Post-calibration Uncertainty Analysis for Travel Times at the Naval Weapons Industrial Reserve Plant

Bulbul Ahmmed[1], Scott C. James *[2], and Joe Yelderman[1]

[1]Baylor University, Department of Geosciences
[2]Baylor University, Departments of Geosciences and Mechanical Engineering

Article Impact Statement: Necessity of parameter estimation and uncertainty analyses are demonstrated for a contaminated site.

Abstract

The Naval Industrial Reserve Plant (NWIRP) in McGregor, Texas began manufacturing explosives in 1980 and several hazardous chemicals were discovered in lakes and streams surrounding the plant in 1998. This research demonstrates the importance of using a numerical study to support remedial decision making by investigating post-calibration linear and Null-space Monte Carlo (NSMC) nonlinear uncertainty analyses. Based on MODFLOW and MODPATH models, which simulated hydraulic heads and tracer travel times at the site, the following measures were quantified: parameter uncertainties, parameter identifiabilities, observation worth, and predictive uncertainties. Parameter uncertainties were reduced by up to 92%; a total of 19 parameters were at least moderately identifiable (>10%); travel-time uncertainties were reduced up to 92%. An observations-worth analysis found that additional data (11 more measurements) could reduce travel-time uncertainties by factors from 1.04 to 4.3 over existing data if collected at targeted locations. Finally, travel-time predictions and post-calibration parameter distributions were generated using the NSMC technique. NSMC predicted that conservative tracers exited the flow system within a year, which agrees well with field data.

Key words: parameter estimation, uncertainty analysis, forecasting, numerical modeling, Monte Carlo

Introduction

The Naval Weapons Industrial Reserve Plant (NWIRP) occupies about 40 km^2 in southwest McGregor, Texas on a topographic divide underlain by a shallow groundwater system within

*Corresponding author: sc_james@baylor.edu

fractured limestone bedrock (Figure 1). The NWIRP began manufacturing explosives in 1980 (Hare, 2000; Moore and McSpadden, 2009) and several hazardous chemicals including ammonium perchlorate were discovered in lakes and streams surrounding the plant in 1998 (Ensafe Inc., 1999; Craig and Burdick, 2007) defying estimates from groundwater velocities that suggested the contamination should not have entered streams and migrated offsite (Clark, 2000). However, those estimates did not consider the increased fluxes and hydraulic heads affecting groundwater flow velocities during storm periods. Ensafe Inc. (1999) estimated groundwater flux in Georgetown Limestone at 2 m/year using average gradients, hydraulic conductivity from slug-test data, and porosity while assuming homogeneous and steady groundwater velocities throughout the area. But groundwater in Georgetown Limestone also flows through more conductive features when the water table rises during storm periods. Such recharge also increases groundwater velocity. The Georgetown Limestone, similar to other fractured carbonates like the Austin Chalk, exists as an upper, highly fractured and unsaturated zone overlying a low-permeability, moderately fractured zone (Barquest, 1989; Bingham, 1993; Mace, 1998; Ashworth and Hopkins, 1995; Chowdhury et al., 2010). The highly fractured upper layer facilitates fast fluid flow during storms. Although no study was performed on variance of flux but an inference can be drawn from Figure 2 about the variability of the flow rate. As shown in Figure 2, the hydrograph from an unnamed spring, the mean flow rate from May through September, 1999 was 3.3 L/s with a spike to 18.7 L/s in July. The sharp response of the flow rate to the recharge event reflects the water table entering the upper, highly fractured layer, which was then quickly drained. Although the shallow groundwater is not used locally, the streams flow into two major water supply reservoirs in the region (Lake Belton and Lake Waco).

Recharge from precipitation significantly increases lateral flow through the Georgetown Limestone when the water table rises into the upper, highly fractured zone. The water table is sensitive to recharge (storms) as shown in Figure 3. A rising water table can mobilize dissolved perchlorate such that it enters the upper zone where it is more easily transported off site. Even though some dissolved perchlorate is transported off site, the source persists as residual perchlorate in the lower fractured zone awaiting remobilization during the next storm (see Figure 4).

Tracer (or contaminant) transport times from the NWIRP to surrounding streams, rivers, and lakes are the primary concern for stakeholders. Clark (2000) built a piecewise-homogeneous MODFLOW model of the NWIRP, which was updated here to a pilot-point-based, heterogeneous MODFLOW model to predict flow fields while MODPATH estimated tracer transport times. MODFLOW is a groundwater modeling tools developed by United States Geological Society and MODPATH (Pollock, 2016)is a particle-tracking code that uses boundary conditions, hydraulic heads, and parameters from MODFLOW to simulate flow streamlines. The conceptual model included the upper and lower fractured layers, topography, and heterogeneous hydraulic conductivities. The predictions of interest were the times for tracers from NWIRP administrative areas to exit the model domain as discharge to streams. Predictions were conditioned through calibration against measured water levels.

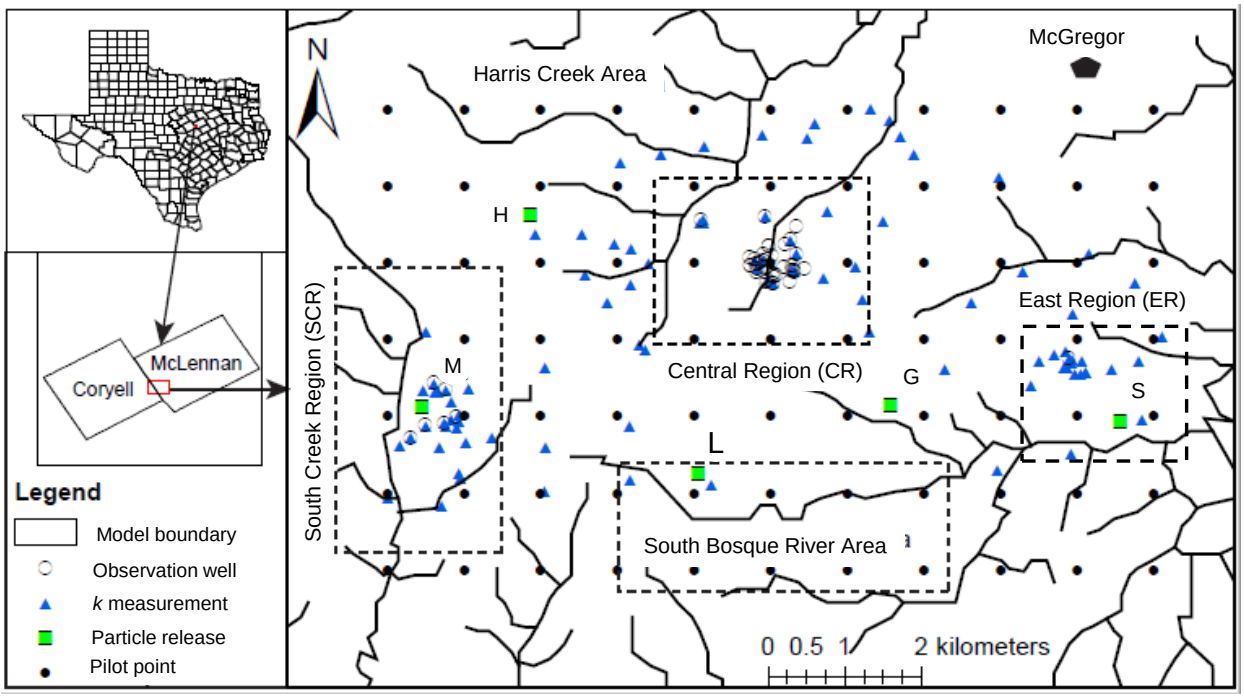

Figure 1: Study location including streams, 43 observation wells (open circles), 99 hydraulic conductivity measurements (blue triangles), 77 pilot points (filled circles), and particle-release locations (green squares) with associated NWIRP administrative designations G, H, L, M, and S.

Uncertainties in these predictions were quantified and the most important parameters and observations identified. Moreover, the Null-space Monte Carlo (NSMC) nonlinear uncertainty analysis technique was used to generate probability distributions of calibrated parameters and commensurate travel-time predictions (Tonkin et al., 2007; Tonkin and Doherty, 2009; Moore, 2006). Briefly, the NSMC technique uses subspace approaches like singular value decomposition (Moore and Doherty, 2005) to identify only those model parameters informed by the observation dataset (Moore and Doherty, 2006). This facilitates inversion of over-parametrized models by only calibrating those variables about which the dataset has information while relegating the rest to their user-preferred initial guesses through Tikhonov regularization (Tonkin and Doherty, 2005).

Theoretical Background

The level of parametrization of an environmental model should be commensurate with the quality and quantity of data used in its calibration to ensure confidence in the range of predictive possibilities (Moore and Doherty, 2006; Hunt et al., 2007). Calibration is constrained by the information content of the calibration data set (plus expert judgment) and linear predictive uncertainty can be assessed even before a calibration exercise. Parameter

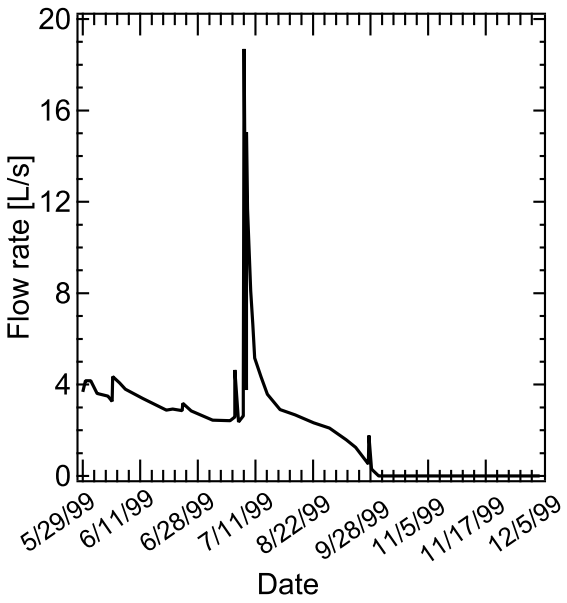

Figure 2: Flow hydrograph from an unnamed spring in the study area near McGregor, Texas with no contribution from surface runoff (Clark, 2000, Fig. 2). Flow was measured with a wire and a pressure transducer data logger. The sharp peak in discharge indicates an increase in saturated thickness and gradient due to recharge and water rising into the upper, more conductive layer.

uncertainty and identifiability along with observation worth can be quantified (James et al., 2009; Doherty, 2016). Post-calibration, the NSMC method facilitates a nonlinear assessment of parameter and prediction uncertainties (Doherty, 2016), but even with the use of super parameters (linear combinations of estimable parameters) to reduce the number of model calls, the approach can still be computationally intensive.

Predictive uncertainty analyses can be undertaken with a calibrated model using methods based on the propagation of variance (Doherty, 2016), which acknowledges that historic observation can be replicated with many non-unique parameter combinations. Predictive uncertainty reduction is calculated based on the pre- and post-calibration parameter uncertainties, where pre-calibration parameter variances (uncertainties) are specified according to measurements and expert judgment while post-calibration uncertainties are revealed through the calibration process. In addition, each parameter contributes to uncertainty in model predictions. Reduction in predictive uncertainty is predicated upon enhanced knowledge of the parameter space. Parameter uncertainty can be divided into solution- and null-space components. Solution-space uncertainty, usually the smaller of the two, is due to uncertainty in the calibration data (i.e., measurement error). Null-space uncertainty is due to shortcomings in the data or model that preclude precise identification of the parameter (i.e., many parameter combinations can calibrate the model about equally well). The mathematical process of distinguishing solution- from null-space uncertainty is achieved through singular

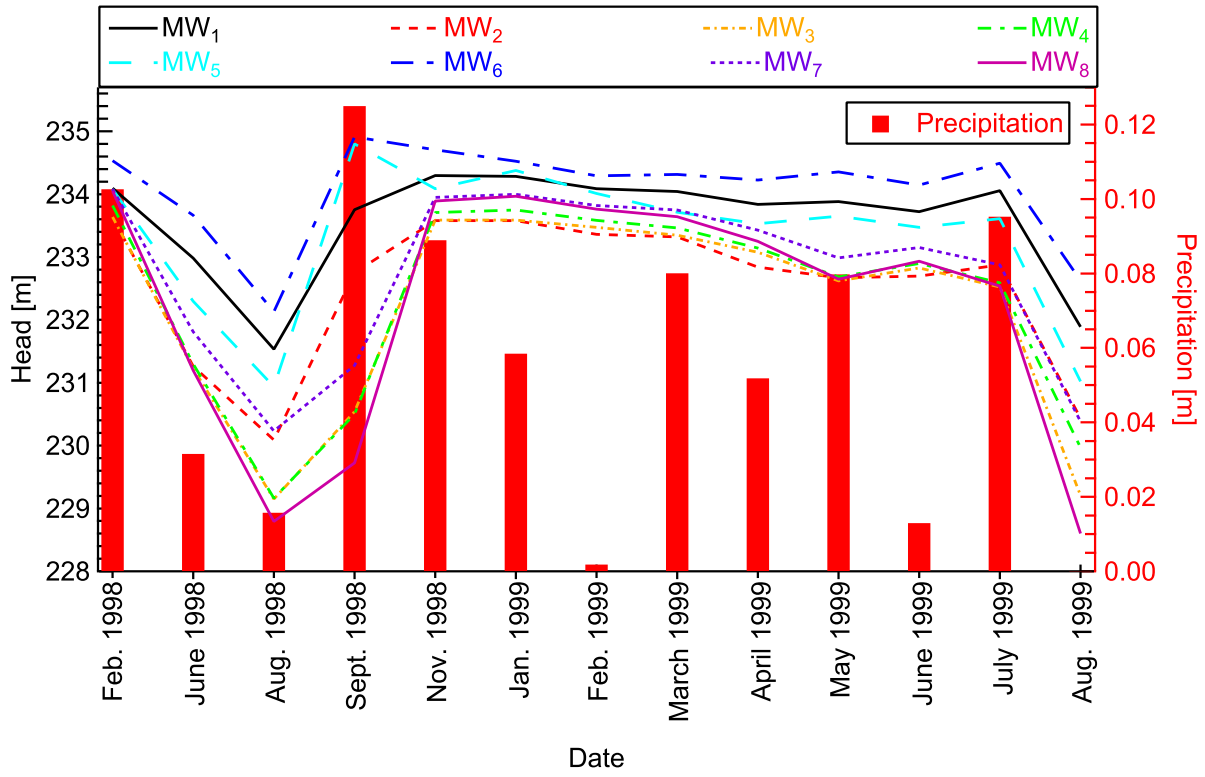

Figure 3: Hydraulic heads in eight observation wells near administrative area M demonstrating the correlation with precipitation. Precipitation data: Waco Regional Airport (2016).

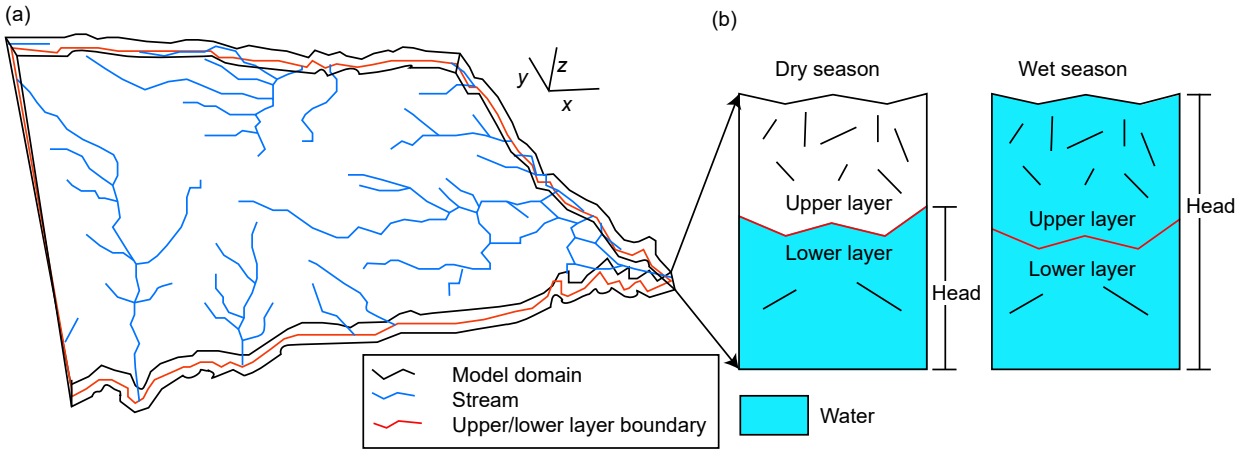

Figure 4: (a) Schematics of the model domain with undulating layers and streams. (b) Cross sections during dry seasons, contaminants travel slower in the less-fractured, lower layer and during the wet season, they travel faster in the more fractured, upper layer.

100 value decomposition (SVD), which is conducted with straightforward mathematical vector

and matrix manipulations (James et al., 2009; Doherty, 2016).

Measurement errors (observation noise) can never be eliminated and these impact predictive uncertainties. The calibration process minimizes the weighted-sum-of-squares differences, the objective function, between site observations and their corresponding model predictions. Both quantitative (observation noise, measurement accuracy, number of measurements comprising an observation, etc.) and qualitative (expert judgment) metrics should be used to specify weights in the objective function.

Identifiability is a metric indicating the calibration data's ability to constrain a model parameter (Doherty and Hunt, 2009). Quantitatively, it is the direction cosine between a parameter and its projection onto solution-space uncertainty. Identifiability can be used in both model design and implementation to assess whether a model needs more calibration data to reduce parameter uncertainty while also quantifying the uncertainties in predictions that depend specifically upon that parameter.

Observation worth is quantified based on the reduction in uncertainty in a parameter or prediction that is accrued through the acquisition of that datum (Lotti and Doherty, 2016). Reduction in these uncertainties below their pre-calibration level is a measure of the worth of an observation (or observation group) with respect to that parameter or prediction.

The NSMC technique generates multiple, unique parameter fields that satisfy both the model-to-measurement misfit (i.e., a sufficiently low objective function) and parameter-reality constraints (i. e., parameters cannot be assigned unrealistic values) and it quantifies post-calibration parameter and prediction uncertainties (Welter et al., 2015). It generates a suite of equally likely and realistic parameter fields that are used to make predictions. Generating parameter fields involves three steps: (1) generating random parameter fields according to pre-calibration uncertainty, (2) perturbing pre-calibrated parameters by adding null-space uncertainty, and (3) a brief model calibration (usually three optimizations) using fewer parameters or (super parameters), which are linear combinations of those parameters that have their pre-calibration uncertainty reduced by more than an insignificant amount. Uncertainty in a prediction can thereby be assessed through construction of an empirical probability density function (PDF) assembled by running the model using each NSMC parameter field realization to generate PDFs of predictions.

Methods

Conceptual Model

The conceptual model was built using four digital elevation maps (DEMs) from the Texas Natural Resources Information System website (TNRIS, 1999) to create the model topography. GIS (ESRI, 1996) and SURFER (Golden Software, 1997) were used to mosaic and grid the DEM data (Clark, 2000). The conceptual model also included local rivers, streams, creeks, and spring as shown in Figure 4 (a). The model comprised two 4-m-thick layers representing the upper, weathered layer and the lower, less-permeable limestone. The NWIRP

model domain and terrain-following layers were adjusted according to the topographic elevation (see Figure 4 (a)). The top of the upper model layer corresponded to the water level, which was initially assumed 2 m below the land surface. The bottom of the model was established uniformly 10 m below surface (Gburek et al., 1999; Verma et al., 1980). Average precipitation from 1960 to 2015 was 2.7×10^{-8} m/s. A base-flow study conducted in a nearby similar geologic setting indicated that 7% of total precipitation infiltrated to the aquifer (Cannata, 1988; Myrick, 1989), so recharge was 1.8×10^{-9} m/s.

Model Development

Because the rock is densely fractured (Figure 5), it was approximated as an equivalent porous medium. Therefore, this effort started by developing MODFLOW and MODPATH models of the NWIRP site (Figure 1), calibrating to available data, assessing observation worth and parameter identifiability, as well as quantifying uncertainties in parameters and tracer travel-time predictions. Next, results from the NSMC approach were used to generate PDFs of pilot points (hydraulic conductivities) and travel times. The model had two layers, 126 columns, and 97 rows with $100 \times 100\,\mathrm{m}^2$ cells. Recharge through precipitation was specified at the top of the model while the lateral sides along with the bottom of the model were specified as no-flow boundaries. Because all streams in the domain are gaining reaches, there was no need to specify streambed properties distinct from the hydraulic conductivities. No-flow boundaries can affect results near the edges of the model, so the NWIRP regions of interest were always at least 1.5 km from the model edges. The model was executed with MODFLOW-NWT (Niswonger et al., 2011), which admits the drying and re-wetting nonlinearities of an unconfined aquifer (Hughes et al., 2012; Niswonger et al., 2011). Single particles were released at the midpoint of each layer from the five administrative areas and tracked until they exited the model domain at streams.

Parameters

Groundwater flow is fundamentally governed by the distribution of hydraulic conductivities. In conventional calibration methods, property uniformity or pilot-point distributions are used as the basis for spatial parameter distribution (Doherty, 2003). In the absence of data, piecewise-homogeneous zones are often specified. If geologic zones are not piecewise-uniform, pilot points are distributed throughout such zones. Pilot-point property values were estimated during this calibration exercise and the hydraulic conductivities at model cells were assigned according to a kriging algorithm (Doherty, 2003). Pilot points facilitate a smooth but realistic distribution of hydraulic properties over a geologic unit, which cannot be achieved using piecewise-uniform methods. The upper model layer had only a single hydraulic conductivity measurement, but one parameter over such a large region would give false confidence in the solution because it would yield unrealistic homogeneity (Parker, 1977). Instead, a total of 77 pilot points (Figure 1) were used in each layer such that hydraulic conductivity fields were developed with heterogeneity commensurate with the information

Figure 5: Picture of the fractured formation in the study area (scale is $0.5\,\mathrm{m}^2$).

available in the observation data set. The initial value of the horizontal hydraulic conductivity in the upper layer, k_u, was $3.048 \times 10^{-3}\,\mathrm{m/s}$ for all 77 pilot points with a 1.5% porosity (Cannata, 1988). Just like the upper layer, a pilot-point-based heterogeneous distribution of hydraulic conductivities were also assigned to the lower layer. Ensafe Inc. (1999) conducted 99 slug tests measuring horizontal hydraulic conductivities in the lower layer, k_l, ranging from 10^{-8} to $10^{-4}\,\mathrm{m/s}$ with mean $10^{-7}\,\mathrm{m/s}$. These hydraulic-conductivity measurements were used in an exponential variogram with specified range and sill (variance) (Deutsch and Journel, 1992). The range and variance of log of hydraulic-conductivity measurements were 700 m and 1.52, respectively, and using the 99 measured hydraulic conductivities, they were kriged (interpolated) onto each model cells. The vertical anisotropies and porosity of the lower layer were specified as one tenth of horizontal hydraulic conductivities and 0.5% (Cannata, 1988), respectively.

The calculated range and sill were used to generate kriging factors for pilot points in the lower layer using the `PPK2FAC` utility in the PEST suit. Later, these factors were used to interpolate k onto the model grid using the `FAC2REAL` utility in the PEST suit. Because of exposure to weathering and erosional process, the upper layer is more heterogeneous even though it comprises similar rock types, so a larger variance is appropriate. Thus, a variance of 3.04 (twice that of the lower layer) was assigned to the upper layer with the same 700-m range as the lower layer. Parameter uncertainties and observation worth were calculated based on propagation of variance. Initially, a pre-calibration covariance matrix was calculated for the pilot points using the `PPCOV` (Doherty, 2016). The diagonal elements of the covariance matrix were the variance while off-diagonal elements were non-zero numbers based on 99 measured

hydraulic conductivities and their geospatial characteristics. Later, this covariance matrix was used to generate pre-calibration pilot point realizations.

A total of 156 parameters were adjusted during calibration. Parameters were subdivided into three groups: (1) 77 pilot-point-based horizontal hydraulic conductivities for the upper (k_{u_1}–$k_{u_{77}}$), (2) and lower (k_{l_1}–$k_{l_{77}}$) layers, and (3) horizontal anisotropies for the upper, h_u, and lower, h_l, layers. Each hydraulic conductivity was assigned a pre-calibration lognormal probability distribution with mean of 3.048×10^{-3} m/s and 10^{-7} m/s for the upper and lower layers, respectively.

Calibration Data and Predictions of Interest

Because of model complexity and under-determinacy, regularization was used during calibration (Doherty, 2016) to reduce bias and to decrease the required number of model calls. A total of 156 log-transformed regularized parameters were calibrated. The calibration was performed against 43 steady-state heads (average water levels if multiple measurements were available) at the monitoring wells indicated with open circles in Figure 1. Each observation was assumed equally important (equal weight). Predictions of interest were travel times for 10 particles released at the midpoints of the upper and lower layers in the five administrative hazardous storage sites in Figure 1 (Moore and McSpadden, 2009). Travel times for the particles were simulated and their uncertainties were quantified.

Calibration and NSMC

The work flow for calibration and uncertainty analyses was: (1) parametrization and calibration of the NWIRP MODFLOW model; (2) parameter identifiabilities were calculated along with observation worth; (3) the effects of additional monitoring wells on travel-time predictions and observation worth were estimated; (4) uncertainties in travel-time predictions were then explored upon consideration of the new hypothetical observations; and (5) post-calibration uncertainties of pilot points and particle travel times were assessed using the NSMC technique.

Calibration estimated 156 parameters (154 pilot points and two horizontal anisotropies) with only 43 observations, which made this an ill-posed problem; not all parameters could be uniquely estimated (Moore and Doherty, 2005; Moeck et al., 2019). Using SVD, it was determined that 25 unique linear combinations of parameters (super parameters) could be reasonably identified. In other words, the modeler presents all parameters (here 113 more than could possibly be identified by the 43 data points) to the calibration and PEST's subspace-regularization capabilities determine which parameters can be uniquely identified by the dataset and focuses only on those. Here, the dataset informed 25 parameters, which can be specified as linear combinations if there are strong correlations between some of them. For over-parametrized models, an NSMC analysis affords a more accurate, nonlinear assessment of predictive uncertainty. The NSMC approach was executed in three steps. First, the `RANDPAR` utility in the PEST suite generated a total of 1,000 pilot-point realizations (the mean

of each pilot point stabilized by 1,000 realizations) using a lognormal distribution with mean (from pump tests) and the pre-calibration covariance matrix. Second, each pre-calibrated parameter realization was perturbed by adding null-space uncertainty using the `PNULPAR` utility in the PEST suite. `PNULPAR` calculated the orthogonal differences between calibrated parameters and those produced with `RANDPAR`. Then, these differences were added to each realization of calibrated parameters to generate 1,000 realizations of calibrated parameters with perturbations in the null space. For each of these null-space-projected parameter-field realization, three iterations of a PEST calibration were undertaken (`NOPTMAX = 3`) using 25 super parameters. Each resulted in a calibration-constrained parameter-field realization (NSMC realization) with a corresponding objective function value, Φ, which indicated how closely the simulated heads from that model run matched their corresponding observation. Only those NSMC realizations with objective functions less than 150% of the calibrated objective function were retained to form distributions of parameters and corresponding particle transport times.

Results

Post-calibration Linear Uncertainty Analyses

Figure 6 compares measured (y axis) and calibrated (x axis) heads. There was a slight bias toward underprediction (-0.1 m) while the mean absolute error was 0.7 m and the root-mean-squared error was 0.8 m, all of which indicated a well-calibrated model.

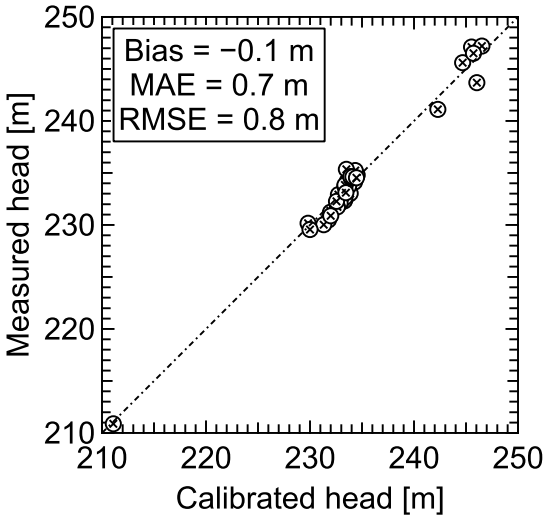

Figure 6: Cross-plot of measured and calibrated heads.

The linear uncertainty analysis identified reductions in parameter uncertainty (i.e., pilot-point hydraulic conductivities, so parameter uncertainty always refers to the reduction in the *a priori* hydraulic conductivity variance while prediction uncertainty refers to reduction in

particle travel times realized through the reductions in hydraulic-conductivity uncertainty) subject to the information content in the calibration dataset. These post-calibration parameter uncertainties reflect the degrees to which the observations reduced the pre-calibration parameter uncertainties as indicated in Figure 7. The size and color of the circles indicate the percent reduction in pre-calibration uncertainty upon application of the observation data set. Pilot-point uncertainty reductions ranged from 0.4 to 92% with greater reductions nearer to observation wells in the upper layer (Figure 7a). The Central (CR), Station Creek (SCR), and East Regions (ER) of the model, indicated on Figure 1, contain 36, six, and one observation well(s), respectively. In the upper layer, observation wells in the CR significantly informed the six nearby pilot points by reducing their uncertainties from 15 to 92%. Similarly, the six wells in the SCR decreased uncertainties in four nearby pilot points from 25 to 90% while the single well in the ER notably decreased uncertainty in the nearest pilot point by 50%. Uncertainty reduction for pilot points in the lower layer ranged from 0.3 to 30%. Observation wells in the CR reduced uncertainties from 1 to 30%; the uncertainty reductions were relatively lower in this layer because of their smaller (by half) variances 1.52 (1.78×10^{-10} m^{2}s^{-2}). The observation wells in SCR informed three nearby pilot points and reduced their uncertainties by >10% (Figure 7b). All told, uncertainty reductions for 120 of the pilot points were <10%, typically for pilot points distant from observation wells. Overall, given that there were 43 unequally distributed monitoring wells and low initial uncertainties in the lower-layer pilot points, it was not surprising that only 36 pilot points had their uncertainties reduced by >10%. Moreover, this is in accord with the 25 super-parameters.

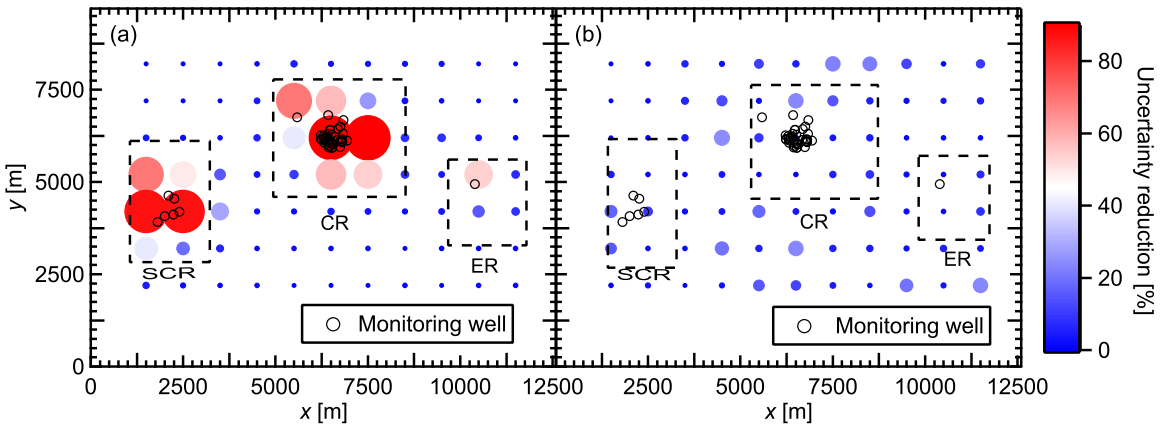

Figure 7: Reductions in pre-calibration uncertainties for the (a) upper- and (b) lower-layer pilot points upon application of the observation data set (43 head measurements).

Parameter identifiability apportions parameter uncertainties into solution and null spaces. An identifiability of zero means that the data set says nothing about that parameter while an identifiability of one means that uncertainty in that parameter is solely due to measurement error. Of the 156 parameters, seven had high (>0.5), 12 moderate (0.1–0.5), and 137 low

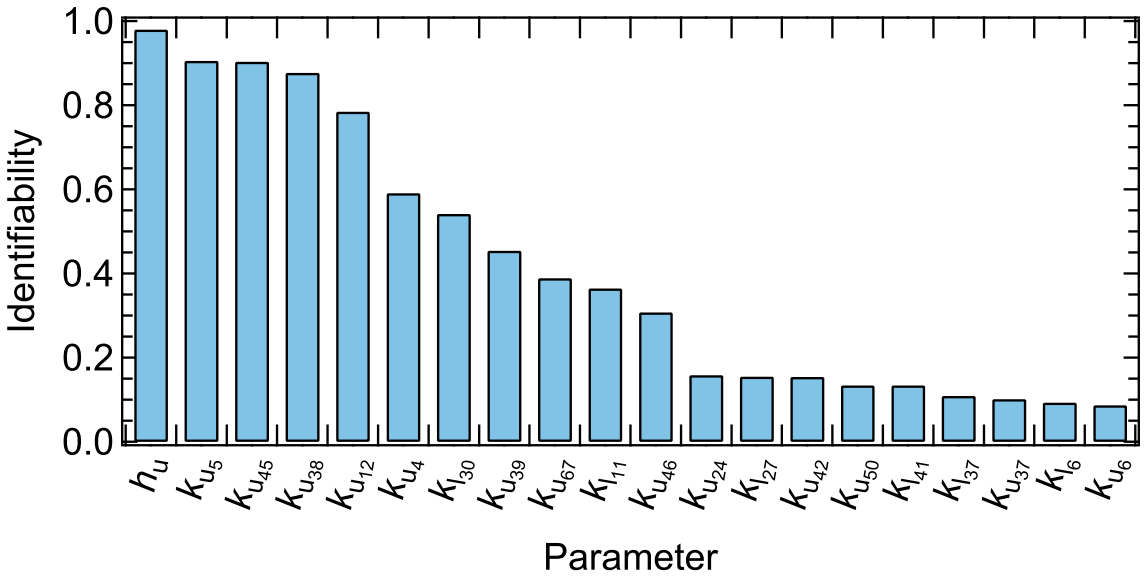

Figure 8: The 20 most identifiable parameters.

(<0.1) identifiabilities (Figure 8).

Initial travel-time variances were calculated by running the 1,000 uncalibrated hydraulic-conductivity-field realizations and these uncertainties were reduced by up to 92% (i.e., standard deviation in travel times reduced from 3.15 to 0.25 years) when the model was run with the calibrated hydraulic-conductivity-field realizations (discussed below in the NSMC section). Later, an observation-worth analysis was performed to assess the contribution of monitoring wells toward reducing travel-time uncertainties. Normalized contributions to uncertainty reduction from each monitoring well are indicated in Figure 9(a). The contributions of observations toward decreasing travel-time uncertainties depended upon the degree to which parameters were constrained along the particle paths and the proximity of the observations to the particle paths. Monitoring wells in the SCR reduced uncertainties for particles released at M and L while one monitoring well in the CR reduced uncertainties for particles released at H and one monitoring well in the ER reduced uncertainties for particles from S. Particles released at G did not travel through a region with monitoring wells, so their uncertainties were not reduced. Overall, only four monitoring wells (large open circles in Figure 9(a)) significantly contributed to decreasing travel-time uncertainties indicating that the existing monitoring well network does not effectively inform contaminant transport times.

Additional targeted well installations would greatly reduce travel-time uncertainties. To this end, an observation-worth analysis was performed by adding 11 hypothetical monitoring wells (Figure 9(b)). The analysis was conducted by placing wells at every third model cell throughout the model domain and selecting optimal locations. The 11 hypothetical wells yielding the greatest reductions in uncertainty were "installed" down gradient of particles

released at administrative area M and H and in the vicinity of hazardous-materials storage sites L, G, and S where no wells exist. The normalized contributions of monitoring plus hypothetical wells in reducing travel-time uncertainties are indicated in Figure 9(b). Because the 11 hypothetical wells were optimally located, they significantly reduced travel-time uncertainties. Interestingly, upon adding the 11 hypothetical wells, the contribution of an existing well toward reducing travel-time uncertainty greatly increased (largest red circle) because it gained important gradient information.

Post-calibration sensitivities of travel-time predictions to observations were calculated and the degree to which observations reduced uncertainties in travel times are indicated in Figure 9(c). M_u and M_l had six nearby monitoring wells and these particles traveled through a well-constrained region of the model and happened to have the shortest travel distances; hence uncertainties were notably reduced by the existing wells. The addition of the 11 hypothetical wells further constrained parameters (and corresponding predictions) such that travel-time uncertainties for particles released at H, G, L, and S decreased by factors from 1.04 to 4.3 (red bars) compared to the existing well network.

NSMC

Of the 1,000 parameter-field realizations generated, 882 had objective functions less than 1.5 times the calibrated (minimum) objective function after three optimization iterations of a NSMC calibration. These calibration-constrained NSMC realizations and the corresponding 882 realizations from `RANDPAR` were used to compute parameter variances to assess the decreases in uncertainty due to the information content in the calibration data set. Figure 10 compares variances of pre-calibration and calibration-constrained parameter distributions. NSMC reduced uncertainties, at least slightly, for the majority of parameters near existing monitoring wells. For both layers, NSMC reduced notably uncertainties for highly and moderately identifiable parameters (see Figure 10 (a)–(b)).

Parameter realizations were also used to generate PDFs and Figure 11 provides six examples. For highly identifiable parameters (e.g., k_{u_5} and $k_{u_{45}}$) in the top row of Figure 11, distributions were significantly narrower for NSMC than from pre-calibration reflecting the information obtained from the null-space projection and three additional optimization iterations. Distributions of the moderately identifiable parameters (e.g., $k_{u_{39}}$ and $k_{u_{67}}$) in the middle row of Figure 11 were also narrower than their pre-calibration equivalents. Finally, the three optimization iterations slightly reduced uncertainties for minimally identifiable parameters (e.g., $k_{l_{37}}$ and $k_{l_{41}}$) in the bottom row of Figure 11, because those parameters were not informed by existing monitoring wells.

The field data from the monitoring wells were collected by a third party and the model was created after the investigation was finished to try to show how contaminates may have migrated off-site and to demonstrate the need to incorporate the NSMC technique earlier in the monitoring process. There were no opportunities to collect additional data nor drill new wells.

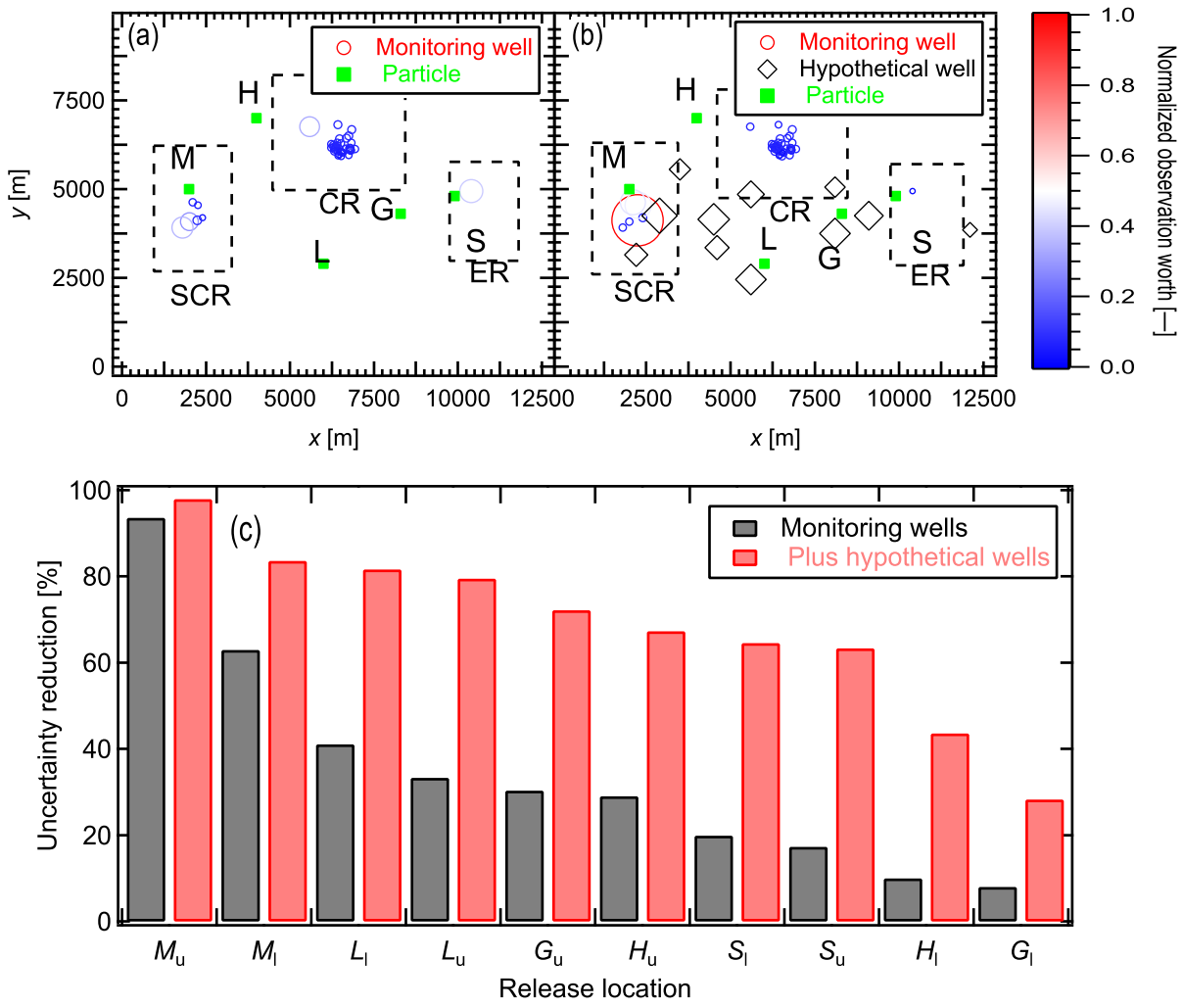

Figure 9: Normalized observation worth of (a) monitoring wells (b) plus 11 hypothetical wells, symbol color and size indicate the observation worth; and (c) uncertainty reductions in travel times due to the existing wells (gray) and as augmented by 11 hypothetical wells (red).

Travel times for particles to reach exit points at surrounding streams were calculated for each of the 882 pre-calibration and NSMC calibration-constrained pilot-point parameter fields and Figure 12 shows distributions of log-transformed travel times. Figure 13 shows 250 of 889 particle tracks released from each administrative designation with a representative hydraulic conductivity field. A particle's travel time depended on distance traveled and hydraulic conductivities along its path. Passing through even a single low-hydraulic-conductivity cells significantly decreased that particle's travel time. The combination of one or more low-hydraulic conductivities along a particle's path in conjunction with lognormally distributed hydraulic conductivities (long tails toward low values) yielded some realizations

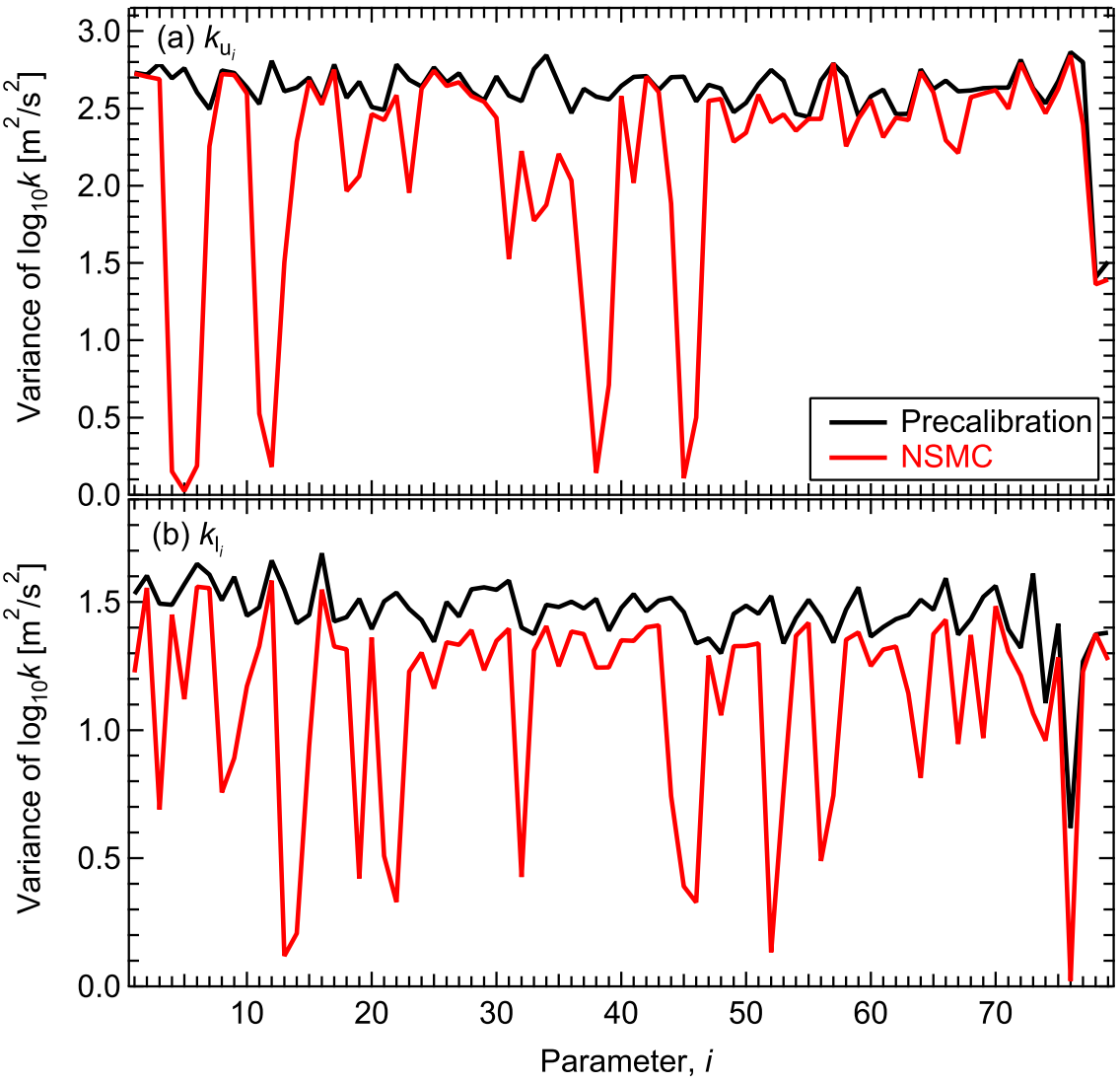

Figure 10: Variances of pilot point, $\log_{10}(k)$, for the (a) upper and (b) lower layers for the pre-calibration (black) and NSMC (red) parameter realizations.

with exceptionally long travel times, hence it was more appropriate to compare median travel times (Table 1).

Particles released at the same location for pre-calibration and NSMC parameter realizations had similar path lengths in both layers, but median travel times in the upper layer were significantly shorter than those through the lower layer. Consistent with the conceptual model, hydraulic conductivity ranges were two to six order of magnitude higher in the upper layer than the lower layer resulting in the travel-time disparities. However, all simulations indicated that particles released in the upper layer reached surrounding streams within a year (consistent with site observations) while particles released in the lower layer took one to five orders of magnitude longer.

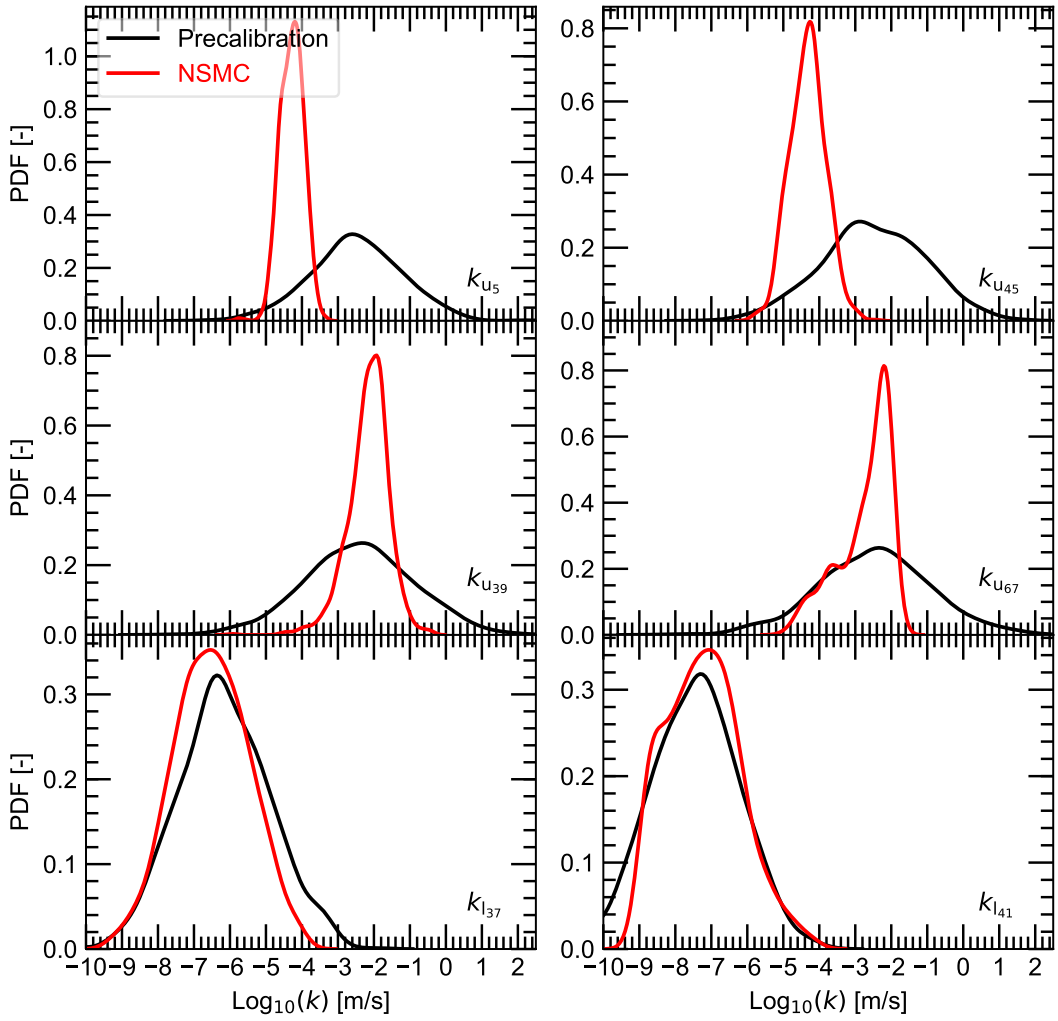

Figure 11: PDFs of highly identifiable parameters k_{u_5} and $k_{u_{45}}$ (top row), moderately identifiable parameters $k_{u_{39}}$ and $k_{u_{67}}$ (middle row), and minimally identifiable parameters $k_{l_{37}}$ and $k_{l_{41}}$ (bottom row).

Although travel times through the lower layer were longer when using pre-calibration parameter realizations than their NSMC counterparts, there was no consistent trend for the upper layer. Nevertheless, corresponding travel-time uncertainties (variances of log-transformed travel times) always decreased from pre-calibration to NSMC (visually evident in Figure 12), which is consistent with the decreases in parameter variances (see Figure 10). Although the NSMC travel times through the lower layer indicated that there may be more

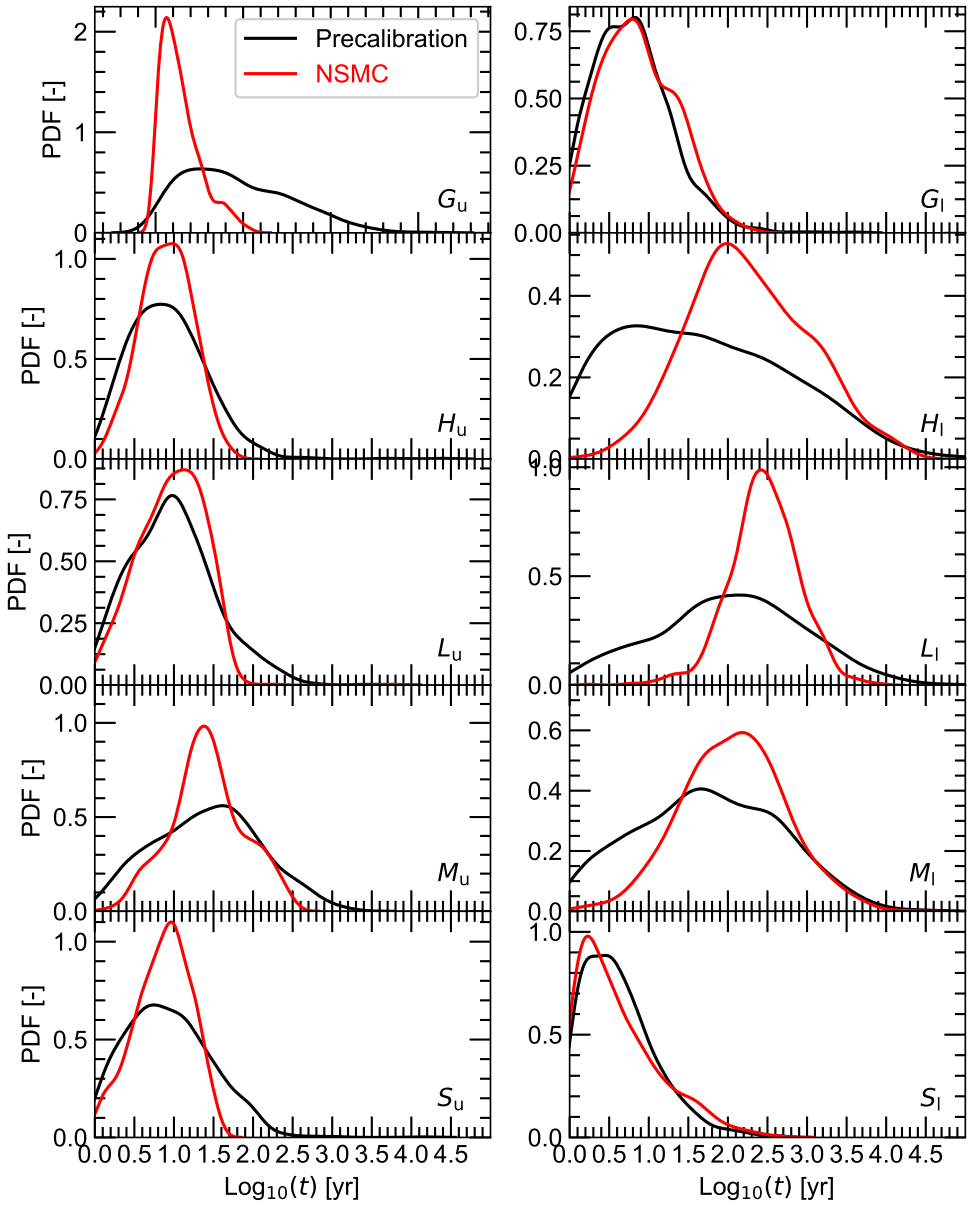

Figure 12: PDFs of particle travel times.

time to remediate the contamination than might have been initially expected, the short travel times through the upper layer offered no such consolation.

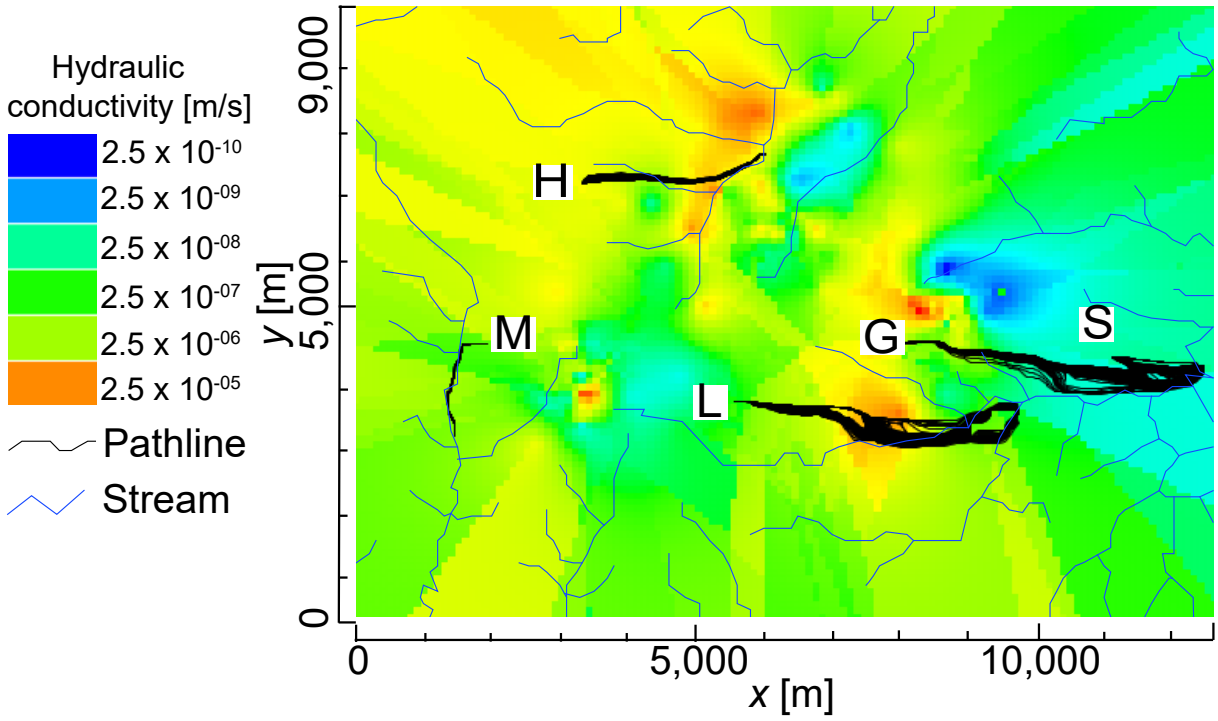

Figure 13: Particle paths for 250 realizations with background color a representative hydraulic conductivity distribution.

Table 1: Median travel times for the 882 pre-calibration and NSMC parameter fields used for particle tracking.

Particle	Pre-calibration [yr]	NSMC [yr]
G_u	0.29	$\uparrow$ 0.65
G_l	5.51	$\uparrow$ 6.65
H_u	0.14	$\downarrow$ 0.12
H_l	39.01	$\uparrow$ 166.34
L_u	0.11	$=$ 0.11
L_l	109.92	$\uparrow$ 303.23
M_u	0.03	$\uparrow$ 0.04
M_l	54.67	$\uparrow$ 122.98
S_u	0.13	$\downarrow$ 0.12
S_l	0.96	$\uparrow$ 2.36

Conclusions

Based on MODFLOW and MODPATH simulations of the NWIRP site, calibration and model interrogation quantified: parameter and predictive uncertainties, parameter identifiabilities, and observation worth for both existing and hypothetical monitoring wells. Using

a linear analysis, pre-calibration parameter uncertainties were reduced up to 92% and 36 of 156 parameters exceeded a 10% reduction when constrained by the calibration data set. An identifiability study revealed that seven parameters were highly identifiable (>0.5) while 12 parameters had identifiabilities between 0.1 and 0.5. Travel-time uncertainties were reduced up to 92%. Using a nonlinear analysis, pre-calibration travel-time uncertainties were reduced by >50% for two particles released at site M and between 5 and 40% for the remaining sites. An observation-worth analysis showed that the existing monitoring well network does not strongly constrain travel times. Targeted data collection, especially at the locations shown in Figure 5, could reduce travel-time uncertainties for all particles by factors from 1.04 to 4.3.

This study generated pre- and post-calibration parameter distributions along with corresponding travel-time PDFs. The decreases in width of these distributions (variances) reflected the information content in the calibration data set. This study also predicted travel times for conservative tracers to reach nearby streams and revealed that conservative tracers exited the flow system through the more conductive upper layer within a year.

Any seriously contaminated site like NWIRP should undergo a rapid and detailed modeling study before further data collection and, of course, before making remediation decisions. For example, the authority collected clustered water-level measurements, which could have been optimized if a comprehensive study was conducted before drilling wells. In addition, no base-flow data were collected even though a single base-flow measurement significantly improves uncertainty quantification (Hunt et al., 2019). This study also demonstrated that if a contaminant reached the upper layer (for example during storm events that raise the water table), it will travel much faster to surrounding streams. The predictive-uncertainty and observation-worth analyses determined the most important parameters and observations contributing to the greatest decreases in predicted travel-time uncertainties. Ultimately, the 43 poorly distributed water-level measurements over such a large model domain and the absence of base-flow data were notable shortcomings. This analysis can support decision making by identifying where additional wells should be located to achieve the greatest reductions in predictive uncertainty.

Looking to the future, transient modeling would be appropriate for this system, but it could not be undertaken because of a lack of water-level time-series data. Although beyond the scope of this study, it would be reasonable to use these calibrated parameters in a transient model to assess other aspects of contaminant transport subject to storm or flood events.

Data Resources and Computer Code Availability

`PEST` source code can be downloaded at http://www.pesthomepage.org/. The `NWIRP` simulation input files and executables used for this manuscript are available in the public Github repository https://github.com/bulbulahmmed/NWIRP. Additional information regarding

the simulation datasets can be obtained from Bulbul Ahmmed (Email: bulbul_ahmmed@baylor.edu).

References

Ashworth, J. B. and Hopkins, J. (1995). Aquifers of Texas. Technical report, Texas Water Development Board.

Barquest, B. A. (1989). *A hydrogeologic assessment of the Austin Chalk outcrop belt, central Texas: Waco, TX*. MS thesis, Baylor University.

Bingham, N. L. (1993). *Runoff, soil water, and groundwater along a first order stream in the Austin Chalk: Waco, Texas*. PhD thesis, Baylor University.

Cannata, S. (1988). *Hydrogeology of a portion of the Washita Prairie Edwards Aquifer, Central Texas: Waco, Texas*. MS thesis, Baylor University.

Chowdhury, A. H., Osting, T., Furdan, J., and Mathews, R. (2010). Groundwater-surface Water Interaction in the Brazos River Basin: Evidence from Lake Connection History and Chemical and Isotopic Compositions. Technical Report 375, Texas Water Development Board.

Clark, B. (2000). *Modeling groundwater velocities in fractured limestone*. MS thesis, Baylor University.

Craig, M. and Burdick, K. (2007). Navy divestiture deep in the heart of Texas: Collaboration leads to successful remediation and redevelopment of McGregor property. *The Navy's Environmental Magazine Currents*.

Deutsch, C. and Journel, A. (1992). *GSLIB: Geostatistical Software Library and User's Guide*. Applied Geostatistical Series. Oxford University Press.

Doherty, J. (2003). Ground water model calibration using pilot points and regularization. *Groundwater*, 41(2):170–177.

Doherty, J. (2016). Model-independent Parameter Estimation User Manual Part I: PEST, SENSAN and Global Optimisers. Technical report, Watermark Numerical Computing.

Doherty, J. and Hunt, R. J. (2009). Two statistics for evaluating parameter identifiability and error reduction. *Journal of Hydrology*, 366(1-4):119–127.

Ensafe Inc. (1999). Draft Groundwater Investigation Report, NWIRP, Mcgregor, Texas. Technical report, EnSafe, Inc.

ESRI (1996). *ArcView GIS, ARC/Info: User's Manual*. Environmental System Research Institute, Inc. version 3.2.

Gburek, W. J., Folmar, G. J., and Urban, J. B. (1999). Field data and ground water modeling in a layered fractured aquifer. *Groundwater*, 37(2):175–184.

Golden Software (1997). Surfer, Surface Mapping System. version 6.04.

Hare, J. P. (2000). Remediation of perchlorate contamination at NWIRP McGregor, Texas. *Federal Facilities Environmental Journal*, 36(8):33–38.

Hughes, J., Langevin, C., Chartier, K., and White, J. (2012). Documentation of the Surface-Water Routing (SWR1) Process for Modeling Surface-water Flow with the U.S. Geological Survey Modular Ground-Water Model (MODFLOW - 2005). Technical report, U.S. Geological Survey.

Hunt, R. J., Doherty, J., and Tonkin, M. J. (2007). Are models too simple? Arguments for increased parameterization. *Groundwater*, 45(3):254–262.

Hunt, R. J., Fienen, M. N., and White, J. T. (2019). Revisiting "An exercise in groundwater model calibration and prediction" after 30 years: Insights and new directions. *Groundwater*, pages 1–15. https://doi.org/10.1111/gwat.12907.

James, S. C., Doherty, J. E., and Eddebbarh, A.-A. (2009). Practical postcalibration uncertainty analysis: Yucca Mountain, Nevada. *Groundwater*, 47(6):851–869.

Lotti, F. and Doherty, J. (2016). The role of numerical models in environmental decision-making. *Acque Sotterranee-Italian Journal of Groundwater*, 5(3):73–75.

Mace, R. E. (1998). *Groundwater flow and solute transport in a fractured chalk outcrop, north-central Texas: Austin, Texas*. PhD thesis, University of Texas, Austin.

Moeck, C., Molson, J., and Schirmer, M. (2019). Pathline density distributions in a Null-Space Monte Carlo approach to assess groundwater pathways. *Groundwater*, pages 1–19. https://doi.org/10.1111/gwat.12900.

Moore, C. (2006). The use of regularized inversion in groundwater model calibration and prediction uncertainty analysis.

Moore, C. and Doherty, J. (2005). Role of the calibration process in reducing model predictive error. *Water Resources Research*, 41(5):W05020.

Moore, C. and Doherty, J. (2006). The cost of uniqueness in groundwater model calibration. *Advances in Water Resources*, 29(4):605–623.

Moore, T. L. and McSpadden, H. J. (2009). From bombs to rockets at McGregor, Texas. In *The New Horizons Forum and Aerospace Exposition*, Orlando, FL. American Institute of Aeronautics and Astronautics.

Myrick, M. (1989). *Aquifer-stream interactions and their hydrologic implications in nonkarstic limestones, Washita Prairie, Central Texas: Waco, Texas*. PhD thesis, Baylor University.

Niswonger, R. G., Panday, S., and Motomu, I. (2011). MODFLOW-NWT, a Newton Formulation for MODFLOW-2005. Technical Report 6-A37, U.S. Geological Survey.

Parker, R. L. (1977). Understanding inverse theory. *Annual Review of Earth and Planetary Sciences*, 5(1):35–64.

Pollock, D. (2016). User Guide for MODPATH Version 7-A Particle-tracking Model for MODFLOW. Report OFR 2016-1086, U.S. Geological Survey.

TNRIS (1999). Digital data. http://www.tnris.state.tx.us/DigitalData/DEMs/dems.htm, Accessed in Aug. 1999.

Tonkin, M. and Doherty, J. (2009). Calibration-constrained monte carlo analysis of highly parameterized models using subspace techniques. *Water Resources Research*, 45(12).

Tonkin, M., Doherty, J., and Moore, C. (2007). Efficient nonlinear predictive error variance for highly parameterized models. *Water Resources Research*, 43(7).

Tonkin, M. J. and Doherty, J. (2005). A hybrid regularized inversion methodology for highly parameterized environmental models. *Water Resources Research*, 41(10).

Verma, R. K., Rao, M. K., and Rao, C. V. (1980). Resistivity investigations for ground water in metamorphic areas near Dhanbad, India. *Groundwater*, 18(1):46–55.

Waco Regional Airport (2016). National Weather Service, Zip code: 76708. http://weather-warehouse.com/WeatherHistory/PastWeatherData_WacoRgnArpt_Waco_TX.html, Accessed on 2016-01-30.

Welter, D. E., White, J. T., Hunt, R. J., and Doherty, J. E. (2015). Approaches in Highly Parameterized Inversion – PEST++ Version 3, A Parameter ESTimation and Uncertainty Analysis Software Suite Optimized for Large Environmental Models. Technical report, US Geological Survey.