Predictions of contaminant transport produced by computer models are now widely used to evaluate health risks, determine remediation strategies, and guide environmental regulatory decisions. But model prediction errors are not generally well-known or quantified. In this study, the results of an extensive tracer test at the Columbus Air Force Base (CAFB), MS, test site are used to quantify error in computer model predictions of groundwater contaminant movement in the subsurface. Our modeling assumes no knowledge of the observed solute transport, uses widely applied modeling methods to generate computer predictions, and uses permeability and water levels as input data. Predicted and observed mass transport differ significantly, despite the relative abundance of aquifer data. It is not possible to accurately predict mass transport at CAFB using our approach because permeability cannot be characterized at the necessary resolution. Prediction errors at Superfund sites and other contaminated groundwater sites may be even higher because data is less abundant. This suggests that widespread use of predictive mass transport models that depend strictly on measured permeability and head values to predict transport should receive greater scrutiny and that methods of model parametrization based on early transport behavior should be more widely applied.

Introduction

Prediction of mass transport in the subsurface is critical to assessing groundwater contamination hazards. Groundwater models, which incorporate basic physical and chemical processes governing mass transport, are commonly used in conjunction with field and laboratory measurements to make quantitative predictions of subsurface pollution transport. For regulatory purposes, predictions are often based on approximations to the advection dispersion equation (1) or on simple analytical solutions to the advection-dispersion equation that assumes homogeneity of hydrologic and geochemical properties (2).

Computer-based transport predictions are routinely used as an aid to calculate health risks (3), determine cleanup strategies (4), guide environmental regulatory policy (5), and in post-audit studies, to determine culpable parties in lawsuits (6). Collectively, computer-based predictions of subsurface mass transport influence legal and policy decisions involving the allocation of at least 1 billion dollars annually in the United States alone (7, 8).

Implicit in the use of these models is the assumption that their predictions possess some reasonable degree of accuracy and that the error bounds on the predictions are meaningful. However, given the inherent heterogeneity of the subsurface, the accuracy of these models is very site specific and is unknown a priori. Despite the heavy use of these models in decision making, few post-audits of their value appear in the literature, and these find significant differences between predicted and observed mass transport (9–11).

In this study, through the use of one of the most well-studied cases of subsurface mass transport in the world, we explore limitations in the prediction of mass transport in the subsurface and question the validity of decision making based upon transport predictions produced by commonly applied groundwater modeling methods.

Test Site Conditions and Behavior

We use results from an aquifer test site at Columbus Air Force Base (CAFB), MS (Figure 1). The CAFB aquifer lies 2–3 m below land surface and consists of unconsolidated sand and gravel 10–12 m thick of fluvial origin. In 1986, a bromide tracer plume was injected and then monitored for 1.5 yr while it moved with natural groundwater flow (12). Figure 2 shows observed depth-maximum bromide concentrations. The study provided an unprecedented detailed examination of subsurface mass transport and its controls. Measurements at the site, 24 ha in area, include 2451 tests of permeability, water level measurements in 78 wells, 38 test holes for sediment analysis, and detailed geophysical testing (13–15). The 1986 tracer test involved collecting 11 126 groundwater samples from 225 multi-level wells and analyzing bromide, tritium, and fluoric acid concentrations. Subsequent tracer tests studied hydrocarbon transport (16).

It has been argued that spatial variability in permeability at this site is unusually large (12), and if so, the generality of the results obtained from this site would be limited. The results summarized in Table 1 indicate that permeability variation at the CAFB site (ln (K) variance = 4.4) is in the upper range of 20 North American aquifers considered. Given that field-based research of subsurface mass transport is, for ease of testing and interpretation of results, biased toward field sites with relatively homogeneous permeability structure (17–19), comparison of values in Table 1 likely overestimates the variability of the Columbus aquifer relative to other aquifers. We estimate that the results of this study are directly applicable to 10–20% of aquifers in North America.

* Corresponding author phone: (919)684-5245; fax: (919)684-5833; e-mail: jre4@duke.edu.
Mass transport of bromide in this aquifer did not follow a Gaussian pattern as might be expected from common solutions to the advection–dispersion equations (20) but instead was patchy in character with the plume’s leading edge moving 10–20 times faster than the center of mass (Figure 1). The field test of bromide movement, while carefully performed, encountered complications typical of field work. Water levels and head gradients fluctuated significantly during the bromide tracer test (Figure 3); the saturated thickness of the aquifer varied by up to 40%, and the direction of head gradients shifted by up to 40° (12).

Predicting Subsurface Mass Transport

Prior Studies. Previous transport models of the CAFB site have used observed plume behavior to constrain predictions. Adams and Gelhar (13) prescribe a steady diverging flow field and dispersion parameters to recreate spatial moments of the bromide plume. Others have made hindcasts of mass transport by assuming that the aquifer contains two domains of contrasting permeability (21). Harvey and Gorelick (22) use a rate-limited mass transfer that model that assumes the aquifer has low permeability material with mostly diffusive transport and high permeability material with mostly advective transport. The model provides a good explanation of the declining mass recoveries and plume migration but requires the assignment of three mass-transfer parameters and assumes that flow velocities are completely known. Zheng and Jiao (23) use a steady-state flow field and two dispersivity values to recreate the tritium tracer plume. They had difficulties in fitting observed transport behavior, which they attributed to inadequate characterization of permeability. The four prior studies above all use observed plume behavior to calibrate their models and therefore are “predicting” solute transport as we attempt to do in this study. A prior modeling study by Eggleston and Rojstaczer (24) found large-scale spatial permeability trends did not have primary control over transport behavior at the CAFB site. The study presented here differs from the 1998 study in that it focuses on groundwater model prediction accuracy rather than the effect of large-scale trends, uses a transient model rather than a steady-state model, and includes permeability variation at scales <10 m in the model.

FIGURE 2. Measured depth–maximum bromide concentrations after 503 days of transport. Warmer colors indicate higher [Br]. Note non-Gaussian plume behavior. Depth-averaged permeability in gray; darker shading indicates lower permeability. Curved solid line is the 1 ppm [Br] contour. Black circles are locations of permeability measurement wells.

TABLE 1. Variability in Aquifer Permeabilitya

<table>
<thead>
<tr>
<th>site and reference</th>
<th>no. of measurements</th>
<th>variance of ln (K)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Glatt Valley, Switzerland (38)</td>
<td>18</td>
<td>1.43</td>
</tr>
<tr>
<td>2 Chalk River Nuclear Laboratories (39)</td>
<td>18</td>
<td>0.32</td>
</tr>
<tr>
<td>3 Twin Lakes (19)</td>
<td>89</td>
<td>0.07</td>
</tr>
<tr>
<td>4 Borden, Ontario (18)</td>
<td>1279</td>
<td>0.38</td>
</tr>
<tr>
<td>5 Columbus, MS (14)</td>
<td>2451</td>
<td>4.40</td>
</tr>
<tr>
<td>6 Cape Cod, MA (17)</td>
<td>668</td>
<td>0.24</td>
</tr>
<tr>
<td>7 Richland, WA (40)</td>
<td>46</td>
<td>3.24</td>
</tr>
<tr>
<td>8 Havana, IL (41)</td>
<td>31</td>
<td>0.15</td>
</tr>
<tr>
<td>9 Central IL (42)</td>
<td>38</td>
<td>0.25</td>
</tr>
<tr>
<td>10 East Central IL, Wedron formation (43)</td>
<td>60</td>
<td>0.99</td>
</tr>
<tr>
<td>11 East Central IL, Henry formation (43)</td>
<td>23</td>
<td>1.16</td>
</tr>
<tr>
<td>12 East Central IL, Banner formation (43)</td>
<td>71</td>
<td>1.37</td>
</tr>
<tr>
<td>13 East Central IL, Gladford formation (43)</td>
<td>118</td>
<td>0.95</td>
</tr>
<tr>
<td>14 Northeast AR (44)</td>
<td>19</td>
<td>0.18</td>
</tr>
<tr>
<td>15 West NV (45)</td>
<td>64</td>
<td>1.83</td>
</tr>
<tr>
<td>16 Colorado River, AZ and CA (46)</td>
<td>30</td>
<td>3.61</td>
</tr>
<tr>
<td>17 Corpus Christi, TX (47)</td>
<td>15</td>
<td>4.58</td>
</tr>
<tr>
<td>18 Yakima Reservation, WA (48)</td>
<td>35</td>
<td>3.47</td>
</tr>
<tr>
<td>19 Castle Valley, UT (49)</td>
<td>15</td>
<td>1.39</td>
</tr>
<tr>
<td>20 Kalamazoo, MI (50)</td>
<td>22</td>
<td>0.64</td>
</tr>
</tbody>
</table>

a Permeability variation is indicated by ln (K) variance, where K is the hydraulic conductivity. Permeability measurement techniques include well pumping tests and air permeability tests. Sampling scales vary from 2 to 100 km (updated from ref 37).
Mass Transport Model. For this study, we constructed a predictive groundwater model of the CAFB bromide plume. Groundwater flow was modeled with the software program MODFLOW (25), and bromide mass transport was modeled with the software program MT3D (26). Similar forward predictive models are commonly employed in regulatory settings. In setting up the model, we only included data that are commonly collected during a groundwater remediation project, such as the cleanup of a Superfund site, with the proviso that we had orders of magnitude more measurements of the aquifer than a typical site. We did not include any information that assumed knowledge of the actual plume behavior.

Model Parameters. The model domain was 96 m × 260 m × 10.5 m (X, Y, Z) with a total of 109,200 grid cells, each measuring 4 m × 4 m × 0.15 m. Uniform porosity 0.35 was a best estimate from Adams and Gelhar (13). Porosity can be expected to vary spatially and may exert significant control over transport variability at the site, but there is not enough data to accurately assess that variability over the model domain.

Values for permeability at the many computer grid nodes without measurements (106,749 out of 109,200) were interpolated using four different, commonly used, methods: homogeneous assignment, dual homogeneous zones, kriging, and simulated annealing. All four interpolation methods use the 2451 flowmeter measurements as the basis of generating interpolated model permeability values. For the homogeneous assignment, hydraulic conductivity at all model grid nodes was set equal to the global geometric mean of 0.0044 (cm/s). For the dual homogeneous zone assignment, two zones were delineated using grain-size analyses from six sediment cores and the flowmeter measurements of permeability. The test site aquifer can be roughly described as a region of predominantly coarse sediments within a region of predominantly fine sediments (14). Model permeability values for each grid node were calculated as the geometric mean of all flowmeter measurements within the surrounding sediment region (0.042 or 0.00176 cm/s). For assignment by kriging, ordinary kriging was implemented with a search radius of 120 m, a maximum of 16 conditioning data values, and an exponential variogram having a sill of 4.4 m, a horizontal range of 12.48 m, and a vertical/horizontal range ratio of 0.116. Figure 4 shows a vertical slice of the kriged permeability field. For simulated annealing assignment, hydraulic conductivity values were drawn from the measured...
distribution and spatially rearranged to minimize an objective function \( E \) (eq 1):

\[
E = \sum_{h} \left[ \frac{y'(h) - \gamma(h)}{\gamma(h)^2} \right]^2
\]  

(1)

There are many other methods that could be used to interpolate permeability values (27). But without looking to the observed bromide transport, it is impossible to know a-priori which interpolation methods will give permeability estimates yielding accurate model results. The four interpolation methods used here were chosen to represent common interpolation approaches (global average and kriging) and alternate approaches (hydrofacies and stochastic). Hydrogeologic field data other than the flowmeter measurements could also be used to infer model permeability values. In exclusively using the flowmeter data, we are following the conclusions of Rehfeldt et al. (28), who before the tracer test began examined hydraulic conductivity values obtained from slug tests, grain-size correlation, surface geophysics, borehole geophysics, pumping tests, hydrofacies mapping, laboratory core testing, and flowmeter tests before concluding that the flowmeter tests provide the most direct and successful measure of hydraulic conductivity.

Constant heads were assigned on the side boundaries, and no-flow boundary was specified on the bottom corresponding to an observed low permeability marine clay/sand unit. For each stress period, constant head values along the boundaries were changed to reflect measured water levels. Temporal variability in water levels were included in the model by reassigning the constants heads after 0, 170, 200, 300, and 410 days. Because the available water level data was extensive and there was little reliable recharge data, we did not explicitly include recharge in the model. The measured heads show vertically downward gradients as would be expected in a shallow surficial aquifer receiving recharge. The model heads are consistent with this pattern. Initial bromide concentrations were assigned to approximate field injection of the bromide solution. No dispersion or diffusion bromide concentrations were assigned to approximate field injection of the bromide solution. No dispersion or diffusion in initial assignment of concentration can be expected to be too small a spread as compared to the observed plume while high bromide concentrations have a spread that is much wider than the observed plume.

Because the available water level data was extensive and there was little reliable recharge data, we did not explicitly include recharge in the model. The measured heads show vertically downward gradients as would be expected in a shallow surficial aquifer receiving recharge. The model heads are consistent with this pattern. Initial bromide concentrations were assigned to approximate field injection of the bromide solution. No dispersion or diffusion in initial assignment of concentration can be expected to be too small a spread as compared to the observed plume while high bromide concentrations have a spread that is much wider than the observed plume.

Results and Discussion

Profiles of predicted and observed plumes are shown in Figure 5. Modeled plumes do not reproduce the dilute front stretching far ahead of peak concentrations. Low concentrations are predicted to have too small a spread as compared to the observed plume while high bromide concentrations have a spread that is much wider than the observed plume.

The base case simulation, which employed 5 time periods, kriged permeability values, and assignment of bromide concentrations on day 0, gave a prediction of bromide transport having \( \chi^2 = 0.98 \). All other simulations deviate from the base case in just one aspect; either the number of time periods, method of permeability assignment, or bromide concentration updating. \( \chi^2 \) values for all simulations are given in Table 2.

Predictions of bromide movement are highly dependent on the interpolation algorithm used to assign permeability values. The poor predictive performance of assuming a uniform permeability is expected, yet at most other contaminated groundwater sites there is insufficient sampling to adequately estimate spatial variation of permeability. Although numerous algorithms for generating realistic permeability fields exist (29), more work is needed to quantify the accuracy of their generated fields for different hydrogeologic settings.

Stochastic approaches for assigning permeability, such as simulated annealing, offered little benefit in our predictive modeling efforts. Of the four permeability interpolation methods used, only simulated annealing allows for stochastic predictions of mass transport. But with all permeability fields generated, simulated annealing consistently led to simulations that over-predicted plume velocities by 1–3 orders of magnitude, apparently because simulated annealing produced greater fine scale permeability contrasts with enough spatial continuity to short circuit less permeable regions.

Frequently, in practical use of physical models, updating is used to reduce predictive error, as for example in predicting meteor impacts on spacecraft (30) and modeling vehicle dynamics (31). To test if the poor model predictions were caused by propagation of error in initial concentration assignment, we updated model concentrations to the observed concentrations after 49, 126, 202, and 370 days and then ran the model to the 503 day stopping point. Figure 6 shows the reduction in error from updating bromide concentrations. Error is, as expected, reduced but differences between the observed and predicted behavior grow substantially after each updating, indicating that the error is coming from a source other than incorrect assignment of initial bromide concentrations. Because our measurements of concentration have exceptional spatial resolution, it appears that error in assignment of initial bromide concentrations has little influence on degradation of the predictions. At contaminated aquifer sites that typically have less data, spatial resolution of concentration is much poorer and errors in initial assignment of concentration can be expected to be significant and propagate over time.
Although it is outside the scope of this study to do so, early bromide concentration measurements could be used to refine model permeability, boundary conditions, or porosity and to justify the inclusion of other nonadvective transport processes such as dual porosity diffusion. We do not use the bromide data in these ways because we want the model to be strictly predictive (not conditioned to concentration data) and because honoring measured data is a guiding principle in the construction of predictive groundwater transport models. Also, there is no pre-tracer-test evidence that justifies letting secondary data take precedence over flowmeter measurements in generation of model permeability values (28). Results of the tracer test and analyses performed after the tracer test can be used to infer dual permeability domains (21, 22) or to get an inverse solution for permeability and improve model results (32, 33). Inverse methods generally require that some aspects of the measured data be ignored and introduce considerable subjective decision-making during parametrization.

Predictive models of mass transport in groundwater most often infer mass flux from hydraulic head gradients, porosity, and permeability. Head gradients and porosity vary by several orders of magnitude less than permeability, hence it is expected that errors in permeability estimation are the main source of predictive error in our mass transport simulations. If permeability were very well characterized so that its
standard error were comparable to the standard error for porosity and hydraulic gradient, then porosity and hydraulic gradient would become relatively important sources of prediction error.

It has been shown that permeability variations with length scales less than 1 m affect mass transport at CAFB (22–24). Effective porosity may also vary at scales less than 1 m, possibly in correlation with permeability, and affect mass transport. Conventional field methods for measuring permeability (slug tests, pumping tests) average out such fine scale variations. Permeability measurement was performed at CAFB using flowmeter techniques (34) in which the length scale of the measurement is dependent on the permeability. For the CAFB aquifer, flowmeter tests represent sampling volumes with length scales of 4–50 m (35). Unless fine-scale permeability measurements are performed on site, it is not possible a priori to estimate the magnitude of error associated with conventional large-scale testing. Even then, successful mass transport prediction would likely require such extensive field testing of small scale variability as to be prohibitively expensive and would substantially alter the nature of the aquifer materials.

Our results indicate that, unless small-scale variations in mass flux are generally absent, it is likely that useful prediction of subsurface mass transport is not possible using forward techniques that rely largely upon permeability and head measurements. How likely are small-scale variations to be present in typical settings? This is not an easy question to answer because of the limited information on small-scale variations that have been obtained elsewhere. However, assuming that high variability in conventional permeability tests are an indication of high variability in mass flux, comparison of this site to other sites suggests that difficulties in prediction of subsurface mass transport will be common.

### TABLE 2. Groundwater Model Parameters and Results

<table>
<thead>
<tr>
<th>K assignment</th>
<th>no. of time periods</th>
<th>day of bromide assignment</th>
<th>( \chi^2 ) value (at 503 days)</th>
</tr>
</thead>
<tbody>
<tr>
<td>kriging</td>
<td>5</td>
<td>0</td>
<td>0.98</td>
</tr>
<tr>
<td>homogeneous</td>
<td>5</td>
<td>0</td>
<td>7.11</td>
</tr>
<tr>
<td>2 zones</td>
<td>5</td>
<td>0</td>
<td>7.84</td>
</tr>
<tr>
<td>simulated</td>
<td>5</td>
<td>plume left</td>
<td></td>
</tr>
<tr>
<td>annealing</td>
<td>model domain</td>
<td></td>
<td></td>
</tr>
<tr>
<td>kriging</td>
<td>1</td>
<td>0</td>
<td>1.60</td>
</tr>
<tr>
<td>kriging</td>
<td>6</td>
<td>0</td>
<td>0.78</td>
</tr>
<tr>
<td>kriging</td>
<td>10</td>
<td>0</td>
<td>0.78</td>
</tr>
<tr>
<td>kriging</td>
<td>5</td>
<td>49</td>
<td>1.46</td>
</tr>
<tr>
<td>kriging</td>
<td>5</td>
<td>126</td>
<td>0.85</td>
</tr>
<tr>
<td>kriging</td>
<td>5</td>
<td>202</td>
<td>0.79</td>
</tr>
<tr>
<td>kriging</td>
<td>5</td>
<td>370</td>
<td>0.39</td>
</tr>
</tbody>
</table>

FIGURE 6. (a–c) Spatial moments of bromide in the direction of groundwater flow. Effects of updating bromide concentrations are shown by comparing updated model runs and observed plume after 503 days of transport. Spatial moments calculated along the direction of groundwater flow. (a) First spatial moment, position of the bromide center of mass. (b) Second spatial moment, bromide variance. (c) Third spatial moment, bromide skewness.

VOL. 34, NO. 18, 2000 / ENVIRONMENTAL SCIENCE & TECHNOLOGY • 4015
Our results also indicate that the common approach of using permeability and water level measurements as input to a computer transport model allows only the direction of mass transport to be predicted accurately. Possible alternative approaches are being developed, and based on our results, we recommend them as areas of future research. One alternative approach is to use early time transport observations to predict later time transport behavior. Distributions of observed transport velocities can be calculated from two plume snapshots measured at different times, and these velocity distributions can be used to predict plume migration over time. Another approach is to subdivide measured permeability values to concentration measurements and other data, use inverse solution techniques to determine updated model permeability values, and then run the model to predict later-time behavior. The mass-transfer approach of Harvey and Gorelick (22), which assumes strong permeability contrasts at scales <1 m such as are found at CAFB, has the potential to improve transport predictions. But first flow velocities must be known at the model grid scale and methods need to be developed for estimating the required transport parameters. Another potential method for better characterizing mass flux is direct measurement of porewater velocities (36).

While this site is more heterogeneous than most other field sites where mass transport has been observed in detail, the hydrogeologic description of the CAFB site is probably better than for any other currently active site in the world. When data used for predictive modeling are less comprehensive or less reliable, greater prediction errors can be expected. It is likely that decision making based on predictive models of subsurface mass transport like the one in this study will often be arbitrary. Better decision making will likely require better methods for characterizing subsurface mass flux or better methods for estimating permeability such as those mentioned above.

Literature Cited

(4) Begey, M. B.; Carnelutti, M.; Piran, E. Groundwater model for management and remediation of a highly polluted aquifer (organochlorine compounds) in an urban area, using radioactive tracers (14C) for hydrodynamic parameters and dispersivity measurements; Isotopes in Water Resources Management, Vol. 2; IAEA: Vienna, Austria, 1996.
(11) Hanson, R. T. Postaudit of head and transmissivity estimates and groundwater flow models of Avra Valley, Arizona; USGS Water-Resources Investigative Report 96-4405; USGS: Denver, CO, 1996.


Received for review January 14, 2000. Revised manuscript received May 30, 2000. Accepted June 7, 2000.