Journal of Hydrology 531 (2015) 96–110

Contents lists available at ScienceDirect

Journal of Hydrology

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

Finding the right balance between groundwater model complexity and experimental effort via Bayesian model selection

Anneli Schöniger [a,*], Walter A. Illman [b], Thomas Wöhling [c,d], Wolfgang Nowak [e]

[a] Center for Applied Geoscience, University of Tübingen, Tübingen, Germany
[b] Department of Earth and Environmental Sciences, University of Waterloo, Waterloo, Ontario, Canada
[c] Water & Earth System Science (WESS) Competence Cluster, Institute for Geoscience, University of Tübingen, Tübingen, Germany
[d] Lincoln Environmental Research, Lincoln Agritech Limited, Hamilton, New Zealand
[e] Institute for Modelling Hydraulic and Environmental Systems (LS^3)/SimTech, University of Stuttgart, Stuttgart, Germany

ARTICLE INFO

Article history:
Available online 15 August 2015

Keywords:
Groundwater modeling
Hydraulic tomography
Geostatistics
Bayesian model averaging
Model selection
Model calibration

SUMMARY

Groundwater modelers face the challenge of how to assign representative parameter values to the studied aquifer. Several approaches are available to parameterize spatial heterogeneity in aquifer parameters. They differ in their conceptualization and complexity, ranging from homogeneous models to heterogeneous random fields. While it is common practice to invest more effort into data collection for models with a finer resolution of heterogeneities, there is a lack of advice which amount of data is required to justify a certain level of model complexity. In this study, we propose to use concepts related to Bayesian model selection to identify this balance. We demonstrate our approach on the characterization of a heterogeneous aquifer via hydraulic tomography in a sandbox experiment (Illman et al., 2010). We consider four increasingly complex parameterizations of hydraulic conductivity: (1) Effective homogeneous medium, (2) geology-based zonation, (3) interpolation by pilot points, and (4) geostatistical random fields. First, we investigate the shift in justified complexity with increasing amount of available data by constructing a *model confusion matrix*. This matrix indicates the maximum level of complexity that can be justified given a specific experimental setup. Second, we determine which parameterization is most adequate given the observed drawdown data. Third, we test how the different parameterizations perform in a validation setup. The results of our test case indicate that aquifer characterization via hydraulic tomography does not necessarily require (or justify) a geostatistical description. Instead, a zonation-based model might be a more robust choice, but only if the zonation is geologically adequate.

1. Introduction

Groundwater models are built for various types of investigations, both in science and in practice. They can serve as a basis for hypothesis testing, risk assessment, and management of resources. To provide reliable predictions for these objectives, models must be calibrated sufficiently well. However, in light of limited budgets, modelers have to cope with small calibration data sets. For physically-based models that consider the fundamentally important processes, the calibration procedure aims at finding appropriate parameterizations and then constraining the plausible parameter ranges. In groundwater modeling, the most effort is typically spent on characterizing the heterogeneity of the subsurface parameters hydraulic conductivity and specific storage. Under steady-state assumptions, only the spatial distribution of hydraulic conductivity influences the flow conditions.

Several approaches are available to characterize the heterogeneity in hydraulic conductivity, which differ in effort and scale. Traditionally, a large number of hydraulic conductivity estimates is obtained from collecting core samples and performing permeameter tests (Sudicky, 1986; Sudicky et al., 2010), or from performing slug or pumping tests. The local-scale information obtained from such campaigns is then regionalized to larger scales by upscaling, zonation, interpolation, or geostatistical simulation. Alternatively, more detailed measurements can be obtained from geophysical investigations (e.g., Hubbard and Rubin, 2000) or hydraulic tomography (e.g., Gottlieb and Dietrich, 1995; Butler et al., 1999; Yeh and Liu, 2000; Straface et al., 2007; Li et al., 2007; Illman et al., 2010).

Hydraulic tomography has been developed to investigate the heterogeneity in aquifer properties in a fine spatial resolution. A number of pumping tests is performed sequentially in different

* Corresponding author.
E-mail address: anneli.schoeniger@uni-tuebingen.de (A. Schöniger).

http://dx.doi.org/10.1016/j.jhydrol.2015.07.047
0022-1694/

wells at various locations throughout the aquifer. Pumping induces a spatial distribution of drawdown, which is captured by observation wells throughout the domain. These drawdown data are then used to derive via numerical inversion the spatial distribution of hydraulic conductivity and related properties such as connectivity. Further, the uncertainty attached to the inferred parameters can be quantified. The spatial resolution of the derived parameter distribution depends on the horizontal well spacing and the vertical packer intervals (Yeh and Liu, 2000).

Several approaches exist for the analysis and interpretation of the data obtained from all these aquifer characterization methods, and for the representation of the observed spatial heterogeneity in groundwater models. In general, a groundwater model with a specific spatial structure of hydraulic conductivity must be assumed. These assumptions vary in their conceptualization and their complexity (e.g., the number of parameters involved). Please note that the definition of model complexity is not unique, ranging from pure parameter counting over factor analysis to concepts that take into account probability distributions of parameters, data-parameter sensitivity and predictive variance. In principle, any parameterization ranging from the simple homogeneous case with an effective conductivity value to a geostatistical random field could be used. For the inversion of hydraulic tomography data, geostatistics-based inverse modeling methods are most frequently applied, such as the quasi-linear geostatistical approach (QL) (Kitanidis, 1995) and the sequential successive linear estimator (SSLE) (e.g., Yeh and Liu, 2000).

Eventually, the adequacy of the inferred hydraulic conductivity field and the overall groundwater model will depend on both the aquifer characterization technique and the chosen parameterization. The more data are available for calibration, the more detailed heterogeneities can be resolved. While it is common practice to invest more effort into data collection for geostatistical models (e.g. in form of hydraulic tomography data) than for simpler, effective conductivity models (e.g. in the form of core samples, slug tests or single-hole tests), there is a lack of advice, which amount and information content of data is required to justify a certain level of model complexity. We therefore see a need for a method that balances calibration effort (meaning both the effort for data collection and the computational effort to perform the inversion with the model) with model complexity and, implicitly, with model predictive performance. Assuming that the calibration effort increases with data set size, we use the amount of available data as proxy for the calibration effort in the following.

The formal statistical approach of Bayesian model averaging (BMA) (Draper, 1995; Hoeting et al., 1999) qualifies as such a method. It objectively ranks a number of competing models based on their fit to available data. Starting from a prior belief about the plausibility of each considered model, BMA updates this belief with knowledge from observed data via Bayes' theorem, and yields posterior model probabilities that reflect the updated plausibility. These probabilities allow for a quantitative ranking of the competing models and provide a basis for model selection. If more than one model obtains a significant model probability, their predictions can be combined in a weighted average that uses the probabilities as model weights. Finally, the uncertainty caused by the inability to uniquely choose only one of the considered models can be quantified as between-model variance.

BMA has been used in various disciplines as a statistical tool for model averaging (e.g., Ajami and Gu, 2010; Najafi et al., 2011; Seifert et al., 2012), model selection (e.g., Raftery, 1995; Huelsenbeck et al., 2004), quantification of model choice uncertainty (e.g., Rojas et al., 2008; Singh et al., 2010; Troldborg et al., 2010; Ye et al., 2010), data worth analysis (e.g., Rojas et al., 2010; Neuman et al., 2012; Xue et al., 2014; Wöhling et al., 2015), and model component dissection (Tsai and Elshall, 2013; Elshall and Tsai, 2014). In groundwater modeling, it has been applied to choose between different parameterizations of aquifer heterogeneity, e.g. by Ye et al. (2004), Tsai and Li (2008), Rojas et al. (2008), Morales-Casique et al. (2010), Seifert et al. (2012), and Elsheikh et al. (2013), to name only a few selected examples. Refsgaard et al. (2012) provide a review of strategies, including BMA, to address geological uncertainty in groundwater flow and transport modeling.

In the context of groundwater model selection and calibration, finding a balance between performance and complexity is of great interest (e.g., Yeh and Yoon, 1981; Fienen et al., 2009; Elsheikh et al., 2013). BMA is ideally suited to guide this search, because it implicitly honors the principle of parsimony or "Occam's razor" (Jeffreys, 1939; Gull, 1988). The BMA ranking reflects an optimal tradeoff between goodness-of-fit and model complexity, with model complexity being encoded in the prior probability distributions of the model parameters. The prior uncertainty in parameters is propagated through the model to the predictions, which are then compared to the observed data. A wide predictive distribution will be penalized by BMA, whereas a precise and accurate predictive distribution will be favored.

Although this optimal tradeoff is a main result of BMA, BMA has not yet been used to find the data amount required to justify a given level of complexity. In a certain sense, this reverses the direction in which BMA is usually applied, i.e. to rank models of different complexity for a given data set. We intend to fill this gap by isolating the complexity component of the tradeoff from its performance counterpart. We achieve this in a synthetic setup for BMA, where the models are mutually tested against their own predictions, instead of against real data. We introduce the concept of a *model confusion matrix*, which expresses how likely it is to identify the respective true model given the current experimental setup. We refer to this analysis as *model justifiability analysis*, because it reveals whether any specific level of complexity can be justified by the available amount and type of data (independent of the actually measured values) through the eyes of BMA. The question of justifiability is hence detached from the observed data values and becomes a function of the calibration effort only. Note that the calibration effort does not depend on the information content in the data (the effort for data collection is the same, no matter if the data turn out to be informative or not). The sensitivity of the model parameters to the data, on the other hand, has an impact on the outcome of BMA results and on the justifiability analysis.

While the *justifiability* analysis is based on the experimental design but not the actually measured data values, the *adequacy* of a model with regard to a specific prediction goal is defined by the tradeoff between complexity and performance in predicting the actually observed data values. The observations serve simultaneously as training and testing data for the specified model purpose. Hence, model adequacy as opposed to justifiability is assessed by the standard BMA routine based on the observed data. We therefore propose to perform BMA in a two-step procedure, running the synthetic justifiability analysis for the experimental setup first and determining the adequacy of each model in light of the observed data values in a second step that consists of the conventional BMA method. The results of the first step will then help to decide whether (a) the identified most adequate model is really the best choice given the current set of models, or (b) whether the identified model is only optimal given the currently too limited amount and information content of the data. The latter could occur when the available data do not allow to identify a more complex model among the model set, although the more complex model would actually be closer to the observed response of the system.

Further, the justifiability analysis can uncover the reasons for two models obtaining almost the same weight in the conventional