Big Data and Environmental Remediation: Gaining Predictive Insights

by Suthan Suthersan, Margaret Gentile, Caitlin Bell, Joseph Quinnan, and John Horst

Introduction

Since the invention of the world-wide web 25 years ago, a lot of data sources collected in the past, and currently being collected, have been digitized and become easily accessible. Today it is easy to locate, collect, and analyze digitized data and information that would have taken weeks or months in the past. The expanding access to huge datasets and emergence of powerful and real-time technologies represent opportunities for analysis that were never possible in the past—thus leading to the term "*big data*." In a digitized world, *big data* refers to the things one can do at a large-scale that cannot be done at a smaller one, to extract new insights or create new forms of value, in ways that can change markets, organizations, the relationships between various stakeholders, and more. It has been transformative within many industrial sectors and has been heralded as the next frontier for innovation and competitiveness.

Large volumes of data have accumulated within the remediation industry primarily driven by the requirements of regulatory compliance and risk management. Although this topic might once have interested only a few "data geeks," *big data* are now relevant within our industry, and all stakeholders stand to benefit from its application. In this column, we explore what *big data* means for environmental restoration, and the potential it has to yield insights that support more streamlined and sustainable approaches to benefit all stakeholders—regulatory agencies, responsible parties, researchers, remediation practitioners, and local communities. In the past, we learned slowly and gradually about natural attenuation of petroleum hydrocarbons and chlorinated solvents from data collected from thousands of sites. The concept of *big data* did not exist at that time. The insights gained took several years to observe mainly because of the poor access to the necessary data points across many sites.

The Evolution of Remediation Related Data Management

The gravity and impacts of *big data* may not be transparent to remediation practitioners yet, due to the fact that the mere concept of "*environmental big data*" may still be an abstract or amorphous concept. Remediation data management is fragmented, and many different kinds of data are managed in different formats and configurations. In addition, emphasis was, and is, always placed on detailed evaluation and analysis of data collected specific to individual sites.

What has been overlooked and ignored in our industry so far is the vast trove of environmental data and information related to but beyond the immediate purview of individual site specific needs (Table 1). *Big data* relevant to remediation can be classified into structured and unstructured data. Examples of structured data are cleanup standards, performance monitoring results, groundwater and soil chemistry, geologic and hydrogeologic parameters, air quality measurements, weather conditions, and compliance records. Good examples of unstructured data are scientific papers, regulatory changes, site history, USGS maps, aerial photographs, spill records, and permit renewals. Although evaluation of unstructured data is possible, it presents more challenges than evaluation of structured data. Therefore, *big data* evaluation of structured data is the focus of this discussion.

Environmental remediation data usually have multiple dimensions (contaminant type, time, location, depth, different variables, and so on) to be relevant for decisions. As a result, there is a tremendous amount of data generated that can be tapped into to develop insights. However, the evolution of how these data are managed has been slow and unplanned. Even though we are yet to see the *big data* revolution enter the remediation domain, the industry is ready for harnessing the available data, information and the acquired knowledge to be at the heart of gaining predictive insights and smart decision-making. Important steps in this direction include: benchmarking the data for the remediation industry, developing standardized reporting standards, mapping the available data and information, and prioritizing the analytical tools.

Table 1
Existing Publically Available and Relevant Environmental Datasets

Database	Description	Type of Data
U.S. EPA Toxicology Testing in the 21st century (Tox21)	Develop better toxicity assessment methods to quickly and efficiently test whether certain chemical compounds have the potential to disrupt processes in the human body that may lead to negative health effects. Toxicity data that can then be evaluated within the ToxCast tool	10,000 environmental chemicals and approved drugs
U.S. EPA EPA's Toxicity Forecaster (ToxCast)	Bring together various toxicity data to provide rapid and efficient methods to prioritize, screen, and evaluate thousands of chemicals. iCSS ToxCast Dashboard that allows for easy access to related data	Chemical properties and structure
U.S. EPA Enforcement and Compliance History Online (ECHO)	Online real-time database of (e.g., CAA, CWA) permits and enforcements/violations	Facility list; violations list
U.S. EPA Aggregated Computational Toxicology Resource (ACToR)	Online warehouse of all publicly available chemical toxicity data	Aggregates data from over 1000 public sources on over 500,000 chemicals
U.S. EPA Integrated Risk Information System (IRIS)	Identifying and characterizing the health hazards of chemicals found in the environment	Toxicity values for use in risk assessment (e.g., RfD, RfC, oral slope factor)
EJSCREEN	EJSCREEN is an environmental justice mapping and screening tool that provides EPA with a nationally consistent dataset and approach for combining environmental and demographic indicators. A good example of the power of integrating datasets in geographical analysis	Twelve environmental indicators (e.g., air quality data, proximity to NPL sites) Six demographic indicators (e.g., percent low income) Twelve environmental justice indices
Global Earth Observation System of Systems (GEOSS)	The Group on Environmental Observation, GEO, is a voluntary partnership of governments and organizations which created the GEOSS platform for integration and analysis of global environmental data for study in multiple areas of societal benefit	Global environmental datasets, including agriculture, biodiversity, climate, disasters, ecosystems, energy, health, water management and climate
Earth Resources Observation and Science (EROS) Center	EROS is a remotely sensed data management, systems development, and research field center for the U.S. Geological Survey's (USGS)	Remotely sensed, for example satellite, datasets
National Wetlands Inventory	U.S. Fish and Wildlife maintains the NWI to provide public access to wetlands mapping and to provide a tool for reporting and predicting trends	Geospatial wetlands data
National Water Information System	U.S. Geological survey maintains this database of surface and groundwater data for use in managing water resources	Water resource data including surface water flow and levels, groundwater levels, water quality, and water use
State site databases	In some states, contaminated site data and water supply well data are housed in state databases. Examples are the Geotracker database of environmental data for regulated facilities and the Groundwater Ambient Monitoring and Assessment Program databases in California	Concentration data over time, well construction data, and so on

Another area to improve is connection of different methods used for the collection and storage of environmental datasets. Older datasets may have relied on pen and paper for collection and storage. Recently, datasets housed solely in Microsoft Excel® seem to have become antiquated. Currently, relational databases (e.g., Microsoft Access®, EQuIS) are a common method for storing data and querying focused datasets. These are complemented by data collection devices, many times handheld, that are configured to sync flawlessly with the designed database. Some specialized relational databases are designed to report certain metrics based on predetermined *big data* evaluations. Conversely, many data repositories were developed to house only the documents in which the data are reported, such as state regulatory repositories. Several states, however, have the raw data stored in online databases readily accessible to the public (e.g., GeoTracker in California). As *big data* evaluations become more and more desirable, storage of the data in its raw form is recognized as being useful. The quality of the data is equally important as its format. Different data types are held to different levels of quality assurance. Use of data with variable levels of quality assurance can lead to unreliable conclusions, and use of high quality data is an important consideration when preparing a *big data* evaluation.

There are multiple ways to approach evaluation of *environmental remediation big data* sets. The most straightforward practice in the past was simple correlation plots between two data types that may be related (e.g., organic carbon concentration and total chlorinated solvent concentration in engineered reductive dechlorination [ERD] systems). If there is a strong correlation between datasets, a simple linear relationship can provide meaningful predictions. More sophisticated evaluation techniques are also available and may include statistical analyses (e.g., *t*-test, path analysis, Spearman rank correlations) or data mining. Data mining looks for hidden patterns in data that can be used to predict future behavior and has been used in the fields of communications, insurance, retail, banking, manufacturing, education, and bioinformatics. Techniques such as clustering, principal component analysis, decision trees, Bayesian networks, and neural networks are used in data mining to evaluate relationships among datasets, most significant variables and develop models for prediction. If strong relationships are found between datasets, the correlations and casual relationships can be identified. This can be very powerful in developing strategies for future decision-making. A good example of this in environmental remediation is the reduction in operating and maintenance costs by monitoring only the relevant and required parameters for enhancing the performance and compliance of remediation systems.

What makes the emerging field of *big data* analysis in remediation interesting is that the full value is not always evident. In addition to compiling and maximizing value from existing data, evaluators can integrate new data collected from emerging technologies, industry trends and regulatory, and other recent developments. Data analysis capabilities in the remediation arena has been lagging and a modern approach for exploiting information from multiple data streams is essential for us to extract valuable insights. The remediation industry should begin to embrace techniques derived from web-based analytics, visualization, and other computational tools that deliver clear, concise portrayals of the insights gained from the data. The concept of *big data fusion* provides the foundation for making the most of *big data* and it is more than statistical evaluations. Data fusion requires an understanding of the underlying information contained in the data, so that one can combine different datasets to identify key relationships that lead to game changing insights of the process or concept of interest.

The Universe of Environmental Remediation Big Data

Big data in environmental remediation can come from many sources and may be related to data from a single site (site level), a group of related sites (portfolio level), or across unrelated sites (global level). In some cases, particularly large complex sites, significant amount of data and information may be available collected over many years (decades) and from hundreds of monitoring wells which could be considered as a stand-alone *big data* set. Figure 1 shows the relationship between these different levels of *big data*. Evaluation of site-level datasets can provide insights into the relationships between natural site conditions (e.g., ambient geochemistry, hydrogeologic complexities), engineered systems (e.g., substrate addition), remedy performance (e.g., concentration reductions), and process trouble-shooting (e.g., formation of by-products). Examples of *big data* at the site level are smart characterization data (e.g., high-resolution investigation results), automated sensor data, and large plume remediation data which are discussed in more detail below. *Big data* evaluations at a specific portfolio level (e.g., hundreds of petroleum retail sites) can be used to articulate insights gained and predict performance and conditions across an entire portfolio. Looking at data relationships across unrelated sites (e.g., National Priority

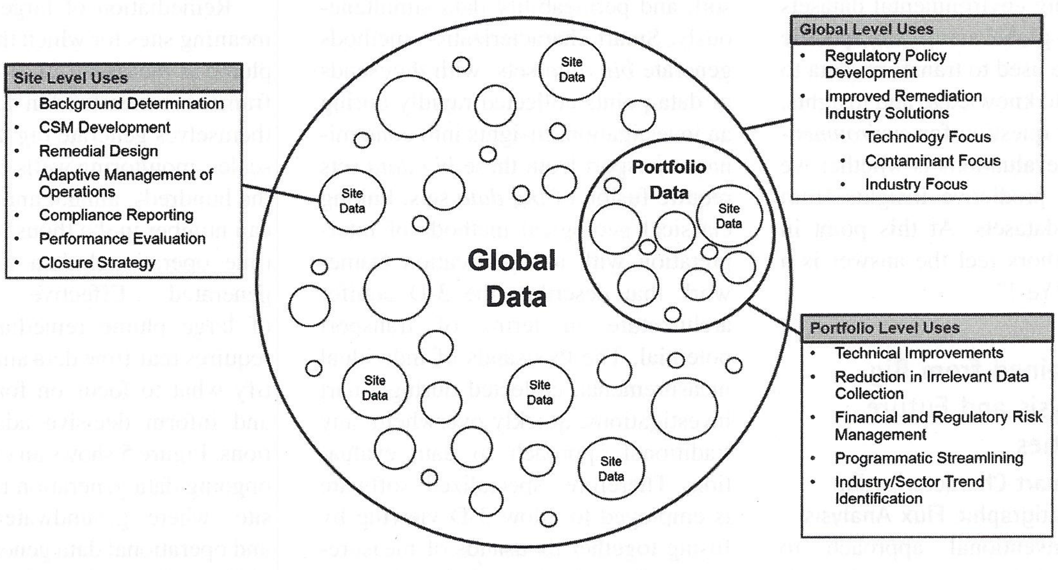

Figure 1. Scales of environmental datasets and uses from data analysis.

List sites) can provide insights related to specific industry-wide trends and practices. Examples of how global level datasets can be used to turn data into insights for improving remediation practices is provided in Table 2. Examples of global datasets used to refine our understanding contaminant mass flux, develop and validate benchmark values, improve light non-aqueous phase liquid (LNAPL) management strategies and understand the implications of bioremediation as a relevant mechanism in natural attenuation of 1,4-dioxane are discussed in more detail below.

Environmental big data evaluations can provide an introspective look into the remediation "black box," remove the dependence and bias toward specific remedial technologies, and provide the technical knowledge and insights for effective decision-making. At the portfolio level, it can lead to an advanced understanding of remediation effectiveness, or areas for remedial optimization, and can provide direction for future efforts within the portfolio. At the global level, these types of evaluations can lead to a paradigm shift in understanding the effectiveness and limitations of specific technologies and thus developing robust technical impracticability guidelines for certain types of complex sites. Predictive insights gained can also help in refinement of existing or development of new regulatory guidelines (e.g., alternate clean-up standards and risk-based regulations). Figure 2 illustrates how analysis of big environmental datasets derived from global sets of site specific data could be used to transform data to information to knowledge and insights. The ultimate question for *environmental big data* evaluations is whether we can develop predictive insights from all existing datasets. At this point in time, the authors feel the answer is a resounding "Yes!"

Insights Gained from Big Data Analysis and Future Opportunities

Site Level Smart Characterization Data for Stratigraphic Flux Analysis

The conventional approach to viewing and interpreting site investigation data relies on two-dimensional data-box figures, which were adequate when characterization datasets only consisted of a small number of permanent monitoring wells. When only groundwater data are considered and interpolated, the resulting picture provides a distorted view of the mass distribution in the source zone as well as the entire plume, because groundwater sampling is inherently biased toward the most permeable zones. We now recognize that our understanding of contaminant transport needs to go beyond the traditional distorted view of mass distribution and include a more detailed analysis of the mass flux, as experience has repeatedly shown that the majority of the mass flux occurs in a small portion of the plume cross-section within the aquifer. As such, characterization methods have to integrate multiple types of data to find the zone where mass flux is occurring.

Recent developments in smart investigations have the potential for providing a systematic approach that focuses on developing a three-dimensional (3-D) interpretation of the aquifer architecture in conjunction with contaminant concentrations for the evaluation of mass flux through the application of *big data* (Figure 3). The use of smart characterization methods, in place of traditional investigation methods, integrate dynamic, real-time, high-density soil, and groundwater sampling data with hydro-stratigraphic interpretations and permeability mapping in three-dimensions including groundwater, whole-core saturated soil, and permeability data simultaneously. Smart characterization methods generate *big data* sets, with thousands of data points collected rapidly during an investigation. Insights into contaminant transport from these *big data* sets require fusion of *big data* sets, linking classical geological methods of interpretation with a classification framework that describes the 3-D aquifer architecture in terms of transport potential. The thousands of individual measurements, collected during smart investigations, quickly overwhelm any traditional approach to data evaluation. Therefore, specialized software is employed to allow 3-D viewing by fusing together thousands of measurements from high-resolution borings (Figure 4). *Big data* fusion is needed because the high-density data collected must be organized into the stratigraphic flux framework, where permeability and concentration data are combined to provide an indication of mass flux. This approach builds on sequence stratigraphy methods and provides insights regarding how permeability distributions within elements of the aquifer architecture control contaminant transport. This process is different than the conventional stratigraphy approach by the fact that the impacted aquifer system can be classified by the relative flux within the different segments.

Past methods of interpolation which ignore the permeability structure will interpolate mass into storage zones even though the mass is present almost exclusively in transport zones. These limitations are overcome using *big data* fusion and interpretation based on concepts of plume maturing. In the source zone, the approach will correctly identify mass present in the slow advection and storage zones, as well as the architecture of the zones. In the downgradient plume, the data fusion approach will correctly identify the mass flux zones where contaminant mass is present. Understanding these key aspects of contaminant transport can be used to develop quantitative metrics to rank and prioritize remediation efforts and provides insights into clean-up time frames and realistic end points.

Evaluation of Site Level Big Data in Effective Large Plume Management

Remediation of large plume sites, meaning sites for which the scale of the plume is measured in miles and time-frames are measured in decades, can in themselves generate *big data*. At these scales, monitoring wells can number in the hundreds, annual analytical counts can number in the thousands, and real-time operational data are constantly generated. Effective management of large plume remediation projects requires real-time data analysis to identify what to focus on for remediation and inform decisive adaptive operations. Figure 5 shows an example of the ongoing data generation from one such site, where groundwater monitoring and operational data generated from 12 different hexavalent chromium remedial systems, including large-scale in

Table 2

Examples of Big Data and Their Evaluations in Environmental Restoration

Reference(s)	Description	Type of Data	Evaluation Approach	Outcome
Arcadis internal evaluation	Catalog measurements of flame retardants in indoor dust in peer-reviewed literature and evaluate spatial and temporal variances	Concentrations of flame retardants in dust samples in a variety of indoor environments	Basic statistics based on different environmental relationships	Scientific study of flame retardants in indoor dust has been increasing rapidly since 2001 with focus on polybrominated diphenyl ethers. Concentrations of polybrominated biphenyl ethers in North American residential settings appear to be decreasing, while concentrations of replacement flame retardant chemicals increase
Schnobrich et al. (2011)	Evaluation of 85 enhanced reductive dechlorination projects	Groundwater concentrations of chlorinated solvents, total organic carbon, pH, and sulfate	Linear regression; correlations	Results indicated that the maximum rate of degradation was independent of total organic carbon concentration as long as reducing conditions dominate and TOC remains approximately 10 mg/L above background
Rice et al. (1995)	Early evaluation of data from 271 sites in California with leaking underground storage tank (LUST) to evaluate petroleum hydrocarbon plume length and stability	Transcription of California state file information from 271 sites into electronic database based on standard checklist	Statistics and nonparametric graphical displays	Early work on plume length and stability that supported the eventual establishment of the low threat closure policy for LUST sites in California in 2012
Adamson et al. (2014) and Adamson et al. (2015)	Evaluation of the co-occurrence and natural attenuation of 1,4-dioxane and chlorinated solvents in groundwater	Groundwater concentrations of chlorinated solvents and 1,4-dioxane	GIS-based algorithms, Wilcoxon rank sum test, linear regression, Spearman rank correlation	Industry-wide paradigm shift in understanding of 1,4-dioxane fate and transport in groundwater
Maguire et al. (2006)	Remediation performance and rebound of intensive source depletion technologies were evaluated where treatment targeted dense nonaqueous phase liquid (DNAPL) source zones	Performance data from 59 sites where chemical oxidation, enhanced bioremediation, thermal treatment, or surfactant/co-solvent flushing were implemented	Quantification and comparison of datasets	Set expectations on achievable percent reduction and permanence of source zone NAPL depletion
Connor et al. (2015)	Review and summary of 13 studies on benzene, MTBE and/or TBA plumes	Compilation of multiple-site datasets from the United States containing data from 22 to more than 4000 sites each on plume length and stability	Statstical analysis	Demonstrated that MTBE, TBA, and benzene plume lengths are comparable at most sites. Found consistency among datasets in different geographic regions and in a variety of hydrogeologic regimes
Lahvis et al. (2013)	Evaluation of soil-gas data at sites impacted with dissolved phase petroleum hydrocarbons to determine vapor intrusion risk	Soil-gas measurements at hundreds of petroleum UST sites spanning a range of environmental conditions, geographic regions, and a 16-year time period (1995 to 2011)	Nonparametric Kaplan–Meier statistics	This is an example of *big data* analysis informing regulatory policy: U.S. EPA Office of Research and Development recently cited conclusions from this study, among others, to provide technical recommendations to the Office of Underground Storage Tanks for the development screening criteria to identify structures that are at risk from petroleum vapor intrusion (Wilson et al. 2012)

Table 2 (Continued)

Reference(s)	Description	Type of Data	Evaluation Approach	Outcome
Microbial Insights Inc. Molecular Biology Results Database	An agglomeration of results generated by Microbial Insights and relevant environmental data from project sites	Constituent concentrations, geochemical data, molecular data (e.g., qPCR DNA results)	In progress	In progress, but could provide predictive values for a variety of microbial genetic targets
Schnobrich et al. (2010)	Comparative analysis between 20+ active enhanced reductive dechlorination sites to evaluate the overall rates of dechlorination and length of remedial time observed at the field scale	Groundwater concentrations of chlorinated solvent, total organic carbon, nitrate, sulfate, and pH	Linear regression; correlations	Identified half-lives of various carbon substrates. Correlated elevated sulfate (more than 1000 milligrams per litre [mg/L], elevated nitrate (more than 500 mg/L), and decreased pH (less than 6 s.u.) with decreased chlorinated solvent degradation half-lives
Arcadis internal evaluation	Evaluate risk from soil gas at 91 fuel service stations with free product	Soil-gas results for TO-15	Comparison to Michigan regulatory standards	Concluded the vapor intrusion pathway does not pose a risk at 90 of the 91 sites. Paradigm shift in ability to close petroleum sites under a risk-based program in Michigan

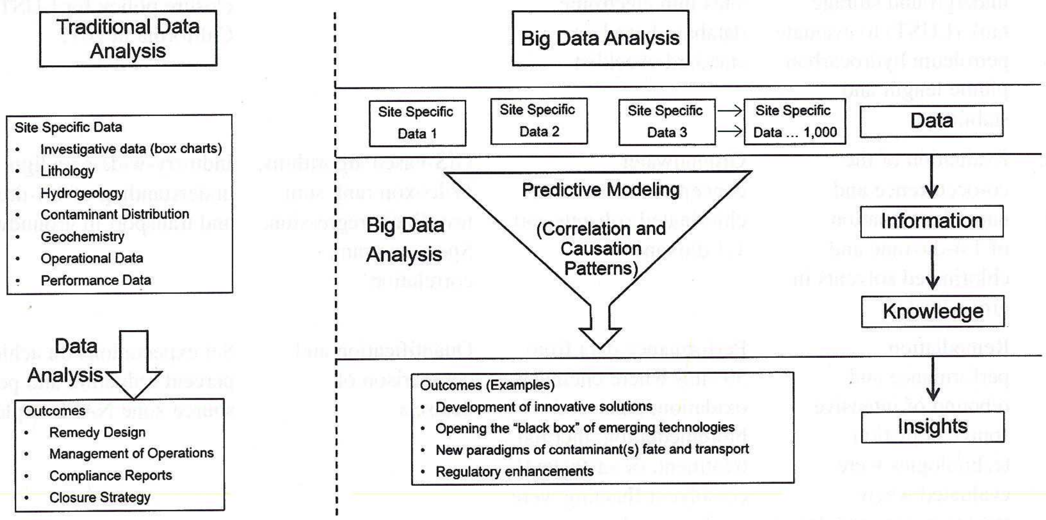

Figure 2. Path of big data analysis in remediation.

situ reactive treatment recirculation systems and groundwater extraction and agricultural treatment systems. The example in Figure 5 represents a small fraction of the data collected at this site, which is approximately 1 mile by 2 miles and includes operation of over 180 remediation wells with data collection from over 700 monitoring wells. The datasets generated are evaluated on a monthly to quarterly frequency to adapt the remedy, guide which injections and extraction wells to operate, make changes to remedial well flow rates and changes to organic carbon injection rates. Advanced data management and analytical techniques could greatly improve the ability to quickly assimilate and analyze large accumulating *big data* at large plume sites, providing insights to more efficiently manage remediation and reduce overall remedial timeframes.

Analysis of Global Datasets for Developing and Validating Benchmarks

Benchmarking within the context of remediation practices can be interpreted as improving the performance of technologies by continuously identifying, understanding, and adapting new information and knowledge for developing best practices and processes. Many times *big data* evaluation can provide the benchmarks, but more importantly can validate the benchmarks on a continuous basis. For example, after the development of molecular targets to quantify the subsurface population of *Dehalococcoides* (DHC) microbial species, project teams started collecting environmental samples at individual sites. The commonly accepted benchmark of 1×10^7 cells/L for a healthy microbial population of DHC capable of complete dechlorination came from academic research (Lu et al. 2006). In the case of DHC, it will be worthwhile to conduct a *big data* analysis to correlate

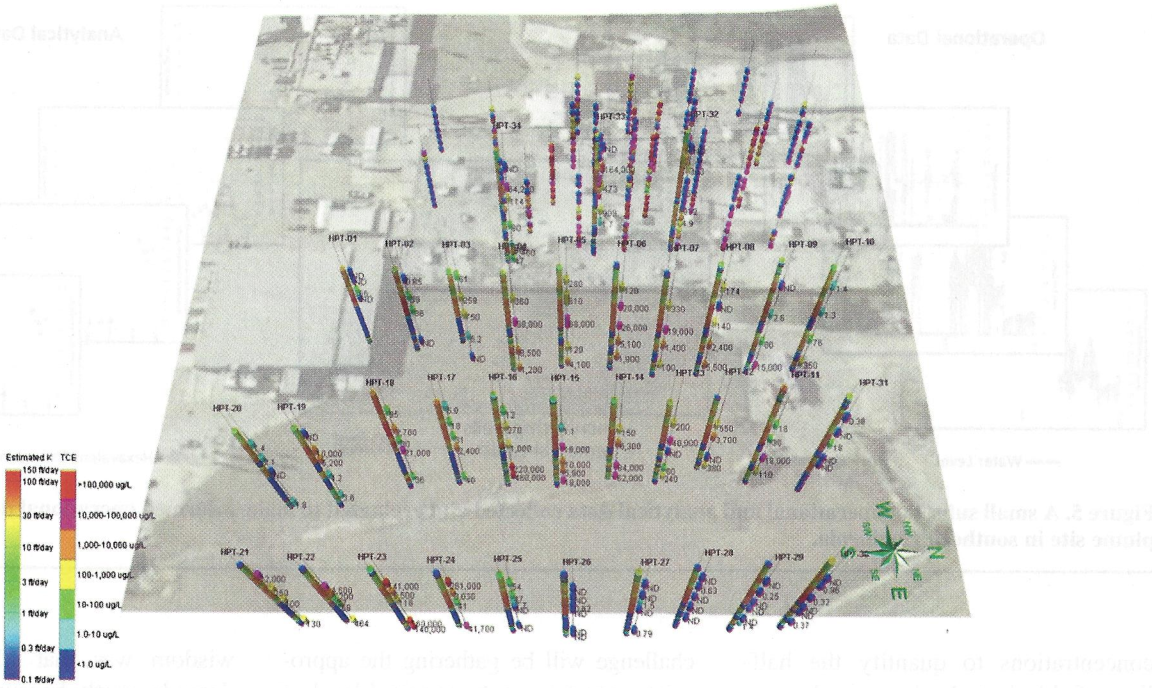

Figure 3. Real-time smart characterization results at a chlorinated solvent site. The color-ramp columns at the left of each sounding depict relative hydraulic conductivity (K) derived from HPT; the color ramp symbols to the right show TCE concentrations measured through vertical aquifer profile groundwater sampling. Nearly 200 groundwater samples and 400 soil samples were combined with more than 10,000 HPT estimates of relative K to generate the relative flux map shown below.

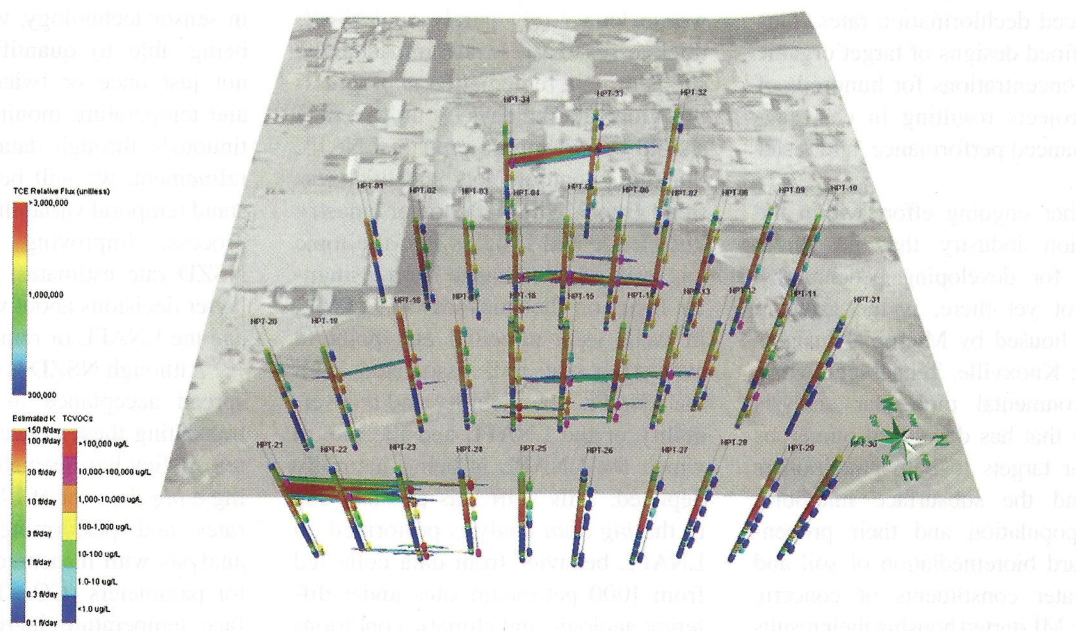

Figure 4. Relative flux results based on the product of relative hydraulic conductivity and TCE groundwater concentrations. The color-ramp shaded areas on the flux transects high-light more than 90% of the relative flux at the site.

these population counts with dechlorination performance across hundreds of sites. The insights gained either would validate or refute the commonly accepted benchmark. It will be a worthwhile effort to refine and reinforce this benchmark over time by evaluation of *big data* from multiple sites with successful enhanced reductive dechlorination systems.

Another example of benchmarking is from the insights gained from analysis of performance data from more than 100 ERD sites. The authors performed a *big data* evaluation of these ERD systems by evaluating chlorinated solvent concentrations, total organic carbon, pH, and sulfate

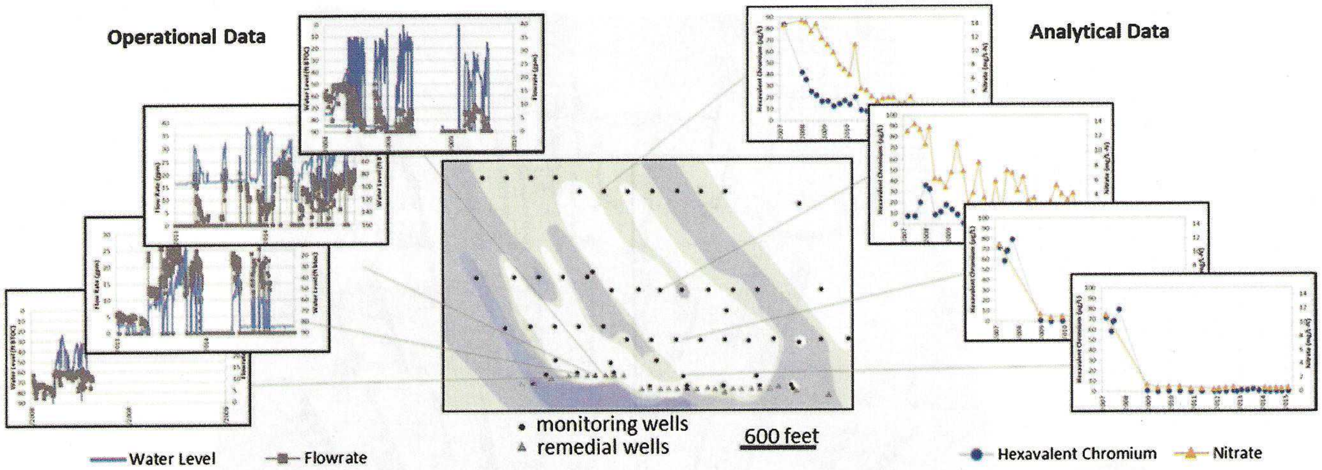

Figure 5. A small subset of operational and analytical data collected and evaluated to make informed operational decisions at a large plume site in southern California.

concentrations to quantify the half-lives of chlorinated solvents and assess what factors affect the first order decay rate (Schnobrich et al. 2010, 2011). The rate of chlorinated solvent degradation did not exhibit a linear relationship with total organic carbon concentration; however, a threshold of 10 mg/L above background was related to enhanced dechlorination rates. This led to refined designs of target organic carbon concentrations for hundreds of future projects resulting in cost savings, enhanced performance, and faster cleanups.

Another ongoing effort within the remediation industry that has great potential for developing benchmarks, but is not yet there, is the growing database housed by Microbial Insights Inc. (MI; Knoxville, Tennessee). MI is an environmental molecular analysis company that has developed numerous molecular targets to help practitioners understand the subsurface microbiological population and their propensity toward bioremediation of soil and groundwater constituents of concern. Recently, MI started housing their results in a relational database and asking practitioners to help provide additional site data (e.g., constituent concentrations, geochemistry, and performance data). With these datasets together, they may be able to identify relevant benchmarks (similar to the DHC example above). This would allow practitioners to apply future microbial results more directly in development of site conceptual models and remedial technology screening. The challenge will be gathering the appropriate raw data, at the required level of quality, from practitioners across the industry on a continuing basis.

Analysis of Global Datasets to Improve LNAPL Management

How we view and manage sites with LNAPL has changed recently, and we no longer rely purely on LNAPL thicknesses to assess how impacted the site is or how remediation is progressing. Gone are the days (or at least they should be) where we prepare LNAPL thickness contour plots and thickness trend graphs. The evolution of industry knowledge and insights into the topic has resulted in a change from a singular focus on the thickness of LNAPL in wells as a remedial end point to evaluating the risk associated with the LNAPL, the mobility and recoverability of the LNAPL and the rate at which the LNAPL is being naturally depleted. This shift was possible due to the *big data* analysis performed on LNAPL behavior from data collected from 1000 petroleum sites under different geologic and climatic conditions across the country and the insights gained from that analysis.

The most important insight gained from the *big data* analysis is that the mass removed by natural source zone depletion (NSZD) is far greater than the mass recoverable by any of the most efficient removal technologies. It is also apparent that the rate of LNAPL depletion in the subsurface has been greatly underestimated. Conventional wisdom was that LNAPLs did not degrade, partly because it was thought that the LNAPL conditions impeded degradation. This paradigm shifted as correlations were made between the generation of carbon dioxide (CO_2), elevated temperatures (Figure 6), and microbial degradation of LNAPL. As a result of *big data* fusion and advances in sensor technology, we can envision being able to quantify NSZD rates not just once or twice through CO_2 and temperature monitoring, but continuously through data logging. With refinement, we will be able to understand temporal variability of the NSZD process, improving the reliability NSZD rate estimates, and then make better decisions about whether to manage the LNAPL or remediate it.

Although NSZD is receiving widespread acceptance in concept, fully translating these concepts into regulatory policy has been slow. Accumulating a *big data* set of LNAPL depletion rates and performing a correlation analysis with measured NSZD indicator parameters (CO_2 flux and subsurface temperature increase) is crucial for the next step. Developing an established methodology, from the insights gained, to quantify the NSZD rates for the regulatory community will increase the strength of these arguments and inclusion as an acceptable technology in lieu of active remediation. It has been rewarding to participate in the industry-wide effort to make a step change in how we manage LNAPL sites.

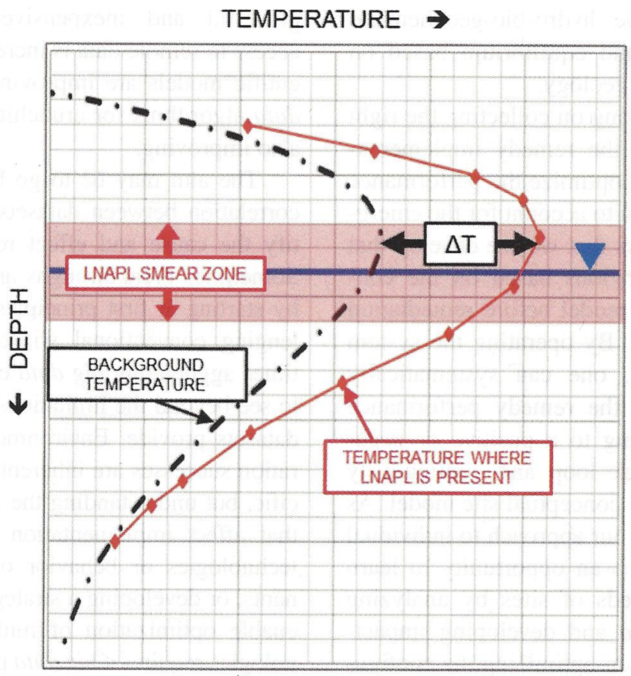

Figure 6. Conceptual model of subsurface temperature increase within the LNAPL zone. The temperature profile data are cost-effective to collect at high resolution and over time. Current initiatives are ongoing to develop reliable methods that convert the temperature signal into a rate of petroleum depletion via calculation of the biogenic heat generated during hydrocarbon degradation and assessment of thermodynamic heat flux from zones in which LNAPL and soil-gas methane are observed.

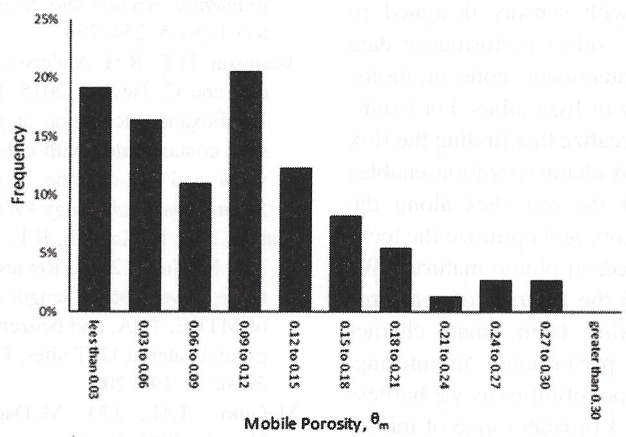

Figure 7. Frequency distribution of mobile porosity (θ_m) values from tracer tests completed at dozens of different sites across the United States.

Evaluation of Global Datasets of Tracer Study Data Provides Insight into Realistic Values of Porosities

We recognized early on that conventional hydrogeologic approaches to designing in situ remediation systems, particularly those relying on reagent injections, was not reliable. We came to this understanding through a review of many in situ chemical oxidation (ISCO) and ERD projects and through performance monitoring and optimization of systems. For a long time, practitioners relied on design assumptions that incorporated the concept of effective porosity, often assumed to be 20%, and transport based on the advection–dispersion equation. The industry was faced with reagent vendors claiming that 5 gallons of reagent per foot of injection well would lead to widespread reagent distribution and enable injection well spacings in excess of 30 and even 50 feet. Many early in situ remedy designs applied this high concentration, low-volume injection strategy with less than satisfactory results. Often we would see little indication of the injected reagent or resulting degradation reactions in downgradient performance monitoring wells in addition to other adverse impacts.

To understand the disparity between vendor claims and observed results, we undertook a review of a global set of performance data from dozens of sites and compared the results to first principles on hydraulics and contaminant transport. Based on this analysis, we recognized that injection tracer testing was essential to actually measure the volume–radius relationship between injected reagents and radius of influence and then determining what degree of post injection expansion could reliably be obtained during the drift phase. We would also monitor tracer concentrations at wells downgradient to verify flow direction and transport velocity, based on breakthrough curve analysis. The observed focusing of flow within the permeable zones of the aquifer led to observations that were consistent with transport zone porosities of 5% to 10% at most; many sites exhibited transport zone porosities that were less than 5% (Figure 7). These insights gained from tracer testing data from hundreds of sites were documented (Payne et al. 2008), and we adopted the term mobile porosity, owing to the fact that majority of the transport occurs in a small segment of the aquifer. Through tracer testing, we have developed a new paradigm on in situ remediation via injection.

Understanding Biodegradation as a Relevant Mechanism in Natural Attenuation of 1,4-Dioxane

Our experience with 1,4-dioxane natural attenuation provides a great example of how *big data* analytics can provide insights, if not paradigm shifting perspectives on remediation of emerging contaminants. Until recently, conventional wisdom has been that 1,4-dioxane is not subject to the biodegradation mechanisms of natural attenuation. But as we analyze more and more data from dozens of sites, it is obvious that stable, if not receding, 1,4-dioxane plumes are identified in our portfolio of sites.

With the development of applicable molecular biology tools, we can begin to connect decreasing concentrations and geochemical data with potential biodegradation mechanisms at play in the subsurface. With the help of commercial molecular labs, like MI, we can understand the relative contributions of metabolic biodegradation of 1,4-dioxane under aerobic conditions and cometabolic biodegradation of 1,4-dioxane in the presence of appropriate substrates (such as methane, ethane or propane) using monooxygenase enzymes.

However, there are still challenges in understanding why 1,4-dioxane biodegradation appears to be a relevant mechanism at some sites, but not others, and which substrate(s) or functional monooxygenase enzymes are driving 1,4-dioxane biodegradation. Where 1,4-dioxane degradation is apparent, the mechanism remains unclear. For sites with no expression of 1,4-dioxane biodegradation, it may be that the particular monooxygenase enzymes capable of 1,4-dioxane oxidation under environmental conditions have not yet been characterized, or the appropriate genetic targets have not been established. We think that data fusion across many sites will provide answers. By digging deeper into the active microbial communities at many sites we can identify which monooxygenase enzymes are responsible and how they can be enabled, allowing more 1,4-dioxane sites to be cost-effectively managed with natural and/or enhanced attenuation.

The Road Ahead

A robust debate is ensuing over whether *big data* will, or should, result in a decision-making paradigm shift that emphasizes correlation identification over causation analysis. It can be argued that correlation is an adequate basis for action in many situations because correlations can be found much faster and cheaper than causation. Many also think that for most every day needs, knowing *what* and not *why* is good enough. A real-world example of this could be the propensity for people to want to understand what causes by-products formation within ERD reactive zones in certain geographic locations without wanting to know the hydro-bio-geochemical conditions and equilibrium based on the regional geology.

By focusing on collecting the right data during the remedy implementation, we can optimize the performance of the system to account for the emerging behaviors that we see develop that were not obvious based on the conceptual site model before remediation commenced. By operating the system dynamically, one can systematically learn about the remedy performance by responding to a positive or negative feedback loop and continuously calibrate the conceptual site model. As we improve our approach to individual sites, there is an opportunity to learn from hundreds of sites by analyzing the *big data* and developing impactful insights on optimizing the configuration and operation of remediation systems. This could lead to the development of expert systems which could break down the barriers specialized knowledge and provide decision support to practitioners at multiple levels.

The next frontier in remediation can be harnessing the power of automated monitoring with sensors designed to continuously collect performance data such as contaminant concentrations, geochemistry or hydraulics. For example, now we realize that finding the flux through smart characterization enables one to tailor the remedies along the plume trajectory and optimize the technologies based on plume maturity. We could extend the benefits of real-time decision-making from smart characterization to performance monitoring. Imagine the possibilities as we harness the power of a broader range of microsensors capable of transmitting data on contaminant concentrations and reagent distribution during in situ remediation? Or, what if we were able to track targeted microbial populations and growth rates without the time consuming specialized laboratory analyses? Or, what if we mixed virtual reality solution with high-resolution characterization that would allow all stakeholders to take a site walk from anywhere in the world? Streaming technology can take the concept of investigation and performance data collection and evaluation to new dimensions by developing instrumentation from the macrolevel down to the microlevel. Today's sensors are powerful and inexpensive, network access to remote data is increasing, scientific models are improving, and *big data* algorithms for crunching data are also improving.

The aim may be to go beyond the correlation between datasets and identify the cause and effect relationship. Sometimes breakthroughs are obtained by starting at first principles and challenging conventional wisdom. Sometimes aggregating *big data* enables one to see beyond the limitations that small datasets provide. Environmental restoration successes are inherently site specific, but understanding the key drivers that affect implementation of various technologies or behavior of contaminants, or developing a strategy that will enable optimization of multiple technologies requires a *big data* perspective.

References

Adamson, D.T., S. Mahendra, K.L. Walker Jr., S.R. Rauch, S. Sengupta, and C. Newell. 2014. A multisite survey to identify the scale of the 1,4-dioxane problem at contaminated groundwater sites. *Environmental Science and Technology Letters* 1, no. 5: 254–258.

Adamson, D.T., R.H. Anderson, S. Mahendra, and C. Newell. 2015. Evidence of 1,4-dioxane attenuation at groundwater sites contaminated with chlorinated solvents and 1,4-dioxane. *Environmental Science and Technology* 49: 6510–6518.

Connor, J.A., R. Kamath, K.L. Walker, and T.E. McHugh. 2015. Review of quantitative surveys of the length and stability of MTBE, TBA, and benzene plumes in groundwater at UST sites. *Groundwater* 53, no. 2: 195–206.

McGuire, T.M., J.M. McDade, and C. Newell. 2006. Performance of DNAPL source depletion technologies at 59 chlorinated solvent-impacted sites. *Groundwater Monitoring and Remediation* 26, no. 1: 73–84.

Lahvis, M.A., I. Hers, R.V. Davis, J. Wright, and G.E. DeVaull. 2013. Vapor intrusion screening at petroleum UST sites. *Groundwater Monitoring and Remediation* 33, no. 2: 53–67.

Lu, X., J.T. Wilson, and D.H. Kampbell. 2006. Relationship between Dehalococcoides DNA in ground water and rates of reductive dechlorination at field scale. *Water Research* 40, no. 16: 3131–3140.

Payne, F.C., J.A. Quinnan, and S.T. Potter. 2008. *Remediation Hydraulics*. Boca Raton, Florida: CRC Press.

Rice, D.W., R.D. Grose, J.C. Michaelsen, B.P. Dooher, D.H. MacQueen, S.J. Cullen, W.E. Kastenberg, L.G. Everett, M.A. Marino. 1995. California leaking underground fuel tank (LUFT) historical case analyses, UCRLAR-122207. Livermore, California: Lawrence Livermore National Laboratory (LLNL).

Schnobrich, M., D. Nelson, M. Logan, A. Auffermann, and J. Martin. 2010. Field-scale correlations between degradation rates, TOC and biogeochemistry in ERD. Battelle Conference 2010, Platform Presentation, Monterey, California.

Schnobrich, M., D. Nelson, M. Logan, A. Auffermann, and J. Martin. 2011. Optimized design of large plume enhanced reductive dechlorination (ERD) systems: an analysis of 85 sites. RemTEC 2011, Platform Presentation , Chicago, Illinois.

Wilson, J.T., J.W. Weaver, and H. White. 2012. An approach for developing site-specific lateral and vertical inclusion zones within which structures should be evaluated for petroleum vapor intrusion due to releases of motor fuel from underground storage tanks. Groundwater Issue.

Biographical Sketches

S.S. Suthersan, *Ph.D., P.E., Corresponding Author, is Executive Vice President and Chief Technical Officer at ARCADIS N.A., 10 Friends Lane, Suite 200, Newtown, PA 18940; (267) 685-1800; suthan.suthersan@arcadis.com*

M. Gentile, *Ph.D., P.E., is a Principal Engineer at ARCADIS N.A., 100 Montgomery Street, Suite 300, San Francisco, CA 94104; (415) 432-6920; margaret.gentile@arcadis.com*

C. Bell, *P.E., is a Senior Engineer at ARCADIS N.A., 100 Montgomery Street, Suite 300, San Francisco, CA 94104; (415) 432-6944; caitlin.bell@arcadis.com*

J. Quinnan, *P.E., P.G., is Senior Vice President and Director of Site Investigation Practice at ARCADIS N.A., 10559 Citation Drive, Suite 100, Brighton, MI 48116; (810) 225-1943; joseph.quinnan@arcadis.com*

J. Horst, *P.E., is Executive Vice President of Technical Services at ARCADIS N.A., 10 Friends Lane, Suite 200, Newtown, PA 18940; (267) 685-1800; john.horst@arcadis.com*

Groundwater
Monitoring&Remediation

Evaluation of Long-Term Performance and Sustained Treatment at Enhanced Anaerobic Bioremediation Sites

by Travis M. McGuire, David. T. Adamson, Michael S. Burcham, Philip B. Bedient, and Charles J. Newell

Abstract

This study evaluated the long-term performance of enhanced anaerobic bioremediation (EAB) at chlorinated solvent sites to determine if sustained treatment processes were helping to prevent concentration rebound. A database of groundwater concentration versus time records was compiled for 34 sites, with at least 3 years of posttreatment monitoring data (median = 4.7 years, range = 3.0 to 11.7 years). Long-term performance was evaluated based on order-of-magnitude (OoM) changes in parent compound concentrations during various monitoring periods. Results indicate that, relative to the pretreatment concentration, a median concentration reduction for all 34 sites of 1.0 OoM (90% reduction) was achieved by the end of the posttreatment monitoring period. No rebound was observed at 65% of the sites between the first year of posttreatment monitoring and the final year. During this posttreatment period, Mann-Kendall trend analysis indicated that the concentration was stable or decreasing at 89% of the sites where a trend could be established ($n = 27$; 33% decreasing, 56% stable, 11% increasing). Statistical analysis indicates there is no evidence that the distribution of median concentration reductions after the first year of posttreatment monitoring was different than the distribution of median reductions 2 to 11 years later at the end of the monitoring period ($p = 0.67$). Similarly, statistical analysis indicates that there is no evidence that the distribution of median reductions for a larger set of sites ($n = 84$) with less than 3 years of posttreatment monitoring data (1.1 OoM; 92% reduction) was different than the distribution of median OoM reductions for the 34-site dataset with longer monitoring periods ($p = 0.80$). This suggests that, at a typical site, a 3-year monitoring period should be sufficient for evaluating performance. The results of this study indicate that, in the long term, after the end of active treatment, sustained treatment processes contribute to relatively modest concentration reductions but do mitigate rebound at the majority of EAB sites.

Introduction

Managing sites with contaminated soil and groundwater frequently requires a decision about whether remediation is necessary, and if so, what type of technology is appropriate to address site-specific risks. Given this, there is considerable interest in understanding how well various treatment technologies will perform, both in terms of making progress toward cleanup objectives and in performing in a cost-effective manner (Stroo et al. 2003; McDade et al. 2005; McGuire et al. 2006; Triplett Kingston et al. 2010; National Research Council (NRC) 2012; Leeson et al. 2013). For sites where chlorinated volatile organic compounds (CVOCs) are the primary contaminant, there are a number of widely-used technologies for in situ treatment of source zone groundwater (NRC 2012; US EPA 2013). This includes enhanced anaerobic bioremediation, chemical oxidation or reduction, and thermal treatment. Most of these technologies are fairly mature in the sense that extensive guidance and vendor support exists for those who wish to implement them (Leeson et al. 2013). There is also a relatively high level of regulatory comfort with these technologies, and one or more is commonly included in feasibility studies for treating CVOCs in groundwater (US EPA 2013). As a result, these technologies have been implemented at a large number of sites in the past several decades, and the monitoring data collected during the posttreatment period provide a means for assessing long-term performance.

Multi-site evaluations of posttreatment monitoring data have highlighted that the concentration reductions achieved during active treatment are not always sustained. At some sites, a rebound in aqueous-phase concentrations is observed in the period after treatment is over (McGuire et al. 2006; Mundle et al. 2007; Krembs et al. 2010; Scheutz et al. 2010; Hadley and Newell 2012). McGuire et al. (2006) reported that 25% of sites where a source depletion technology was

doi: 10.1111/gwmr.12151