Journal of Contaminant Hydrology 193 (2016) 74–85

Contents lists available at ScienceDirect

Journal of Contaminant Hydrology

journal homepage: www.elsevier.com/locate/jconhyd

Contaminant point source localization error estimates as functions of data quantity and model quality

Scott K. Hansen*, Velimir V. Vesselinov

Computational Earth Science Group, Earth and Environmental Sciences Division (EES-16), Los Alamos National Laboratory, Los Alamos, NM87545, United States

ARTICLE INFO

Article history:
Received 5 April 2016
Received in revised form 23 August 2016
Accepted 6 September 2016
Available online 9 September 2016

Keywords:
Solute transport
Inverse problems
Environmental forensics
Source identification
Error quantification
Model error

ABSTRACT

We develop empirically-grounded error envelopes for localization of a point contamination release event in the saturated zone of a previously uncharacterized heterogeneous aquifer into which a number of plume-intercepting wells have been drilled. We assume that flow direction in the aquifer is known exactly and velocity is known to within a factor of two of our best guess from well observations prior to source identification. Other aquifer and source parameters must be estimated by interpretation of well breakthrough data via the advection-dispersion equation. We employ high performance computing to generate numerous random realizations of aquifer parameters and well locations, simulate well breakthrough data, and then employ unsupervised machine optimization techniques to estimate the most likely spatial (or space-time) location of the source. Tabulating the accuracy of these estimates from the multiple realizations, we relate the size of 90% and 95% confidence envelopes to the data quantity (number of wells) and model quality (fidelity of ADE interpretation model to actual concentrations in a heterogeneous aquifer with channelized flow). We find that for purely spatial localization of the contaminant source, increased data quantities can make up for reduced model quality. For space-time localization, we find similar qualitative behavior, but significantly degraded spatial localization reliability and less improvement from extra data collection. Since the space-time source localization problem is much more challenging, we also tried a multiple-initial-guess optimization strategy. This greatly enhanced performance, but gains from additional data collection remained limited.

1. Introduction

Inverse problems of contaminant source identification are an essential part of environmental engineering practice, relevant to both design of remediation schemes and assignment of responsibility. A goal of the inverse analysis might be, for example, to determine the location of a source, its time of release, or both, based on measurements downgradient of the source. This problem is confounded by two factors: the subsurface is a highly heterogeneous environment, and it is also an information-poor one, in which the heterogeneity is inevitably only partially characterized. Thus, even if it were possible computationally to model the subsurface at a high resolution, data would not constrain the model. As a consequence, in practice one is always attempting to estimate quantities of interest (along with a number of nuisance parameters), using a model that is simplified relative to reality. A schematic of this situation is shown in Fig. 1. In this regard, inverse analysis in contaminant hydrogeology is converse to the situation in a number of other disciplines in which a process model is assumed to be reliable, but data to be limited, poor and "noise-corrupted". Here, measurements are comparatively accurate, but the assumed process model is at best a gross simplification. The practitioner's hope is that, by collecting more data, a more accurate prediction can be made, even though all data will be interpreted through a systematically incorrect model. Given that subsurface contamination puts human health at risk and costs for those found liable for remediation may be large, it appears important to not only make optimal predictions, but to understand of how severe the errors in these predictions may be, given a certain amount of data. Looked at another way: we may want to understand the marginal value of further data collection expense; how much will this reduce uncertainty, and will this be worth the cost?

In light of the importance of inverse analysis to contaminant hydrogeology, many authors have attempted to address aspects of the problem, using a variety of methods. These techniques notably include classic regularization methods (e.g. Skaggs and Kabala, 1994), statistical methods (e.g. Snodgrass and Kitanidis, 1997), and nonlinear simulation-optimization methods (e.g. Mahar and Datta, 2000). While a full review of methods employed for this problem is out of scope, the reader is referred to the survey paper by Bagtzoglou and Atmadja (2005), and to Table 1 of Michalak and Kitanidis (2004) for summary of what had been accomplished as of the

* Corresponding author.

http://dx.doi.org/10.1016/j.jconhyd.2016.09.003
0169-7722/Published by Elsevier B.V.

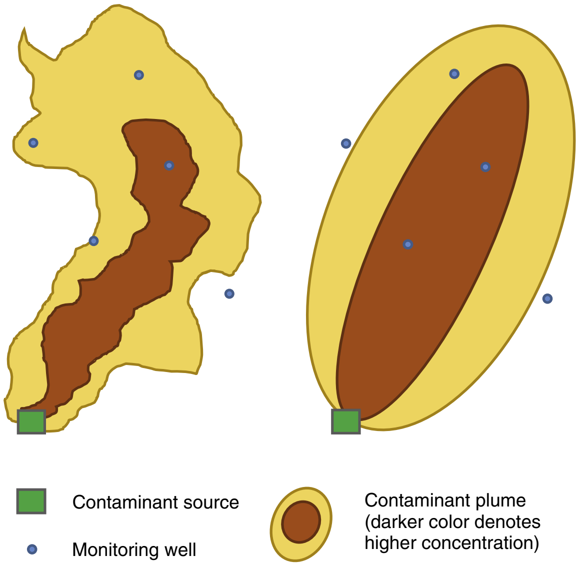

Fig. 1. Schematic diagram comparing the true contaminant plume developing in a heterogeneous environment (left) with a potential best-fit plume (right) generated by assuming subsurface transport is described by an advection-dispersion equation with spatially homogeneous parameters.

middle of the last decade. In subsequent literature, broadly the same types of inverse techniques have been used, although a notable recent conceptual development is the introduction of Markov Chain Monte Carlo (MCMC) methods to the source identification problem by Hazart et al. (2014) and Zhang et al. (2015). As indicated by Michalak and Kitanidis, much of the early literature was focused on identification of contamination histories at known-location point-sources given transport in previously-characterized homogeneous flow fields. In recent literature, simulation-optimization methods have gained prominence, as more complicated scenarios featuring, e.g., multiple dimensions, potentially unknown source locations, and flow-field uncertainty, have come to be considered (Alapati and Kabala, 2000; Aral et al., 2001; Ayvaz, 2010; Bashi-Azghadi et al., 2010; Datta et al., 2009; Guan et al., 2006; Jha and Datta, 2013; Mahar and Datta, 2001; Yeh et al., 2007).

Error estimation has also been considered in the literature. To some extent, analytical adjoint techniques (Cheng and Jia, 2010; Huang et al., 2008; Milnes and Perrochet, 2007; Neupauer and Lin, 2006; Neupauer and Wilson, 1999, 2005), and their particle tracking analogs (e.g. Bagtzoglou et al., 1992) directly solve for uncertainty estimates, but only to the extent that all uncertainty is captured by a Fickian dispersion overlain on a known velocity field. Statistically-oriented methods (Michalak and Kitanidis, 2004; Snodgrass and Kitanidis, 1997; Wagner, 1992; Wagner and Gorelick, 1986; Woodbury et al., 1998) incorporate a covariance matrix for the parameters, and from its diagonal entries produce confidence intervals, assuming independent Gaussian errors. However, this is assumed known a priori, and methods are not given for grounding this covariance matrix in physics. Similarly, Bayesian techniques (e.g., Hazart et al., 2014; Koch and Nowak, 2016), generate a posterior probability distributions on the parameter of interest, which may be considered as error envelopes on maximum a posteriori point estimates.

Despite the large literature on optimal identification, as well as uncertainty analysis once an error structure has been posited, there appears to be comparatively little in the literature regarding the development of error bounds from the interplay of physics, model and data inaccuracies. In our review, we found only the following papers addressing by parametric study the connection between data quality and prediction error: Skaggs and Kabala (1998) considered the recovery of an upgradient contaminant impulse from downgradient point breakthrough in a 1D advective-dispersive transport problem. They considered how signal strength and noise level combined to affect source history identification accuracy. A simulation-optimization study by Datta et al. (2009) considered how fixed-noise-level head and concentration measurement errors

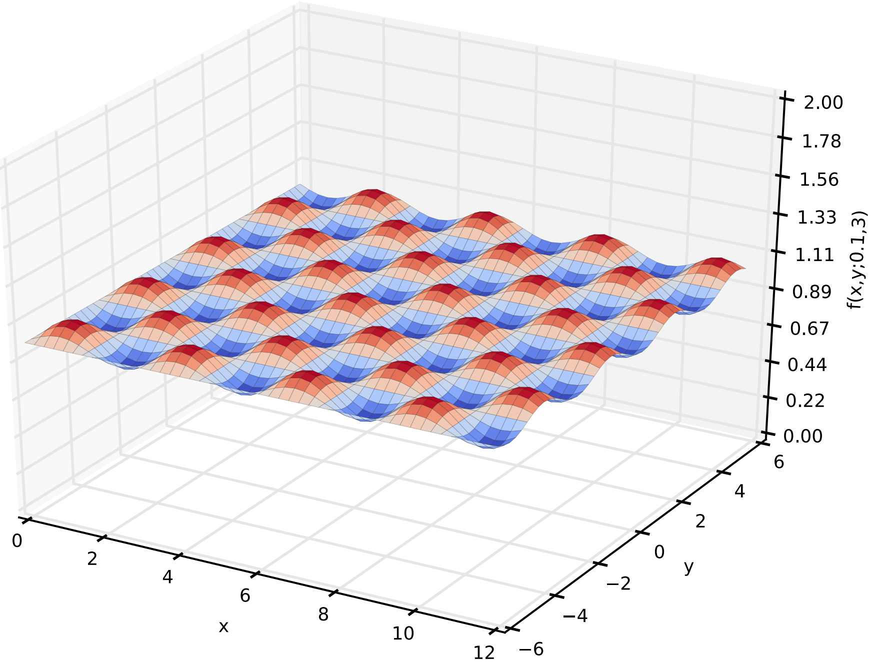

Fig. 2. Plot of the fluctuation (model infidelity) function, f, with parameters $L = 3$ m, $m = 0.1$.

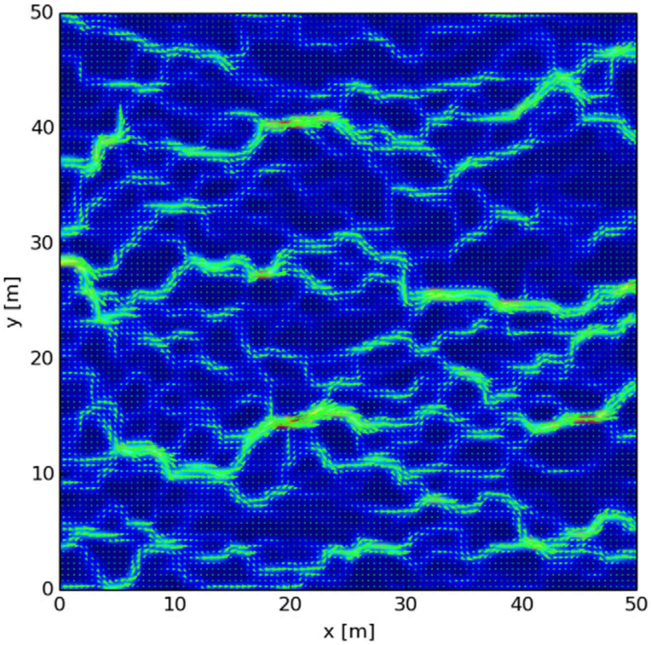

Fig. 3. Example flow field induced by mean head gradient in the negative *x*-direction across a heterogeneous hydraulic conductivity field (multi-Gaussian with $\sigma^2_{\ln K} = 2$), illustrating semi-regular flow channeling. Each green arrow reflects the flow direction and magnitude at its base.

combine to affect source-strength estimation error statistics for three levels of subsurface parametric uncertainty. (The number of realizations used was not revealed.) A similar study (Datta et al., 2011) instead considered propagation of concentration measurement errors for five combinations of known porosity and longitudinal dispersivity through to source strength estimation error. (Ten to twenty realizations were used for each of the five scenarios.) Neither of these two studies considered the impact of different noise levels the way that Skaggs and Kabala (1998) had, and both used fixed numbers of monitoring wells. Jha and Datta (2013) qualitatively considered the accuracy of source identifications from five different arbitrary configurations of five monitoring wells, but abstracted no general principles and considered no other numbers of wells.

Data quantity does not appear to have been considered explicitly in any study we found, and model quality/infidelity was only explored as outlined in the previous paragraph: by treatment as random, non-systematic noise. In light of the importance of physically-grounded error analysis for source identification, it appears timely to directly consider quantification of the combined effects of model infidelity and data quantity in the context of a quasi-realistic solute transport model. This is the topic of our paper.

Naturally, there are a huge amount of possible forensic contaminant transport problems that environmental engineers may face in practice, and it is not possible to address all of them systematically in one place. For our purposes, we select a simple problem: the space-time localization of a point contaminant source in a 2D aquifer with simple heterogeneity, by means of an advection-dispersion equation (ADE) model and breakthrough curves from a number of randomly located wells intersecting the plume.

2. Methodology

We perform three separate, related source localization studies:

1. Purely spatial localization, employing plausible single initial parameter guesses and a local optimization algorithm.
2. Spatio-temporal localization, employing plausible single initial parameter guesses and a local optimization algorithm.
3. Spatio-temporal localization, employing a pseudo-global search algorithm (i.e., multiple random initial guesses).

2.1. General configuration

All our Monte Carlo analyses involve repeated random generation of 2D “aquifer” realizations with different transport parameters

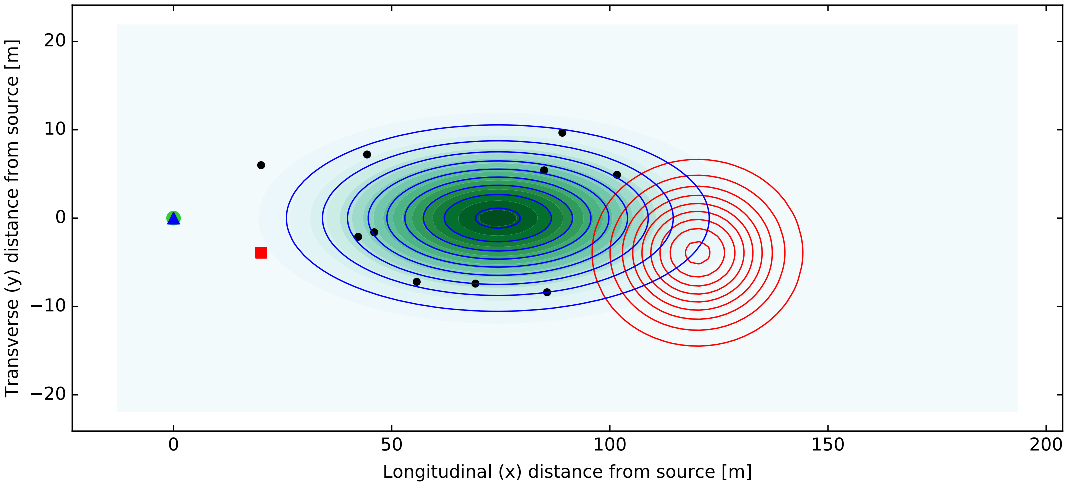

Fig. 4. Results of a single optimization run, with no model infidelity ($m = 0$), and 10 wells, which was optimally successful (all parameters were correctly estimated to several significant figures: $x_s = 0.00\text{m}, y_s = 0.00\text{m}, v = 1.487\text{m/y}, \alpha_l = 3.347\text{m}, \alpha_t = 0.135\text{m}$). The plume described by the initial parameter guess is indicated by the red square (release location at $t = 0$ y), and the red contours (plume extent at $t = 50$ y). The plume described by the optimized parameters is indicated by the blue triangle (release location at $t = 0$ y), and the blue contours (plume extent at $t = 50$ y). The true plume is indicated by the green circle (release location at $t = 0$ y), and elsewhere its concentration at $t = 50$ y is illustrated by the green intensity. Black points indicate monitoring well locations. Note unequal axis scales.

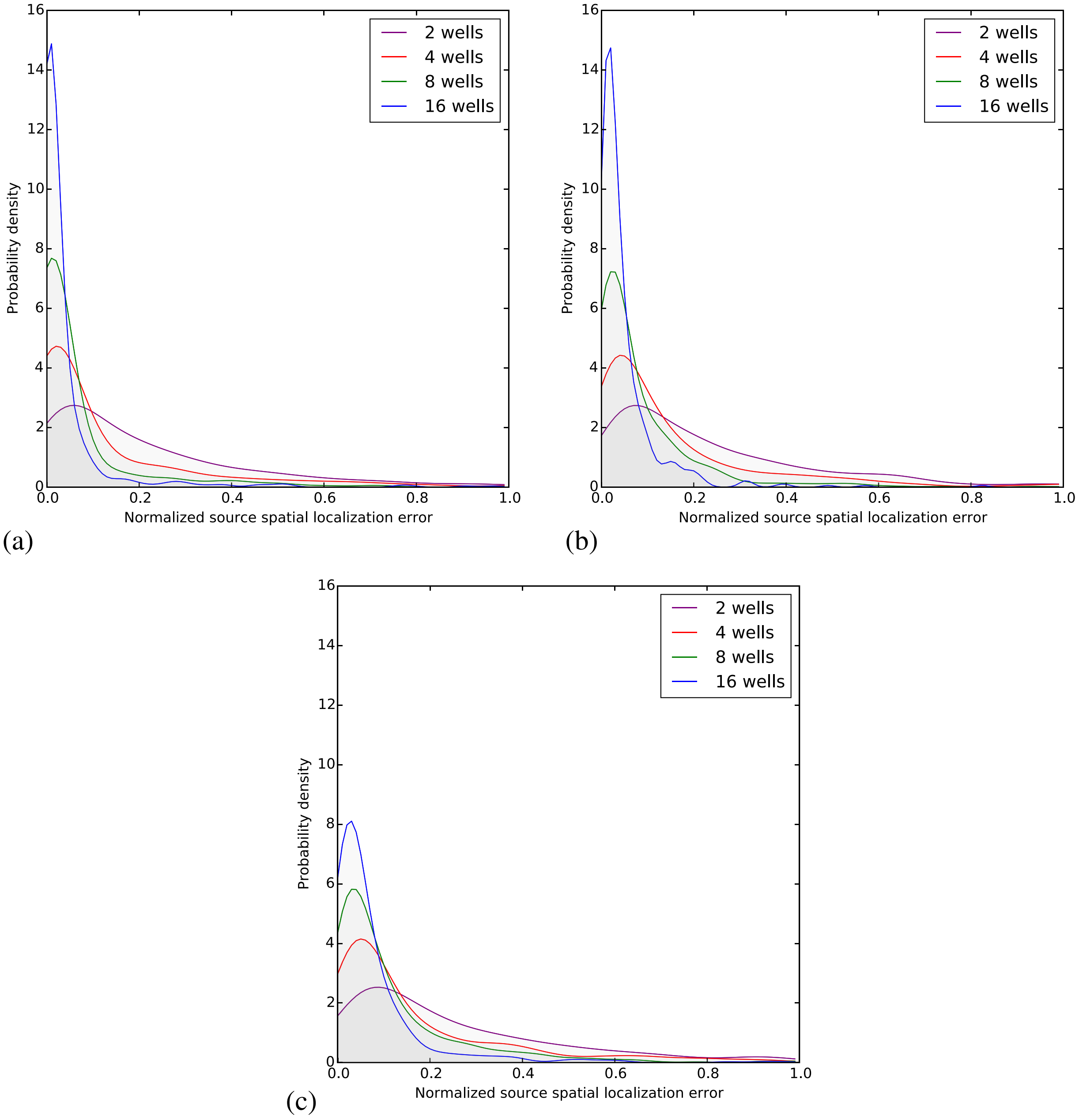

Fig. 5. Empirical probability distribution functions for the normalized source spatial localization error, ϵ (Eq. (10)), for given numbers of available well breakthrough curves, for each of three degrees of model infidelity: (a) $m = 0.1$, (b) $m = 0.5$, (c) $m = 0.9$.

and well locations. The random generation of realizations is performed automatically via scripts written in the Julia language. Each realization is defined by a mean velocity, $\mathbf{v}$, with known direction (without loss of generality, always in the positive x-direction) and uniformly distributed random magnitude, $v \equiv \| \mathbf{v} \|$, between 1 and 4 m/y, longitudinal dispersivity, α_l, between 0.2 and 5 m, and transverse dispersivity, α_t, between 0.02 and 2 m. The instantaneous contaminant source is always at the point $(x_s, y_s) = (0, 0)$. For spatial identification the release event is at $t_{\mathrm{rel}} = 0$ y (assumed known), and for space-time identification is chosen randomly (discussed below). In all studies, the monitoring interval runs from $t_{\min} = 0$ y until $t_{\max} = 50$ y, with measurements recorded every 1 y. It is worth noting that with such a long monitoring interval, it is unlikely that (at least in an unconfined aquifer), the flow field would be free of seasonal transients. However, if we view $\mathbf{v}$ as the mean velocity over time as well as space, then short-duration zero-mean flow fluctuations may be subsumed into the Fickian dispersion appropriate to long range forecasting at such a site. Thus, the procedure outlined here does not need to explicitly account for seasonal transients.

The concentration taken to be the truth for the purposes of the study, $c(x, y, t)$, is determined in three stages:

First, a dummy concentration, $d(x, y, t)$, is computed by means of the advection-dispersion equation (ADE),

$$\frac{\partial d(x,y,t)}{\partial t} = v\left[-\frac{\partial d(x,y,t)}{\partial x} + \alpha_l \frac{\partial^2 d(x,y,t)}{\partial x^2} + \alpha_t \frac{\partial^2 d(x,y,t)}{\partial y^2}\right], \qquad (1)$$

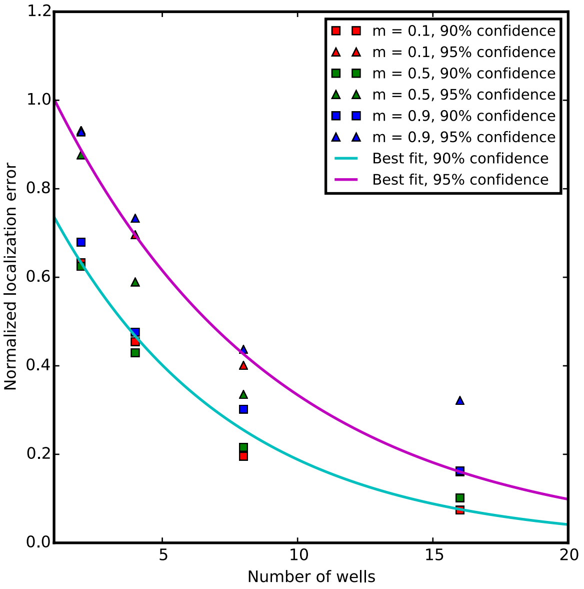

Fig. 6. Empirical 90% and 95% confidence intervals for the normalized source spatial localization error, ϵ (Eq. (10)). Data points are plotted for the 90% and 95% error thresholds for each of the four quantities of wells, for each of the three values of m (model infidelity) considered explicitly in the study. Best-fit exponential curves are plotted through the data points for each of the two confidence thresholds.

subject to the initial condition

$$d(x, y, 0) = \delta(x)\delta(y). \tag{2}$$

This initial value problem, solved on an infinite domain, yields the following Gaussian solution:

$$d(x, y, t) = \frac{1}{4\pi vt\sqrt{\alpha_l \alpha_t}} \exp\left(-\frac{(x - vt)^2}{4\alpha_l vt}\right) \exp\left(-\frac{y^2}{4\alpha_t vt}\right). \tag{3}$$

Second, a fluctuation function, $f(x, y; m, L)$, is defined as

$$f(x, y; m, L) \equiv 1 + m \sin\left(\frac{2\pi x}{L}\right) \sin\left(\frac{2\pi y}{L}\right), \tag{4}$$

where m and L represent fluctuation magnitude and period, respectively, and are fixed for each batch of realizations. An example form of this function is shown in Fig. 2.

Third, the true concentration is computed:

$$c(x, y, t) = f(x, y; m, L) d(x, y, t). \tag{5}$$

While its exact form is arbitrary, f qualitatively reflects the effect of flow channeling in heterogeneous hydraulic conductivity fields (an example is shown in Fig. 3) and the result that, under advection in heterogeneous velocity fields, concentration is much higher in the high velocity regions (Edery et al., 2014). The fluctuation function we have chosen to generate model infidelity also has the desirable properties of conserving mass in the limit of large uniform plumes, and being spatially smooth.

Wells are randomly located on a rectangle whose edges are aligned with the coordinate axes, with corners at $(x_{\min}, y_{\min})$ and $(x_{\max}, y_{\max})$, where

$$x_{\min} = v(t_{\min} - t_{\text{rel}}) - \sqrt{12\alpha_l v(t_{\min} - t_{\text{rel}})} \tag{6}$$

$$x_{\max} = v(t_{\max} - t_{\text{rel}}) + \sqrt{12\alpha_l v(t_{\max} - t_{\text{rel}})} \tag{7}$$

$$y_{\min} = -\sqrt{12\alpha_t v(t_{\max} - t_{\text{rel}})} \tag{8}$$

$$y_{\max} = \sqrt{12\alpha_t v(t_{\max} - t_{\text{rel}})}. \tag{9}$$

(Note again that $t_{\text{rel}} = 0$ for the spatial localization study.) These conditions ensure that the well field used for inversion is spatially coextensive with the plume, and prevent use of an excess of wells with zero readings for all time. We posit that this is realistic, and corresponds to a strategy of generally not calibrating against remote wells with null readings. Note also that because the well field used for calibration scales with the plume, the actual units chosen for space (m) and time (y) are immaterial to the analysis, and the same reliability statistics will hold regardless of their choice. Breakthrough curves for well w, located at (x_w, y_w) are computed by evaluating $c(x_w, y_w, t)$ over the interval $t = [0, t_{\max}]$. Fig. 4 shows an example point source and randomly generated well field in relation to the plume.

For a given realization (i.e. set $\{v, \alpha_l, \alpha_t\}$, for spatial localization or set $\{v, \alpha_l, \alpha_t, t_{\text{rel}}\}$, for space-time localization), interpretation is performed by repeated simulation and optimization. Candidate values of parameters $\tilde{x}_s$, $\tilde{y}_s$, $\tilde{v}$, $\tilde{\alpha}_l$, $\tilde{\alpha}_t$ and, if relevant, $\tilde{t}_{\text{rel}}$, are chosen (the ~ overbar indicates an estimate of the variable beneath), substituted into Eq. (1), and breakthrough curve estimates are made at each well. Different sets of parameters are tried, in an attempt to reduce breakthrough curve fitting error. During the optimization, the upper and lower bounds on v, α_l, and α_t are assumed known, and the estimated source location is allowed to vary within the box $[-x_{\max}, x_{\max}] \times [y_{\min}, y_{\max}]$. Note that the true data is generated by c, but simulation is performed with the overly simple model, d, with the degree of model infidelity controlled by m.

For both the spatial and space-time localization studies, three different values of m are considered: $m = 0.1$, $m = 0.5$, and $m = 0.9$. For each m, four different data quantities (i.e. numbers of wells, n) are considered: $n = 2$, $n = 4$, $n = 8$, or $n = 16$ wells are used for identification. For each of these twelve combinations, the simulation-optimization problem is repeated for hundreds of different realizations, with the number of realizations selected to produce an acceptable trade-off between computation time and empirical probability distribution (pdf) robustness. In all cases, $L = 3$ m is used as the fluctuation length scale. Fig. 4 presents the results of optimization on a single characteristic realization.

2.2. Spatial localization

For each realization, r, the initial guess for the x-coordinate of the source location is the x-coordinate of the farthest upgradient well, the y-coordinate guess is randomized, and the initial guesses $v = 2$ m/y, $\alpha_l = 0.5$ m, and $\alpha_t = 0.1$ m are used. The Levenberg–Marquardt iterative algorithm (Levenberg, 1944; Marquardt, 1963) is used to attempt to find the optimal parameter set, using the MADS (Model Analysis and Decision Support) software. The Levenberg–Marquardt algorithm is a popular method for model calibration used in many software tools. During the optimization process, the algorithm interpolates between the second-order

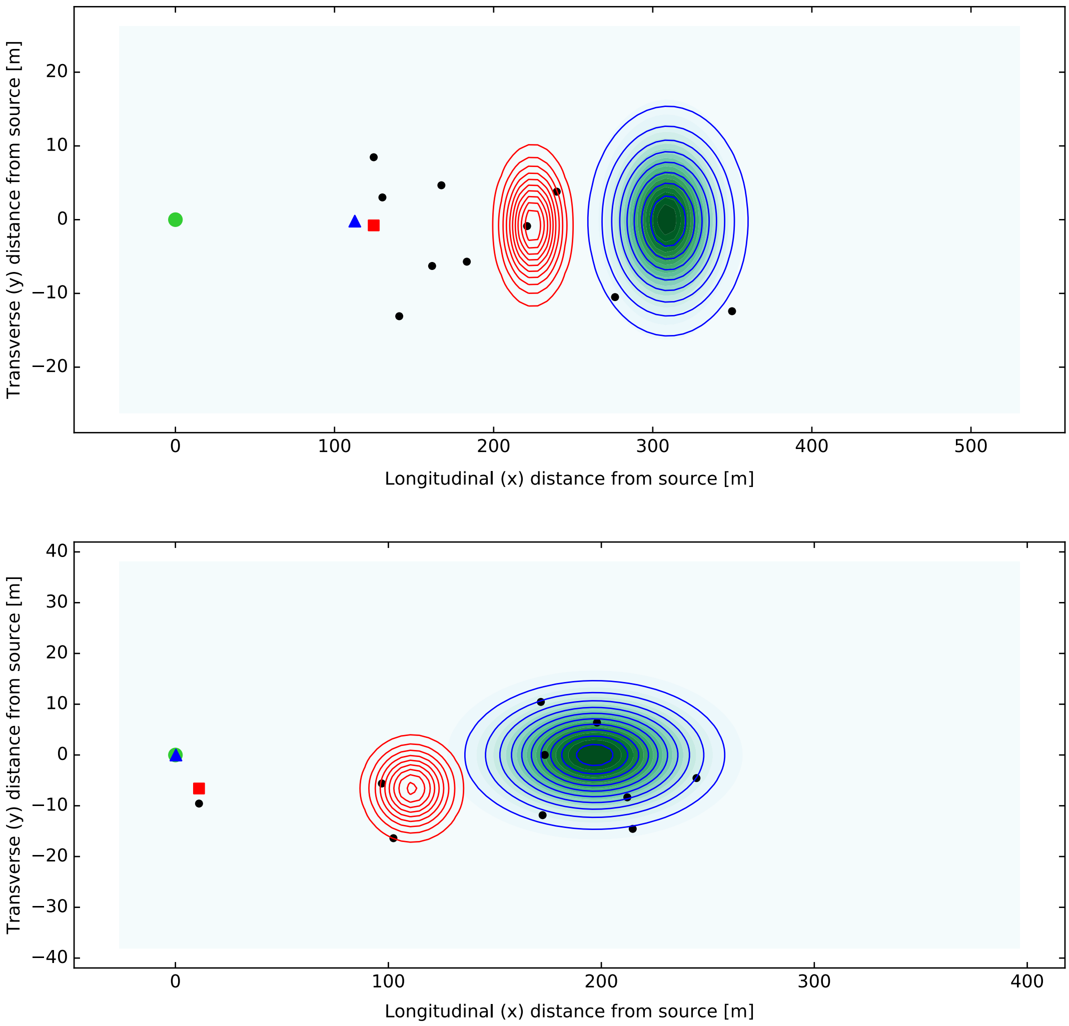

Fig. 7. Two examples of realizations with excellent coherence of true and fitted plumes at late time, with greatly different error metrics ϵ and τ. (Note that spatial scales are not the same.) Color and shape coding is as in Fig. 1. Top: true release time was −49.90 y and fitted release time was −13.09 y ($\tau = 0.736$, $\epsilon = 0.723$). Bottom: true release time was −0.50 y and fitted release time was −0.44 y ($\tau = 0.001$, $\epsilon = 0.001$); green point is covered by blue point due to near-perfect fit.

Gauss–Newton algorithm and the first-order method of steepest descent in a way which makes the algorithm more robust than either the Gauss–Newton or the steepest descent algorithms by themselves. As with many other optimization algorithms, the Levenberg–Marquardt algorithm is efficient for finding local minima without guaranteeing that these are global minima. MADS is an open-source high-performance computational framework written in the Julia language and available as a Julia distribution package. Information about MADS including documentation, examples, source code, and applications can be found in Vesselinov and Harp (2013), Vesselinov et al. (2016a,b,c). The identified location of the source in space after 100 Levenberg–Marquardt iterations (informally, we observed that the algorithm typically converged long before this many iterations had elapsed) is taken as the prediction of the optimization technique, unless the optimized location is downgradient of its initial guess. In this case, the optimization is repeated with further upgradient initial guesses until the optimized $\tilde{x}_s$ is upgradient of all the wells. The magnitude of the distance from the estimated source location to the true source location (at the origin) is determined, and is normalized relative to the distance traveled by the plume centroid over the 50 year monitoring period. The resulting statistic,

$$\epsilon_r \equiv \frac{\sqrt{\tilde{x}_s^2 + \tilde{y}_s^2}}{v(t_{\max} - t_{\min})}, \qquad (10)$$

of relative spatial source localization error is recorded. The value of ϵ_r for each realization is considered to be a draw of a random variable, ϵ. From the ensemble of values of ϵ_r, by computing a smoothed histogram of the values of ϵ_r corresponding to a given ordered pair of values (m, n), empirical probability distribution functions of ϵ are generated for each of these 12 ordered pairs. (For clarity, pseudocode for this procedure is shown in Algorithm 1.) Also a relationship for 90% and 95% confidence envelopes on the normalized spatial localization error magnitude is computed. For this study, 500 realizations are used.

Algorithm 1. Pseudocode for generating error envelopes in the spatial localization study.

```
$L = 3$;
// for each infidelity level
for $m$ = 0.1, 0.5, 0.9 do
    // for each quantity of wells
    for $n$ = 2, 4, 8, 16 do
        // for each of 500 realizations
        for $r$ = 1:500 do
            Randomly select subsurface parameters $v$, $\alpha_l$, and $\alpha_t$ defining $c(x, y, t)$;
            // for each well
            for $w$=1:$n$ do
                Randomly select $x_w$ and $y_w$ for well locations in the vicinity of the plume;
                // for each of 50 yearly measurements
                for $t$=1:50 do
                    Record well concentration as $c(x, y, t) = f(x_w, y_w; m, L)d(x_w, y_w, t)$;
                end
            end
            Run MADS on the well concentration data, calibrating for $\tilde{x}_s$, $\tilde{y}_s$, $\tilde{v}$, $\tilde{\alpha}_l$, and $\tilde{\alpha}_t$;
            Compute and store $\epsilon_r$;
        end
        Compute smoothed histogram from the set of values $\{\epsilon_r\}_{r=1:500}$;
    end
end
```

2.3. Space-time localization (single initial guess)

This space-time localization study is nearly identical to the spatial localization study described above. It differs in that for each realization, t_{rel} is uniformly randomly selected as an *unknown* random time in the interval $[-50, 0]$, rather than having *known* value $t_{\text{rel}} = 0$, as in the spatial previous set of simulations. That $t_{\text{rel}} \in [-50, 0]$ is taken as the only known prior information about it. Using the same Levenberg–Marquardt algorithm used for the spatial localization simulations, t_{rel} is estimated (as $\tilde{t}_{\text{rel}}$) in parallel with all the parameters fit in the purely spatial identification problem. (The initial guess for t_{rel} is always 0 y.) Because wells were still placed according to Eqs. (6)–(9), wells only cover the plume trajectory from $t = t_{\text{min}}(= 0)$, onwards. Samples also are only taken, as before, beginning at $t_{\text{min}} = 0$ *y*. This set of space-time identification simulations is thus an extrapolation problem in a way that the previous set is not. For each realization, *r*, the statistic ϵ_r is computed, as before, and empirical pdfs of the spatial localization error, ϵ, are again computed for each ordered pair (m, n). In addition, for each realization, another diagnostic statistic, τ_r, summarizing the temporal localization error is tabulated:

$$\tau_r \equiv \frac{\left|\tilde{t}_{\text{rel}} - t_{\text{rel}}\right|}{t_{\text{max}} - t_{\text{min}}}. \tag{11}$$

These are treated as samples of a random variable, τ, and empirical pdfs for τ are computed for each (m, n), as above. For this study, 1000 realizations are used to compute the source localization error, and a subset of 500 are used to compute the temporal localization error.

2.4. Space-time localization (pseudo-global search)

This space-time localization study uses the same essential setup and parameter ranges as the single-initial-guess space-time localization study. It differs from it only in the number of realizations used (300, on account of increased computational complexity), and in the optimization approach used. For this study, for each realization, twenty sets of initial parameter guesses, $\{\tilde{x}_s, \tilde{y}_s, \tilde{v}, \tilde{\alpha}_l, \tilde{\alpha}_t, \tilde{t}_{\text{rel}}\}$, are chosen randomly from the full range of allowable parameters. For each set of initial parameters, a Levenberg–Marquardt space-time optimization, identical to that used for the single-initial-guess study, is performed. Of all twenty sets of locally optimized parameters, the set which produces the smallest value of the objective function is selected as the optimal solution. The values of ϵ_r and τ_r corresponding to the optimal solution for each realization, *r*, are recorded, and pdfs of ϵ and τ are generated, as before.

3. Results and discussion

3.1. Spatial localization

Results for the spatial localization simulations are summarized in two figures. Fig. 5 shows empirical pdfs for the spatial localization error, ε, under each of three degrees of model infidelity. Fig. 6 plots the amount of error at two different confidence intervals for each of the pdfs shown in Fig. 5, and illustrates how these are affected by both data quantity and model quality. Qualitatively, it is apparent that increasing data collection increases estimation reliability, both at the 90% and 95% confidence thresholds. A perhaps surprising result was that relatively extreme differences in model fidelity had only moderate impact on the accuracy of prediction, for a given quantity of sampling locations. Furthermore, the impact of model infidelity appeared to be easily exceeded by the impact of more sampling (for the modest number of wells that we considered, similar to what one might find at a real site). Another observation was that the increase in confidence interval size (i.e. uncertainty) with increasing model infidelity was often small.

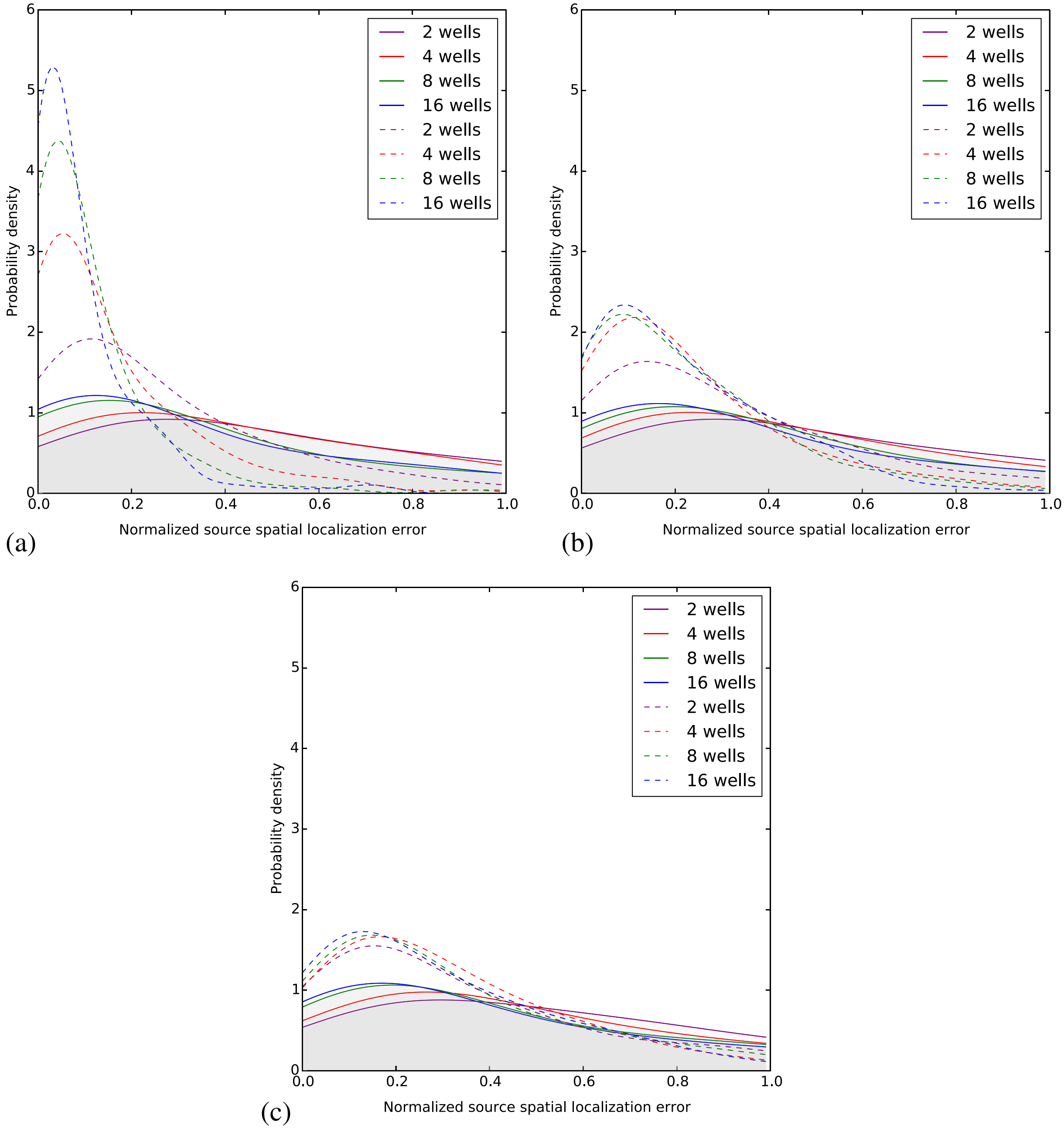

Fig. 8. Empirical probability distribution functions, in the case of unknown release time, for the normalized source spatial localization error, ϵ (Eq. (10)), for given numbers of available well breakthrough curves, for each of three degrees of model infidelity: (a) $m = 0.1$, (b) $m = 0.5$, (c) $m = 0.9$. Solid curves with shading underneath represent the single-initial-guess optimization procedure. Dashed curves represent the twenty-initial-guess pseudo-global optimization procedure.

3.2. Space-time localization

For space-time localization, as indicated in Section 2, the source was spatially and temporally outside the data region. Thus, identification involved space-time extrapolation, and was consequently more prone to error. An example of the difficulties is seen in Fig. 7, in which fitted plumes for two realizations of the same space-time identification problem are shown. Despite both having excellent coherence to the true plume when $t = 50$ y, they vary greatly in the accuracy of source identification, with one featuring error metrics $\tau = 0.736$, $\epsilon = 0.723$, and the other featuring error metrics $\tau = 0.001$, $\epsilon = 0.001$. This is indicative of the difficulties that are faced with unsupervised identification when both location and time of release are unknown, with data unavailable for the early part of the plume's trajectory. Approximately, errors in both spatial and temporal localization can counteract each other to generate similar data: since dispersivity is also a fitting parameter, a recent source located near the observation field can produce breakthrough data similar to that from an earlier source which is located farther upgradient. A hypothesis tested was that by adding additional measurement points (i.e. wells), the non-uniqueness would be diminished and space-time identification improved.

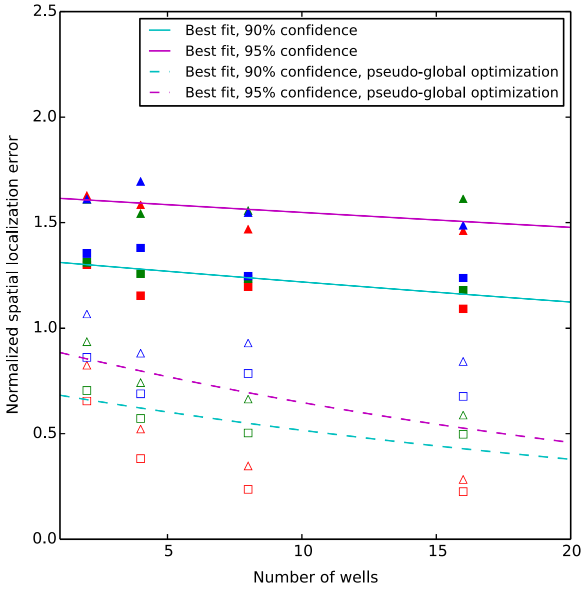

Fig. 9. Empirical 90% and 95% confidence intervals for the normalized source spatial localization error, ϵ (Eq. (10)), in the case of unknown release time. Data points are plotted for the 90% (square markers) and 95% (triangular markers) confidence thresholds for each of the four quantities of wells, for each of the three values of model infidelity considered explicitly in the study: $m = 0.1$ (red markers), $m = 0.5$ (green markers), and $m = 0.9$ (blue markers). Both the single-initial-guess optimization (solid markers) and pseudo-global optimization (empty markers) procedures are represented. Best-fit exponential curves are plotted through the data points for each of the two confidence thresholds and for each of the optimization procedures.

We found that this was generally not the case, and because the results obtained appeared to be too poor for practical use, we ran a second study which employed a pseudo-global optimization strategy, hypothesizing that the poor performance of the optimization might be accounted for by the presence of many local minima of the objective function (note that in Fig. 7, the better identification features a source closer to the well field than the poor fit).

Fig. 8 shows empirical probability distribution functions (pdfs) for ϵ, for each level of model infidelity, m, and number of wells, n, in the case when t_{rel} is unknown. The pdfs have qualitatively the same behavior as seen in Fig. 5: they are asymmetric, with mean location moving towards the true location as data quantity increases. However, for the single-initial-guess identification procedure it is clear that the spatial localization is generally much worse when t_{rel} is unknown; indeed the lack of knowledge about release time is seen as a greater confounding factor than lack of knowledge about the subsurface transport parameters and even high model infidelity in the spatial localization study. Furthermore, despite more data generally improving localization, for 90% and 95% confidence thresholds, additional sampling had only a modest impact, as seen in Fig. 9. This contrasts strikingly with the study of source localization in which the release time is known. The pseudo-global (twenty-initial-guess) optimization approach performed significantly better at all model infidelity levels, and particularly for low model infidelity, when $m = 0.1$. However, there remained limited value to additional data collection, except at the lowest infidelity level, where it was apparent as strongly as in the purely spatial identification task.

For temporal localization in the same circumstances, the single-initial-guess localization performance was relatively poor, as seen in Fig. 10. Here, the effect of additional data was modest, although larger numbers of wells tended to have a better probability of a localization close to the true release time. Increasing model infidelity (i.e., m) also had a modest degrading effect on localization. However, it is easy to analytically compute the pdf for τ for a naive identification scheme in which $\tilde{t}_{\text{rel}}$ is simply selected from a uniform distribution on the interval $[-50, 0]$, and this is also shown in Fig. 10. Remarkably, this approach performs as well as or better than the Levenberg-Marquardt optimization approach for temporal identification in this case. The 90% and 95% confidence thresholds, shown as a function of n in Fig. 11 reinforce this observation. They showed no improvement with additional information, were of approximately of the same scale as the 50 y fitting window, and did not improve upon their naive estimates. This is to say: for error envelopes corresponding to large degrees of confidence, the optimization (or model calibration) process was not seen to add value to the *temporal* identification beyond the prior information about the unknown model parameters. The pseudo-global optimization procedure substantially improved temporal localization and produced better-than-random, although still not outstanding, results. As for the spatial component of the space-time localization task, at the lowest infidelity level the positive impact of additional data collection re-emerged strongly.

4. Summary and conclusion

In this paper, we addressed the near absence from the contaminant source identification literature of empirically-grounded relationships between meta-parameters representing the data quantity and model quality in an inverse estimation exercise and the quality of predictions from that exercise. We approached this task by means of Monte Carlo studies: running multiple forward models to generate simulated data and then attempting to localize the contaminant source by inverse analysis of each set of simulated data, using a simplified model. Three related studies were performed, one considering spatial localization of a point source and two considering space-time localization of a point source. In all studies, only temporal breakthrough data at randomly-located wells was considered for the identification; nuisance parameters describing the subsurface transport regime had to be simultaneously identified from the same data set as part of model calibration. Prior information was restricted to flow direction and rough bounds on its speed and very rough, one-to-two-order-of-magnitude, bounds on the other subsurface parameters. In the space-time identification problem, a 50 y interval in which the source event was known to have occurred was also assumed.

For all studies, we employed an unsupervised, unregularized simulation-optimization approach for the source localization. We developed relationships between localization accuracy, data quantity (represented by number of wells), and model quality (represented by magnitude of divergence between the model used to generate the synthetic observations and the simpler model used for calibration against them). A key question considered was the degree to which additional data could counteract the inevitable over-simplification of the model used for calibration.

For the purely spatial localization problem, we found that the unregularized source identification algorithm was qualitatively sufficient, and that additional sampling points were capable of making up for model over-simplification, even in the case of high model infidelity ($m = 0.9$). We generally found a strong relationship between use of more wells and increased identification performance. When 90% and 95% confidence intervals were computed, the error decreased quasi-linearly with the number of wells used for calibration, up to about 10 wells, with diminishing, but still significant, returns beyond this point. It is worth noting that the modal (i.e., most likely) error was small, relative to final plume extent, even with large model infidelity and small numbers of wells. However, the 95% confidence localization error (relevant to responsibility assignment

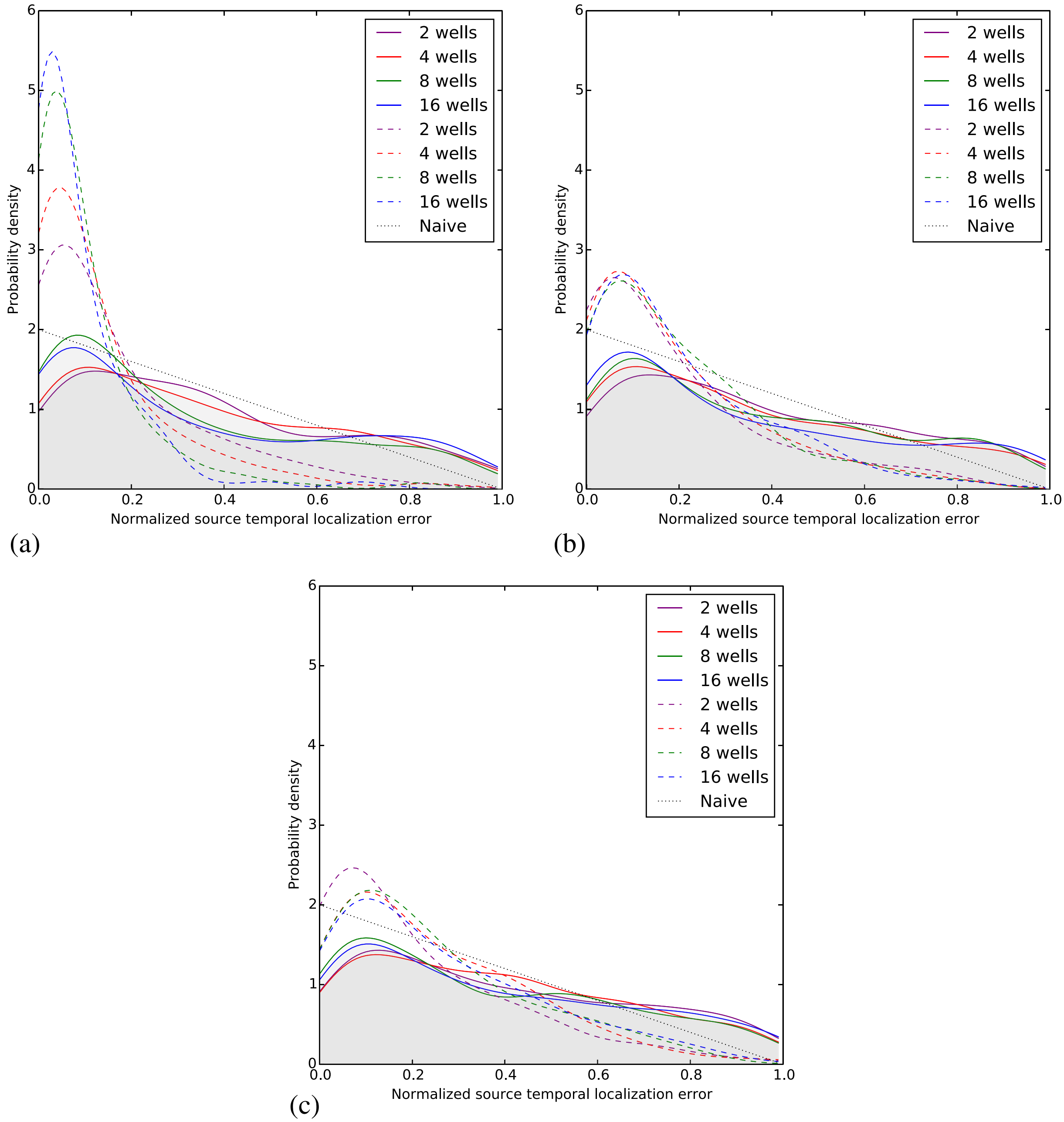

Fig. 10. Empirical probability distribution functions, in the case of unknown release time, for the normalized source temporal localization error, τ (Eq. (11)), for given numbers of available well breakthrough curves, for each of three degrees of model infidelity: (a) $m = 0.1$, (b) $m = 0.5$, (c) $m = 0.9$. For comparison, the naive distribution function for τ, arising from a uniform random selection of $\tilde{t}_{\mathrm{rel}}$ is shown as a dotted line on all plots. Solid curves with shading underneath represent the single-initial-guess optimization procedure. Dashed curves represent the twenty-initial-guess pseudo-global optimization procedure.

or interventions targeted at human health) remained comparatively substantial.

For the space-time localization problems, identification performance was worse, particularly when optimization proceeded from a single initial parametric estimate. Our results are suggestive that, when neither the source location nor source release time are known, results of unregularized simulation-optimization inverse analysis may not be reliable enough for responsibility assignment, and that one should be cautious about forward predictions stemming from the optimized parametrization. These results underscore the importance of bringing all available information about the source to bear on the problem (easy to do in a simulation-optimization framework) to regularize the results.

When twenty random starting points were used for the space-time optimization approach, rather than a "best guess", and the best fit of the twenty local optimizations was selected, performance was significantly improved, suggesting that the objective function has a relatively complicated arrangement of multiple local minima for this problem. However, performance remained relatively poor compared to the purely spatial identification, particularly at the 90% and 95% confidence levels, and did not appear to be reliable enough for responsibility assignment. We also observed that, with multiple starting points (i.e. initial guesses for unknown model parameters in the optimization process) and low levels of heterogeneity, the pattern of additional data leading to improved localization re-emerged. Because of computational time limitations, it was not possible to test

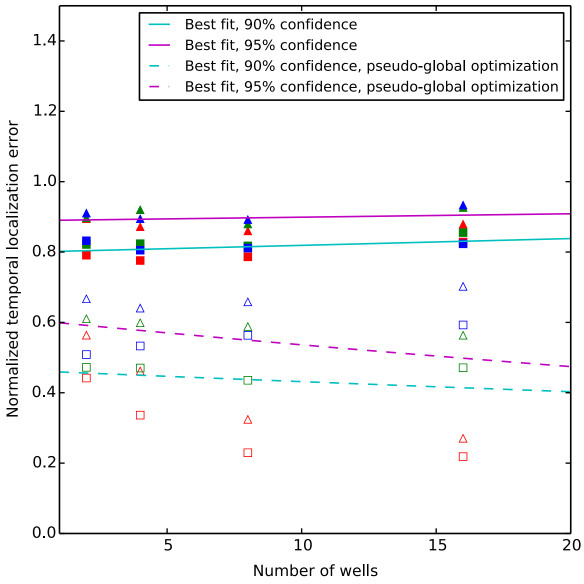

Fig. 11. Empirical 90% and 95% confidence intervals for the normalized source temporal localization error, τ (Eq. (11)), in the case of unknown release time. Data points are plotted for the 90% (square markers) and 95% (triangular markers) confidence thresholds for each of the four quantities of wells, for each of the three values of model infidelity considered explicitly in the study: $m = 0.1$ (red markers), $m = 0.5$ (green markers), and $m = 0.9$ (blue markers). Both the single-initial-guess optimization (solid markers) and pseudo-global optimization (empty markers) procedures are represented. Best-fit exponential curves are plotted through the data points for each of the two confidence thresholds and for each of the optimization procedures.

space-time optimization with very large numbers of initial guesses. It would be interesting to see if, using near-true global optimization, with thousands of random initial guesses, this pattern re-emerged fully.

In all cases, it bears recognition that the data simulated for the study was generated in a relatively idealized fashion (regular channeling was imposed on a plume that otherwise was precisely described by the 2D ADE with spatially homogeneous parameters), the point-like nature of release was taken as known, and intact well records were available at all of the relevant locations. Furthermore, some sites of remedial interest might have a shorter observation window than assumed here (while the units of time are arbitrary, simulations were designed so that substantial plume evolution occurred in the well field over the monitoring interval—something that is not assured in practice). Thus, the computed confidence intervals should be considered as baselines that are suggestive of qualitative trends, and practitioners should be aware that real-world uncertainty is apt to be greater.

Since the computation of error envelopes as functions of data quantity and model quality has been largely unaddressed in the contaminant source identification literature, there are many opportunities to both refine these results and to undertake related work of both practical and theoretical significance. A particularly interesting topic for follow-up study is an extension of the space-time localization study which showed comparatively weak localization performance. We believe that consideration of the size of error envelopes as functions of the tightness of the prior constraint on the source release time (i.e. $t_{\mathrm{rel}} \in [t_1, t_2]$), simultaneously with the value of $|t_{\min} - t_{\mathrm{rel}}|$ is apt to be enlightening. This follow-up study would reveal how much regularization is required to counteract a certain amount of extrapolation (i.e., ill-posedness of the inverse problem). It is left to be addressed in future work.

Acknowledgments

The authors acknowledge support of the LANL Environmental Programs.

References

Alapati, S., Kabala, Z., 2000. Recovering the release history of a groundwater contaminant using a non-linear least-squares method. Hydrocarb. Process. 14, 1003–1016. http://dx.doi.org/10.1002/(SICI)1099-1085(20000430)14.

Aral, M.M., Guan, J., Maslia, M.L., 2001. Identification of contaminant source location and release history in aquifers. J. Hydraul. Eng. 6, 225–234. June.

Ayvaz, M.T., 2010. A linked simulation-optimization model for solving the unknown groundwater pollution source identification problems. J. Contam. Hydrol. 117 (1-4), 46–59. ISSN 1873-6009. http://dx.doi.org/10.1016/j.jconhyd.2010.06.004.

Bagtzoglou, A., Atmadja, J., 2005. Mathematical methods for hydrologic inversion: the case of pollution source identification. Water Pollution 5, 65–96. http://dx.doi.org/10.1007/b11442.

Bagtzoglou, A.C., Dougherty, D.E., Tompson, A.F.B., 1992. Application of particle methods to reliable identification of groundwater pollution sources. Water Resour. Manag. 6 (1), 15–23. ISSN 09204741. http://dx.doi.org/10.1007/BF00872184.

Bashi-Azghadi, S.N., Kerachian, R., Bazargan-Lari, M.R., Solouki, K., 2010. Characterizing an unknown pollution source in groundwater resources systems using PSVM and PNN. Expert Systems with Applications, 37 (10), 7154–7161. ISSN 09574174. http://dx.doi.org/10.1016/j.eswa.2010.04.019.

Cheng, W.P., Jia, Y., 2010. Identification of contaminant point source in surface waters based on backward location probability density function method. Adv. Water Resour. 33 (4), 397–410. ISSN 03091708. http://dx.doi.org/10.1016/j.advwatres.2010.01.004.

Datta, B., Chakrabarty, D., Dhar, A., 2009. Simultaneous identification of unknown groundwater pollution sources and estimation of aquifer parameters. J. Hydrol. 376 (1-2), 48–57. ISSN 00221694. http://dx.doi.org/10.1016/j.jhydrol.2009.07.014.

Datta, B., Chakrabarty, D., Dhar, A., 2011. Identification of unknown groundwater pollution sources using classical optimization with linked simulation. J. Hydro Environ. Res. 5 (1), 25–36. ISSN 15706443. http://dx.doi.org/10.1016/j.jher.2010.08.004.

Edery, Y., Guadagnini, A., Scher, H., Berkowitz, B., 2014. Origins of anomalous transport in heterogeneous media: structural and dynamic controls. Water Resour. Res. 50, 1490–1505. ISSN 00431397. http://dx.doi.org/10.1002/2013WR015111.

Guan, J., Aral, M.M., Maslia, M.L., Grayman, W.M., 2006. Identification of contaminant sources in water distribution systems using simulation-optimization method: case study. J. Water Resour. Plan. Manag. 132 (4), 252–262.

Hazart, A., Giovannelli, J.F., Dubost, S., Chatellier, L., 2014. Inverse transport problem of estimating point-like source using a Bayesian parametric method with MCMC. Signal Process. 96 (PART B), 346–361. ISSN 01651684. http://dx.doi.org/10.1016/j.sigpro.2013.08.013.

Huang, C.H., Li, J.X., Kim, S., 2008. An inverse problem in estimating the strength of contaminant source for groundwater systems. Appl. Math. Model. 32 (4), 417–431. ISSN 0307904X. http://dx.doi.org/10.1016/j.apm.2006.12.009.

Jha, M., Datta, B., 2013. Three-dimensional groundwater contamination source identification using adaptive simulated annealing. J. Hydrol. Eng. 18 (3), 307–317. ISSN 10840699. http://dx.doi.org/10.1061/(ASCE)HE.1943-5584.0000624.

Koch, J., Nowak, W., 2016. Identification of contaminant source architectures-a statistical inversion that emulates multiphase physics in a computationally practicable manner. Water Resour. Res. 52, 1009–1025. ISSN 00431397. http://dx.doi.org/10.1002/2014WR015716.

Levenberg, K., 1944. A method for the solution of certain non-linear problems in least squares. Q. Appl. Math. 2, 164–168.

Mahar, P.S., Datta, B., 2000. Identification of pollution sources in transient groundwater systems. Water Resour. Manag. 14 (3), 209–227. ISSN 0920-4741. http://dx.doi.org/10.1023/A:1026527901213.

Mahar, P.S., Datta, B., 2001. Optimal identification of ground-water pollution sources and parameter estimation. J. Water Resour. Plan. Manag. 127 (February), 20–29.

Marquardt, D., 1963. An algorithm for least-squares estimation of nonlinear parameters. SIAM J. Appl. Math. 11 (2), 431–441. http://dx.doi.org/10.1137/0111030.

Michalak, A.M., Kitanidis, P.K., 2004. Estimation of historical groundwater contaminant distribution using the adjoint state method applied to geostatistical inverse modeling. Water Resour. Res. 40 (8), 1–14. ISSN 00431397. http://dx.doi.org/10.1029/2004WR003214.

Milnes, E., Perrochet, P., 2007. Simultaneous identification of a single pollution point-source location and contamination time under known flow field conditions. Adv. Water Resour. 30 (2007), 2439–2446. ISSN 03091708. http://dx.doi.org/10.1016/j.advwatres.2007.05.013.

Neupauer, R.M., Lin, R., 2006. Identifying sources of a conservative groundwater contaminant using backward probabilities conditioned on measured concentrations. Water Resour. Res. 42, 1–13. ISSN 00431397. http://dx.doi.org/10.1029/2005WR004115.

Neupauer, R.M., Wilson, J.L., 1999. Adjoint method for obtaining backward-in-time location and travel time probabilities of a conservative groundwater contaminant. Water Resour. Res. 35 (11), 3389–3398.

Neupauer, R.M., Wilson, J.L., 2005. Backward probability model using multiple observations of contamination to identify groundwater contamination sources at the Massachusetts Military Reservation. Water Resour. Res. 41 (2), 1–14. ISSN 00431397. http://dx.doi.org/10.1029/2003WR002974.

Skaggs, T.H., Kabala, Z.J., 1994. Recovering the release history of a groundwater contaminant. 30 (1), 71–79.

Skaggs, T.H., Kabala, Z.J., 1998. Limitations in recovering the history of a groundwater contaminant plume. J. Contam. Hydrol. 33 (3-4), 347–359. ISSN 01697722. http://dx.doi.org/10.1016/S0169-7722(98)00078-3.

Snodgrass, M.F., Kitanidis, P.K., 1997. A geostatistical approach to contaminant source identification. Water Resour. Res. 33 (4), 537–546.

Vesselinov, V.V., Harp, D., 2013. Model analysis and decision support (MADS) for complex physics models. XIX International Conference on Water Resources-CMWR.

Vesselinov, V.V., O'Malley, D., et al. 2016. Model analysis and decision support (MADS) documentation. http://mads.readthedocs.org.

Vesselinov, V.V., O'Malley, D., et al. 2016. Model analysis and decision support (MADS) source code. http://github.com/madsjulia.

Vesselinov, V.V., O'Malley, D., et al. 2016. Model analysis and decision support (MADS) web site. http://mads.lanl.gov.

Wagner, B.J., 1992. Simultaneous parameter estimation and contaminant source characterization for coupled groundwater flow and contaminant transport modelling. J. Hydrol. 135 (1), 275–303.

Wagner, B.J., Gorelick, S.M., 1986. A statistical methodology for estimating transport parameters: theory and applications to one-dimensional advective-dispersive systems. Water Resources Research 22 (8), 1303–1315.

Woodbury, A., Sudicky, E., Ulrych, T.J., Ludwig, R., 1998. Three-dimensional plume source reconstruction using minimum relative entropy inversion. J. Contam. Hydrol. 32 (1-2), 131–158. ISSN 01697722. http://dx.doi.org/10.1016/S0169-7722(97)00088-0.

Yeh, H.D., Chang, T.H., Lin, Y.C., 2007. Groundwater contaminant source identification by a hybrid heuristic approach. Water Resour. Res. 43 (9).

Zhang, J., Zeng, L., Chen, C., Chen, D., Wu, L., 2015. Efficient Bayesian experimental design for contaminant source identification. Water Resour. Res. 51, 576–598. ISSN 00221694. http://dx.doi.org/10.1002/2014WR016259.