Alternatives to the Current Uncertainty Methodology

Introduction

EPA during their completeness review of the CRA-2019 wrote in Section 7.5 of their Technical Support Document [1] "*An evaluation of the relative advantages and disadvantages of other approaches for addressing +III and +IV actinide solubility uncertainties may improve confidence in the current approach.*" In contrast, EPA wrote in 1998 [2], "*DOE used the outermost limits of the differences between model results and experimental results for all data…";" This procedure greatly expands the size of the uncertainty bounds beyond what might be calculated from statistical treatment of the distribution of the differences…*" Hence, it's clear that in 1998 EPA considered the possibility of treating the differences between model and experimental results in a statistical way. This appendix covers an alternative methodology that is not intended for use in the CRA-2024. This statistical approach is an alternative to the current approach to uncertainty analysis in the actinide concentrations used in the near field [3, 4] in PA.

Current Methodology

The current uncertainty methodology calculates the difference between the experimental results and the model prediction of those same experimental results as the so-called *D* values on a log scale,

$$D_i = \log m_{\text{exp},i} - \log m_{\text{mod}}(\xi_i, \theta), \tag{1}$$

where $m_{\text{exp},i}$ is the experimentally measured concentration of the actinide or analog species of interest in the i'th experiment, $m_{\text{mod}}(\xi_i, \theta)$ is the model of the *i*'th experiment which depends on input independent variable vector ξ_i associated with experiment *i*, and θ is the vector of model parameters that don't change from experiment to experiment (Pitzer parameters, log K(s), etc.). In equation (1), the log is taken to be the base-10 logarithm (natural logarithm would not change the conclusions herein). The usual assumption in formulating a difference, as in equation (1), is that the *D* values correspond to an error term that is distributed in some way, e.g.

$$D_i \approx \varepsilon(\xi_i, \theta), \tag{2}$$

where for example we might find through an analysis that

$$\varepsilon(\xi_i, \theta) \sim N\left(\mu(\xi_i, \theta), \sigma^2(\xi_i)\right), \tag{3}$$

where $N(\mu, \sigma^2)$ is a normal distribution of mean μ and variance σ^2. Here we write the mean value as a function of the independent variables and the model parameters, $\mu(\xi_i, \theta)$, as this represents the deterministic bias between the experiment and the model, which depends on the model through θ and depends on the experimental conditions through ξ_i. The variance $\sigma^2(\xi_i)$ is written as a function of the independent parameters as well, to account for possible heteroskedasticity, meaning that the variance may be large for some experimental conditions (e.g., when the concentration approaches detection limits).

WIPP:4.2.1:PA:QA-L:574938

Information Only

1 of 7

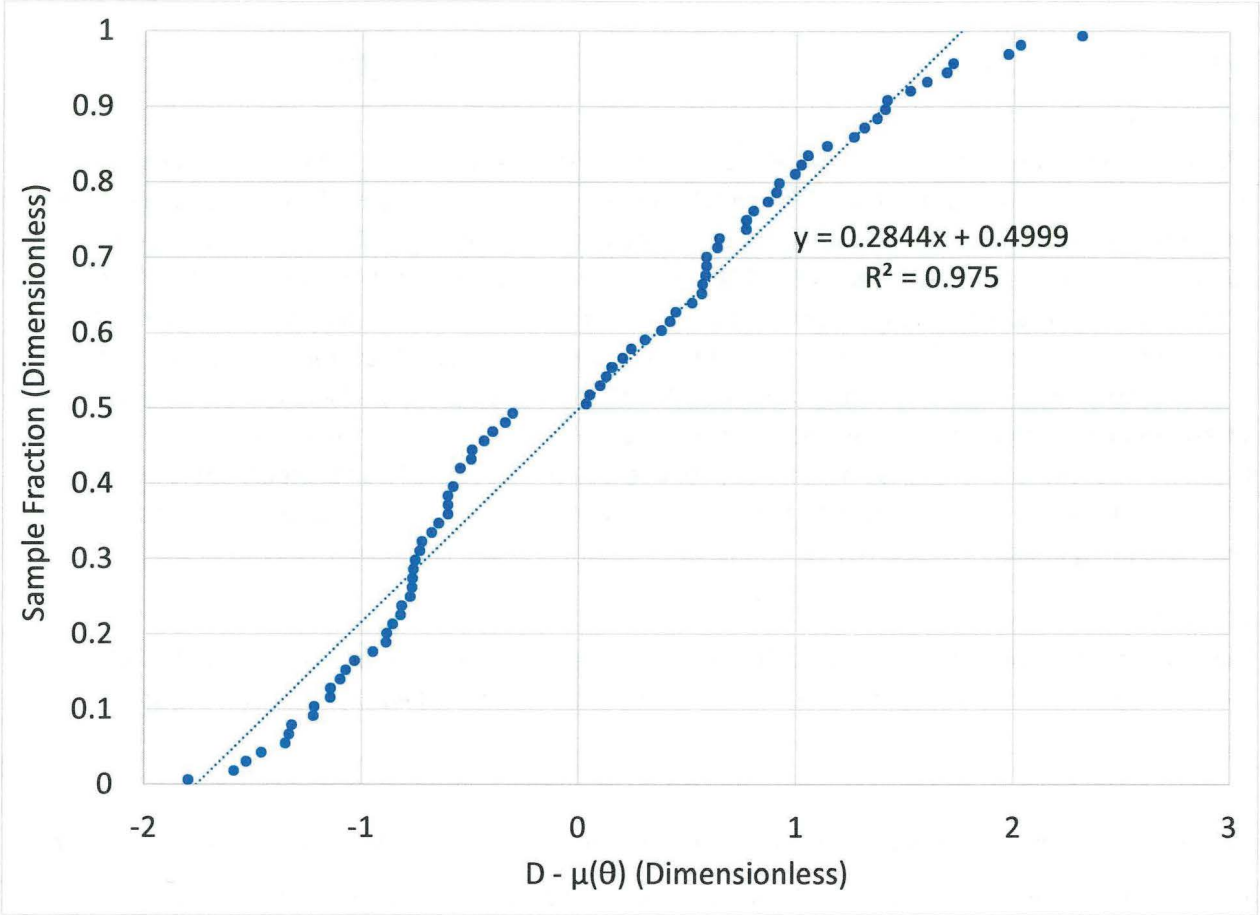

Figure 1. Sample Fraction (dimensionless) versus mean corrected bias values from the An(III) test data of the CRA-2019.

In the previous and current CRA uncertainty methodology, the bias term and the variance are both assumed to be independent of ξ, i.e. $\mu(\xi_i,\theta) \approx \mu(\theta)$ and $\sigma^2(\xi_i) \approx \sigma^2$. This assumption results in an error term that is nearly a uniform distribution,

$$\varepsilon \sim U(-2,2), \tag{4}$$

where the bounds -2 to 2 here are approximate and meant for this discussion only. It should be noted that PA doesn't sample equation (4). Instead, they fit a CDF curve to the *D* values (i.e. an empirical CDF) and directly sample that fitted CDF [3, 4].

Conceptual and Mathematical Issues with the Current Methodology

Philosophically, one could argue that Figure 1 and equation (4) represent an empirical sampling from the epistemic distribution of possible model deviations. The validity of this argument remains unclear to the authors. However, as will be shown, the *D* values calculated via equation (1) are significantly affected by non-stochastic bias. This implies that the form, shape, and width of any distribution fit to the *D* values calculated via (1) (e.g. equation (4)) will depend strongly on the data sets selected to calculate the *D* values and the current state of the model.

A distinct issue with using Figure 1 and equation (4) is that they marginalize over ξ, and hence marginalize out the chemical conditions for which the model is used in PA. This is problematic because

Information Only

the purpose of the uncertainty analysis is to determine the uncertainty in the model under the chemical conditions for which the model is used in PA. The purpose of the uncertainty analysis is not to estimate global epistemic uncertainty in the model, which is difficult to mathematically justify at best, and at worst would require an expert elicitation.

Alternative Methodology

As an alternative, one can separate the bias from the error term by modeling the bias as deterministic and the remainder as stochastic

$$R_i \equiv D_i - \mu(\xi_i, \theta) \approx \varepsilon(\xi_i). \tag{5}$$

where we now require the expected value of ε to be zero (un-biased) for a given fixed value of ξ_i,

$$E[\varepsilon(\xi_i)] = 0. \tag{6}$$

One way to obtain the decomposition shown in equation (5) is to fit the *D* values as a function of ξ = (*pc*H, *m*Na, *m*Mg, *m*Ca, mCO_3), using standard linear regression in Excel,

$$\mu(\xi, \theta) \approx C(\theta) + A(\theta) \cdot \xi + \xi^T \cdot B(\theta) \cdot \xi, \tag{7}$$

where $C(\theta)$ is a scalar, $A(\theta)$ is a vector, and $B(\theta)$ is a matrix of coefficients. This ensures that the mean value of R_i is zero (property of linear least squares regression) for the fitted dataset, maintaining the constraint given by equation (6). The results of that regression are shown in Table 1, where we (subjectively) only fit to the terms shown in Table 1.

Table 1. Coefficients of the linear regression of the CRA-2019 D values with independent variables ξ.

Term	Coefficient	Standard Error	t Stat	P-value
Constant	15.08	3.83	3.93	0.00
*pc*H	-1.40	0.43	-3.26	0.00
*m*Na	-3.16	0.74	-4.29	0.00
*m*Mg	-2.54	1.47	-1.73	0.09
*m*Ca	-0.95	1.43	-0.67	0.51
mCO_3	-59.39	23.28	-2.55	0.01
*pc*H *x* *m*Na	0.30	0.08	3.61	0.00
*pc*H *x* *m*Mg	0.24	0.17	1.38	0.17
*pc*H *x* *m*Ca	0.04	0.16	0.26	0.80

The *P*-values shown in Table 1 indicate most of the terms (with perhaps the exception of *pc*H x *m*Ca) are somewhat important. Interpreting the *P*-values in Table 1 relies on the assumption of normality, i.e.,

$$\varepsilon(\xi_i) \sim N\left(0, \sigma^2(\xi_i)\right). \tag{8}$$

This can be checked by plotting the resulting R values on a probability plot. A probability plot of R, based on the linear regression shown in Table 1, is shown in Figure 2 versus both a normal distribution

Information Only

and a Student-T distribution; in both cases we assumed a constant variance $\sigma^2(\xi_i) \approx \sigma^2$, i.e. ignoring any heteroskedasticity for now. A perfect fit is represented by the black line; the Student-T distribution is perhaps a slightly better fit than the normal distribution. Thus, we have shown that bias-correcting the *D* values results in a residual that, in our opinion, is adequately modeled by a normal distribution and perhaps slightly better modeled by a Student-T distribution.

If we are to argue that Figure 2 represents epistemic (and aleatory) uncertainty, one could counter that we should have used a Bayesian regression rather than a frequentist regression. This is a valid criticism; however, the purpose of this appendix is to demonstrate the effect of deterministic bias in the *D* values on the resulting uncertainty distribution. For that purpose, the linear regression shown herein is just an example of a means to bias correct the *D* values. Furthermore, the calculations provided herein were all performed in Excel for the ease of the reader to follow. This can be revisited in future work.

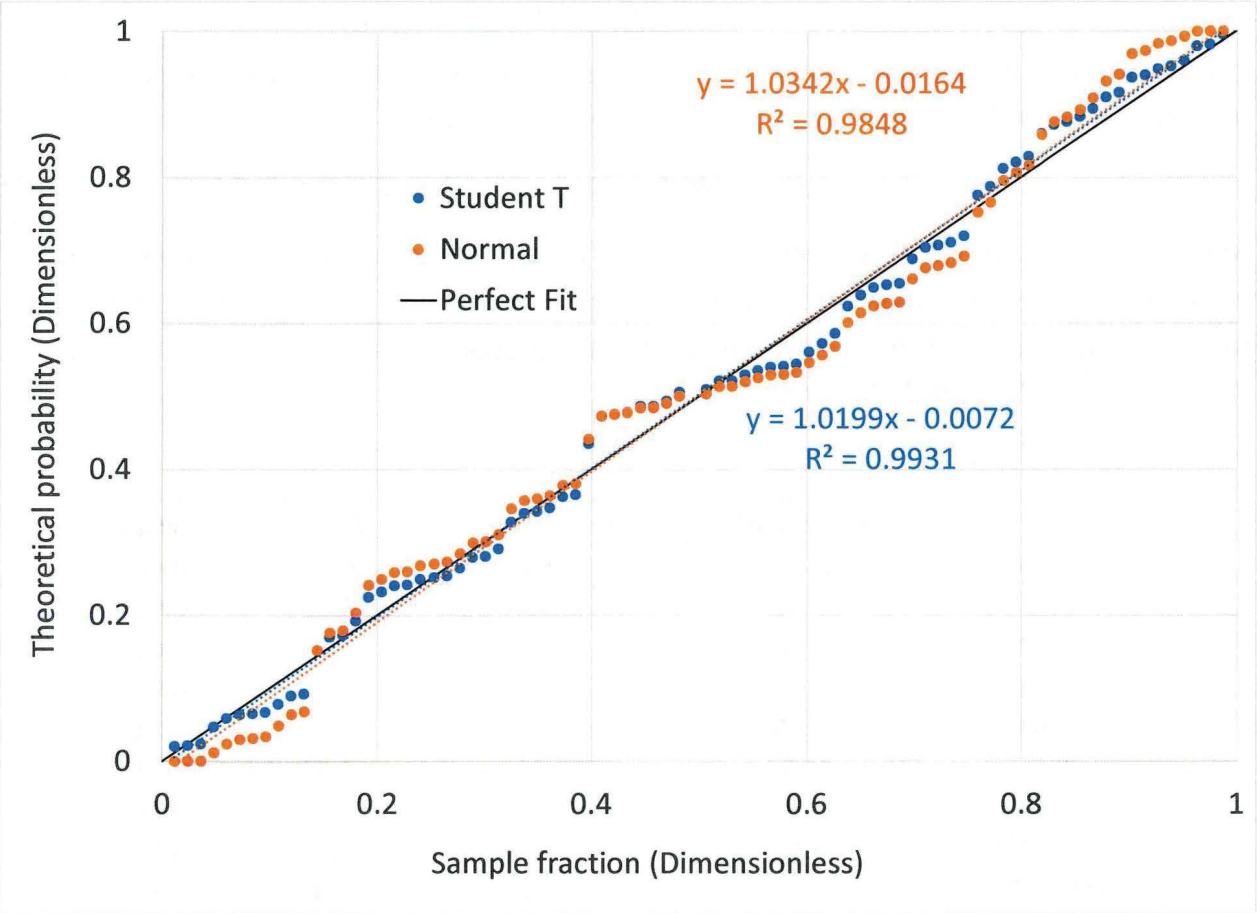

Figure 2. Bias-corrected D values for An(III) from the CRA-2019. Here the black solid line is a line of slope = 1 and intercept = 0, representing perfect agreement between the distribution and the data. Here, the blue dots are the theoretical probability values assuming a Student-T distribution versus the sample fractions plotted; orange dots are the theoretical probability values assuming a Normal distribution versus the sample fractions; the small blue and orange dots are trendlines of the data using the Add Trendline feature in Excel.

As discussed in the previous paragraph, a frequentist linear regression is not the only way to separate the deterministic (bias) and stochastic portions of the *D* values. Linear regression, in general, has the limitation that a model-form for the bias term $\mu(\xi,\theta)$ must be assumed, e.g. equation (7), which will

Information Only

affect (bias) the shape of the resulting distribution on ε, particularly in the absence of abundant experimental data. There are many model-free methods such as a Gaussian process [5], and/or other types of locally co-variant fits (e.g. a piece-wise continuous spline). We don't expect ε to change dramatically using different fits to remove the bias (particularly when uncertainty in the fit is taken into account), but we acknowledge that there will be differences.

Given a decomposition of the *D* values into a deterministic bias $\mu(\xi,\theta)$, and a stochastic component ε, the next question is how these terms can be utilized in WIPP PA. The simplest usage is to calculate $\mu(\xi,\theta)$ in the neighborhood of WIPP conditions used in PA, and then sample ε for each of the PA vectors, similarly to how *D* is sampled in PA. Finally add $\mu(\xi,\theta)$ plus the sampled value of ε to the modeled actinide solubilities in every vector in PA.

Figure 2 also indicates that if the geochemistry model better fit the test data (smaller bias), the remainder R_i could be adequately modeled by simple distributions such as a normal distribution or a Student-T distribution. This provides evidence for the future use of a data likelihood function

$$p(\log m_{\exp,i} \mid \theta) = \frac{1}{\left(2\pi\sigma^2\right)} \exp\left(\frac{-R_i^2}{2\sigma^2}\right), \tag{9}$$

where $p(\log m_{\exp,i} \mid \theta)$ is the probability density of the experimental data point $\log m_{\exp,i}$, given model parameters θ. This then leads to the possibility of utilizing equation (9) in a Bayesian inverse analysis to determine the uncertainty in the geochemistry model parameters θ through Bayes' rule [6, 7],

$$p(\theta \mid \log m_{\exp,i})\, p(\log m_{\exp,i}) = p(\log m_{\exp,i} \mid \theta)\, p(\theta), \tag{10}$$

where $p(\theta \mid \log m_{\exp,i})$ is the probability density of the model parameters given the experimental data $\log m_{\exp,i}$. That distribution could then be forward propagated through the geochemistry model to produce uncertainty in the model predictions for use in PA. We will not go deeper into this discussion here, only to state that Figure 2 provides evidence for the use of equation (9) in future model parameter uncertainty analysis.

Conclusion

Herein we have provided evidence that the An(III) test data from the CRA-2019 can be adequately modeled by standard probability distributions after bias correction. However, data in the neighborhood of WIPP conditions (e.g. in GWB, ERDA-6, and in statistical perturbations of these brines) would provide the most relevant test data for the model, and would greatly strengthen the alternative methodology shown herein. We do not intend to implement any bias-correction method described herein in the CRA-2024, and we continue to believe that the distribution of the un-bias-corrected *D* values is at least bounding.

References

1. U. S. Environmental Protection Agency, *Technical Support Document For Section 194.24; Evaluation Of the Compliance Recertification Application (CRA-2019) Actinide Source Term, Gas Generation, Backfill Efficacy, Water Balance, and Culebra Dolomite Distribution Coefficient*

Information Only

Values. 2022, U. S. Environmental Protection Agency; Office of Radiation and Indoor Air; Center for Waste Management and Regulations: Washington, DC. DOCKET: EPA-HQ-OAR-2019-0534.
2. Environmental Protection Agency, *Criteria for the Certification and Recertification of the Waste Isolation Pilot Plant's Compliance With the Disposal Regulations: Certification Decision*, 1998. Federal Register, 63(95), 27354-27406. United States of America: Washington, DC. FR Doc. 98–13100.
3. Zeitler, T.R., *Analysis Package for Parameter Sampling in the 2019 Compliance Recertification Application Performance Assessment (CRA-2019 PA)*. 2019, Sandia National Laboratories: Carlsbad, NM. ERMS 571367.
4. United States Department of Energy; Waste isolation Pilot Plant, *Title 40 CFR Part 191 Subparts B and C Compliance Recertification Application 2019 for the Waste Isolation Pilot Plant; Appendix PA-2019 Performance Assessment*. 2019, US DOE, Carlsbad Field Office: Carlsbad, NM.
5. Gramacy, R.B., *Surrogates: Gaussian Process Modeling, Design, and Optimization for the Applied Sciences*. Chapman & Hall/CRC Texts in Statistical Science. 2020, Boca Raton, FL: CRC.
6. Silva, D.S., *Data Analysis; A Bayesian Tutorial*. Second Edition ed. 2007, Norfolk, Great Britain: Oxford University Press.
7. Hoff, P.D., *A First Course in Bayesian Statistical Methods*. 2009, Dordrecht, Netherlands: Springer.
8. Domski, P.S., *Uncertainty Analysis of Actinide Solubilities for CRA 2019*. 2019, Sandia National Laboratories: Carlsbad, NM. ERMS# 571179.

QA information on the Calculations shown Herein

All of the figures and calculations shown herein were performed in the spreadsheet "An(III)_meas-sim_V3_Raw_Simple_Linear_Regression_V5.xlsm" archived at cvs/CVSLIB/WIPP_EXTERNAL/uncertaintymethod. Columns A-S, Rows 1-83 are the raw experimental data and model results for An(III) obtained as part of the uncertainty analysis from the CRA-2019. These values are traceable to the CRA-2019 Uncertainty Analysis [8].

The linear regression shown in Table 1 is shown in the spreadsheet in Columns C-J, Rows 90-200. This regression was performed using the Excel Regression Data-Analysis toolbox. The input Y-range for this analysis was Column S, Row 1-83. The input X-range for this analysis was Columns C-J, Rows 1-83. Admittedly, the X-range for the linear regression are outputs from EQ3/6 rather than direct experimental measurements. For the purpose of demonstrating the alternative statistical analysis shown herein, we do not believe this difference is important as the differences between EQ3/6 and the experimental values for these columns are small. We refer the reader to the Excel instructions for understanding the other elements of the regression analysis shown in Columns C-J, Rows 90-200.

Column T, Rows 1-83 calculates the bias in the *D* values using the linear regression discussed in the previous paragraph. Column U then subtracts that bias from the *D* value in column S. Column W calculates the empirical percentile (as a fraction) of each data point using the PERCENTRANK.EXC() function.

Column Y re-scales the data in column U by shifting and rescaling the data according to the values given in cells AA2 and AA3, according to the following formula

Information Only

$$\overline{R}_i = \frac{R_i - \overline{\mu}}{\overline{\sigma}}, \tag{11}$$

where $\overline{\mu}$ and $\overline{\sigma}$ should be understood as fitting parameters to the probability distribution. Instead of the method of moments (e.g. calculating the mean and the standard deviation), we opted for a global approach by calculating the sum-of-square difference between the theoretical probabilities for a given distribution and the calculated empirical percentiles given in column W. The effect is to give the points in the tails the same weight as points closer to the middle of the distribution. For the Student-T distribution, the probabilities are calculated in Column AB via the T.DIST() function. Then in column AC, the square difference between the theoretical probability in Column AB and the empirical "percentile" in Column W is calculated. The sum of that square difference is then given in Cell AA5. The Solver in Excel was then used to minimize the sum of square error in Cell AA5, by manipulating Cells AA2 – AA4. Excel appears to be unable to accommodate non-integer values of the degrees of freedom, ν, in the Student-T distribution; we are using the parameter ν as a fitting parameter, not the number of degrees of freedom in the problem. Hence, the fit is likely not optimum for a continuous $\nu \in (1, \infty)$. Columns AE through AI, Rows 1 – 83 perform the same analysis as in the above text but for a Normal distribution. Figure 2 plots Column AB and Column AH versus Column V. Trendlines are added using the Add Trendline feature in Excel.

Column AK, Rows 1-83, are the values in Column U after subtracting the mean value from Column U. Column AM is a copy of the values in Column AK, sorted from smallest to largest. Column AN is then the percentiles of the sorted values in Column AK, again using the PERCENTRANK.EXC() function. Figure 2 is then plotted as Column AN versus Column AM. A trendline is plotted using the Add Trendline feature in Excel.

Information Only