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.

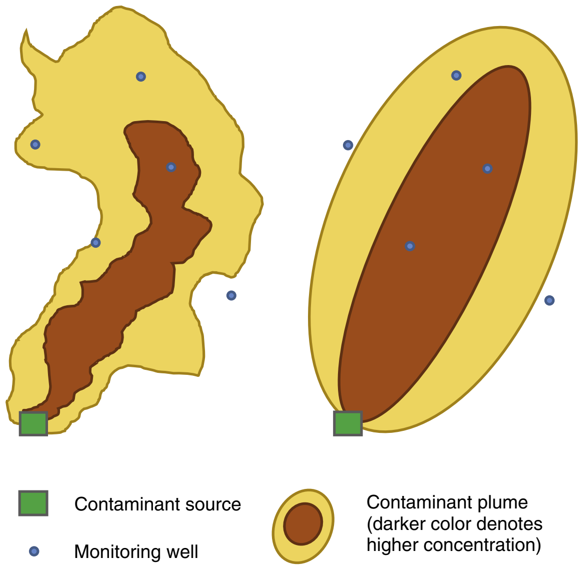

Fig. 1. Schematic diagram comparing the true contaminant plume developing in a heterogeneous environment (left) with a potential best-fit plume (right) generated by assuming subsurface transport is described by an advection-dispersion equation with spatially homogeneous parameters.

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

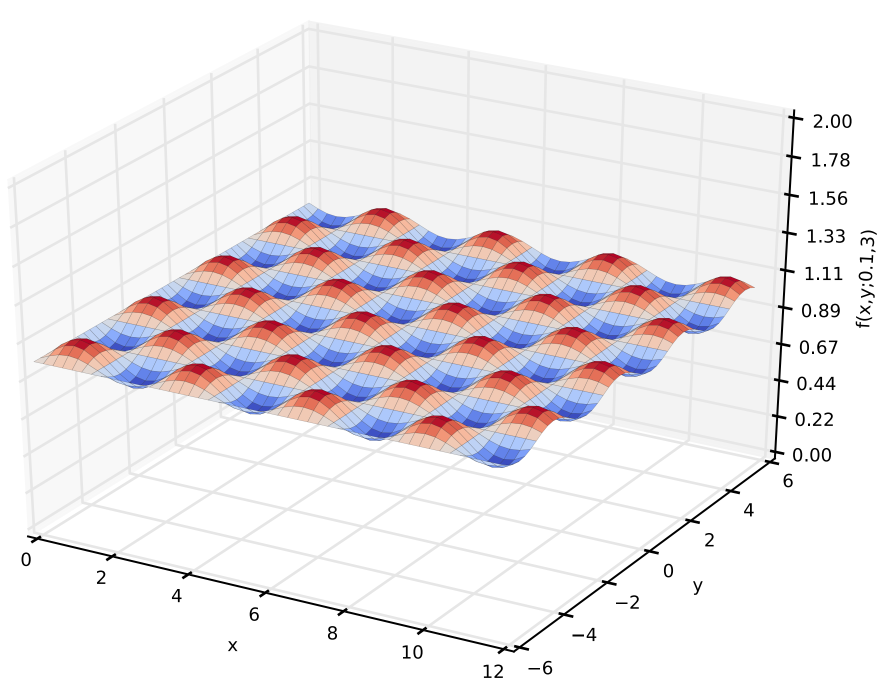

Fig. 2. Plot of the fluctuation (model infidelity) function, f, with parameters $L = 3$ m, $m = 0.1$.