Journal of Hydrology 501 (2013) 13–24

Contents lists available at ScienceDirect

Journal of Hydrology

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

Augmentation of groundwater monitoring networks using information theory and ensemble modeling with pedotransfer functions

A. Yakirevich [a], Y.A. Pachepsky [b,*], T.J. Gish [c], A.K. Guber [d], M.Y. Kuznetsov [a], R.E. Cady [e], T.J. Nicholson [e]

[a] Zuckerberg Institute for Water Research, J. Blaustein Institutes For Desert Research, Ben-Gurion University of Negev, Sede Boqer Campus 84990, Israel
[b] USDA-ARS, Environmental Microbial and Food Safety Lab., Beltsville, MD 20705, United States
[c] USDA-ARS, Hydrology and Remote Sensing Lab., Beltsville, MD 20705, United States
[d] Michigan State University, Dep. of Plant, Soil and Microbial Sciences. East Lansing, MI 48824, United States
[e] Office of Regulatory Research, U.S. Regulatory Commission, Rockville, MD, United States

ARTICLE INFO

Article history:
Received 25 September 2012
Received in revised form 22 July 2013
Accepted 25 July 2013
Available online 2 August 2013
This manuscript was handled by Peter K. Kitanidis, Editor-in-Chief, with the assistance of Roseanna M. Neupauer, Associate Editor

Keywords:
Groundwater monitoring network
Contaminant transport model
Ensemble modeling
Sequential design

SUMMARY

Improving understanding of chemical transport in the subsurface commonly employs evolving groundwater monitoring networks. The objective of this work was to apply the information theory to propose an objective algorithm for augmenting a subsurface monitoring network (SMN) with the purpose of discrimination of conceptually different subsurface flow and transport models. This method determines new monitoring locations where the Kullback–Leibler total information gain is maximized. The latter is computed based on estimates of the uncertainty in modeling results and uncertainty in observations. The method was applied to discriminate models in (1) a synthetic case of groundwater contamination from a point source; (2) the tracer experiment conducted at the USDA-ARS OPE3 research site where a pulse of KCL solution was applied with irrigation water and CL^- concentrations were subsequently monitored. Models were compared that included or ignored the effect of subsurface soil lenses on chemical transport. Pedotransfer functions were used to develop the ensemble of models for estimating the uncertainty in modeling results obtained with the numerical 3D flow and transport model. Peak tracer breakthrough concentrations were used to define the information gains. The determination of the new locations to augment existing ones was conducted on a 2-D grid. The information gain peaked in small area, and additional observation locations were very well spatially defined. Well-calibrated models provided a single optimal location, whereas, if models were not calibrated well, the Bayesian estimates of the new observation location depended on the activation sequence assumed for existing locations. The information gain maximization can suggest data collection locations to reduce uncertainties in the conceptual models of subsurface flow and transport.

Published by Elsevier B.V.

1. Introduction

Significant effort has been invested in the development of techniques to design effective groundwater monitoring networks (GMN). To this end, several state-of-the art reviews and guidance documents have been published during the past decade (Bloomfield, 2000; Hassan, 2003; Minsker, 2003; U.S. EPA, 2005; U.S. DOE, 2004; Kollat et al., 2011). To date, it is generally agreed that there is no single "best" method to optimize a long term groundwater monitoring network. The most significant advantage conferred by any optimization approach is that they are used to apply consistent, well-documented procedures, which incorporate formal decision tools, to the process of evaluating and optimizing monitoring programs (U.S. EPA, 2005).

The dynamic nature of GMN is an important factor in network design in many groundwater monitoring programs. Network design may be an iterative process, where initial sampling programs are often revised or updated as a result of collected data. Thus network augmentation or reduction is the characteristic feature of dynamic GMNs (Hudak and Loaiciga, 1992). In addition, the objectives of the monitoring network may also change with time. As a result, the dynamic GMN design includes the iterative validation–monitoring–refinement cycle (Hassan, 2003). The purposes of the GMN augmentation may include improvements in parameter estimation, source identification, plume delineation, as well as improvement and discrimination of conceptual models and their mathematical implementations.

GMN augmentation for model discrimination was addressed in early works by Knopman and Voss (1989), Usunoff et al. (1992) and Nordquist and Voss (1996). Usunoff et al. (1992)

* Corresponding author. Address: USDA-ARS, Environmental and Food Safety Lab., 10300 Baltimore Avenue, Building 173, Beltsville, MD 20705, United States. Tel.: +1 301 504 7468; fax: +1 301 504 6608.
E-mail address: yakov.pachepsky@ars.usda.gov (Y.A. Pachepsky).

0022-1694/$ - see front matter Published by Elsevier B.V.
http://dx.doi.org/10.1016/j.jhydrol.2013.07.032

noted that conceptual model uncertainties are often the main source of prediction uncertainties. They emphasized that simulating experiments with all available models is necessary but not sufficient since several models can be successfully fitted to measurements obtained from solute transport experiments. It was shown that up to one order of magnitude differences in peak concentrations and arrival times were found when performing long-term predictions (Usunoff et al., 1991). Knopman and Voss (1989) postulated that points of greatest difference in predictions can contribute the most information to the discriminatory power of a sampling design. They suggested that three objective functions be used in the design and optimization process: (1) the sum of squared differences in predicted vs. observed concentrations; (2) squared scaled difference; and (3) minimum squared difference. Nordquist and Voss (1996) developed an approach to model discrimination and GMN design based on the hypothesis that measurements in regions where alternative models produce the most divergent predictions are best suited for deciding which of the candidate models is the most appropriate (Knopman et al., 1991). The regions of high sensitivity for important system parameters can be considered as areas where measurements may be made during field tests aiming at efficient estimation of parameters and model discrimination. James and Gorelick (1994) developed a Bayesian data worth framework to improve cost-effectiveness of data collection in groundwater remediation problem. Recently the concept of the value of information as the context-specific metric of uncertainty has been introduced in groundwater remediation (Liu et al., 2012).

The argument has been made that comparisons between results in simulations with different models reflect uncertainties in both modeling results and experimental data (Himmelblau, 1970). Generally speaking, it may be difficult to distinguish between models where the difference between average estimated results from two models is large and the uncertainty of each average is also large. It may also be important to determine where the data values are uncertain; low confidence in data may make model comparisons with that data inappropriate.

A fruitful approach to evaluate the usefulness of a measurement for distinguishing between two hypotheses was proposed in the seminal paper (Kullback and Leibler, 1951) in which the information in a measurement for discrimination between two hypotheses was first defined. The mean value of this Kullback–Leibler information represented the information gain that could be encountered if the true hypothesis were accepted rather than the wrong one. This mean value eventually was termed Kullback–Leibler divergence, information gain, relative entropy, or information divergence, and was computed for model predictions an estimate of the information loss when full truth is approximated by the model (Poeter and Anderson, 2004). This measure was proven to be useful in (a) evaluation of predictive capabilities of hydrological models when observations presented the 'true' distribution that was approximated by model predictions (Wejss et al., 2010), (b) improving inverse solutions of groundwater flow models (Szucs et al., 2006), and (c) assessment of improvement in modeling results with data assimilation (Bulygina and Gupta, 2009). Various approximations of the Kullback–Leibler divergence resulted in development of the family of model discrimination criteria, such as Akaike criterion AIC, and later AICc (Burnham and Anderson, 2004) that were used in groundwater modeling to rank calibrated models (Poeter and Anderson, 2005; Foglia et al., 2007; Ye et al., 2008), and in inverse groundwater modeling for hydraulic conductivity estimation using Bayesian model averaging (Tsai and Li, 2008). All applications of the Kullback–Leibler divergence to model discrimination, however, relied on the existing set of observations and did not attempt to seek additional measurements.

Kullback (1959) showed that Kullback–Leibler information could be also applied for the selection of a new observation location to better discriminate between models without making an assumption that one of models generates the 'true' distribution whereas another one does not. Himmelblau (1970) implemented this suggestion for the case of non-linear models. The objective of this work was to apply the Kulback–Himmelblau sequential design method (KHSD) to discriminate models for the case study of inert tracer transport in variably saturated soil and shallow ground water. Specifically, this approach was used to determine the location where additional observations are needed to discriminate models in (1) a synthetic case of groundwater contamination from a point source; (2) the tracer experiment conducted at the USDA-ARS OPE3 research site where a pulse of KCL solution was applied with irrigation water and tracer transport was monitored afterwards.

2. Kullback–Himmelblau method to select new observation location for model discrimination

Different new observation locations will provide different additional information about model performance and different possibilities to discriminate between models. The Kullback–Himmelblau methodology quantifies these differences. The method is based on the value of information in a measurement to distinguish between two hypotheses as introduced by Kullback and Leibler (1951). Let Y be a random variable that is distributed with a probability density $p_1(y)$ when hypothesis H_1 is true (Model 1 is correct) and distributed with a probability density $p_2(y)$ when hypothesis H_2 is true (Model 2 is correct).

The quantity

$$\ln[p_1(y)/p_2(y)] \tag{1}$$

is defined as the 'information in y for discrimination between H_1 and H_2' (Kullback and Leibler, 1951). This is a measure of the odds in favor of choosing H_1 over H_2 or, from the information theory viewpoint, of the information in favor of hypothesis H_1 as opposed to hypothesis H_2. The expected information in favor of choosing H_1 over H_2, or information gain due to choosing H_1 over H_2 is

$$I(1:2) = \int_{-\infty}^{\infty} p_1(y) ln \frac{p_1(y)}{p_2(y)} dy \tag{2}$$

Similarly, the expected information in favor of choosing H_2 over H_1, or information gain due to choosing H_2 over H_1

$$I(2:1) = \int_{-\infty}^{\infty} p_2(y) ln \frac{p_2(y)}{p_1(y)} dy \tag{3}$$

Kullback (1959) proposed that total information gain due to selecting one model instead of another, i.e. value

$$J(1,2) = I(1:2) + I(2:1) = \int_{-\infty}^{\infty} [p_1(y) - p_2(y)] ln \frac{p_1(y)}{p_2(y)} dy \tag{4}$$

be maximized to distinguish between two models. It means that the new monitoring point has to be selected where the $J(1,2)$ value for model predictions reaches a maximum.

Explicit expression for $J(1,2)$ can be found under normality assumption for probability distribution function p_1 and p_2, i.e. for the case when two models are considered. Assume that n observations has been made and the sought new observation location is the location number $(n+1)$, and the model prediction in this point is denoted $Y^{(n+1)}$. Assume that the observations in this location $(n+1)$ are normally distributed about the expected value for the model, $\varepsilon\{Y_r^{(n+1)}\} = y_r^{(n+1)}$, $r = 1,2$, with a variance of σ_Y^2 Furthermore, $y_r^{(n+1)}$ is distributed in a local (linearized) region about a predicted value, $Y_r^{(n+1)}$, with a variance of σ_r^2. Consequently, $Y^{(n+1)}$

Download English Version:

https://daneshyari.com/en/article/6413510

Download Persian Version:

https://daneshyari.com/article/6413510

Daneshyari.com