+} and Na^{+} concentrations. In that case we might simply leave the Cl^{-} concentration out of the sum-of-square function, since it adds little-to-no additional information. A weighting, for example, could be applied in this case to include the effect Cl^{-}.

In the analysis herein, we reduce the dataset to only *pm*H and *pm*OH measurements and the only model parameter being determined is the scalar value of $\theta = \log_{10}K$.

An Achilles heel of equation (18) is the assumption that $\mu(\xi,\theta)$ is small and can be neglected. The model will become biased if the model parameters θ depart too far from the "good" solution. The best remedy is to fix the model, and/or obtain enough data such that the inferred posterior uncertainty in the model parameters θ is small enough such that the assumption that $\mu(\xi,\theta)$ is small remains valid. Regardless, given mild assumptions on continuity, there should exist a neighborhood of θ where equation (24) holds, even if the window of probability for which it holds is small. As will be shown in Section 4.2, the data model appears to the authors to adequately model the distance between the experimental data from the model.

1.3.1.1 *Linearized Bayesian Inference*

For the purposes of computational speed, convenience, and simplicity, the sum-of-square error $S(\theta)$ in equation (24) is sometimes linearized around the solution point θ_{min}. Expanding $S(\theta)$ in a Taylor series about θ_{min} yields,

$$S(\theta) \approx S(\theta_{\min}) + \left.\frac{\partial S}{\partial \theta}\right|_{\theta=\theta_{\min}} \bullet (\theta - \theta_{\min}) + \frac{1}{2}(\theta - \theta_{\min}) \bullet \left.\frac{\partial^2 S}{\partial \theta^2}\right|_{\theta=\theta_{\min}} (\theta - \theta_{\min}) . \tag{26}$$

The first term in equation (26) is a constant, and can be absorbed into the integration constant for the probability density in equation (14). The second term in equation (26) is assumed to be zero, assuming the solution point accounts for a local minimum and not on an boundary of the parameter set that is being explored. Then only the third term remains, which yields

$$p(y_{meas} \mid \theta,\sigma) \approx p_{lin}(y_{meas} \mid \theta,\sigma) \sim \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}(\theta - \theta_{\min}) \bullet \left.\frac{\partial^2 S}{\partial \theta^2}\right|_{\theta=\theta_{\min}} (\theta - \theta_{\min})\right]. \tag{27}$$

Readers with experience in multivariate gaussian distributions will recognize the term inside the square brackets as the square of the Mahalanobis distance [25] between θ and θ_{min}, implying that

$$\frac{1}{\sigma^2}\left.\frac{\partial^2 S}{\partial \theta^2}\right|_{\theta=\theta_{\min}} = \Sigma^{-1} \tag{28}$$

where Σ is the covariance matrix as customarily defined. Thus, for regions where $\left.\frac{\partial^2 S}{\partial \theta^2}\right|_{\theta=\theta_{\min}}$ is a nearly singular matrix, we would expect the condition number of the covariance matrix to be correspondingly large. This would seem to be a useful tool to investigate whether the inverse is ill-posed in the neighborhood of θ_{min}, and in general the authors agree with that sentiment. However, in our experience a near-singular $\left.\frac{\partial^2 S}{\partial \theta^2}\right|_{\theta=\theta_{\min}}$ does not always imply the non-linear inverse is completely ill-posed, as higher order terms often will become significant even if $\left.\frac{\partial^2 S}{\partial \theta^2}\right|_{\theta=\theta_{\min}} \bullet (\theta - \theta_{\min})$ or $(\theta - \theta_{\min}) \bullet \left.\frac{\partial^2 S}{\partial \theta^2}\right|_{\theta=\theta_{\min}}$ is small. The influence of higher order terms makes relying on linear Bayesian analysis alone questionable,

Information Only

particularly when $\left.\frac{\partial^2 S}{\partial \theta^2}\right|_{\theta=\theta_{\min}}$ is ill-conditioned. Rather in the authors' opinion it's often useful to use the linearized version of $S(\theta)$ to help visualize and understand why the inverse in a particular region of the parameter space appears to be ill-posed. On the other hand, in the cases shown herein we find that a constant second term $\left.\frac{\partial^2 S}{\partial \theta^2}\right|_{\theta=\theta_{\min}}$ captures nearly all the variance in the problem and that $\left.\frac{\partial^2 S}{\partial \theta^2}\right|_{\theta=\theta_{\min}}$ and the inverse are well posed.

Another advantage of equation (27) is that it produces a multivariate gaussian distribution with known properties. That same advantage forces the posterior distribution to be gaussian when the priors are uniform, which forces the model parameters θ to follow a gaussian distribution.

1.3.2 Prior Distribution on θ

The next term that is needed is the prior $p(\theta)$, and because we allow σ to be fit in the present analysis we also need a prior for $p(\sigma)$. The choice of priors tends to be of one the more controversial element of Bayesian inference, as it involves subjective choice and hence a human-biased choice. For a deeper discussion of prior distributions, we refer the reader to pg. 11 of [14], chapters 1 and 2 of [26], and the seminal paper by Richard Cox [27]. An important point obtained from this literature is that there is often some element of subjectivity and bias in statistics, however Bayesian inference at least makes those biases explicit.

In this report, we have tried to choose prior distributions that span the range of possible parameter values, while not biasing any particular value of θ. Thus, for θ we use uniform distributions over a range that we (subjectively) believe covers the possible range of parameter values, given the information we currently have. For σ, we use the so-called un-informed distribution (see pg. 31 of [28]) for the same reasons, although the form of the prior is different from that for θ because of the fact that σ^{nk} resides in the denominator of equation (20).

1.3.3 Determination of the Posterior $p(\theta,\sigma|y)$

The dimensionality of the posterior distribution $p(\theta,\sigma|y)$ for the work herein is only two but for multi-parameter posterior distributions, determining the integral in the denominator of equation (14) can be difficult. For that reason, algorithms have been developed to determine $p(\theta,\sigma|y)$ by sampling the data model and the prior distributions. The method used herein is the Markov Chain Monte Carlo method (MCMC) [26], which produces samples of the posterior distribution given a data model and prior distributions. In this work, we used the pyMC4 package [4] in Python to perform the sampling of the posterior distribution. The pyMC4 [4] package requires the programmer to define the prior distributions and define a subroutine that calculates the log-likelihood function. The log likelihood function used herein is just the natural log of equation (24),

$$L(y\,|\,\theta,\sigma) = \ln\left[p(y\,|\,\theta,\sigma)\right] = -\frac{n}{2}\ln\left[2\pi\sigma^2\right] - \frac{S(\theta)}{2\sigma^2}. \tag{29}$$

Because many samples are required to obtain statistics on the posterior distribution, it is currently infeasible to use EQ3/6 directly in the pyMC package to determine $L(y|\theta,\sigma)$. Instead, $S(\theta)$ is determined over a grid of possible parameter values θ using EQ3/6, and then a polynomial fit is used to approximate $S(\theta)$, as closely as needed/desired. This fit is often called the surrogate model, a reduced order model, or sometimes a response surface. Regardless of the name, it's a fit of the model results as a function of θ. It should be noted that the MCMC algorithm provides only the integration of the normalization constant for determining the posterior density distribution. In this case the MCMC algorithm could have been replaced by standard integration since there is only one parameter $\theta = log_{10}K$, and σ are being fit, and

Information Only

there is no deterministic relationship between θ and σ, only a probabilistic one through the log-likelihood function.

1.3.4 Posterior Predictive Distribution

Once the posterior distribution $p(\theta,\sigma|y)$ is determined, a predicted dataset y_{pred} can be calculated via the posterior predictive distribution $p(y_{pred}|y_{meas})$ and compared to the experimentally obtained dataset y_{meas}. This is obtained by integrating the product of the posterior distribution the data distribution over θ and σ,

$$p(y_{pred} \mid y_{meas}) = \int p(\theta,\sigma \mid y_{meas}) p(y_{pred} \mid \theta,\sigma) d\theta d\sigma . \tag{30}$$

Information Only

2 Experimental Data

The experimental data used in the analysis is described in Subsections 2.1 and 2.2. We do not claim that this represents an entirely inclusive set of experimental data available in the literature. Inclusion of additional data could be performed in future work.

2.1 Phase-3 Data

The experimental data used herein for phase-3 came from two semi-independent sources. The first set of data was published by Altmaier et. al (2003) [29]. The experimental data was collected for the equilibrium solubility of phase-3 in various concentrations of $MgCl_2$(aq) at 298K and at one atmosphere of pressure. The dependent variable collected in these experiments was the $pm\mathrm{H} = -\log_{10}(m\mathrm{H}^+)$, which was obtained using a Ross pH electrode and Gran titrations to calibrate the probe to high ionic strengths. From Altmaier et. al (2003) [29], "*The analytical uncertainty of the ICP-MS analysis (ELAN 6100, Perkin Elmer) is 5–10%.*" and "*Log aHCl was measured at the end of each experiment and at different times in 0.5 and 5.6 m NaCl and in 0.25, 2.67 and 5.15 m $MgCl_2$. H^+ concentrations measured by the two methods described above were generally consistent within ± 0.04 log-units, with the exception of the values in concentrated $MgCl_2$ which showed a certain systematic deviation. The average deviation between the log mH^+ values determined with ROSS electrodes and those calculated from aHCl measurements was 0.01 +- 0.03 in 0.5 m NaCl, 0.00 ± 0.03 in 5.6 m NaCl, 0.02 ± 0.04 in 0.25 m $MgCl_2$, 0.04 ± 0.03 in 2.67 m $MgCl_2$, and 0.08 ± 0.04 in 5.15 m $MgCl_2$.*" The Altmaier et. al (2003) data are shown in Table 1.

Table 1. Experimental data for phase-3 from Altmaier et. al (2003) [29]

Simulation ID	*pm*H (mol/kg H_2O)	*m*$MgCl_2$ (mol/kg H_2O)
altm_1	8.87	2.11
altm_2	8.80	2.67
altm_3	8.74	3.26
altm_4	8.72	3.87
altm_5	8.74	4.50
altm_6	8.75	5.15

The second set of experimental data used herein was obtained from Pannach et. al (2017) [30]. In this more-recent experimental work, phase-3 was equilibrated with $MgCl_2$ solutions of varying strength. Unlike the Altmaier work, in Pannach $pm\mathrm{OH}^- = -\log_{10}(m\mathrm{OH}^-)$ was determined in the final equilibrium solutions by removing an aliquot and performing a potentiometric titration. The Pannach work is notable because they claim that phase-3 is the only stable phase in the solutions that they tested. From Pannach et. al (2017) [30], "*For each solution sample, the OH− concentration determination was carried out at least three times (uncertainty ur ± 0.5%).*" and "*The magnesium concentration was determined by complexometric titration with 0.05 M sodium ethylendiaminetetraacetate (Na-EDTA) in a NH_3/NH_4Cl buffered solution (pH 9−10) using Erio T as the indicator (uncertainty u_r ± 0.3%). The chloride concentration was determined by applying the Mohr method with 0.1 M $AgNO_3$ in a*

Information Only

$Na_2CO_3/NaHCO_3$ buffered solution (pH 8−9) using K_2CrO_4 as the indicator (uncertainty $u_r \pm 0.2\%$)." The experimental data is shown in Table 2.

Table 2. Pannach et. al (2017) [30] data for phase-3 solubility.

Simulation ID	*m*$Mg(OH)_2$ (mol/kg H_2O)	*m*$MgCl_2$ (mol/kg H_2O)
pann_1	0.00134	1.521
pann_2	0.00165	2.042
pann_3	0.00143	2.556
pann_4	0.00175	3.066
pann_5	0.00187	3.512
pann_6	0.00183	3.584
pann_7	0.00205	3.973
pann_8	0.00202	4.131
pann_9	0.00237	5.057
pann_10	0.00264	5.070
pann_11	0.00294	5.456

2.2 Phase-5 Data

The experimental data used herein for phase-5 came from only one study, Xiong et. al (2010) [31]. In this experiment *pm*H was measured analogously to Altmaier et. al [29]. In the Xiong et. al (2010) paper there is a discrepancy between the measured magnesium and chloride values and the values form the preparation of the solutions used. Herein, we use the prepared values. Determining the cause and impact of this discrepancy is left to potential future work. We do not see any experimental methodology uncertainty estimates given in Xiong et. al (2010) [31], however full experimental replicates are reported. The data used is shown in Table 3.

Information Only

Table 3. Xiong et. al (2010) [31] data for phase-3 solubility.

Simulation ID	Experimental ID used in Xiong et. al (2010)	Sample #	*pm*H (dimensionless)	*m*$MgCl_2$ (mol/kg H_2O)	*m*NaCl (mol/kg H_2O)
p5_1	P5-MgCl-A-1	1	9.50	0.5	5.0
p5_2	P5-MgCl-A-1	2	9.52	0.5	5.0
p5_3	P5-MgCl-A-2	1	9.48	0.5	5.0
p5_4	P5-MgCl-A-2	2	9.53	0.5	5.0
p5_5	P5-MgCl-B-1	1	9.32	0.8	3.5
p5_6	P5-MgCl-B-1	2	9.36	0.8	3.5
p5_7	P5-MgCl-B-2	1	9.30	0.8	3.5
p5_8	P5-MgCl-B-2	2	9.35	0.8	3.5
p5_9	P5-MgCl-C-1	1	9.16	1.25	3.0
p5_10	P5-MgCl-C-1	2	9.22	1.25	3.0
p5_11	P5-MgCl-C-2	1	9.16	1.25	3.0
p5_12	P5-MgCl-C-2	2	9.22	1.25	3.0
p5_13	P5-MgCl-D-1	1	9.10	1.5	2.0
p5_14	P5-MgCl-D-1	2	9.13	1.5	2.0
p5_15	P5-MgCl-D-2	1	9.06	1.5	2.0
p5_16	P5-MgCl-D-2	2	9.13	1.5	2.0
p5_17	P5-MgCl-E-1	1	9.09	1.75	1.5
p5_18	P5-MgCl-E-1	2	9.13	1.75	1.5
p5_19	P5-MgCl-E-2	1	9.10	1.75	1.5
p5_20	P5-MgCl-E-2	2	9.13	1.75	1.5
p5_21	P5-MgCl-F-1	1	9.00	2.0	0.5
p5_22	P5-MgCl-F-1	2	9.06	2.0	0.5
p5_23	P5-MgCl-F-2	1	8.98	2.0	0.5

Information Only

p5_24	P5-MgCl-F-2	2	9.04	2.0	0.5

Information Only

3 Model Setup

Setup of the probability model is discussed in the following subsections. The objective function $S(\theta)$ construction is described in Subsection 3.1. Development of surrogates for the geochemical model and the objective function are described in Subsection 3.2. Construction of the prior distributions is described in Subsection 3.3. Finally, the calculation of the posterior distribution is discussed in Subsection 3.4.

3.1 Objective Function

The objective functions $S(\theta)$ are given in the following Subsections for both phase-3 and phase-5. As discussed in Section 1.3.3., herein we model the log-likelihood as linearly proportional to the objective function. This assumption corresponds to an exponential probability density for the likelihood (data model).

3.1.1 Phase-3

The specific objective function $S(\theta)$ used is given as follows. For each experiment we calculated a residual $R_j(\log_{10}K)$ as

$$R_i(\log_{10}K) = \left\{ \begin{array}{l} \left(pm\mathrm{OH}^-_{meas,i} - pm\mathrm{OH}^-_{\mathrm{mod},i}(\xi_i, \log_{10}K) \right) \textit{ for Pannach} \\ \left(pm\mathrm{H}^+_{meas,i} - pm\mathrm{H}^+_{\mathrm{mod},i}(\xi_i, \log_{10}K) \right) \textit{ for Altmaier} \end{array} \right\} \tag{31}$$

Then the sum of squares error is just the sum over all experiments of the square of the residuals

$$S(\theta) = \sum_{i=i}^{n} \left(R_i(\log_{10}K) \right)^2 . \tag{32}$$

Notice that in equation (31) the value of the objective function depends on the experiment, e.g. from Pannach or from Altmaier because the two researchers used different measurements (*pm*OH and *pm*H) to characterize solubility. The log scaling removes the difference in scale between the two measurements (i.e. difference in logs is the log of the ratio). No additional weighting was used. Notice also that one distribution for σ was assigned to the combined Pannach & Altmaier datasets as a whole; this is an admittedly subjective choice based on a Q-Q plot of the residuals given by (31), as discussed in Section 4.2.

3.1.2 Phase-5

The objective function used for the phase-5 analysis is simpler than that of phase-3 given that there's only one set of data and all measurements were performed in the same way,

$$R_j(\log_{10}K) = \left(pm\mathrm{H}^+_{meas,j} - pm\mathrm{H}^+_{\mathrm{mod},j}(\xi_j, \log_{10}K) \right), \tag{33}$$

with the sum-of-square function given by equation (32).

3.2 Surrogates

For both phase-3 and phase-5, the model results for all ions in solution and the total summed objective function were first calculated by running EQ3/6 for each simulation condition in Table 1 - Table 3, for 50 values of $\log_{10}K$ of the respective phase being dissolved. EQ3 was used to setup the simulation, followed by EQ6 to determine the final equilibrium concentrations after titrating in the respective solid (phase-3 or phase-5). The details of the simulation preparation, running, and post-processing are left to an Appendix (Section 9) as this use of python scripting to setup EQ3/6 input files, run EQ3/6, and then

Information Only

collect the results from the EQ3/6 output files has been performed routinely on the WIPP project since at least 2011 [10].

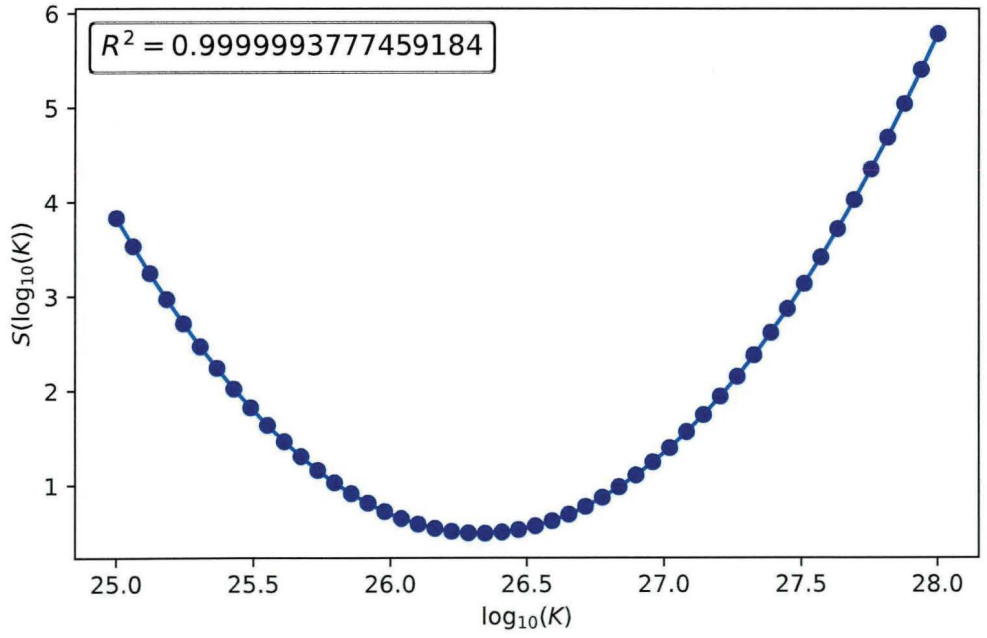
Two surrogates were generated based on the calculations described in the paragraph above. A surrogate for the objective function $S(\theta)$ and surrogates for the individual models $pm\mathrm{OH}^-_{\mathrm{mod},i}(\xi_i, \log_{10} K)$ and $pm\mathrm{H}^+_{\mathrm{mod},i}(\xi_i, \log_{10} K)$. These surrogates are used for computational speed. For the objective function, a parabolic fit to $S(\theta)$ was used; the reasons for which are described in the following subsection (Section 3.2.1). For the models $pm\mathrm{OH}^-_{\mathrm{mod},i}(\xi_i, \log_{10} K)$ and $pm\mathrm{H}^+_{\mathrm{mod},i}(\xi_i, \log_{10} K)$, we used a linear interpolation.

3.2.1 Surrogate Objective Function for Phase-3

The data and the fit are shown below in Figure 1 for the objective function for phase-3. The R^2 value for the fit shows that the sum of squares function is nearly a perfect parabola for the objective function. This implies that the modeled solution ion concentrations (H^+, and OH^-) in the objective function are nearly linear functions of $\log_{10}K$, when holding all other variables fixed. This likely occurred because the activity coefficients of these ions aren't strongly affected by the $\log_{10}K$ value, in the neighborhood of $\log_{10}K$ values explored, but rather are nearly fixed for each background $MgCl_2$ solution that was experimentally measured. This can be seen most clearly by combining equations (5) and (11),

$$2\log_{10}(m_{Mg^{2+}}) + \log_{10}(m_{Cl^-}) - 3\log_{10}(m_{H^+}) = \log_{10}K + (-7\log_{10}(a_{H2O}) - 2\log_{10}(\gamma_{Mg^{2+}}) - \log_{10}(\gamma_{Cl^-}) + 3\log_{10}(\gamma_{H^+})) \quad (34)$$

If all of the terms in parenthesis on the right hand side of equation (34) remain roughly fixed for each experiment, then the left-hand side depends linearly only on $\log_{10}K$. This is reasonable because the solubility of phase-3 (and phase-5) isn't large, and hence within the relatively narrow window of $\log_{10}K$ values that are explored, changes in the $\log_{10}K$ value disturbs total ionic strength and the concentrations of the major ions only slightly. A rigorous sensitivity analysis would be needed to confirm the above hypothesis. Furthermore, this parabolic response wouldn't necessarily be observed for highly soluble phases, or for fits that also include Pitzer parameters.



Information Only

Figure 1. $S(\log_{10}K)$ versus $\log_{10}K$ using equation (31) (blue dots) and quadratic fit (blue line) for phase-3. There is a clear minimum at $\log_{10}K \approx 26.3$.

3.2.1 Surrogate Objective Function for Phase-5

The data and surrogate objective function for phase-5 is shown in Figure 2. This error rises more rapidly away from the minimum than does Figure 1. We would suspect this is partially due to the additional data in this experiment (24 experiments in Xiong et. al (2011) [31]). A minimum occurs for $\log_{10}K \approx 43.8$.

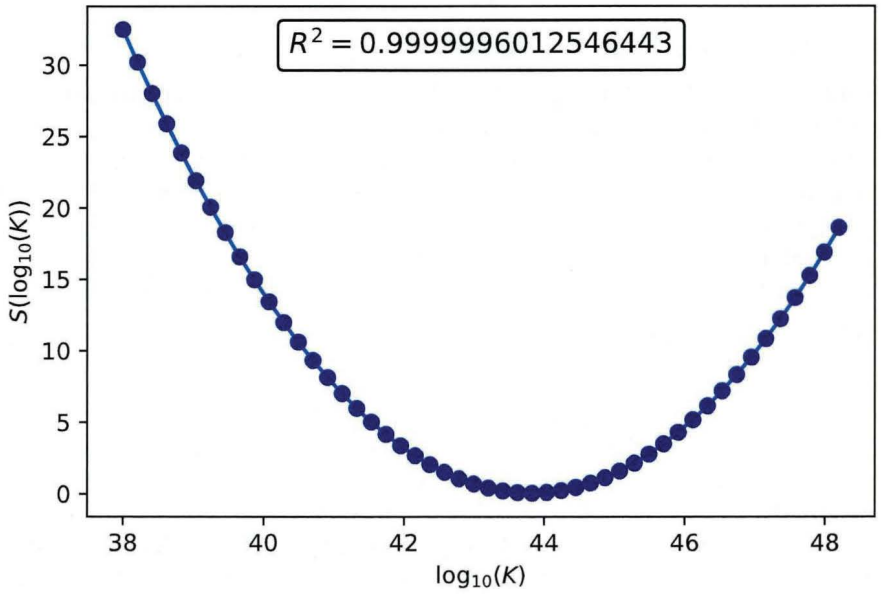


Figure 2. Sum-of-square error given by Equation (32) with residual given by Equation (33) for phase-5.

3.3 Priors on $\log_{10}K$ and σ

As discussed in 1.3.2, the prior on $\log_{10}K$ was chosen to be a uniform distribution. In order to choose an appropriate range for this distribution, we can utilize Figure 1- Figure 2 along with the log-likelihood given by equation (29). The value of σ is, at this point, unknown, however we can find the value of σ that maximizes equation (29) for fixed n, to ensure our range of $\log_{10}K$ values is large enough such that $L(y|\theta,\sigma) \approx C$ at the endpoints of the uniform prior distributions, for all values of σ. Taking the derivative of equation (29) with respect to σ and setting the result equal to zero yields,

$$-\frac{n}{\sigma}+\frac{S(\theta)}{\sigma^3}=0, \tag{35}$$

Hence the maximum occurs for

$$\sigma^2=\frac{S(\theta)}{n}. \tag{36}$$

Notice that equation (36) is just the usual definition of the population standard deviation. Plugging equation (36) back into equation (29) yields,

$$L_{\max}=-\frac{n}{2}-\frac{n}{2}\ln\left(\frac{2\pi S(\theta)}{n}\right). \tag{37}$$

Information Only

Equation (37) can then be inverted for $S(\theta)$ as a function of L_{max}

$$S(\theta) = \frac{n}{2\pi} \exp\left(\frac{-(2L_{\max} + n)}{n} \right). \tag{38}$$

We're interested in the 95'th quantile (corresponding to a 95'th credible interval). Hence, we need to ensure sampling extends at least an order of magnitude below a probability of 0.05 to ensure the posterior distribution within the 95'th quantile is adequately sampled (loosely applying the Nyquist-Shannon sampling theorem). For that reason, we chose a probability of 0.0001 (1 in 10,000), corresponding to a log-likelihood cutoff L_{max} = -9. For the Pannach and Altmaier dataset combined, there are 17 individual experiments, n = 17. Plugging L_{max} = -9 and n = 17 into equation (38) gives $S(\theta) \approx 3$. Looking at Figure 1 with a cutoff of $S(\theta) = 3$, the prior distribution on $\log_{10}$K was chosen to be uniform for $\log_{10}$K values between 25 and 28 (which extends beyond $L(y|\theta,\sigma) < -9$)

$$p(\log_{10} K) \propto Uniform(25, 28), \text{ for phase-3.} \tag{39}$$

For phase-5, there are 24 individual experiments, n = 24. Plugging L_{max} = -9 and n = 24 into equation (38) also gives $S(\theta) \approx 3$. Looking at Figure 2 with a cutoff of $S(\theta) \approx 3$, the prior distribution on $\log_{10}$K was chosen to be uniform for $\log_{10}$K values 40 and 48.5 (which extends well beyond $L(y|\theta,\sigma) < -9$)

$$p(\log_{10} K) \propto Uniform(40, 48.5), \text{ for phase-5.} \tag{40}$$

For both phase-3 and phase-5, for choosing a prior on σ, we chose the so-called uninformed prior (see pg. 31 of [28]) which is accomplished by switching variables to a variable for sigma on a natural-log scale (could have just as easily used $\log_{10}$). Then this transformed variable is given a uniform prior on a range of -10 to 10, which corresponds to possible values of σ between 4.5 x 10^{-5} and 2.2 x 10^{4}, which to the authors subjectively (perhaps objectively) would seem to cover all reasonably expected values of σ in the present study. Hence,

$$p(\gamma) \propto Uniform(-10, 10). \tag{41}$$

Individual values of the randomly sampled σ were recovered via $\sigma = \exp(\gamma)$. Recall from equation (20) that σ (perhaps improperly mixing frequentist and Bayesian terminology) is the standard deviation of the difference between the model and the experiment on a log scale.

Thus, the priors in this study correspond to uniform distributions (in some sense) extending well beyond what the authors' subjectively chose as the relevant region of probability (95'th quantile). Alternatively, one could iteratively set limits on the prior distributions, calculate the posterior, and continue expanding the prior ranges until the posterior is sufficiently converged within the desired probability level. Another alternative is to ignore the prior all together and just set it equal to 1 for the any value of θ (so-called improper prior), although this may not work as well (numerically) for σ.

3.4 Code determination of the Posterior distribution

The posterior distribution was determined using the pyMC4 package in Python [4]. The main loop is shown below in Figure 3, and was performed using the MCMC algorithm without a gradient for two chains consisting of 100,000 samples in each chain. The log-likelihood function code is shown in Figure 4. The burn-in (tuning stage in the language of pyMC4) was 1000 samples (`tune` argument shown in Figure 3). The tuning samples are discarded from the final set of posterior samples (`discard_tuned_samples=True` argument shown in Figure 3).

Information Only

```
with pm.Model():
    # uniform priors on m and c
    lnK = pm.Uniform("lnK", lower=a, upper=b)
    lsig = pm.Uniform("lsig", lower=-10,upper=10)

    # convert parameters to a aesara tensor
    theta = at.as_tensor_variable([lnK,lsig])

    # use a Potential to "call" the Op and include it in the logp computation
    pm.Potential("likelihood", logl(theta))

    # Use custom number of draws to replace the HMC based defaults
    idata_mh = pm.sample(chains=2,cores=1,draws=100000,tune=1000,
discard_tuned_samples=True)
    mean_q = pm.find_MAP()
    sumry = pm.summary(idata_mh,['lnK','lsig'])
```

Figure 3. Main pyMC loop for determining posterior distribution, where the limits of the prior for $\log_{10}$K (written `lnK` in the code above) for phase-3 (a, b) = (25, 28) and (a, b) = (40, 48.5) for phase-5 as given by equation (39) and (40).

```
def my_loglike(theta, x, data, sigma):
    npts = data[0]
    ssf = data[1]
    lnK = theta[0]
    lsig = theta[1]

# Here ssf() is a function passed to my_loglike.  It is a polynomial interpolation of the
sum of square values versus log10K.
    ss = ssf(lnK)
    sig = np.exp(lsig)

    reslt = -(0.5 / sig**2) * ss - 0.5*npts*np.log(sig**2) - 0.5*npts*np.log(2*np.pi)
    return reslt
```

Figure 4. log likelihood function. The variables `x` and `sigma` in the function call are boilerplate and aren't used.

Polynomial interpolation of the sum-of-square values versus $\log_{10}$K is provided by the numpy Polynomial.fit() subroutine [32]. The Python code used to determine the posterior distribution is located in files Posterior_Distribution_Using_Surrogate_P3.py and Posterior_Distribution_Using_Surrogate_P5.py, for phase-3 and phase-5 respectively. The variable `npts` here is equal to $n \times k$, with $k = 1$, and n equal to the number of independent experiments performed.

Information Only

4 Results

The results of the Bayesian analysis are presented in this Section. Subsection 4.1 presents the results of the posterior distributions for $\log_{10}K$ and σ for phase-3 and phase-5. Subsection 4.2 compares the results of the Bayesian analysis to the experimental data.

4.1 Code Results

Graphs are given in Subsections 4.1.1 and 4.1.2 to visually present the posterior distribution for phase-3 and phase-5. These plots were generated in the python files plots_all_P3.py and plots_all_P5.py for phase-3 and phase-5 respectively. The reader is directed to look at these files and their imbedded comments for a full description of how these plots were made

4.1.1 Phase-3 Results

The posterior marginal distributions for $\log_{10}K$ and σ are shown in Figure 5 for phase-3, along with graphs of the samples in all chains. The pair distribution is shown in Figure 6. Notice that as σ increases, so does the width of the distribution in $\log_{10}K$. The summary statistics on $\log_{10}K$ and σ for phase-3 are given in Table 4. The variance and covariance are defined by equation (8) and (9) where the samples z_{im} in equations (8) and (9) are samples of the posterior distribution. The (numerically) small covariance indicates that $\log_{10}K$ and σ are nearly independent in the posterior distribution. This "independence" can be seen from Figure 6, from the observation that the shape of credible region remains roughly the same at each probability level, only the scale changes.

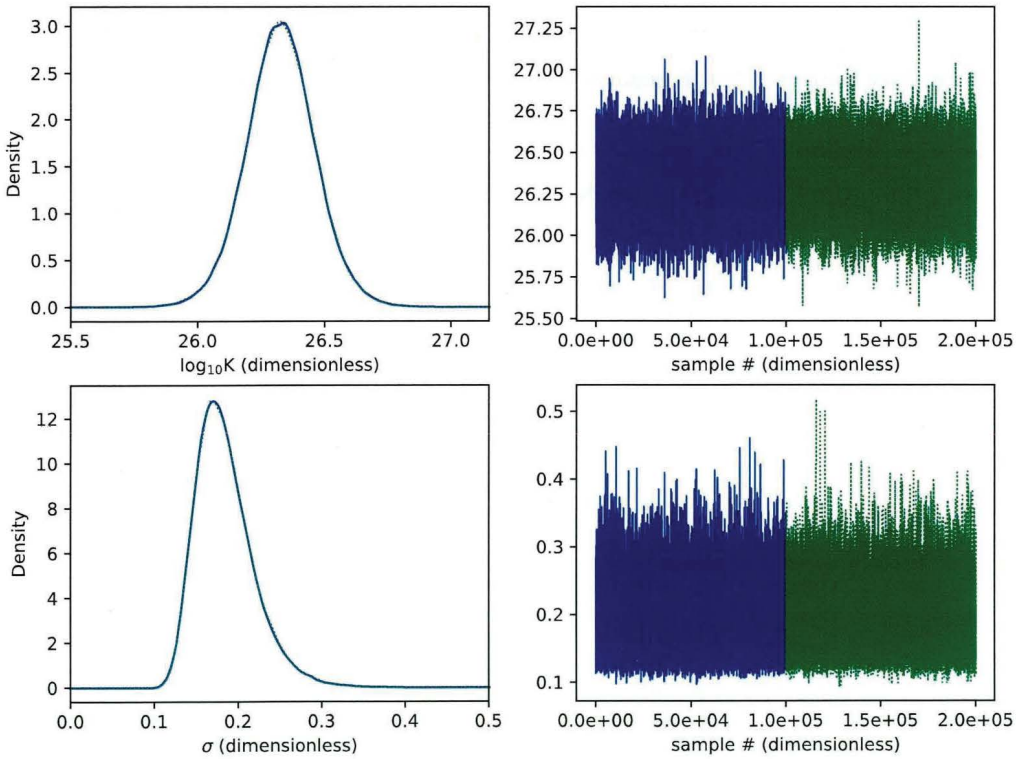


Figure 5. Marginalized distributions for $\log_{10}K$ (top left) and σ (bottom left) for phase-3. Visualization of the sampling chains for $\log_{10}K$ (top right) and σ (bottom right). First chain (blue solid line), second chain (red dashed). Each chain consists of 100,000 samples.

Information Only

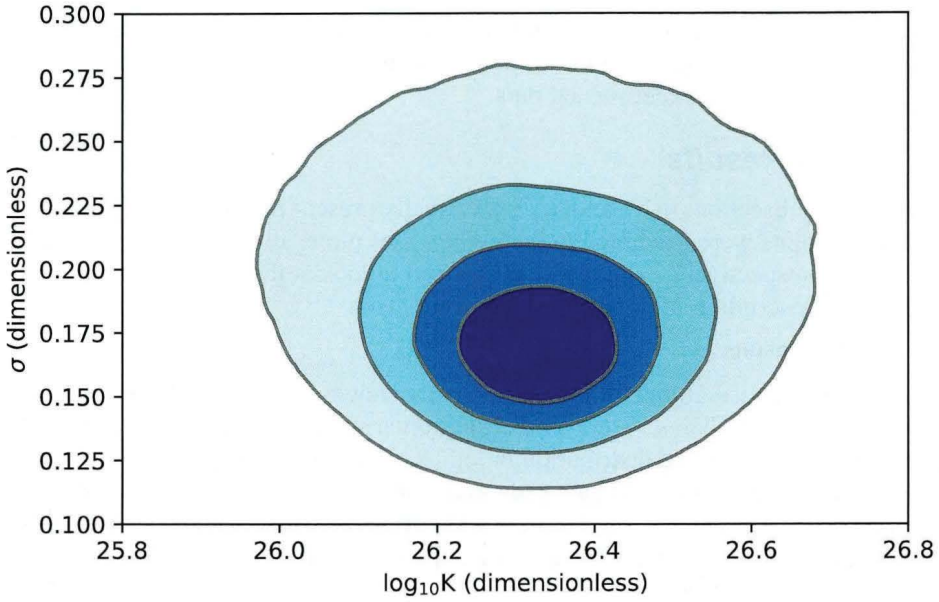


Figure 6. Pair distribution for $\log_{10}K$ and σ, for phase-3. The probability contours are drawn at the 25%, 50%, 75%, and 95% quantiles.

Table 4. Summary statistics for marginalized posterior distributions for phase-3

Variable	**Mean (dimensionless)**	**Variance (dimensionless)**	**5'th q[1] (dimensionless)**	**95'th q[1] (dimensionless)**	**Covariance (dimensionless)**
$\log_{10}K$	26.33	0.019	26.11	26.55	-8×10^{-7}
σ	0.18	0.0012	0.14	0.25	-8×10^{-6}

1. q – quantile on a scale from 0 – 100%.

Information Only

4.1.2 Phase-5 Results

The results for phase-5 are presented in Figure 7 - Figure 8, with the summary statistics given in Table 5. The results are narrower for phase-5 than for phase-3. However, only one experimental study on phase-5 solubility was used in this analysis.

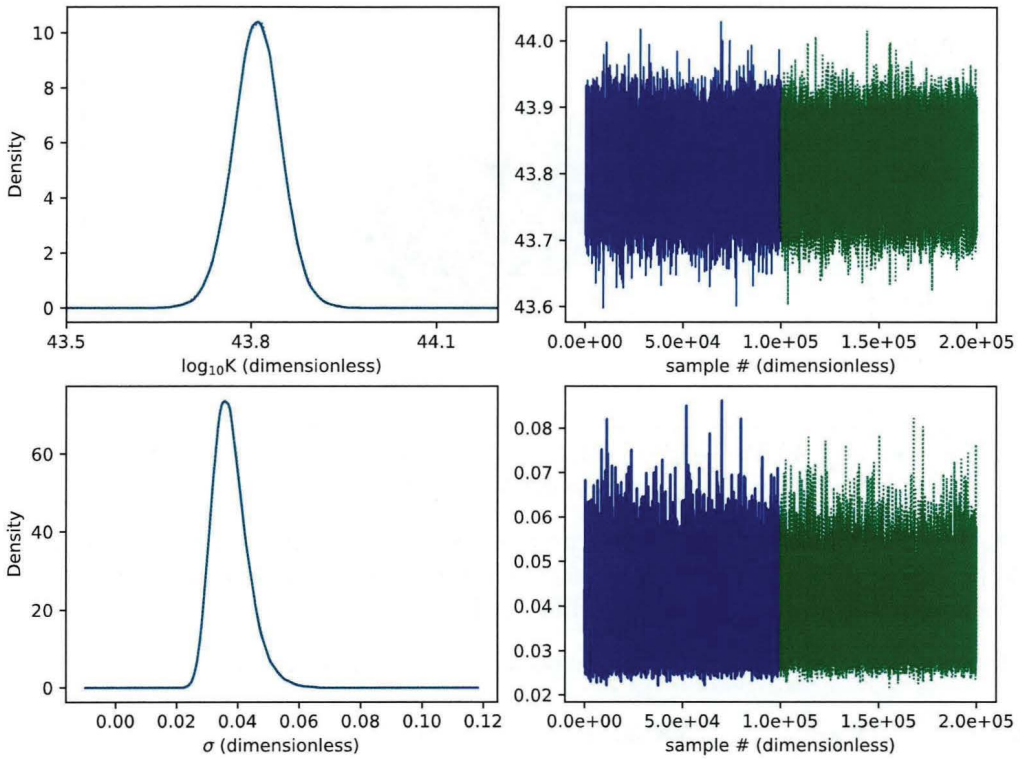


Figure 7. Marginalized distributions for $\log_{10}K$ (top left) and σ (bottom left) for phase-5. Visualization of the sampling chains for $\log_{10}K$ (top right) and σ (bottom right). First chain (blue solid line), second chain (green dashed). Each chain consists of 100,000 samples.

Information Only

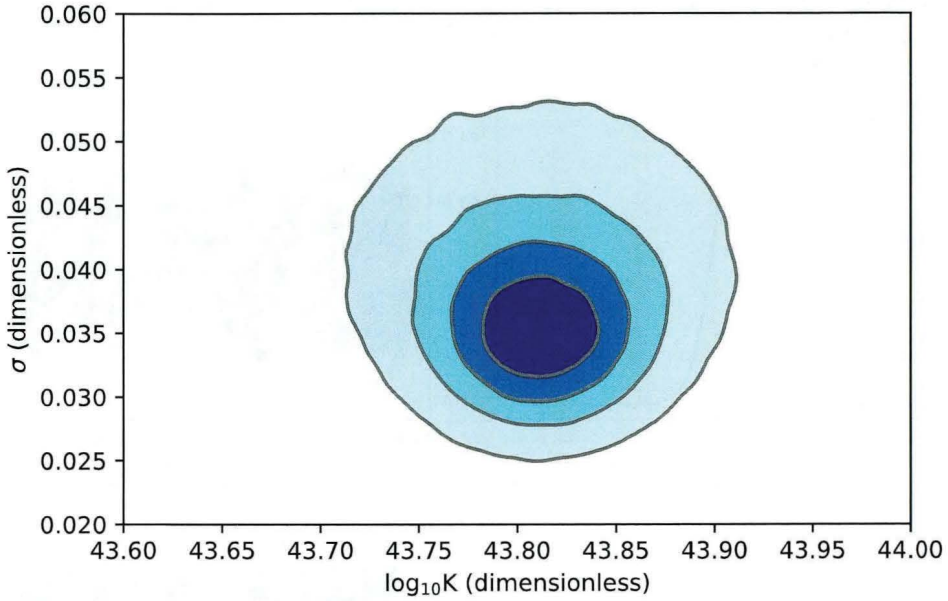


Figure 8. Pair distribution for $\log_{10}K$ and σ, for phase-5. The probability contours are drawn at the 25%, 50%, 75%, and 95% quantiles. Note that while this pair distribution appears more symmetric than phase-3 shown in Figure 6, this is an optical illusion due to the scale.

Table 5. Summary statistics for marginalized posterior distributions for phase-5

Variable	**Mean (dimensionless)**	**Variance (dimensionless)**	**5'th q[1] (dimensionless)**	**95'th q[1] (dimensionless)**	**Co-variance (dimensionless)**
$\log_{10}K$	43.81	1.51×10^{-3}	43.75	43.87	-4.8×10^{-7}
σ	0.04	3.46×10^{-5}	0.03	0.05	-4.8×10^{-7}

1. q – quantile on a scale from 0 – 100%.

Information Only

4.2 Comparison to Experimental Data

In this Subsection, the results of the Bayesian analysis are compared to the experimental data in two ways. The first, in Subsection 4.2.1, compares the data to the model in a probabilistic sense, to present evidence as to whether the likelihood model (data model) adequately describes the stochastic distance of the data from the model. The second, in Subsection 4.2.2, presents forward model calculations compared to the experimental data to provide evidence that credible intervals cover the majority of the experimental observations.

4.2.1 Comparison of the Data Model to the Experimental Data

In this Subsection, Q-Q plots are presented to provide evidence of the adequacy of the chosen likelihood model. The plot is presented in Subsection 4.2.1.1 for phase-3. A similar plot for phase-5 is presented in Subsection 4.2.1.2.

4.2.1.1 Phase-3

Figure 9 shows a Q-Q plot of the residuals, calculated by Equation (31) for the mean value of the posterior marginal distribution for $\log_{10}K = 26.33$. Note that in this version of a Q-Q plot, we have plotted the theoretical Z-score of the data versus the Z-normalized residuals (the data Z-Score). The theoretical Z-score of the residuals are calculated via the so called "percentile-rank" function [33]

$$PR_j = \frac{CF_j - 0.5 \times F}{N} \tag{42}$$

where CF_j is the total number of residuals less than or equal to R_j, F is the frequency of R_j ($F = 1$, there are no exact duplicate differences herein) and N is the number of data points (17 for the combined Pannach & Altmaier dataset). Then the theoretical Z-score, assuming a normal distribution with mean = 0, and standard deviation = 1, is given by inverting the cumulative normal distribution for that value of PR_j,

$$Z - Score_{theoretical,j} = CDF^{-1}(PR_j). \tag{43}$$

The data Z-score is just the Z-score of the residual calculated from the mean and standard deviation of the differences,

$$Z - Score_{data,j} = \frac{R_j - \bar{R}}{\sigma_R} \tag{44}$$

where $\bar{R}$ is the mean of the differences calculated using Equation (31), and σ_R is the associated standard deviation. The above definition of a Q-Q plot may be elementary to some readers, however it's critical to understand because distance of the points in Figure 9 to the straight line of slope = 1 is our only evidence at this time to support the use of the data model given by Equation (20).

Information Only

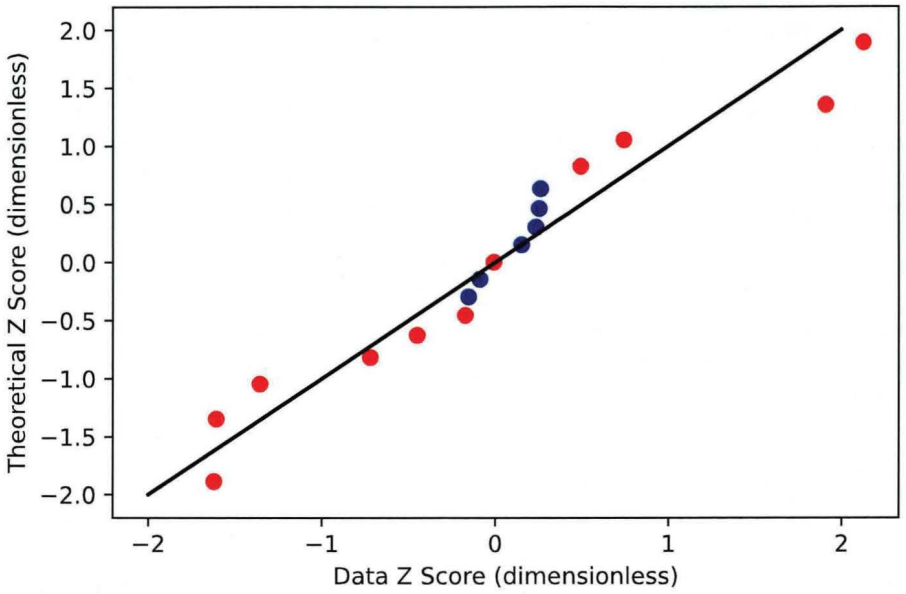


Figure 9. Q-Q plot of the raw difference between experiment and model for $\log_{10}K = 26.33$, the mean value. The red dots correspond to the Pannach et. al (2017) dataset and the blue dots correspond to the Altmaier et. al (2003) data. The straight line is of slope = 1.

Given that the two researchers (Altmaier and Pannach) used different experimental methodologies, it may be reasonable in future analyses to give each dataset its own distribution for σ. In using a single distribution on σ, for the combined Altmaier and Pannach datasets, the authors' have (explicitly) assumed that the Pannach and Altmaier datasets are identically distributed, which is obviously a doubtful assumption. On the other hand, this assumption probably conservatively predicts a value of σ that covers a majority of the data, as will be shown in 4.2.2.2.

4.2.1.2 Phase-5

A Q-Q plot for phase-5 is shown in Figure 10, produced similarly to Figure 9. Notice the chains of data points that straddle the straight line. The authors speculate that this feature could be due to a second lesser source of uncertainty, which is smaller but distributed differently than the main source of uncertainty predicted by Equation (25).

Information Only

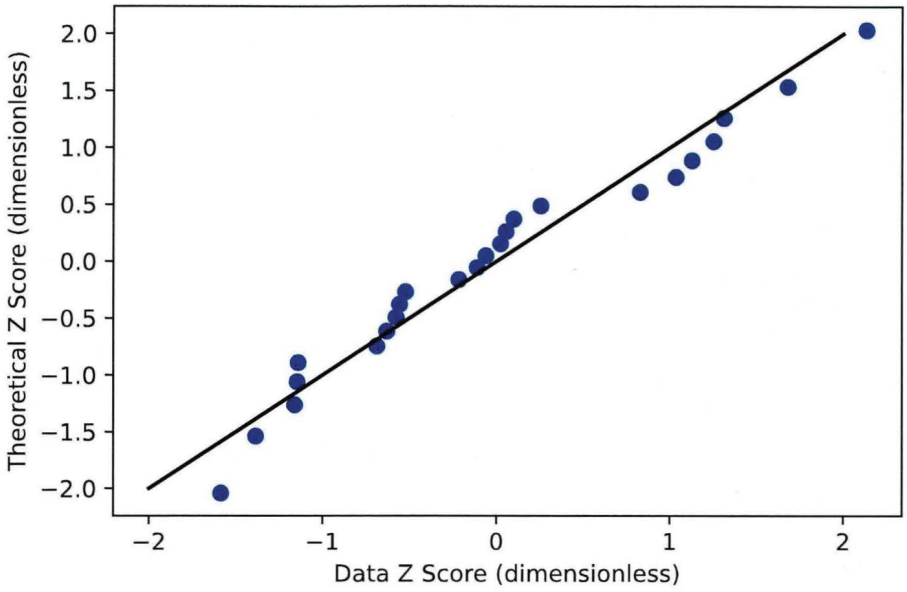


Figure 10. Q-Q plot of the raw difference between experiment and model for $\log_{10}K = 43.80$, the mean value for phase-5. The blue dots correspond to the Xiong et. al (2003) data. The straight line is of slope = 1.

4.2.2 Forward Model Results Compared to Experimental Data

In this Subsection, forward model predictions are compared directly to the experimental Data. Subsection 4.2.2.1 describes the methodology for calculating the posterior predictive distribution. That method is then applied to the phase-3 forward model in Subsection 4.2.2.2. The method is applied to phase-5 in Subsection 4.2.2.3.

4.2.2.1 Calculation of the Posterior Predictive Distribution

The posterior predictive distribution (PPD) shown in Equation (30) was determined through the following algorithm:

1) Randomly sample a pair of values ($\log_{10}K$, σ) from the posterior distribution $p(\log_{10}K,\sigma|y_{data})$
2) Using the pair from 1), sample a simulated experimental dataset from the data distribution $p(y|\log_{10}K,\sigma)$ given by equation (20).
   a) Use the forward-model surrogate (linear interpolation, see Section 3.2) to calculate the experimental dataset for the sampled value of $\log_{10}K$, rather than recalculate from EQ3/6 (for speed).
3) Repeat 1) – 2) 10,000 times.

The number of samples, 10,000, was chosen to ensure simulated experimental dataset samples were obtained significantly beyond 2σ, which is our subjectively chosen range of interest (corresponding to the 95'th quantile). The Python code for performing this sampling is shown below.

Information Only

```
for j in range(nsamp):
    print(j)
# Sample one pair of log10K,Sigma from the posterior distirbution
    df = post.sample(1)
    log10K = df.iloc[0,0]
    sigma = df.iloc[0,1]
# Linearly interpolate to determine the expected concentrations at that log10K
# One could argue whether linear interpolation is the best choice here.  Left to future
work.
    res_list = sth.obtain_res_all(log10K,res_tab)
# Separate out means by Stoichiometry
    res_list = np.log10(res_list)
    means = np.zeros(len(res_list))
    for i in range(len(res_list)):
# Simulation ID
        idi = res_list.index[i]
# Ion concentrations corresponding to idi = Simulation ID, for the given log10K value
        ser = res_list.iloc[i]
# Each Simulation ID will have some ions included and some excluded from the sum of
square calculation
# If it's included it gets a stoichiometry > 0, if not it gets a blank = NaN when Pandas
reads in
# the Excel file.
# This is right now ad-hoc because it only looks at the first two columns of the
Stoichiometry
# DataFrame, which is appropriate only for the present case,
        if np.isnan(sto.iloc[i,0]) == True:
            spec_mean = spec[1]
        else:
            spec_mean = spec[0]
# Chooses the concentration to use as the mean based on the stoichiometry DataFrame
        means[i] = ser[spec_mean]
# The value of Sigma here is from the sample of the posterior distribution
# Then we assume here (also ad-hoc) that sigma and log10K have no covariance,
# which is correct for the phase-3 and phase-5 analysis.  It will, not however
# always be true, and likely not true when we fit data to a larger number of
# parameters (e.g. Pitzer parameters)
    cov = np.identity(len(res_list))*sigma**2
# Here we sample from a multivariate normal distribution with a list of means and
# an associated covariance matrix.  Notice also that if we gave the Pannach and
# Altmaier different sigma values, then the above definition of the covariance
# matrix would need to be changed
    rnd = np.random.multivariate_normal(means, cov)
```

The above python code is taken from the Python files Posterior_Predictive_Using_Surrogate_P3.py and Posterior_Predictive_Using_Surrogate-P5.py for phase-3 and phase-5 respectively. The reader should view these files for a fuller picture of how the posterior predictive was calculated.

4.2.2.2 *Forward Model Comparisons to Phase-3*

A comparison between the Bayesian results and the experimental data for phase-3 is shown in Figure 11 - Figure 12. In these figures the solid blue line represents the surrogate-model results using the mean value of $\log_{10}K$ obtained from the marginal posterior distribution on $\log_{10}K$ shown in Figure 5 and Table 4. The blue-shaded region that covers the solid blue line is the results of the surrogate model obtained at the 5'th and 95'th quantile of the marginal posterior distribution on $\log_{10}K$, corresponding to the values shown in Table 4. The area between the 5'th and 95'th quantile is filled with a single shade of blue to represent a 95'th credible interval. Lastly the green shaded region is the posterior predictive distribution calculated at the 5'th and 95'th quantile, with the region in between these curves filled with a constant shade of green to represent a 95'th credible interval. The means of calculating the posterior predictive

Information Only

distribution is described in Section 4.2.2.1. Repeating what was written in Section 4.2.2.1, the PPD is calculated by randomly sampling from the data model given by equation (20), with the value of $\log_{10}K$ and σ randomly sampled from the posterior distribution shown in Figure 5. Quantiles of the distributions were calculated using the Pandas function DataFrame.quantile() in python directly on the samples of the distributions. Note that the surrogate model used here is a fit of the EQ3/6 results as a function of the parameters under study (herein only $\log_{10}K$). This was used for computational speed. If the reader has concerns as to the fidelity of the fit, the reader could use EQ3/6 directly (this would be computationally expensive), obtain the surrogate at a finer grid of parameter values, or use a higher-order fit.

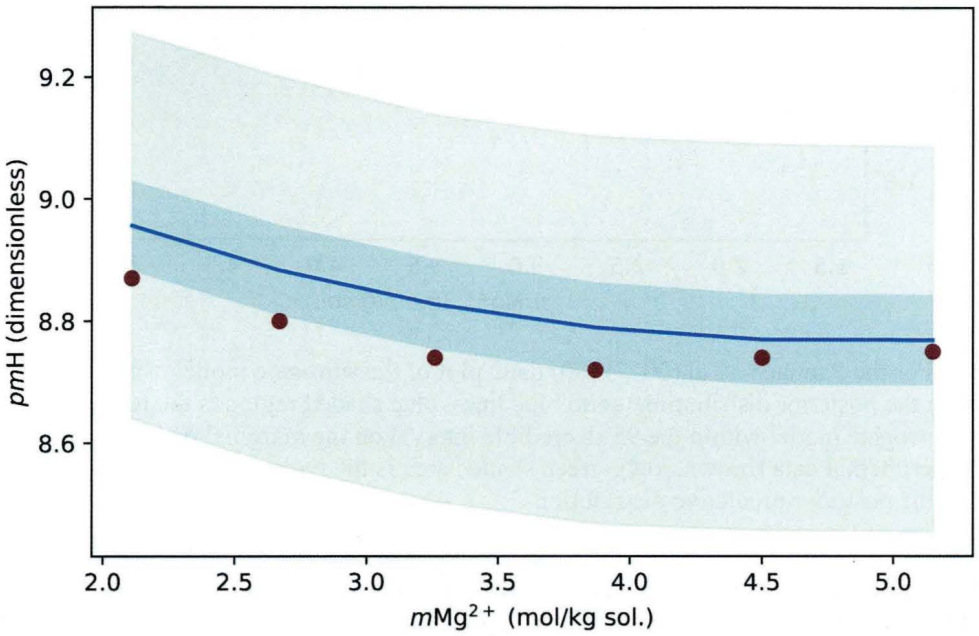


Figure 11. For the Altmaier et. al (2003) [29] data, plot of the surrogate model using the mean value of $\log_{10}K$ from the posterior distribution (solid blue line); blue shaded region is the region that spans the results of surrogate model within the 95'th credible interval on the marginal posterior distribution for $\log_{10}K$; experimental data (brown dots); green shaded area is the region that spans the 95'th credible interval on the posterior predictive distribution.

Information Only

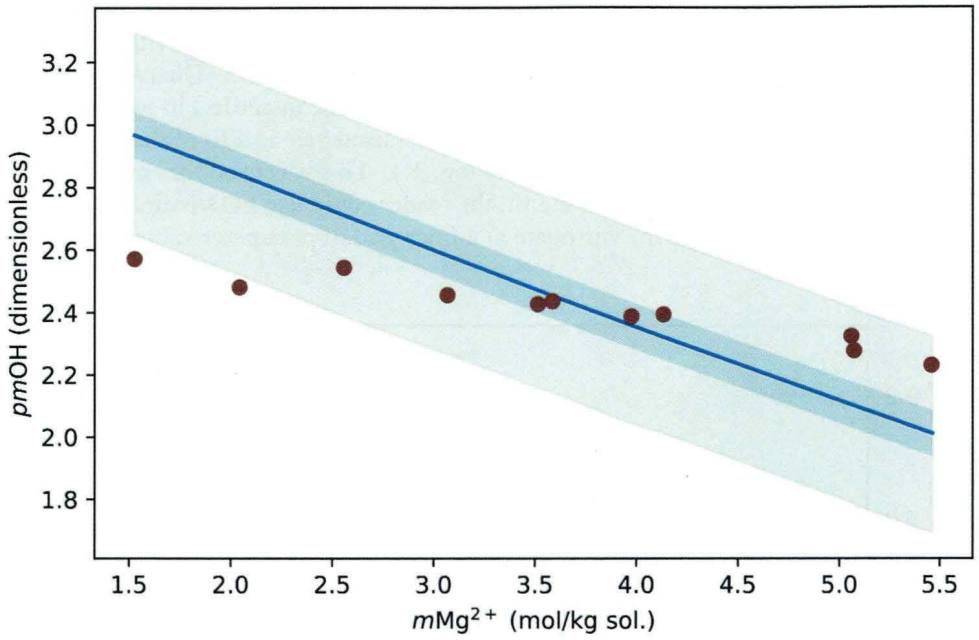


Figure 12. For the Pannach et. al (2017) [30] data, plot of the surrogate model using the mean value of $\log_{10}K$ from the posterior distribution (solid blue line); blue shaded region is the region that spans the results of surrogate model within the 95'th credible interval on the marginal posterior distribution for $\log_{10}K$; experimental data (brown dots); green shaded area is the region that spans the 95'th credible interval on the posterior predictive distribution.

From Figure 11 - Figure 12, it appears the model fits the Altmaier data well, but the slope of the model appears different from the Pannach et. al data; the Pannach et. al data shows a significantly smaller slope with $m\mathrm{Mg}^{2+}$. However, the 95'th credible interval on the PPD captures nearly all of the experimental data and hence may correctly represent the uncertainty in the measurements/technique. Admittedly, the authors subjectively chose to present the 95'th credible interval on the PPD. However, the $\approx \pm 0.5$ *pm*OH uncertainty given by the 95'th credible interval on the PPD is consistent with the authors' experimental experience. Similarly, the $\approx \pm 0.4$ *pm*H uncertainty expressed by the 95'th credible interval on the PPD for the Altmaier dataset shown in Figure 11 is well within the authors' experimental experience. Thus, in the authors' opinion, without further evidence, we have no technical basis to claim that there is anything wrong with the current model, the Bayesian inference method, nor with the Pannach et. al dataset. That is to say, in our opinion it's all within reasonable and expected uncertainty.

Both *pm*H and *pm*OH appear to decrease with increasing $m\mathrm{Mg}^{2+}$, which may appear at first glance to be physically incorrect. This is mostly due to the definition of *pm*OH used in Figure 12, which is calculated as the total OH^- in solution, including that which is complexed as $MgOH^+$(aq), rather than free OH^-(aq), to be consistent with how Pannach et. al (2017) have measured and reported their data. It would not surprise the authors if this definition, combined in some (unknown) way with how Pannach et. al made their measurements, was responsible for the difference in slope between the mean model and the apparent slope in the Pannach et. al data with $m\mathrm{Mg}^{2+}$; unraveling this discrepancy is left to potential future work and would likely require some additional experiments.

One reviewer noted that the Pannach experiments are performed over a wider range of ionic strengths than the Altmaier experiments. Thus, the close resemblance between the model and the fit seen in Figure 11 may be due to the model being (Pitzer parameters) less-well tuned on the wider range of ionic

Information Only

strength shown in Figure 12. We believe this idea speaks to the wider issue with parameter tuning as discussed in Section 5.2.1.

In both the Altmaier and Pannach analyses, the results taken from the papers were likely averaged or processed in some way prior to publication. In that sense (improperly mixing both frequentist and Bayesian terminology), the green-shaded areas in Figure 11 - Figure 12 represent something closer to a standard error in the expected measurements rather than a standard deviation. Thus, this area represents a measure of the variance expected in the processed measurements given uncertainty in both the measurements and in the value of $\log_{10}K$, using the methods that the authors used to processed them. We unfortunately do not know exactly how both researchers processed their measurements prior to publication. This is a topic of active research [34], and the authors intend to revisit the analysis herein in the future by utilizing some of the techniques given in [34]. That future analysis will likely widen the uncertainty bounds in the green shaded area, and perhaps widen the blue area slightly.

4.2.2.3 Forward Model Comparisons to Phase-5

A comparison between the Bayesian results and the experimental data for phase-5 is shown in Figure 13. Note that in this case, while there may have been some data processing, there are several replicates available at each experimental condition. The jaggedness of Figure 13 is due to the surrogate model being fit from the EQ3/6 simulations only at the conditions of each experiment, not at points (magnesium chloride + sodium chloride concentrations) in-between, and no smoothing or fitting was used to produce the plot. Note also that the Xiong et. al (2011) dataset varied $m\mathrm{Mg}^{2+}$ and $m\mathrm{Na}^{+}$ concentrations as shown in Section 2.2. However, in Figure 13 all of the data has been plotted versus $m\mathrm{Mg}^{2+}$, visually marginalizing out the effect of $m\mathrm{Na}^{+}$ (the effect is not marginalized in the calculations).

The cause of the narrowness of the credible interval on $\log_{10}K$ is apparent from Figure 13. The model fits through the center of much of the experimental data, and there appears to be little bias in the model. That lack of bias is consistent with Figure 10. Only one data point lies outside the 95'th quantile on the PPD. To understand how this figure is constructed, we refer the reader to Section 4.2.2.2.

Information Only

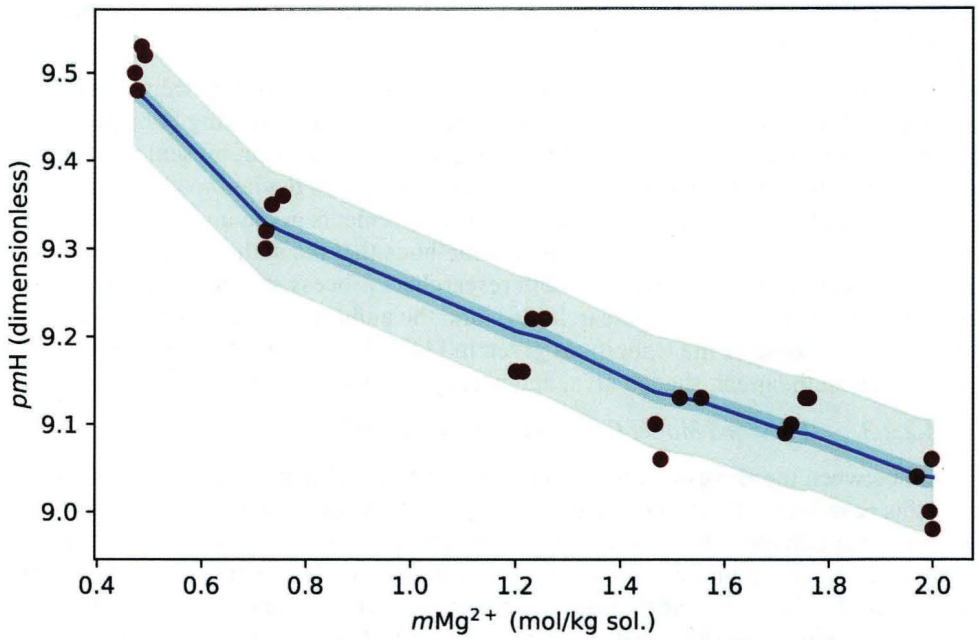


Figure 13. For the Xiong et. al (2011) [31] data, plot of the surrogate model using the mean value of $\log_{10}K$ from the posterior distribution (solid blue line); blue shaded region is the region that spans the results of surrogate model within the 95'th credible interval on the marginal posterior distribution for $\log_{10}K$; experimental data (brown dots); green shaded area is the region that spans the 95'th credible interval on the posterior predictive distribution.

Information Only

5 Conclusions and Future Work

5.1 Conclusions

Overarching goals of this report include: (1) to demonstrate Bayesian inference/analysis in determining the uncertainty in model parameters used in geochemistry; (2) to demonstrate a more statistically based approach to comparing models to experimental data than has been used previously. Each of these topics is discussed below.

The authors believe that sufficient evidence has been provided in this report to support the use of Bayesian inference in determining the uncertainty in geochemistry model parameters. Evidence is given by Figure 9 - Figure 10, which show that the data appear to follow a log-normal distribution from the model, even when the model doesn't visually appear to follow the data (e.g. for the Pannach et. al dataset). Further evidence is given by Figure 11 - Figure 13, which indicate that the posterior-predictive distribution captures nearly all of the experimental data within the 95'th quantile.

It is the authors' opinion that Figure 11 - Figure 13 provide a less human-biased means of comparing the model to the experimental data than current visual methods of looking at whether the model "fits" the data. While both methods are at some point subjective (e.g. we choose to display the 95'th credible interval), the Bayesian inference techniques described herein provide a framework that adds a quantitative step to examining and thinking about the fit, and perhaps more importantly provides an accepted/recommended [35] means of determining the uncertainty in model parameters and forward model calculations.

The Bayesian statistical method shown herein provides another way of comparing EQ3/6 model results to data, rather than supplanting EQ3/6 calculations. All of the figures shown herein rely on the model results of EQ3/6 and the experimental data. The surrogate model is a fit of the EQ3/6 results as a function of the model parameters that were varied (here just $\log_{10}K$). The surrogate model was used to speed up the time required to perform the calculations.

5.2 Future Work

5.2.1 General Themes

In recent years there has been a significant amount of debate concerning the correct model, correct fitting procedure, and best choice of EOS to use in simulating geochemical conditions in the wider nuclear-waste community [36-40]. A notable quote from May [39] reads: *"With nontrivial chemical systems, the chances of independent experts reaching similar, universally applicable conclusions from speciation modeling are accordingly still as bleak as they were a quarter of a century ago."* With Bayesian inference, we are proposing an alternative means of comparing models and model parameters, which in the authors' opinion leads to less researcher bias than traditional methods. Furthermore, the Bayesian inference method has been recommended by the National Academy of Science [35] for uncertainty quantification in high-consequence simulations. In contrast, the NEA, at least for nuclear waste disposal, still recommends linear propagation of error [41]; it is also noteworthy that the date on the NEA document [41] is 1999. A notable quote from Meinrath [42] reads "*Values of chemical quantities are often the result of complicated procedures in which common progression of-error analysis cannot provide satisfactory results.*"

Information Only

5.2.2 Specific Tasks

5.2.2.1 EQ3/6 Code Execution

The EQ3/6 code currently runs as an executable file for each run that is performed. This requires re-loading the input and database files, execution of various operating system tasks, writing of output files, etc. That is, a significant amount of time, if not the overwhelming vast majority, is wasted in completing the runs needed to build a surrogate model. There are several ways of fixing this, each of which requires some thought and wider discussion. EQ3/6 could be re-built as either a Windows COM object library, or a Linux hard-coded library. In this way the library could be called repeatedly without the overhead that is currently slowing it down. Such a task would be helpful for work outside of Bayesian inference as well. But this conversion would require re-qualifying EQ3/6 under the WIPP QA program, which could be a significant undertaking, given that Sandia did not originally write this code. An alternative is to use PHREEQC [43], which has already been converted to a Windows COM object and a Linux hard-coded library, partially for the purpose of sensitivity and Bayesian inference analyses. There is some concern however that PHREEQC may not perform reaction-path simulations as well as EQ6. Furthermore, if we were to work on the PHREEQC code to fix this issue, this would also lead to a requalification.

5.2.2.2 Resolving the Processed Data Issue

The present analysis should be amended in future work to include synthetic generation of experimental data, particularly for the Pannach et. al and Altmaier et. al datasets, by assuming measurement uncertainties that were reported. This was discussed in Section 4.2.2.2. However before embarking on that work, it would be wise to conduct additional experiments to better understand potential sources of uncertainty in the *pm*OH measurement technique used in Pannach et. al (2017), as the posterior predictive distribution predicts far larger uncertainty (Section 4.2.2.2) in the data than was reported by Pannach et al. (Section 2.1).

5.2.2.3 Maintaining Internal Consistency of the Database

In this work we modeled $\log_{10}K$ as a uniquely valued parameter with uncertainty rather than a distributed parameter. However, the authors in general believe that model parameters should be modeled as distributions rather than is single *"true"* values. This distinction becomes important when we look to maintaining constraints and/or relations between various parameters of the model, based on underlying physical principles.

Hence, in general, the authors believe that constraints should be made on metrics of the distributions as a whole (e.g. through covariances) rather than individual points on the distributions. There are of course exceptions (hard constraints) that must be considered in some cases. The consequence of the choice of distributed constraints is the (hypothetical) potential to produce forward simulations that are un-physical, or forward simulations whose parameter values are in a combination that is un-physical. Some researchers have enforced constraints to prevent such an occurrence [44]. Other ideas have been proposed such as making the constraint itself a probability distribution [45].

The authors would argue that the simplest solution, to the concern of producing unphysical results, is to enforce physical meaning on the forward model predictions when using the mean parameters, but relax that meaning for the purpose of determining forward-model uncertainty. I.e., we should not expect perfect physics in the tails of any forward model, rather interpret this as the effect of our lack of knowledge on the parameters.

5.2.2.4 WIPP-Relevant Systems

For other WIPP-relevant systems, in future work we propose to simultaneously fit entire sets of related training data and its parameters ($\log_{10}K$'s, Pitzer parameters, etc.), allowing the mathematical framework

Information Only

shown in Section 1.3 to determine the credible region in the resulting set of model parameters. Such an analysis could focus our resources on developing and performing experiments that would reduce uncertainty, assuming the uncertainty is reducible as discussed further in Section 5.2.2.6. This same argument applies to the cases shown herein as well. That is, if in future work we were to fit the $\log_{10}K$ values for these phases (phase-3 and phase-5) including uncertainty in the underlying Pitzer parameters, it follows that the uncertainty in the $\log_{10}K$ values could be larger and may grow as the dimensionality of the problem increases. Our primary goal in this report is to demonstrate the framework before proceeding with larger and more complex systems of data and parameters. In order to accomplish this goal, the authors believe some of the tasks listed under Section 5.2.2.1 and perhaps 5.2.2.2 will need to be completed first.

5.2.2.5 Comparing Models

As discussed throughout this report, other models that could be considered in the present analysis. For example, the Pannach et. al (2017) and Altmaier et. al (2003) datasets could be given different distributions for σ. This would be a different likelihood model for the data. Additional datasets could be included, perhaps with their own values of σ. Alternative speciation models, in general, could be considered although that doesn't appear to be particularly relevant in the phase-3/phase-5 case. Another model to consider in future work is whether multiple solubility-controlling phases might be present in the Pannach et. al (2017), Altmaier et. al (2003), and Xiong et. al (2011) datasets. E.g., phases including phase-3, phase-5, and brucite that don't show up in XRD. This model has been raised by several technical reviewers.

There are commonly used Bayesian methods to compare different models in terms of which model is more probable based on the calculated posterior distributions [14]. We intend to use these methods in future work, particularly in comparing different models for various WIPP-relevant systems as described in Section 5.2.2.4. We must also state, however, that such an analysis provides a guide to the scientists, and not an absolute rational for choosing a particular model. The scientific method requires that we consider both the physics (chemistry) of the problem and the posterior statistics together holistically.

5.2.2.6 Experimental Design based on Bayesian Inference

A natural question is whether the present analysis indicates where additional data would be useful in reducing uncertainty, assuming its reducible. The details of Bayesian experimental design will be left to future work, but the short answer is in general yes. The method depends on whether reducing parameter uncertainty or forward model prediction uncertainty is the goal. For WIPP, the authors' would argue that forward-model uncertainty reduction is the primary goal for forward model predictions of actinide solubilities in WIPP-relevant brines. For non-actinide species, parameter uncertainty reduction is the goal.

Information Only

6 References

1. Box, G.E.P., W.G. Hunter, and J.S. Hunter, *Statistics for Experimenters; An Introduction to Design, Data Analysis, and Model Building*. Wiley Series in Probability and Mathematical Statistics, ed. R.A. Bradley, et al. 1978, New York, USA: John Wiley & Sons.
2. Pizter, K.S., *Activity Coefficients in Electrolyte Solutions*. 1991, CRC Press: Boca Raton, FL.
3. Wolery, T. *EQ3/6 A Software Package for Geochemical Modeling*. 2010 Version: 8.0a. Available from: https://github.com/llnl/eq3_6.
4. Wiecki, T., et al. *PyMC*. 2022 Version: 4.1.7. Available from: https://www.pymc.io/welcome.html.
5. Weber, C.L., J.M. VanBriesen, and M.S. Small, *A Stochastic Regression Approach to Analyzing Thermodynamic Uncertainty in Chemical Speciation Modeling.* Environmental Science & Technology, 2006. **40**(12): p. 3872-3878.
6. Vielma, T., L. Hnedkovsky, and G. Hefter, *Chemical speciation effects on the volumetric properties of aqueous sulfuric acid solutions.* Journal of Chemical Thermodynamics, 2021. **158**: p. AN 106408.
7. Sarkar, S., et al., *Bayesian calibration of thermodynamic parameters for geochemical speciation modeling of cementitious materials.* Cement and Concrete Research, 2012. **42**(7): p. 889-902.
8. Draper, N.R. and H. Smith, *Applied Regression Analysis*. Wiley Series in Probability and Statistics. 1998, New York, USA: John Wiley & Sons, INC.
9. Seber, G.A.F. and C.J. Wild, *Nonlinear Regression*. 1989, New York, USA: John Wiley & Sons.
10. Nemer, M.B., et al., *Solubility of $Fe_2(OH)_3Cl$ (pure-iron end-member of hibbingite) in NaCl and Na_2SO_4 brines.* Chemical Geology, 2011. **280**(1-2): p. 26-32.
11. De Gaspari, F., et al., *Use of redundant data to reduce estimation errors in geochemical speciation.* Applied Geochemistry, 2015. **55**: p. 184-191.
12. INTERA. *Uncertainty Analysis (PEST++)*. 2022; Available from: https://www.intera.com/pest/.
13. Hoff, P.D., *A First Course in Bayesian Statistical Methods*. 2009, Dordrecht, Netherlands: Springer.
14. Silva, D.S., *Data Analysis; A Bayesian Tutorial*. Second Edition ed. 2007, Norfolk, Great Britain: Oxford University Press.
15. Mishra, P., et al., *Descriptive Statistics and Normality Tests for Statistical Data.* Annals of Cardiac Anaesthesia, 2019. **22**(1): p. 67-72.
16. Ahrens, L.H., *A Fundamental Law Of Geochemistry.* Nature, 1953. **172**(4390): p. 1148-1148.
17. Ahrens, L.H., *The Lognormal Distribution of the Elements (A Fundamental Law of Geochemistry and its Subsidiary).* Geochimica Et Cosmochimica Acta, 1954. **5**(2): p. 49-73.
18. Ahrens, L.H., *The LognormaL Distribution of the Elements .2.* Geochimica Et Cosmochimica Acta, 1954. **6**(2-3): p. 121-131.
19. Ahrens, L.H., *Lognormal-Type Distributions in Igneous Rocks .4.* Geochimica Et Cosmochimica Acta, 1963. **27**(APR): p. 333-343.
20. Ahrens, L.H., *Lognormal-Type Distributions in Igneous Rocks .5.* Geochimica Et Cosmochimica Acta, 1963. **27**(AUG): p. 877-890.

Information Only

21. Buccianti, A., G. Mateu-Figueras, and V. Pawlowsky-Glahn. *Frequency distributions and natural laws in geochemistry*. in *32nd International Geological Congress*. 2004. Florence, ITALY.

22. Ehrlich, G., et al., *On the Valuation of Quantitative Chemical-Analysis - Random Error, Bias, Total Error.* Zeitschrift Fur Chemie, 1984. **24**(6): p. 204-208.

23. Limpert, E., W.A. Stahel, and M. Abbt, *Log-normal distributions across the sciences: Keys and clues.* Bioscience, 2001. **51**(5): p. 341-352.

24. Kuna-Broniowska, I. and H. Smal, *Statistical measures of the central tendency for H^+ activity and pH.* Soil Science Annual, 2017. **68**(4): p. 174-181.

25. Mahalanobis, P.C., *On the generalised distance in statistics.* Proceedings of the National Institute of Sciences of India, 1936. **2**(1): p. 49-55.

26. McElreath, R., *Statistical Rethinking; A Bayesian Course with Examples in R and Stan*. Second Edition ed. 2020, Boca Raton, FL: CRC Press.

27. Cox, R.T., *Probability, Frequency And Reasonable Expectation.* American Journal of Physics, 1946. **14**(1): p. 1-13.

28. Box, G.E.P. and G.C. Tao, *Bayesian Inference in Statistical Analysis*. 1992, New York, USA: John Wiley & Sons.

29. Altmaier, M., et al., *Solid-liquid equilibria of Mg(OH)(2)(cr) and Mg-2(OH)(3)Cl center dot 4H(2)O(cr) in the system Mg-Na-H-OH-O-Cl-H2O at 25 degrees C.* Geochimica Et Cosmochimica Acta, 2003. **67**(19): p. 3595-3601.

30. Pannach, M., S. Bette, and D. Freyer, *Solubility Equilibria in the System $Mg(OH)_2$−$MgCl_2$−HO from 298 to 393 K.* Journal of Chemical & Engineering Data, 2017. **62**: p. 1384-1396.

31. Xiong, Y.L., et al., *Experimental determination of the solubility constant for magnesium chloride hydroxide hydrate (Mg3Cl(OH)(5)center dot 4H(2)O, phase 5) at room temperature, and its importance to nuclear waste isolation in geological repositories in salt formations.* Geochimica Et Cosmochimica Acta, 2010. **74**(16): p. 4605-4611.

32. NumPy. *numpy.polynomial.polynomial.Polynomial.fit*. 2022; Available from: https://numpy.org/doc/stable/reference/generated/numpy.polynomial.polynomial.Polynomial.fit.html#numpy-polynomial-polynomial-polynomial-fit.

33. Wikipedia. *Percentile rank*. 2022; Available from: https://en.wikipedia.org/wiki/Percentile_rank.

34. Chowdhary, K. and H.N. Najm, *Data free inference with processed data products.* Statistics and Computing, 2016. **26**: p. 149-169.

35. National Research Council of the National Academies, *Assessing the Reliability of Complex Models; Mathematica and Statistical Foundations of Verification, Validation, and Uncertainty Quantification*. 2012, The National Academies Press: Washington, DC.

36. Oakes, C.S., et al., *A speciated, conventional Pitzer ion-interaction model for the aqueous $Nd3^+$–H^+–Na^+–K^+–Ca^{2+}–Cl^-–OH^- system at 298 K and 0.1 MPa.* Geochimica et Cosmochimica Acta, 2021. **297**: p. 308-327.

37. Schramke, J.A., E.F.U. Santillan, and R.T. Peake, *Plutonium oxidation states in the Waste Isolation Pilot Plant repository.* Applied Geochemistry, 2020. **116**.

38. Liang, S.Y., et al., *A Review of Geochemical Modeling for the Performance Assessment of Radioactive Waste Disposal in a Subsurface System.* Applied Sciences-Basel, 2021. **11**(13).

Information Only

39. May, P.M. and D. Rowland, *Thermodynamic Modeling of Aqueous Electrolyte Systems: Current Status.* Journal of Chemical & Engineering Data, 2017. **62**(9): p. 2481-2495.

40. Nitzsche, O., G. Meinrath, and B. Merkel, *Database uncertainty as a limiting factor in reactive transport prognosis.* Journal of Contaminant Hydrology, 2000. **44**(3-4): p. 223-237.

41. Wanner, H. and E. Osthols, *Guidelines for the Assignment of Uncertainties*. 1999, AEA NEA OECD: Issy-les Moulineaux, France.

42. Meinrath, G., *Extended traceability of pH: an evaluation of the role of Pitzer's equations.* Annals of Bioanalytical Chemistry, 2002. **374**: p. 796-805.

43. Webb, R. *PHREEQC Version 3*. 2021 Version: Available from: https://www.usgs.gov/software/phreeqc-version-3.

44. Strofer, C.A.M., et al., *Enforcing boundary conditions on physical fields in Bayesian inversion.* Computer Methods in Applied Mechanics and Engineering, 2020. **367**.

45. Wu, J., J.-X. Wang, and S.C. Shadden, *Adding Constraints to Bayesian Inverse Problems.* Proceedings of the AAAI Conference on Artificial Intelligence, 2019. **33**(01): p. 1666-1673.

46. The pandas development team. *pandas-dev/pandas: Pandas*. Version: 1.4.4. Available from: https://github.com/pandas-dev/pandas.

Information Only

7 Acknowledgements

The authors would like to acknowledge Habib Najm for his review of this report.

Information Only

8 Funding Statement

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Information Only

9 Appendix A: Scripting of EQ3/6 runs

Below, the Python scripts and sequence of tasks is laid out and explained in some but not complete detail. Hence all files are retained for future viewing/checking etc. The complete set of files is included in this ERMS package. The production EQ3/6 runs were performed on Paul Domski's computer, as he is the official WIPP EQ3/6 run master as of September 2022. The resulting Excel files containing the EQ3/6 results were then passed back to Martin Nemer for completion of the Bayesian analysis. The EQ3/6 input and output files from all runs were saved and marked with the simulation ID (see Table 1-Table 3) plus the $\log_{10}$K value that was used. These files have been retained for records purposes, spot checking, etc., in cvs/CVSLIB/WIPP_EXTERNAL/phase35production.

9.1 Setup of the EQ3 input files

EQ3 input files were generated via a Python script along with a Microsoft Excel file containing the relevant solution-composition information. The first step is the subroutine read_Excel, which is a method member of the InputFiles class, located in the file EQInputFiles.py . This subroutine expects two worksheets in the Excel file, Solution_Data and Suppression. The Solution_Data worksheet format includes a unique simulation id number, and the initial concentration of cations and anions. An example of this worksheet is shown in Table 6 for the Altmaier et. al (2003) [29] data. The first row after the Case column contains the identity of the ions in solution as they should appear in the EQ3 input file. An arbitrary small value of bicarbonate was used rather than suppress this anion. The second row contains the units of the ions in solution, as they should appear in the EQ3 input file.

Table 6. Worksheet Solution_Data from the file Altmaier_2003_EQ36_V1.xlsx

Case	Mg++	Cl-	H+	HCO3-
NA	Molality	Molality	pmH	Molality
1	2.11	4.22	8.87	1.00E-10
2	2.67	5.34	8.8	1.00E-10
3	3.26	6.52	8.74	1.00E-10
4	3.87	7.74	8.72	1.00E-10
5	4.50	9	8.74	1.00E-10
6	5.15	10.3	8.75	1.00E-10

Table 7. Worksheet Suppression from the file Altmaier_2003_EQ36_V1.xlsx

Species	Option	Alter Value
O2(aq)	Suppress	0
H2(aq)	Suppress	0
Mg3Cl(OH)5.4H2O	Suppress	0
Brucite	Suppress	0
Mg2Cl(OH)3.4H2O	Replace	26.2

An example of the Suppression worksheet is shown in Table 7. This table contains the phases that will be suppressed or altered as part of the analysis, as they should appear in the EQ3 input files. The value here for the $\log_{10}$K for $Mg_2Cl(OH)_3.4H_2O$ is a placeholder. The subroutine read_Excel places the Solution_Data worksheet information into a list of Python dictionaries, with each list item corresponding to a row in Table 6 and a copy of Table 7, both as Pandas Dataframes [46].

Information Only

Next the subroutine create_files is called, which is also a method member of the InputFiles class. This subroutine takes three inputs, a template string that will be appended with a number to be used as the EQ3 file names, an EQ3 template file (see altm_temp.3i, pann_temp.3i, Xiong_temp.3i), and the list of solution data from read_Excel. Note that EQ3/6 has a rather small character-length limit on file names. For each item in the list of solution data, create_files creates a new object of class InputFile, assigns it a file name by appending the case number to the template string, and then calls the member function Create of the class InputFile with the solution data, suppression data, and template file to create the EQ3 input file. The Create method of the InputFile is based on the previous script written by Tom Kirchner [CITE]. It has been modified for clarity and writes both the Aqueous Basis Species block and the Alter/Suppression block to the EQ3 input file. This subroutine assumes that there is a line in both sections of the template file labeled Nothing, as shown below, which is removed and replaced with the contents of the dictionary.

```
|-------------------------------------------------------------------------|
|Aqueous Basis Species/Constraint Species        |Conc., etc. |Units/Constraint|
| (uspeci(n)/ucospi(n))                          | (covali(n))|(ujf3(jflgi(n)))|
|-------------------------------------------------------------------------|
|Nothing                                         | 0.00000E+00|Molality        |
|-------------------------------------------------------------------------|

...

|-------------------------------------------------------------------------|
|Alter/Suppress Options  | (nxmod)                                        |
|-------------------------------------------------------------------------|
|Species                                         |Option          |Alter value |
| (uxmod(n))                                     |(ukxm(kxmod(n)))| (xlkmod(n))|
|-------------------------------------------------------------------------|
|Nothing                                         |Supress         |0.000000E+00|
|-------------------------------------------------------------------------|
```

9.2 Running of EQ3

Before EQ3 (or EQ6) can be run, EQ3 requires certain environmental variables to be set. This is accomplished via the routine AddEQPATH which is part of the EQ3Helpers.py file. AddEQPATH sets the environmental variables PATH, EQ36CO, and EQ36DA. EQ3/6 requires the path of the EQ3 and EQ6 executables be placed in the Windows environmental variable PATH and this same path be placed in the EQ36CO environmental variable. EQ3/6 requires the location of the database folder be placed in the EQ36DA environmental variable.

Next the EQ3 input files can be run, which is performed by the subroutine RUNEQ_Simp_Mult which is also part of the EQ3Helpers.py file. This subroutine takes 4 input parameters, the EQ3/6 work directory, the executable to be run (runeq3 for EQ3), the database "three letter name", and a list of input files to be run. The subroutine RUNEQ_Simp_Mult then calls the subroutine RUNEQ_Simp (also in EQ3Helpers.py) with the subprocess.Popen() subroutine in the subprocess package (standard in Python 3.9) to execute each run. RUNEQ_Simp_Mult waits until each run is done before moving on to the next run, to avoid file conflicts.

9.3 Setup of the EQ6 input files

9.3.1 Generation of EQ6 input files from the EQ3 pickup file

Each EQ3 run from the steps listed above produces an EQ6 pickup file. These files have an extension .3p and contain the bottom half of the EQ6 input file. The corresponding EQ6 input files are then created by the subroutine create_files6 which a member method of the InputFiles class. This subroutine takes as its argument a EQ6 input file template (see altm_temp.6i), which contains only the top half of the EQ6 input file. The subroutine create_files6 then calls the subroutine Create6 which is a member

Information Only

method of the InputFile class. This subroutine simply pastes the top and bottom portions of the template and pickup file together to create the initial EQ6 input file.

In the EQ6 template files (altm_temp.6i, pann_temp.6i, xiong_temp.6i), we have added a block to titrate in the respective phase (phase-3 or phase-6), as shown below for phase-3. The amount added is arbitrarily large, such that the solution is always saturated with the phase in question.

```
|-------------------------------------------------------------------------------|
|Reactants (Irreversible Reactions) | (nrct)                                    |
|-------------------------------------------------------------------------------|
|Reactant        |Mg2Cl(OH)3.4H2O         | (ureac(n))                           |
|-------------------------------------------------------------------------------|
|->|Type         |Pure mineral            | (urcjco(jcode(n)))                   |
|-------------------------------------------------------------------------------|
|->|Status       |Reacting                | (urcjre(jreac(n)))                   |
|-------------------------------------------------------------------------------|
|->|Amount remaining (moles) | 5.00000E+00| (morr(n))                          |
|-------------------------------------------------------------------------------|
|->|Amount destroyed (moles) | 0.00000E+00| (modr(n))                          |
|-------------------------------------------------------------------------------|
|->|Surface area option (nsk(n)):                                               |
|->|  [x] ( 0) Constant surface area:                                           |
|->|             Value (cm2)       | 0.00000E+00| (sfcar(n))                   |
|->|  [ ] ( 1) Constant specific surface area:                                  |
|->|             Value (cm2/g)     | 0.00000E+00| (ssfcar(n))                  |
|->|  [ ] ( 2) n**2/3 growth law- current surface area:                         |
|->|             Value (cm2)       | 0.00000E+00| (sfcar(n))                   |
|-------------------------------------------------------------------------------|
|->|Surface area factor      | 0.00000E+00| (fkrc(n))                          |
|-------------------------------------------------------------------------------|
|->|Forward rate law          |Relative rate equation  | (urcnrk(nrk(1,n)))     |
|-------------------------------------------------------------------------------|
|--->|dXi(n)/dXi (mol/mol)      | 1.00000E+00| (rkb(1,1,n))                     |
|-------------------------------------------------------------------------------|
|--->|d2Xi(n)/dXi2 (mol/mol2)   | 0.00000E+00| (rkb(2,1,n))                     |
|-------------------------------------------------------------------------------|
|--->|d3Xi(n)/dXi3 (mol/mol3)   | 0.00000E+00| (rkb(3,1,n))                     |
|-------------------------------------------------------------------------------|
|->|Backward rate law         |Partial equilibrium     | (urcnrk(nrk(2,n)))     |
|-------------------------------------------------------------------------------|
```

9.3.2 Run of the initial EQ6 input files

The initial EQ6 input files are then run as described above in Section 9.2, but in this case the executable command is now runeq6 rather than runeq3.

9.3.3 Manipulation of EQ6 input files to generate further runs with different parameter values

To alter the $\log_{10}$K value of the respective phase (phase-3 or phase-5), the subroutine Change_LogK_All is used, which is a member method of the InputFiles class. This subroutine takes as input a list of the phases and $\log_{10}$K values to be placed into the Suppression block of the EQ6 input file. Change_LogK_All then calls Change_logK for each file in the InputFiles class object. Change_logK is a member method of the InputFile class, which searches for the relevant line in the input file to modify with the altered value of $\log_{10}$K. An example of the result of that subroutine is shown below for a phase-3 run with a $\log_{10}$K = 28.0.

Information Only

Alter/Suppress Options (nxmod)		
Species (uxmod(n))	Option (ukxm(kxmod(n)))	Alter value (xlkmod(n))
Mg2Cl(OH)3.4H2O	Replace	2.8000E+01
Brucite	Suppress	0.00000E+00
Mg3Cl(OH)5.4H2O	Suppress	0.00000E+00
H2(aq)	Suppress	0.00000E+00
O2(aq)	Suppress	0.00000E+00

9.3.4 Running of EQ6

With the modified input files from Section 9.3.3, the EQ6 input files were repeatedly re-run, always using the pickup file from the previous value of $\log_{10}K$ as the template for the next set of runs. The use of the previous pickup file is accomplished by the subroutine create_files6fp, which is a member method of the InputFiles class. This subroutine just copies each .6p file to its corresponding .6i file.

9.4 Collection of the data from the EQ6 output files

Collection of the output data from the EQ6 *.6o files is performed by several routines in the OutputFiles class, located in the EQ3OutputFiles.py file. The original input conditions of all of the *.3i files is obtained by the inputs_df member method of the OutputFiles class. This method then calls the ReloadInput member method of the OutputFile class, which loads all of the solution ions from the Aqueous Basis Species block of the associated *.3i file. This functionality belongs in the InputFile and InputFiles class, and will potentially be moved there in future work.

The output from the *.6o files is obtained from the results_df member method of the OutputFiles class. This method then calls the ReadOutput6_last member method of the OutputFile class. The ReadOutput6_last reads the last output step of the *.6o file, and places the results into a dictionary of dataframes, for different datablocks of the output file including the “Numerical”, “Sensible”, “Distribution”, and “Saturation” blocks. The method also obtains the Mesmer pmH. Note that in this analysis we only used the ‘Sensible’ and the pmH blocks of the *.6o output files.

Information Only