CONSID: A Toolbox for Contaminant Source Identification

by Alexander Y. Sun

Abstract

Model-based contaminant source identification plays an important role in effective site remediation. In this article, a contaminant source identification toolbox (CONSID) is introduced as a framework for solving contaminant source identification problems. It is known that the presence of various types of model uncertainties can severely undermine the performance of many existing source estimators. The current version of CONSID consists of two robust estimators for recovering source release histories under model uncertainty; one was developed in the deterministic framework and the other in the stochastic framework. To use the robust estimators provided in CONSID, the user is required to have only modest prior knowledge about the model uncertainty and be able to estimate the bound of model deviations resulting from the uncertainty. The toolbox is designed so that other source estimators can be added easily. A step-by-step guidance for using CONSID is described and an example is provided.

Introduction

Model-based contaminant source identification refers to identifying locations and release histories of contaminant sources through solving an inverse problem using a set of ground water models and sampled concentrations of dissolved contaminants. It is well known that solutions to inverse problems are sensitive to uncertainties in both model and data because of the inherent ill-posedness of inverse problems. In this context, the causes of model uncertainties can be multifaceted, such as inadequacy of model conceptualization (including boundary conditions), spatial and temporal variability of model parameters, scale dependency of dispersion coefficients, and errors resulting from the numerical solution process. In a recent white paper on source zone assessment and remediation, the National Research Council (2005) emphasized the need to obtain a better treatment of uncertainty. The same report also points out that quantitative uncertainty analysis is rarely practiced at hazardous waste sites.
estimators and the NNLS. The design of CONSID is modular so that new estimators can be readily added.

The main prerequisites for using CONSID are as follows:

- The user has developed a set of ground water flow and transport models (referred to herein as the nominal models) for the site under study and is interested in reducing the impact of model uncertainty on the estimated source release history. Note that CONSID itself is not a preprocessor for preparing ground water models. Instead, it calls external models [i.e., MODFLOW2000 (Harbaugh et al. 2000) and MT3DMS (Zheng and Wang 1999)] to perform the forward modeling.
- The user has modest a priori knowledge about model uncertainty (e.g., the mean values and the ranges of variations of uncertain parameters).

The latter assumption in the aforementioned actually relieves the end user from having to know the full probability distribution of model uncertainty, which often constitutes one of the main roadblocks to conducting quantitative uncertainty analysis in practice. The robust estimators in CONSID essentially require knowing the upper bound of model deviations that are caused by all types of model uncertainties of interest. Sun et al. (2006a) and Sun (2007) suggested several strategies for conducting the worst-case analysis and, therefore, identifying the upper bound.

Identifying the bounds of uncertainty in input parameters and subsequently propagating them through the model is not a trivial task, and its complexity is case dependent. Some strategies for finding the worst-case scenario are as follows:

- The combinatorial search or greedy search approach, which identifies the worst case through a systematic and complete evaluation of all possible combinations of uncertain parameters. Except for relatively simple cases involving only a few uncertain parameters, the computing effort associated with this strategy can be enormously high.
- The Monte Carlo approach, which randomly samples the parameter space to search for the worst case. The efficiency of this approach can be limited and the convergence is usually hard to prove.
- Various cleverly designed sampling approaches [e.g., the Latin Hypercube Sampling method (McKay et al. 1979)] that can reduce computational effort significantly.

Although in principle it is possible to automate the worst-case search for a generic model, it is not implemented in CONSID at this time. In most situations, however, the search can be done with the help of some simple scripts and a tool included with CONSID for generating model matrices. This should become clear in the Example section. In the following, a step-by-step guide for problem solving using CONSID is provided.

Problem Formulation and Solution Using CONSID

The use of CONSID for contaminant source identification can be divided into two stages, namely, a preparation stage and a solution stage.

Preparation Stage

1. Define the nominal flow and transport models for the site. The nominal model is often associated with a model assigned with mean model parameters, but it can simply be the baseline model used in sensitivity studies. If the nominal model is deemed exact, CRLS and RGS are identical to their classical counterparts [i.e., the NNLS for CRLS and the geostatistical inversion method (Kitanidis 1995) for RGS].
2. Define potential source locations through the user interface. CONSID can be used to identify source release histories at multiple locations. In cases where the prior knowledge is not sufficient for defining good initial guesses of source locations, either CRLS or RGS can be coupled with a global optimization routine to search for source locations in a predefined range (cf. Sun et al. 2006b).
3. Define the time horizon structure for source release history. The time horizon should span the longest release history of interest. The discretization of time horizon can be uniform or nonuniform, both of which are supported by the CONSID user interface.
4. Specify the concentration measurements by providing the sample locations, sampling times, and sample values. This information will be used by the backend engine to invert the source release history.

Solution Stage

1. Generate the flow field for the nominal model by calling MODFLOW. Both steady-state and transient flow fields are supported.
2. Formulate the estimation problem using the unit-pulse response matrix method described in Sun et al. (2006a). Using the standard notations in control systems theory, the resulting estimation problem can be written as follows:

\[
\mathbf{z} = \mathbf{H}_0\mathbf{s} + \mathbf{w}
\]

where \(\mathbf{s} \in \mathbb{R}^m\) (\(\mathbb{R}\) denotes the real space) is the unknown vector of dimension \(n\), \(\mathbf{z} \in \mathbb{R}^n\) is the measurement vector of dimension \(m\), \(\mathbf{H}_0 \in \mathbb{R}^{mn}\) represents the model matrix corresponding to the nominal model and is called the nominal model matrix hereafter, and \(\mathbf{w} \in \mathbb{R}^m\) is the measurement error vector. CONSID calls MT3DMS multiple times during this time to calculate the elements of \(\mathbf{H}_0\).
3. Identify the upper bound of model deviations if CRLS or RGS is chosen. Mathematically, the deviation is defined as the matrix norm of the difference between a perturbed model matrix (from an alternative model) and a nominal model matrix. This part of the analysis is often case dependent but can be done with some simple scripts in most cases. Besides strategies mentioned in the Introduction section, many other techniques used in the experimental design literature can be employed at this step for quantifying model deviations caused by either parameter uncertainty or model structure uncertainty or both. The user is advised to carefully set up his/her case for the uncertainty analysis, for example, by deciding which uncertain parameters are the most important to include in the
analysis. As an example, consider the hydraulic conductivity zonation pattern shown in Figure 1, where the numbers indicate the estimated mean conductivity values for the four conductivity zones. Assume the hydraulic conductivity is the main cause of model uncertainty. Then, the mean conductivities can be used to construct the nominal model matrix, whereas the bounds of variations of zonal values can be used to quantify the upper bound of model deviations. As another example, if the hydraulic conductivity is modeled as a stochastic random field, the deviations from the nominal model can be found by generating an ensemble of scenarios and estimating the maximum deviations using the finite set of scenarios (cf. Sun 2007). The analyses can be carried out in a similar manner if transport parameters are also uncertain. An example is given in Sun et al. (2006a) where the value of dispersivity is assumed uncertain. Numerical experiments have shown that both estimators tolerate inaccuracies in the estimated bound to a certain degree.

4. Obtain solutions and postprocessing. CONSID has the capability to plot identified source release histories as well as reference release histories specified by the user.

Example

In this section, CONSID is demonstrated through identifying the release history of a contaminant source in a two-dimensional confined aquifer. Details of the original problem are given in Sun (2007). Here, we further illustrate how CONSID can be used to solve the problem. The domain is 1000 \times 500 m and is discretized uniformly into a 100 \times 50 grid (Figure 2). The flow is nonuniform because of the boundary conditions. A source is located near the upper left corner of the domain. The “true” source release history \( s(t) \) is a multimodal function given by:

\[
s(t) = 100 \left\{ \exp \left( -\frac{(t-130)^2}{50} \right) + 0.3\exp \left( -\frac{(t-150)^2}{200} \right) + 0.5\exp \left( -\frac{(t-190)^2}{98} \right) \right\}
\]

Figure 1 of Sun 2007). The analyses can be carried out in a similar manner if transport parameters are also uncertain. An example is given in Sun et al. (2006a) where the value of dispersivity is assumed uncertain. Numerical experiments have shown that both estimators tolerate inaccuracies in the estimated bound to a certain degree.

Recovering \( s(t) \) is challenging because the system is sensitive not only to model uncertainty and data error but also to sampling frequencies and locations. The total release period is 360 d and is divided into 120 equal intervals. In reality, the start and end times of the source release history are never known exactly, so the user should choose a period that best covers the time horizon of interest.

For illustration, the uncertainty in hydraulic conductivity is assumed to be the main source of model uncertainty. The uncertain hydraulic conductivity field is specified by its mean (35 m/d) and spatial covariance function as is often done in geostatistical applications. An ensemble of 300 hydraulic conductivity realizations was generated using an isotropic exponential covariance function with correlation length of 100 m. One of the realizations was selected as the synthetic truth to generate concentration samples, and all others were used as “alternative models.” The same approach can be expanded to more general cases, where the alternative models would correspond to different uncertain factors such as the transport parameters. The Latin Hypercube Sampling method mentioned previously can be used to efficiently explore the parameter space to generate alternative models.

The nominal flow and transport models were first set up in a preprocessor and exported as standard MODFLOW and MT3DMS files. The nominal flow model was set up using the mean hydraulic conductivity. After obtaining the nominal model matrix using CONSID, the user needs to determine the bound of model deviations. Typically, this involves calculating deviations between all alternative models and the nominal model and selecting the maximum. The uncertainty analysis step can be the most time consuming and yet the most rewarding part of the whole solution process. In the next step, CONSID is used to obtain estimates. Figure 3 shows the sensitivity of recovered source release histories to different bounds, where CRLS was used in all cases. The deviation between the true model matrix and the nominal model matrix is approximately 0.02. Note that a small bound does not necessarily mean variations in input variables are small. Also, different alternative models may yield similar bounds. Figure 3 indicates that the robust estimator is actually robust to misspecifications in bound to a certain degree.
Being able to use a coarse-resolution model for source identification can also have profound practical meanings because it is impossible for site-scale models to capture the details of geology at all scales. Mass transfer processes can occur at much smaller scales than the resolution of flow models. The consequence is that samples contain signatures of spatial variability that are not necessarily captured by the flow and transport models. In the current problem, the nominal model matrix was generated by simulating mass transport in a homogeneous aquifer, whereas the samples were actually ‘taken’ from a heterogeneous aquifer. Thus, the samples contain effects of physical heterogeneity that are not reflected in the nominal model. Many other source estimators are based on the principle of minimizing the residuals between the measured and the predicted concentrations by assuming that the underlying model is exact. Ignoring the model uncertainty can have significant impact on the estimates. For example, the combination of NNLS and the nominal model gave completely erroneous results for this case. Therefore, the capability to allow model mis-specifications during the source estimation process should be considered a critical property of contaminant source estimators.

Summary

Contaminant source identification under model uncertainty is of genuine interest to ground water modelers because inaccurate information on source locations and release histories can hamper the implementation of effective remediation strategies. The CONSID toolbox has been developed to provide a systematic tool for promoting the use of robust source identification in practice. Existing site models developed in MODFLOW and MT3DMS can be readily converted for conducting source identification with CONSID. The current version of CONSID allows the user to identify the release histories of multiple sources (given locations), which are either nonreactive tracers or subject to linear equilibrium reaction. The CONSID is under active development and more features will be added in the future.

A trial version of CONSID can be downloaded from the following ftp link, ftp://anonymous@ftp.swri.edu/pub/incoming/consid_0_9a.zip.

Acknowledgments

The author appreciates A. Bagtzoglou and two other anonymous reviewers for their constructive comments.

References


