1   **Reduced Order Models for Assessing CO₂ Impacts in Shallow**
2   **Unconfined Aquifers**

3
4
5
6   Keating, Elizabeth H.; Harp, Dylan H.; Dai, Zhenxue; Pawar, Rajesh
7
8
9

10
11
12
13   Computational Geoscience Group
14   Earth and Environmental Sciences Division
15

16   Los Alamos National Laboratory, Los Alamos, NM 87545
17
18
19

20
21   *For submission to the NRAP special issue*
22   *International Journal of Greenhouse Gas Control.*
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40

41
42   Corresponding author: Elizabeth Keating, MS T003, Earth and Environmental
43   Sciences Division, Los Alamos National Laboratory, Los Alamos, NM 87545
44   ekeating@lanl.gov, 505-500-7213 (voice), 505-665-8737 (fax).
45

Abstract

Risk assessment studies of potential CO₂ sequestration projects consider many factors, including the possibility of brine and/or CO₂ leakage from the storage reservoir. Detailed multiphase reactive transport simulations have been developed to predict the impact of such leaks on shallow groundwater quality; however, these simulations are computationally expensive and thus difficult to directly embed in a probabilistic risk assessment analysis. Here we present a process for developing computationally fast reduced-order models which emulate key features of the more detailed reactive transport simulations. A large ensemble of simulations that take into account uncertainty in aquifer characteristics and CO₂/brine leakage scenarios were performed. Twelve simulation outputs of interest were used to develop response surfaces (RSs) using a MARS (multivariate adaptive regression splines) algorithm (Milborrow, 2011).

A key part of this study is to compare different measures of ROM accuracy. We show that for some computed outputs, MARS performs very well in matching the simulation data. The capability of the RS to predict simulation outputs for parameter combinations not used in RS development was tested using cross-validation. Again, for some outputs, these results were quite good. For other outputs, however, the method performs relatively poorly. Performance was best for predicting the volume of depressed-pH-plumes, and was relatively poor for predicting organic and trace metal plume volumes. We believe several factors, including the non-linearity of the problem, complexity of the geochemistry, and granularity in the simulation results, contribute to this varied performance.

The reduced order models were developed principally to be used in probabilistic performance analysis where a large range of scenarios are considered and ensemble performance is calculated. We demonstrate that they effectively predict the ensemble behavior. However, the performance of the RSs is much less accurate when used to predict time-varying outputs from a single simulation. If an analysis requires only a small number of scenarios to be investigated,

1 computationally expensive physics-based simulations would likely provide more
2 reliable results. If the aggregate behavior of a large number of realizations is the
3 focus, as will be the case in probabilistic quantitative risk assessment, the
4 methodology presented here is relatively robust.

5
6 Keywords: response surface, CO₂ leakage impact, shallow groundwater quality

7 **1 Introduction**

8
9 One of the potential risks associated with geologic CO₂ sequestration is
10 environmental damage caused by leaks. One scenario of concern is that injected CO₂
11 or displaced brine might leak upward through a caprock breach, a previously
12 undetected fault, or a leaky wellbore. The ultimate impact of such a leak on shallow
13 groundwater quality is unknown. Although important insights into key processes
14 have been gained through laboratory experiments (10⁻¹ – 10⁰m scale) and
15 controlled release experiments (10⁰ – 10²m scale), the only option for predicting
16 large-scale (10² – 10³m) impacts is to rely on numerical simulations. Creating
17 large-scale predictions of the impact of leaking CO₂ or brine on aquifers requires
18 computationally expensive multi-phase, multi-component simulations. The
19 predictions will be inherently uncertain, due to a number of factors including
20 unknown leak rates, unexpected leak locations, hydrogeochemical variability in the
21 aquifer. To make useful forecasts that fully account for these uncertainties, it may
22 be necessary to derive computationally-efficient surrogate models from the set of
23 available simulations. The inputs for these models could include
24 hydrogeochemical characteristics of the shallow aquifer (e.g. permeability, porosity,
25 mineral content, background flow rate, etc.) and characteristics of the brine and CO₂
26 leak. The outputs of these models would include spatial and temporal
27 characteristics of leakage-induced plumes in the aquifer.

28 The surrogate modeling approach has been widely used in engineering
29 applications (Simpson et al., 1997 and references within), and somewhat less often

1 in the hydrologic sciences (Sharda et al., 2008; Kourakos and Mantoglou, 2009;
2 Keating et al., 2010, M Chen et al. ,2012, M Chen et al. ,2013, Jordan et al., 2015, M
3 Chen et al., 2015). Typical applications for surrogate models are optimization
4 (Regis and Shoemaker, 2007; Keating et al. 2010; Dai et al., 2014a), sensitivity
5 analysis, uncertainty quantification, and risk assessment (Viswanathan et al., 2008).
6 The primary application for this study is risk assessment. The project goal is to have
7 a fast surrogate model of groundwater impacts due to leakage embedded in a
8 systems-level carbon sequestration risk assessment model (Viswanathan et al.,
9 2008; Pawar et al., 2013). For this risk assessment purpose, simulations have been
10 developed for two major classes of aquifers (Carroll et al., 2014), each with separate
11 hydrogeologic and geochemical submodels. Here we only discuss one of these, the
12 hydrogeologic submodel of the unconfined carbonate aquifer model (Bacon et al.,
13 2014).

14   There are many surrogate model methodologies available, including
15 polynomials (Fen et al. 2009), radial basis functions (Regis et al., 2007), kriging
16 (Simpson et al., 2001), support vector machines (Zhang et al., 2009, Lucas et al.,
17 2013), artificial neural networks (Behzadian et al., 2009; Dowla and Rogers, 2003),
18 sparse grid interpolation (Zeng et al., 2012) and random forest techniques (Boyle et
19 al., 2015). Two broad categories are 1) response surface methods, which are
20 trained on simulation ‘data’ (inputs and outputs) and 2) reduced fidelity models,
21 which are simplified physics-based models. Unfortunately, as emphasized by Jin et
22 al. (2001), the selection of a surrogate modeling approach for a particular
23 application is rarely systematic. Furthermore, the criteria used to measure
24 surrogate model adequacy can be lacking. Often, the only criterion used is
25 goodness-of-fit of the RS to the training dataset. This criterion is inadequate if the
26 purpose of the model is to make predictions for parameter combinations not
27 included in the training dataset (Moore and Doherty, 2005). Jin et al. (2001)
28 propose that multiple criteria be used to assess performance including accuracy,
29 robustness, transparency, efficiency and simplicity. They define robustness to be a
30 measure of how well the model performs on parameter combinations that were not
31 part of the training data set, or novel problems. Transparency is a measure of how

1 easily useful statistics such as parameter importance, parameter interactions, and
2 cross-validation can be deduced from the model.

3 For this application we chose to use the MARS (Multivariate adaptive
4 regression splines; Freidman (1991)) methodology, primarily because of its
5 advantages with respect to transparency. The algorithm, as implemented in the
6 'earth' package (Milborrow 2015) for R (R Core Team, 2014), allows facile control of
7 the balance between overfitting and accuracy, extraction of parameter importance
8 information, and assessment of predictive performance (robustness) through cross-
9 validation studies. The method has been used in hydrologic applications previously
10 (Sharda et al., 2008; Jordan et al., 2015; Chen et al., 2012; Chen et al., 2013; Chen et
11 al., 2015). The method does not require independent variables to be uniformly or
12 normally distributed. This is advantageous in situations such as analysis of a
13 simulation dataset with 'missing values' due to simulation crashes, or opportunistic
14 analysis of a simulation set developed for another purpose.

15 The simulations were developed to predict the impact of brine and/or CO₂
16 leaking from a hypothetical point source at the base of a shallow 3-D heterogeneous
17 unconfined aquifer (Carroll et al., 2014; Dai et al., 2014b). The computations
18 include multi-phase (CO₂ and water) flow, pressure and temperature dependent CO₂
19 solubility in water, and reactive transport of solutes dissolved in the brine. For
20 computational efficiency, this complex problem was divided into two separate sets
21 of simulations. The first set focused on multi-phase transport of CO₂, carbonate
22 geochemistry, and non-reactive transport of solutes introduced into the aquifer with
23 the brine. Uncertainty analysis was conducted on this set of simulations to capture
24 the effect of hydrogeologic uncertainty on model outputs. The second set focused
25 on multi-phase transport of CO₂, carbonate geochemistry, and a complex set of
26 aqueous and water/rock reactions affecting trace metal and organics (Bacon et al.,
27 2014, Bacon et al., 2015). Uncertainty analysis was conducted to capture the effect
28 of geochemical uncertainty on model outputs. The two sets of simulations were
29 used to develop two sets of reduced order ROMs, which are then combined when
30 used for risk assessment (Keating et al., 2015). This paper describes only the *first*
31 set of simulations (simple geochemistry) and associated ROM development.

2 Methods

2

3 This section describes the detailed process-level simulations, the simulation
4 input parameter variations used to create the training data set for MARS, and the
5 response surface methodology.

6

2.1 Simulations of brine/CO₂ impacts in a shallow unconfined aquifer

8

9 The FEHM reservoir simulator (Zyvoloski, 2011) was used for this set of
10 calculations.

2.1.1 Leakage scenarios and aquifer models

12

13 A wide range of time-varying brine and CO₂ leak rate scenarios were developed
14 using a heuristic model (Figure 1) derived from physically-based wellbore leakage
15 simulations (Carroll et al., 2014; Jordon et al., 2015). The input parameters for this
16 heuristic model are shown in Table 1. The scenarios last 200 years; however, in
17 most of scenarios the leak is only active during a small fraction of this time. In some
18 scenarios the leak slowly subsides to zero, as might be expected if pressure
19 management is used to mitigate the leak, or if the leak itself reduces reservoir
20 overpressurization (passive mitigation). In other scenarios the leak is assumed to
21 stop instantaneously at a 'mitigation time (T_m)' by plugging of a leaky wellbore
22 (active mitigation). Hereafter we refer to the time period after the leak rate is
23 reduced to zero as 'post leak', regardless of whether active or passive mitigation has
24 occurred. The chloride concentration of the leaking brine is an input parameter,
25 held constant in time.

26 Models of the aquifer were developed using the FEHM simulator (Zyvoloski,
27 2011). Aquifer characteristics were meant to reflect hydrogeochemical conditions
28 at an unconfined portion of the carbonate Edwards aquifer (EA) (Lindgren et al.,

2004), however, some input parameters were varied beyond the range of site-specific conditions so the results could be applied to other sites (see Table 1). Initial (pre-leak) conditions were presumed to be single phase (water) steady-state horizontal 1D flow. The horizontal head gradient was controlled using specified time-invariant hydrostatic pressures imposed at the upstream and downstream boundaries; the gradient was varied from $2.8\text{E-}4$ to $1.9\text{E-}2$ m/m. Pressures at the upper boundary (water table) were maintained at 0.1 MPa. Aquifer thickness was allowed to vary between 100 and 500m. Spatially heterogeneous porosity and permeability values were assigned using a random Gaussian model (Dai et al., 2014b); the mean values, total variance, and correlation length were variable input parameters (Table 1). For the range of leak rates, lateral head gradient, and permeabilities considered, the leak rates (max 0.4 kg/s) were always much smaller than the background flow rate (median = 134 kg/s; range 1.4 – 1018 kg/s). The lateral extent of the model domain was designed to be small enough to allow for fine grid discretization, yet larger than any CO₂ plume. A preliminary set of simulations showed that a lateral domain size of $dx=5\text{km}$, $dy=2\text{km}$ was adequate for most scenarios.

The lower boundary was no-flow except at the location of the point-source leak ($X=500\text{m}$; see Figure 2). We used very fine horizontal grid spacing near the presumed leak location ($dx=dy=20\text{m}$) and coarser spacing elsewhere ($dx=dy=80\text{m}$). The fine grid spacing near the leak was necessary to accurately simulate free-phase CO₂ with the consequent buoyancy. Vertical grid spacing was constant at 20m throughout. At the node at the base of the model representing the leak, both brine and CO₂ were introduced as time-varying specified flow rates (kg/s). At all nodes (including the inlet node) CO₂ is assumed to dissolve completely into the water; if the CO₂ solubility limit is exceeded then free phase will appear.

We simulated transport of four dissolved components (C, H, Ca, and Cl), six aqueous ions/complexes (H^+ , OH^- , H_2CO_3^0 , HCO^- , Ca^{2+} , Cl^-) and one solid phase (calcite). Calcite precipitation/dissolution was included because of its importance in the limestone Edwards Aquifer and because of its buffering effect on pH. Complexation and precipitation/dissolution reactions are listed in Table 2.

1 Because this set of simulations only considers non-reactive transport of solutes
2 originating in the leaking brine, trace metals and organic concentrations could be
3 assumed to co-vary with chloride at fixed ratios. These ratios were derived from
4 analysis of brine water quality data (Last et al., 2013).

5

6 2.1.2 Aquifer impact plumes

7

8 Nine measures of water quality impact were considered: changes in pH, TDS, As,
9 Pb, Cd, Ba, benzene, naphthalene, and phenol. Two alternative definitions of ‘impact’
10 were considered: one, changes that cause an exceedance of a drinking water
11 standard or maximum contaminant level (MCL), and two, changes that are above
12 and beyond ‘natural background variability’ in the aquifer (Last et al., 2013). For
13 the remainder of the paper, we refer to these criteria as ‘thresholds’ and focus on
14 the variability threshold. An ‘impact’ plume for a particular water quality metric is
15 defined as the total volume of groundwater exceeding the threshold. All other
16 factors being equal, impact plume volume will be inversely proportional to the
17 difference between the initial concentration of a solute in the aquifer and the
18 corresponding threshold value.

19 Twelve metrics were catalogued from the simulations at 22 discrete times
20 (1,5,10,20,30..200 years). Those metrics were: plume volumes for each of the nine
21 water quality metrics, the Δx and Δy footprint of the plume (for the pH water
22 quality metric only), and the CO₂ flow rate across the water table. For each time, the
23 instantaneous CO₂ and brine flow rate and the cumulative mass of CO₂ and brine
24 leaked was catalogued. In the training data set used to develop the response
25 surfaces, these four variables replaced the eight inputs for the heuristic leakage
26 scenario model.

27 The relationship between the size of the CO₂ plumes (dissolved and free
28 phase), the brine plumes, and the twelve water quality impact plumes was
29 surprisingly complex. For example, although even a small amount of brine can
30 cause relatively large water quality changes in a dilute drinking water aquifer such
31 as the Edwards Aquifer, we found that the TDS impact plumes were dominated by

CO₂ leaks, not brine leaks. This is because even slight pH depression (caused by CO₂ dissolution) causes sufficient calcite dissolution so that TDS increases beyond the 95% background variability threshold. In contrast, the trace metal and organic plumes are correlated with, but are smaller than, the brine plumes. This is because the only source of trace metals and organics considered was the brine itself. The scaling factor ROMs (Bacon et al., 2014; Bacon et al., 2015) account for sources of trace metals within the aquifer. The pH impact plumes co-vary with the dissolved CO₂ plumes.

2.2 Response Surface theory

We define $f(\theta)$ to be a computationally expensive simulation model with multiple input parameters θ and model outputs f . Our goal is to identify a response surface model $\hat{f}(\xi)$ that is a useful approximation to $f(\theta)$ but is very computationally efficient, where ξ is a subset of θ . MARS is a form of regression analysis that seeks to minimize the differences between f and $\hat{f}$ for a ‘training’ data set, created *a priori* by running the simulation model with N_o parameter sets. In this application, model outputs f are the simulation outputs (plume volumes) described above in Section 2.1.2. Model inputs θ are the parameters listed in Table 1.

MARS creates response surfaces (RS) by fitting splined surfaces together at ‘knots’. The models (Friedman, 1991) take the form

$$\hat{f}(\xi) = \sum_{i=1}^{N_t} c_i B_i(\xi)$$

where c_i is the i th coefficient, $B_i(\xi)$ is the i th basis function, and N_t is the number of terms included in the model. The algorithm avoids overfitting the training data by giving preference to models with smaller values of N_t .

The coefficients and basis functions are extracted from the training dataset to capture the nonlinearities and interactions between inputs. The basis functions describe the basic shape of the response surface, while the coefficients are fitted to

1 reduce mismatch between f and $\hat{f}$ for the training data set. Basis functions can be a
2 constant, a hinge function, or a combination of hinge functions, where a hinge
3 function can take the form

$$\max(0, \theta - h)$$

4 or

$$\max(0, h - \theta)$$

5 A knot is introduced at locations (h) where MARS identifies that a nonlinearity
6 will reduce RS mismatch. MARS uses a two-step process of a forward pass followed
7 by a backward pass to first determine where basis functions should be added and
8 subsequently determine which added basis functions can be removed or 'pruned'.

9 In the forward pass, mirrored basis function pairs are added to the model based
10 on their maximum reduction to the sum-of-squares residual error. The backward
11 pass then prunes basis functions based on the reduction of a generalized cross-
12 validation (GCV) defined as

$$GCV = \frac{\sum_{N_o} (f - \hat{f})^2}{N_o \left(1 - \frac{N_p}{N_o}\right)^2}$$

13 where N_p is the effective number of MARS parameters defined as

$$N_p = N_t + m \frac{(N_t - 1)}{2}$$

14

15 where m is a penalty term per knot. After the backward pass prunes the collection
16 of basis functions, a least squares regression is performed to further reduce the RS
17 mismatch by fitting the coefficients. Reduction of the GCV provides a tradeoff
18 between goodness-of-fit and model complexity. GR^2 is the metric commonly used to
19 assess the generalized performance of the MARS model, and is defined as

$$GR^2 = 1 - \frac{GCV}{GCV_0}$$

1 where GCV_0 is the GCV of the intercept-only MARS model (a single constant basis
2 function for each input parameter only, no hinge functions). The GR^2 in effect
3 normalizes the information in the GCV, and the strategy becomes to maximize the
4 GR^2 . We illustrate the MARS process used here in Figure 3, where f contains the
5 simulator (FEHM) responses for input parameter combinations θ .

6

7 **2.3 Simulation dataset and initial MARS implementation**

8

9 In principle the MARS method can analyze a dataset with a large number of
10 independent variables, eliminate unimportant variables, and build RSs using only
11 sensitive parameters. A large number of simulations would be required, however, if
12 many inputs were varied. It was difficult to predict how many total simulations
13 would be achievable, since some parameter combinations led to very long run times
14 (or incomplete runs). Because these simulations are so computationally expensive
15 and run-times can extend up to weeks or more, it was not practical to vary a large
16 number of variables simultaneously. We launched two relatively small sets of
17 simulations varying only a subset of parameters at a time. First, in an initial suite of
18 simulations (Caroll et al., 2014; Dai et al., 2014b), we varied the CO_2 and brine leak
19 rates, and five aquifer parameters all related to permeability: mean and variance,
20 horizontal-to-vertical anisotropy, and lateral correlation length. Aquifer thickness
21 and horizontal gradient were kept constant. In a second suite of simulations the
22 permeability variables were held at their mean values, while aquifer thickness,
23 horizontal gradient, and leak rates were varied (Bacon et al., 2014). This second set
24 of simulations produced a much larger range of plume volumes. Table 1
25 summarizes the range of input parameters for these combined sets of simulations.

26 Initially we applied MARS to the two simulation datasets combined. Matches
27 between the RS and the training dataset were unacceptably poor. However, useful
28 statistics about parameter importance were derived based on MARS basis function
29 selection/rejection criteria described in Section 2.2. For pH plume volumes, the
30 ranking of parameters was as follows: CO_2 mass, aquifer thickness, hydraulic

gradient, CO₂ flow rate, chloride concentration brine, permeability, anisotropy, permeability variance, and correlation length (unused). Given the higher importance of the parameters varied in the second simulation dataset (aquifer thickness, hydraulic gradient), we abandoned the first simulation dataset and applied MARS to the second one alone. This greatly improved MARS performance, but as a consequence we lost our ability to assess the influence of permeability-related variables to plume volume. We believe this limitation is not significant, given the relatively low importance of permeability-related variables in the combined dataset and the larger variability in simulation outputs in the second suite.

2.4 MARS final implementation

The final dataset used for building the response surface included 1766 sets of 7 inputs (aquifer thickness, background gradient, CO₂ flow rate, CO₂ mass, brine flow rate, brine mass, and brine chloride concentration) and 12 outputs. In this final aggregated set, the aquifer parameters retained a uniform distribution. For example, the cdf for aquifer thickness is shown in Figure 4a. The brine and CO₂ flow rates and cumulative mass, however, were heavily skewed toward the small values. One example, CO₂ flow rate, is shown in Figure 4b. This is due to several factors. One factor is that higher flow rates were more difficult for the physics-based simulator and so fewer successful runs were completed at high flow rates. The most important factor, however, is that most of the 200-year leakage scenarios considered included a long post-mitigation time during which leak rates were assumed to be at or near zero.

We specified the maximum degree of input parameter interaction to be 3, indicating that the RS development would consider terms with up to three hinge functions. Models were developed for 30, 50, 75, 100, and 125 MARS terms. RSs were developed for both untransformed (θ, f) data and log-transformed data ($\log(\theta), \log(f)$). Separate RSs were developed before and after the end of the leak.

1 This is because parameter sensitivities are different in these two time frames. For
2 example, before the leak ends, the plume volume is very sensitive to the mass of CO₂
3 injected; while afterwards this value will be constant. The post-leak RSs predict the
4 ratio of plume volume to that calculated at the time the leakage stops. Log-
5 transformed post-leak RSs were not attempted.

6 Twenty cross-validation runs were performed to assess the predictive capability
7 of each RS, where the model is built with 95% and evaluated on 5% of the samples.
8 The cross-validation step is separate from building the actual MARS model, which
9 uses the complete sample set. In general, the r^2 value calculated using the complete
10 training data set increased with the number of MARS terms. However, the cross-
11 validation r^2 values began to degrade as the number of terms increased. For each
12 model output, the 'optimal' number of MARS terms was selected based on a cross-
13 validation analysis.

14 **3 Results**

15 Results are summarized in Tables 3 and 4. Before the end of leakage (Table 3),
16 high values of r^2 (0.97 and 0.94) were obtained for the full training dataset for pH
17 and TDS plume volumes. Lower values, ranging from 0.78 - 0.88, were obtained for
18 the trace metal and organic plume volumes. Slightly lower r^2 values were obtained
19 after leakage ends (Table 4). Initial results indicated that log transformation of the
20 plume volumes improved RS performance, in some cases. However, cross-
21 validation r^2 values for log transformed RSs were often very poor. Close inspection
22 of the results indicated that log transformation generally produced poor results for
23 very small plume volumes. The selection of the appropriate RS, therefore, may be
24 dependent on the goal of the particular analysis. If accuracy is most important at
25 early times/small leak rates, an untransformed (linear) RS will be most appropriate.

26 The maximum number of terms allowed for each output is also shown in Tables
27 3 and 4. An example cross-validation analysis for the RS for pH during leakage is
28 presented in Figure 5. The selected model has 21 terms in order to maximize the
29 GRSq (indicated by vertical dashed line). The RSq, although obscured from view in

1 this example, continues to increase. While some of the CV runs (light red lines) have
2 significantly lower RSq values (0.3-0.7), the majority of CV runs are similar to the
3 RSq of the full model (> 0.9). This indicates that the full model RSq will reasonably
4 characterize the predictive capability of the full model. The mean CV RSq values is
5 around 0.9 (dark red line). The other CV analyses (not shown) are similar.

6 Figure 6 present syn and post leakage mitigation importance ranking of input
7 parameters for 9 outputs determined by the MARS algorithm during RS building.
8 For the syn-leakage RSs, both linear and logarithmic results are shown (all post-leak
9 RSs were linear). If an input parameter is not shown in Figure 6 for a given output,
10 it was deemed as unimportant by the RS build process and was automatically
11 eliminated from the RS. During the leak, almost all the outputs are dominated by
12 either the CO₂ flow rate (logarithmic) or cumulative mass of CO₂ (linear) entering
13 the aquifer. The dependence of TDS plume volume on CO₂ mass is due to pH-driven
14 calcite dissolution, not brine leakage. The second most important parameter for
15 most outputs is the thickness of the aquifer. The chloride concentration in the
16 brine, the background gradient in the aquifer generally had fairly low importance.
17 Importance rankings are quite different after the leak subsides. The background
18 hydrologic gradient is quite important for several outputs, as is the chloride content
19 of the brine. Other relatively important parameters are the cumulative CO₂ mass
20 leaked, the size of the plume at the time of leak cessation, and the length of time
21 since the leak stopped.

22 Although the MARS models fit the total population of simulation data very
23 well for some model outputs as demonstrated through high r^2 (Table 4), it is not
24 clear whether it would reproduce the time-series from any individual simulations
25 adequately. Comparisons are shown for three pH plume volume simulations in
26 Figure 7. The CO₂ flux rates imposed are shown in the upper part of the figure.
27 There are interesting differences between these simulation results. The magnitude
28 of the maximum CO₂ flow rate increases from the first to second to third example.
29 Likewise, the maximum simulated plume volumes increase. The RSs capture the
30 variation in plume volumes reasonably well, although in the first simulation the RS
31 underestimates the maximum plume volume by ~50%. The response of the plume

1 to leak cessation differs significantly between the simulations. In the first scenario,
2 the plume continues to grow long after the CO₂ leak subsides. In the second
3 scenario, the plume begins to shrink as soon as the CO₂ leak subsides. In the third
4 scenario, the plume is still growing at the end of the simulation (200 years). The
5 MARS models reproduce the amount of plume recovery fairly well, but do not
6 reproduce the timing of the recovery very well. The root cause of this is
7 demonstrated in Figure 7. The MARS model is tasked with capturing the total
8 variation in all the plume volumes, rather than coherent variation within any single
9 simulation (shown as colored dots).

10 In Figure 8 we present the cumulative distribution function for the delay
11 between maximum CO₂ leak rate and maximum plume size, over all the simulations
12 (solid line). This figure illustrates that approximately half the simulations show an
13 immediate plume response to leakage cessation (delay = 0). The MARS models
14 show a very similar distribution. This suggests that the MARS models are not able
15 to predict the delay for a specific scenario accurately, but the models are able to
16 predict the character of an ensemble of scenarios quite well.

17 Two additional examples of ensemble-based MARS performance are shown
18 in Figures 9 and 10. In Figure 9, the maximum plume volume for each simulation is
19 compared against the MARS prediction. The correspondence is very good. In Figure
20 10 we present cumulative distribution functions of plume volumes for four
21 categories of leaks. To create these curves, all 1766 sets were binned into classes
22 according to the cumulative mass of CO₂ leaked. The MARS models reproduce the
23 differences between these classes very well.

24 **4 Conclusions**

25
26 The transport of CO₂ and brine in a shallow aquifer is a complex, non-linear
27 process; predicting plume evolution will require solution of coupled non-linear
28 multi-component reactive-transport equations. We have developed sets of
29 simulations that predict 'impact plumes' for nine distinct water quality metrics, for a

1 range of leakage scenarios and aquifer characteristics (Dai et al., 2014b; Bacon et al.,
2 2014). Response surfaces have been trained on simulation input/output datasets,
3 and performance varies amongst the water quality metrics. Performance is
4 reasonably good for pH and TDS, and less satisfactory for trace metals and organics.
5 We believe several factors, including the non-linearity of the problem, complexity of
6 the geochemistry, and granularity in the simulation results, contribute to this
7 uneven performance.

8   The Multivariate Adaptive Regression Spline method, as implemented in the
9 R (R Core Team, 2014) Earth package (Millborrow 2015), provides control over the
10 balance between fidelity to the training data set and predictive performance with
11 datasets not used in the RS build. For the groundwater impact simulation dataset, a
12 reasonable balance between fidelity and predictive performance was achieved for
13 most simulation outputs. Useful parameter importance information was also
14 obtained. During a CO₂ leak, almost all the outputs are dominated by the cumulative
15 mass of CO₂ entering the aquifer. The second most important parameter for most
16 outputs is the thickness of the aquifer. The chloride concentration in the brine, the
17 background gradient in the aquifer generally had fairly low importance. Importance
18 rankings are quite different after the leak subsides (post mitigation). The
19 background hydrologic gradient is quite important for several outputs, as is the
20 chloride content of the brine. Other relatively important parameters are the
21 cumulative CO₂ mass leaked, the size of the plume at the time of leak cessation, and
22 the length of time since the leak stopped.

23   As would be expected for any regression-based method, the RS explains the
24 total variability in the simulation output reasonably well, but does not reproduce
25 the time series derived from any single simulation particularly well. It is possible
26 that other RS methodologies might perform better than MARS, and this is worthy of
27 future research. However, we have shown that the RSs derived in this study, when
28 used in a probabilistic context, can reproduce key statistics from the simulation
29 dataset very well. These include statistics related to the time delay of plume
30 recovery after leakage cessation, the size of plumes for each of three classes of leaks
31 (1-5, 5-10 , and 10-50 KTon), and the maximum plume volume realized for each

leakage scenario. If however, a deterministic prediction of plume growth for a single leakage scenario is desired, physics-based simulations should be used – not these derived response surfaces. Since the intended use of the RSs here is risk assessment, which is inherently a probabilistic analysis, the response surface methods should prove to be quite useful.

5 Acknowledgements

This research was conducted as part of the Department of Energy's National Risk Assessment Partnership (NRAP). Support for this project came from the DOE Office of Fossil Energy's Crosscutting Research program. We appreciate the very helpful contributions of two anonymous reviewers.

6 References

- Bacon, D.H., Keating E.H., Dai, Z., Harp D.H., Pawar, R.J. , 2014. Reduced Order Model for the Impacts of Carbon Dioxide and Brine Leakage into an Unconfined, Oxidizing Carbonate Aquifer, Generation 3.0. April 2014. PNNL-23297
- Bacon, D.H., Qafoku, N.P., Dai, Z., Keating, E.H., Brown, C.F., 2015. Modeling the impact of carbon dioxide leakage into an unconfined, oxidizing carbonate aquifer. International Journal of Greenhouse Gas Control. <http://dx.doi.org/10.1016/j.ijggc.2015.04.008>
- Behzadian K., Kapelan Z, Savic D, Ardeshtir A., 2009. Stochastic sampling design using a multi-objective genetic algorithm and adaptive neural networks. Environ Model Software, 24 (4): 530–541
- Boyle, J.S., Klein A., Lucas D.D., H.-Y. Ma H-Y, Tannahill J, Xie S. 2015. The parametric sensitivity of CAM5's MJO. Journal Of Geophysical Research: Atmospheres. 120(4):1424–1444,
- Carroll, S., Keating, E.H., Mansoor, K., Dai, Z., Sun, Y., Trainor-Guitton, W., Brown, C., 2014. Key Factors for Determining Groundwater Impacts Due to Leakage from Geologic Carbon Sequestration Reservoirs. International Journal of Greenhouse Gas Control, 29:153-168.
- Chen, M., Tompson, A.F.B., Mellors, R.J., Ramirez A.L., Dyer, K.M., Yang, X., Wagoner, J.L. 2012. An efficient Bayesian inversion of a geothermal prospect using a multivariate adaptive regression spline method. Appl Energy;136:619-27,
- Chen, M., Sun, Y., Fu, P., Carrigan, C.R., Lu, Z., Tong, C.H., Buscheck, T.A. 2013. Surrogate-based optimization of hydraulic fracturing in pre-existing fracture networks. Comput Geosci 2013;58:69-79,

- 1 Chen, M., Tompson, A.F., Mellors, R.J., Abdalla, 2015. An efficient optimization of
2 well placement and control for a geothermal prospect under geological
3 uncertainty. *Applied Energy*; 137: 352-363,
- 4 Dai, Z., Stauffer, P.H., Carey, J.W., Middleton, R.S., Lu, Z., Jacobs, J.F., Hnottavange-
5 Telleen, K., Spangle, L., 2014. Pre-site characterization risk analysis for
6 commercial-scale carbon sequestration. *Environ. Sci. Technol.*, 48,
7 3908–3915.
- 8 Dai, Z., Keating, E.H., Bacon, D.H., Viswanathan, H., Stauffer, P., Jordan, A., Pawar, R.,
9 2014b. Probabilistic evaluation of shallow groundwater resources at a
10 hypothetical carbon sequestration site. *Scientific Reports*, 4,
11 doi:10.1038/srep04006
- 12 Dowla, FU, Rogers LL. 2003. Solving problems in environmental engineering and
13 geosciences with artificial neural networks MIT Press, Cambridge, MA.
- 14 Fen, CS, Chan CC, Cheng HC. 2009. Assessing a response surface-based optimization
15 approach for soil vapor extraction system design. *J Water Resour Plan*
16 *Manage*, 135 (3): 198–207
- 17 Friedman, J. H., 1991. Multivariate adaptive regression splines. *Ann. Stat.* 19(1), 1–
18 67.
- 19 Jin, R., Chen, W., Simpson, T.W., 2001. Comparative studies of meta-modelling
20 techniques under multiple modelling criteria. *Struct. Multidiscip. Optim.*,
21 23(1), 1–13.
- 22 Jordan, A.B., Stauffer, P.H., Harp, D., Carey, J.W., Pawar, R., 2015. A response surface
23 model to predict CO₂ and brine leakage along cemented wellbores.
24 *International Journal of Greenhouse Gas Control*. 33, 27-39
- 25 Keating, E. H., Doherty, J., Vrugt, J.A., Kang, Q., 2010. Optimization and uncertainty
26 assessment of strongly nonlinear groundwater models with high parameter
27 dimensionality. *Water Resour. Res.*, 46, W10517,
28 doi:10.1029/2009WR008584
- 29 Kourakos, G., Mantoglou, A., 2009. Pumping optimization of coastal aquifers based
30 on evolutionary algorithms and surrogate modular neural network models.
31 *Adv. Water Resour.*, 32(4), 507–521.
- 32 Last, G.V., Murray, C.J., Brown, C.F., Jordan, P.D., Sharma, M., 2013. No-Impact
33 Threshold Values for NRAP's Reduced Order Models. PNNL-22077, Richland,
34 Washington, p.72. <[http://www.pnnl.gov/main/publications/
35 external/technical_reports/PNNL-22077pdf](http://www.pnnl.gov/main/publications/external/technical_reports/PNNL-22077pdf)>.
- 36 Lindgren, R. J.; Dutton, A. R.; Hovorka, S. D.; Worthington, S. R.; Painter, S.
37 Conceptualization and Simulation of the Edwards Aquifer, San Antonio Region,
38 Texas; Scientific Investigation Report 2004-5277; U.S. Geological Survey,
39 2004; p 143.
- 40 Lucas, D.D. Klein R., Tannahill J., Ivanova D, Brandon S., Domyancic D., and Zhang Y.
41 2013. Failure analysis of parameter-induced simulation crashes in climate
42 models. *Geosci. Model Dev.*, 6, 1157–1171.
- 43 Millborrow, Steve. 2015. Package 'earth', Multivariate Adaptive Regression Splines.
44 <http://CRAN.R-project.org/package=earth>

- 1 Moore, C., Doherty, J. , 2005. Role of the calibration process in reducing model
2 predictive error. Water Resour. Res., 41, W05020, doi:10.1029/
3 2004WR003501.
- 4 Pawar, R.; Bromhal, G.; Dilmore, R.; Foxall, B.; Jones, E.; Oldenburg, C.; Stauffer, P.;
5 Unwin, S.; Guthrie, G., 2013. Quantification of Risk Profiles for Atmospheres
6 and Groundwater; NRAP-TRS-III-003-2013. NRAP Technical Report Series;
7 U.S. Department of Energy, National Energy Technology Laboratory:
8 Morgantown, WV, 2013; p 28.
- 9 R Core Team. 2014. R: A language and environment for statistical computing. R
10 Foundation for Statistical Computing, Vienna, Austria. URL [http://www.R-](http://www.R-project.org/)
11 [project.org/](http://www.R-project.org/).
- 12 Razavi, Saman, Tolson, B.A., Burn D.H. 2012. Review of surrogate modeling in water
13 resources. Water Resources Research. 48, W07401, doi:10.1029/
14 2011WR011527.
- 15 Regis, RG, Shoemaker CA. 2007. A stochastic radial basis function method for the
16 global optimization of expensive functions INFORMS J Comput, 19 (4):497–
17 509
- 18 Sharda , V.N., Prasher, S.O. , Patel, R.M. , Ojasvi, P.R., Prakash, C., 2008. Performance
19 of Multivariate Adaptive Regression Splines (MARS) in predicting runoff in
20 mid-Himalayan micro-watersheds with limited data. Hydrological Sciences
21 Journal, 53:6, 1165-1175, DOI: 10.1623/hysj.53.6.1165
- 22 Simpson, T.W.; Peplinski, J.; Koch, P.N.; Allen, J.K. ,1997. On the use of statistics in
23 design and the implications for deterministic computer experiments. Design
24 Theory and Methodology – DTM’97 (held in Sacramento, CA), Paper No.
25 DETC97/DTM-3881. ASME
- 26 Simpson, TW, Mistree F. 2001. Kriging models for global approximation in
27 simulation-based multidisciplinary design optimization. AIAA J, 39 (12):
28 2233–2241
- 29 Viswanathan, H.S., Pawar,R.J., Stauffer, P.H., Kaszuba, J.P., Carey, J.W., Olsen, S.C.,
30 Keating, G.N., Kavetski, D., Guthrie, G.D. , 2008. Development of a hybrid
31 process and system model for the assessment of wellbore leakage at a
32 geologic CO₂ sequestration site. Environ. Sci. Technol., 42, pp. 7280-7286
33 <http://dx.doi.org/10.1021/Es800417x>
- 34 Zeng, L. Shi, D. Zhang, L. Wu. 2012. A sparse grid based Bayesian method for
35 contaminant source identification. Adv Water Resour, 37:1–9
- 36 Zhang, XS, Srinivasan, R. ,Van Liew M. 2009. Approximating SWAT model using
37 artificial neural network and support vector machine. J Am Water Resour
38 Assoc, 45 (2) (2009), pp. 460–474
- 39
- 40 Zyvoloski, G.A., Robinson, B. A., Dash, Z. V., Trease, L. L. 2011. Summary of the
41 Models and Methods for the (FEHM) Application --- A Finite-Element Heat-
42 and Mass-Transfer Code. LA-13306-MS; Los Alamos National Laboratory, Los
43 Alamos, NM, 2011.
- 44

45 List of Tables

Table 1. Parameters varied in development of simulation dataset

Table 2. Reactions included in simulations

Table 3. R-squared values for each ROM, before the end of leakage

Table 4. R-squared values for each ROM, after leakage

List of Figures

Figure 1. Dissolved CO₂ Mass fraction at time= 40 years. pH plume size at this time is 1.4E8 m³.

Figure 2. Input parameters for the generalized wellbore leak model

Figure 3. Flow diagram for MARS algorithm where f and θ are training outputs and inputs, respectively, B are basis functions, and c are the coefficients. Based on Figure 1 in <http://www.milbo.org/doc/earth-notes.pdf> by Stephen Milborrow with permission.

Figure 4. Cumulative distribution functions for two model input variables: aquifer thickness and CO₂ flow rate.

Figure 5. Cross-validation analysis of MARS RS building algorithm. The RSq line is obscured from view in this example behind the GRSq line.

Figure 6. Post-leak importance ranking of input parameters determined by the MARS RS building algorithm. a) linear model, syn-leak b) logarithmic model, syn-leak, c) post-leak

Figure 7. Three leakage scenarios. Upper figures show time-varying CO₂ leak rate (g/s). Lower figures show simulation output (blue) and ROM output (red).

Figure 8. Simulated versus ROM computed pH plume volume (m³). Colored dots are the three scenarios from Figure 7.

Figure 9. Cumulative probability distributions for time delay between peak leak rate and peak plume volume. Solid line = simulations; dash line = RS.

Figure 10. Maximum pH plume volume for each leakage scenario

Figure 11. Cumulative probability distributions for pH plume volumes for four sizes of CO₂ leaks: a) 1-5 Kton, b) 5-10 Kton, c) 10-50 Kton, and d) 200-1500 Kton. Upper frame represents full simulation results; lower frame represents ROM results.

Table 1. Parameters varied in development of simulation dataset

<i>Generalized wellbore leak model (see Figure 1)</i>						
Parameter	Description	Units	CO ₂		Brine	
			Min	Max	Min	Max

$Q_{\max}$	Maximum leak rate	kg/s	0.001	0.5	0.005	0.075
λ	Ratio of leakage tail to $Q_{\max}$	-	0	0	0.2	0.3
Dt1	Time to reach $Q_{\max}$	years	5	50	0	0
Dt2	Length of time at $Q_{\max}$	years	0	100	1	50
Dt3	Length of time for leakage to end	years	5	50	1	10
Tm	Mitigation time	years	0.5	200	0.5	200

<i>Aquifer parameters</i>				
	Horizontal hydraulic gradient	0.0002878	0.0188950	-
	Aquifer thickness	100	500	m
	Mean permeability	-14	-10.6	(log m ²)
	Permeability variance	0.017	0.79	
	Permeability correlation length	1	3.95	km
	Anisotropy (K_{xy}/K_x)	1.1	49.1	-
<i>Brine chemistry</i>				
	Chloride concentration	.53	5.98	Moles/L

Table 2: Aqueous and solid phase carbonate reactions

Carbonate aqueous equilibrium reactions:	logK
$\text{HCO}_3^- + \text{H}^+ = \text{H}_2\text{CO}_3$	-6.3
$\text{CO}_3^{2-} + 2\text{H}^+ = \text{H}_2\text{CO}_3$	-16.6
$\text{OH}^- + \text{H}^+ = \text{H}_2\text{O}$	-14
Calcite kinetic dissolution or precipitation:	
$\text{CaCO}_3(\text{s}) + 2\text{H}^+ = \text{H}_2\text{CO}_3 + \text{Ca}^{2+}$	8.36

Table 3. R-squared values for each ROM, before the end of leakage

Output	Linear		Log-transformed		# of terms
	Training data	Cross validation	Training data	Cross validation	

dx	0.86	0.83	0.82	0.81	30
dy	0.90	0.80	0.81	0.79	50
flux	0.94	0.78	0.92	-0.78	100
pH	0.97	0.89	0.70	-141.17	30
Tds	0.96	0.53	0.95	0.94	125
As	0.78	0.60	0.80	0.69	100
Pb	0.80	0.63	0.86	0.83	75
Cd	0.81	0.28	0.80	0.10	125
Ba	0.82	0.70	0.92	0.89	100
Benzene	0.88	0.74	0.90	0.87	100
Naphthalene	0.88	0.58	0.75	0.57	100
Phenol	0.84	0.39	0.92	0.90	75

1 Table 4. R-squared values for each ROM, after leakage

Output	Training data	Cross validation	# of terms
dx	0.77	0.71	125
dy	0.77	0.65	125
flux			
pH	0.82	0.72	125
Tds	0.81	0.75	100
As	0.87	0.78	125
Pb	0.87	0.78	100
Cd	0.85	0.76	100
Ba	0.87	0.79	100
Benzene	0.86	0.80	125
Naphthalene	0.83	0.75	125
Phenol	0.87	0.80	100

2
3
4
5
6
7

Figure 1. Input parameters for the generalized wellbore leak model

Figure 2. Heuristic leakage scenario (a) and resulting dissolved CO2 plume at time= 40 years (b). pH plume size at this time is $1.4E8 \text{ m}^3$. The portion of the grid $y < 1000$ is removed from view so the plume is visible.

Figure 3. Flow diagram for MARS algorithm where f and θ are training outputs and inputs, respectively, B are basis functions, and c are the coefficients Based on Figure 1 in <http://www.milbo.org/doc/earth-notes.pdf> by Stephen Milborrow with permission.

Figure 4. Cumulative distribution functions for two model input variables: aquifer thickness and CO₂ flow rate.

Figure 5. Cross-validation analysis of MARS RS building algorithm. The RSq line is obscured from view in this example behind the GRSq line.

Figure 6. Post-leak importance ranking of input parameters determined by the MARS RS building algorithm. a) linear model, syn-leak b) logarithmic model, syn-leak, c) post-leak

Figure 7. Three leakage scenarios. Upper figures show time-varying CO₂ leak rate (g/s). Lower figures show simulation output (blue) and ROM output (red).

Figure 8. Simulated versus ROM computed pH plume volume (m³). Colored dots are the three scenarios from Figure 6.

Figure 9. Cumulative probability distributions for time delay between peak leak rate and peak plume volume. Solid line = simulations; dash line = RS.

Figure 10. Maximum pH plume volume for each leakage scenario

Figure 11. Cumulative probability distributions for pH plume volumes for four sizes of CO₂ leaks: a) 1-5 Kton, b) 5-10 Kton, c) 10-50 Kton, and d) 200-1500 Kton. Upper frame represents full simulation results; lower frame represents ROM results.