AGU100 ADVANCING EARTH AND SPACE SCIENCE

Water Resources Research

RESEARCH ARTICLE
10.1002/2017WR022135

Key Points:
- We develop a new framework to quantify uncertainty in subsurface prediction based on global sensitivity and prediction-focused approach
- We validate the use of a heat tracing experiment to provide useful information on energy recovery through sensitivity analysis
- We directly predict heat storage in the aquifer from heat tracing data without any explicit full model inversion

Supporting Information:
- Supporting Information S1

Correspondence to:
T. Hermans,
thomas.hermans@ugent.be

Citation:
Hermans, T., Nguyen, F., Klepikova, M., Dassargues, A., & Caers, J. (2018). Uncertainty quantification of medium-term heat storage from short-term geophysical experiments using bayesian evidential learning. *Water Resources Research*, *54*, 2931–2948. https://doi.org/10.1002/2017WR022135

Received 26 OCT 2017
Accepted 26 MAR 2018
Accepted article online 30 MAR 2018
Published online 16 APR 2018
Corrected 3 MAY 2018

This article was corrected on 3 MAY 2018. See the end of the full text for details.

© 2018. American Geophysical Union. All Rights Reserved.

Uncertainty Quantification of Medium-Term Heat Storage From Short-Term Geophysical Experiments Using Bayesian Evidential Learning

Thomas Hermans[1,2], Frédéric Nguyen[3], Maria Klepikova[4], Alain Dassargues[5], and Jef Caers[2]

[1]Department of Geology, Ghent University, Gent, Belgium, [2]Department of Geological Sciences, Stanford University, Stanford, CA, USA, [3]Urban and Environmental Engineering, Applied Geophysics, University of Liege, Liege, Belgium, [4]Department of Earth Sciences, ETH Zürich, Zürich, Switzerland, [5]Urban and Environmental Engineering, Hydrogeology and Environmental Geology, University of Liege, Liege, Belgium

Abstract In theory, aquifer thermal energy storage (ATES) systems can recover in winter the heat stored in the aquifer during summer to increase the energy efficiency of the system. In practice, the energy efficiency is often lower than expected from simulations due to spatial heterogeneity of hydraulic properties or non-favorable hydrogeological conditions. A proper design of ATES systems should therefore consider the uncertainty of the prediction related to those parameters. We use a novel framework called Bayesian Evidential Learning (BEL) to estimate the heat storage capacity of an alluvial aquifer using a heat tracing experiment. BEL is based on two main stages: pre- and postfield data acquisition. Before data acquisition, Monte Carlo simulations and global sensitivity analysis are used to assess the information content of the data to reduce the uncertainty of the prediction. After data acquisition, prior falsification and machine learning based on the same Monte Carlo are used to directly assess uncertainty on key prediction variables from observations. The result is a full quantification of the posterior distribution of the prediction conditioned to observed data, without any explicit full model inversion. We demonstrate the methodology in field conditions and validate the framework using independent measurements.

Plain Language Summary Geothermal energy can be extracted or stored in shallow aquifers through systems called aquifer thermal energy storage (ATES). In practice, the energy efficiency of those systems is often lower than expected because of the uncertainty related to the subsurface. To assess the uncertainty, a common method in the scientific community is to generate multiple models of the subsurface fitting the available data, a process called stochastic inversion. However this process is time consuming and difficult to apply in practice for real systems. In this contribution, we develop a novel approach to avoid the inversion process called Bayesian Evidential Learning. We are still using many models of the subsurface, but we do not try to fit the available data. Instead, we use the model to learn a direct relationship between the data and the response of interest to the user. For ATES systems, this response corresponds to the energy extracted from the system. It allows to predict the amount of energy extracted with a quantification of the uncertainty. This framework makes uncertainty assessment easier and faster, a prerequisite for robust risk analysis and decision making. We demonstrate the method in a feasibility study of ATES design.

1. Introduction

The use of groundwater heat pump systems (GWHP) to provide heating or cooling to buildings allows for large primary energy savings and significant CO_2 reductions through aquifer thermal energy storage (ATES) (e.g., Bayer et al., 2013; Saner et al., 2010; Vanhoudt et al., 2011). In such systems, groundwater is extracted from production wells and circulated through a heat exchanger to heat buildings in winter. Cold water is reinjected in the aquifer through injection wells. In summer, the system is reversed: cold water injected during the winter is extracted to cool buildings and warmer water is reinjected in the subsurface. Designing ATES systems requires a multidisciplinary approach where the energy demand of the building must be known to estimate the volume of water to be extracted and the resulting change of temperature induced in the

aquifer. In turn, hydrogeological modeling is performed to ensure that the aquifer is able to meet those criteria in a sustainable way and with an efficient energy recovery.

For ATES systems to work efficiently, several conditions must be met (e.g., Bakr et al., 2013; Lee, 2010). Permeability and porosity of the aquifer must be sufficient to pump large volumes of water without emptying the resources. Efficient recovery of thermal energy stored in the aquifer is only possible in specific hydrogeological conditions such as low natural gradient. In addition, the spatial geological heterogeneity of aquifers makes the efficiency of such systems difficult to predict, since such heterogeneity influences groundwater flow and the spatial distribution of hot and cold groundwater plumes (Bridger & Allen, 2010, 2014; Ferguson, 2007; Sommer et al., 2013, 2014). Furthermore, hydrogeological conditions and spatial uncertainty of the aquifer parameters may interact and affect the recovery of stored thermal energy. For example, Possemiers et al. (2015) showed that the presence of clay drapes in a sandy aquifer may lead to an increase or decrease of energy efficiency depending on the presence of a natural gradient or not.

Therefore, the proper design of ATES systems and their long-term efficiency strongly depend on the subsurface components of the system (e.g., Bloemendal et al., 2014; Chen et al., 2015). Blum et al. (2011) studied more than 1,000 shallow geothermal systems installed in Germany and showed that inappropriate consideration of the subsurface characteristics lead to over- or undersizing the system. The decision for the development of an ATES system and its design should thus be based on realistic predictions. This requires the development of a hydrogeological model able to simulate heat flow and transport within the subsurface and account for uncertainties related to the subsurface. The expected outcome of such a model is to predict the evolution of groundwater temperature at the extraction and re-injection wells, which combined with the pumping/injection flowrates quantifies the amount of thermal energy.

Although the desired prediction seems relatively simple, building reliable ATES models is a challenging and time-consuming task. It must consider many uncertain parameters involved in heat flow and transport such as hydraulic conductivity, porosity, thermal conductivity, or volumetric heat capacity, their associated spatial heterogeneity and external inputs such as aquifer recharge or boundary conditions. In many cases, the lack of available data leads the modeler to consider homogeneous layered conceptual models (Kim et al., 2010; Lo Russo & Civita, 2009; Nam & Ooka, 2010; Yapparova et al., 2014). Ignoring spatial heterogeneity bears the risk of making wrong decisions based on the prediction of those models or altering the robustness of the ATES design.

In practice, to improve the reliability of models, in-situ tests such as push/pull tests (Klepikova et al., 2016a; Park et al., 2015; Vandenbohede et al., 2009), heat storage experiments (Palmer et al., 1992, Vandenbohede et al., 2011), heat tracer tests (Wagner et al., 2014, Wildemeersch et al., 2014; Macfarlane et al., 2002) or other specific tests (e.g., Kuo & Liao, 2012) can be performed to gain knowledge on subsurface parameters. However, these conventional approaches generally lack the spatial coverage required to characterize the heterogeneity of the subsurface. The complementary use of geophysical methods which provide spatial information on the subsurface with a greater coverage than boreholes has been increasing in the past years. Not only geophysics can be used to define the layers and their geometry, it also has been recently demonstrated that time-lapse geophysics using electrical resistivity tomography (ERT) could be used to monitor temperature changes during heat storage and tracing experiments in aquifers (Arato et al., 2015; Hermans et al., 2012, 2014, 2015b).

An inverse problem approach is typically used to estimate the model parameters based on observed data (borehole and/or geophysics). However, for a proper uncertainty quantification of the prediction variables and a robust design, stochastic approaches accounting for realistic patterns of heterogeneity and parameters uncertainty are necessary (e.g., Fu & Gómez-Hernández, 2009; Gómez-Hernández et al., 1997; Linde et al., 2015; Oliver et al., 1997; Vrugt et al., 2013; Yoon et al., 2013). Most stochastic inversion methods are based on a Bayesian formalism where the posterior distribution $f(\boldsymbol{m}|\boldsymbol{d})$, consisting of an ensemble of models fitting the data, is expressed as

$$f(\boldsymbol{m}|\boldsymbol{d}) = kf(\boldsymbol{m})L(\boldsymbol{m}|\boldsymbol{d}) \tag{1}$$

where $\boldsymbol{m}$ represents the model parameters and $\boldsymbol{d}$ the data. $f(\boldsymbol{m})$ is the prior distribution of model parameters, based on the background knowledge we have on the problem and the study site, $L(\boldsymbol{m}|\boldsymbol{d})$ is a likelihood function, generally measuring the misfit between observed and calculated data, and k is a proportionality constant. It is then possible to estimate the desired prediction for each model of the

posterior distribution and make decisions and risk analysis based on the uncertainty of these predictions. However, stochastic methods are computationally expensive with tens to hundreds thousands of forward model runs, often difficult to implement and tune to convergence, and are therefore not widely used in real field applications.

Bayesian Evidential Learning (BEL) recognizes a number of important difficulties that arise when applying inverse and other modeling methodologies in the real world (see Scheidt et al., 2018). The first concerns the statement of the prior distribution on all variables (including spatial variables). Such prior distribution needs to be able to predict the observations. Any prior that cannot predict the data with finite probability is falsified, and requires restating the prior, which could mean (1) changing the distributions, (2) increasing model complexity (adding variables). Second, the ultimate aim of modeling and predicting is not the model itself, but the key decision variables. In the ATES case, this would be the future temperature evolution over time, as the system is put in operation. Third, only data that inform these key decision variables are relevant. BEL uses a collection of machine learning and sensitivity analysis methods by learning from Monte Carlo based on falsified prior uncertainty models on all model variables jointly. In that sense, it aims to circumvent any difficult model inversions through a prediction-focused approach. Since the method is purely based on Monte Carlo (and not Markov Chain Monte Carlo), it has proven efficient in a variety of application in oil/gas, hydrogeology and geophysics (see Scheidt et al., 2018). The terms *Evidential* and *Learning* refer therefore to the central position occupied by the data and the use of prior models to derive a direct data-prediction relationship, respectively.

Our intended contribution is to demonstrate the ability of BEL to assess the uncertainty of the key variable temperature in a prediction problem facing various sources of uncertainty in the case of a specific ATES design. Here, we will use the example of estimating the heat storage capacity of an alluvial aquifer using a real heat tracing experiment but the framework can be applied to any prediction problem. The paper is organized as follows. First, we describe the methodology and present the studied problem, i.e., the prediction of heat storage in an alluvial aquifer. Then, we build a realistic prior model of the studied aquifer, considering uncertainty on the hydraulic conductivity distribution and its spatial heterogeneity, the porosity and boundary conditions. We analyze the sensitivity of these prior parameters on the prediction of heat storage. Then, we consider the acquisition of data during a short-term heat tracing experiment, known to be informative for the prediction. In the next step, we verify if the acquired data are consistent with our prior description of the aquifer. Finally, we forecast the posterior distribution of the prediction considering the acquired data and validate the approach for another experiment with validation measurements.

2. Bayesian Evidential Learning

In this section, we describe the Bayesian Evidential Learning framework (Figure 1a). In this study, the ultimate objective of BEL is to quantify uncertainty on the heat storage capacity of the aquifer using the evolution of temperature at the pumping well. In the following, we will often refer to this specific case to illustrate the methodology. However, the framework can be adapted to handle any kind of predictions related to the subsurface.

BEL can be divided in two main stages: pre- and postfield data acquisition. In the first stage, the prior distribution is stated and sampled using Monte Carlo. A global sensitivity analysis is used to assess the value of acquiring a new data set. After field data acquisition, the prior distribution is falsified, a relationship between data and prediction is derived using machine learning and the posterior distribution of the prediction is computed, and if possible, validated.

2.1. BEL Prefield Data Acquisition

2.1.1. Definition and Sampling of the Prior

The prior aims at identifying the possible range of variation of parameters from existing knowledge. This includes broad information on the geological context and specific information on the study site acquired in previous studies.

We sample the prior distribution of model parameters and prediction variables using Monte Carlo, i.e., model parameters are randomly sampled from the prior and the prediction is simulated using a forward groundwater flow and transport model. Similarly, for any proposed experiment, the same sampled models can be used to simulate the variable corresponding to experimental data to get the prior distribution of

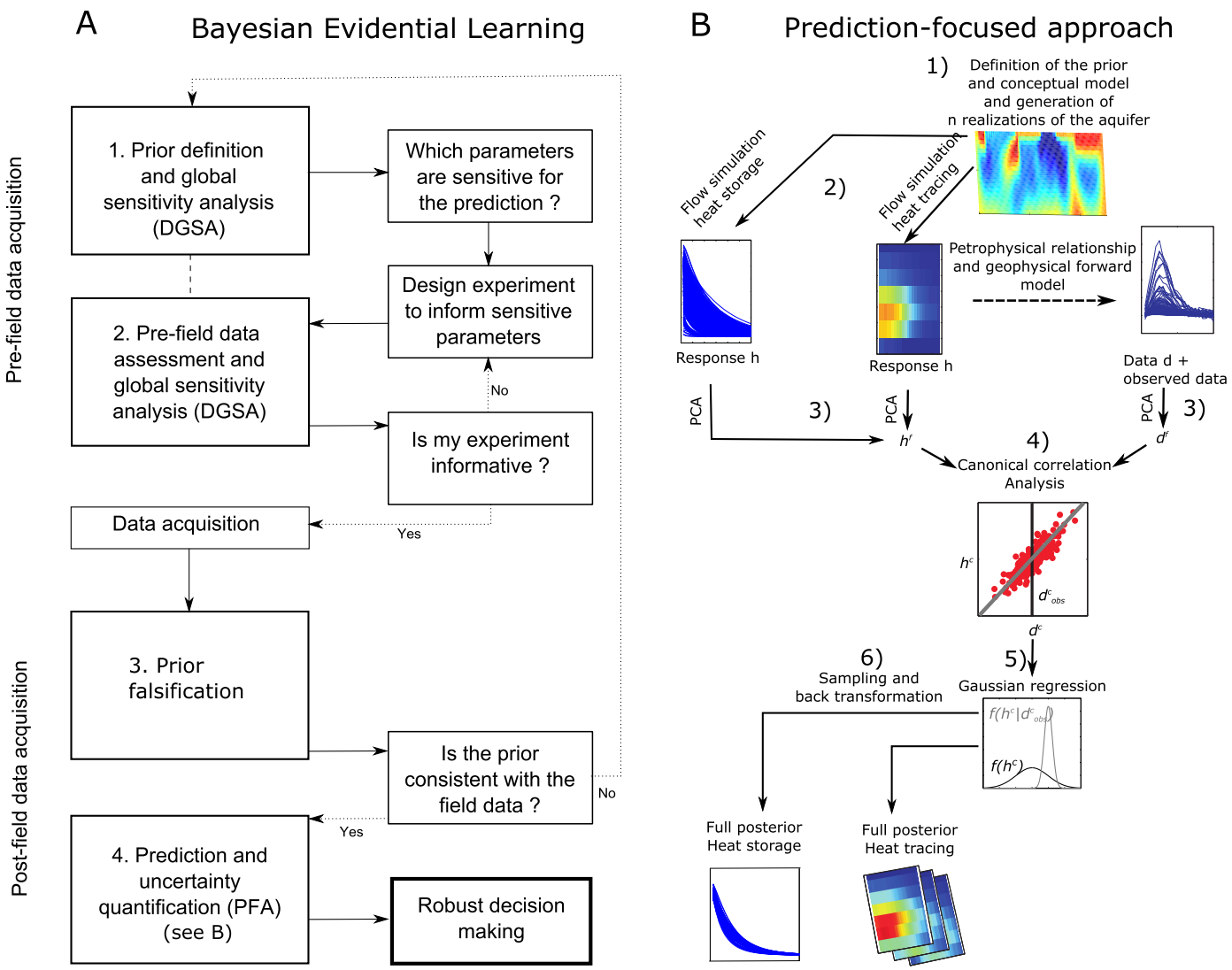

Figure 1. (a) Workflow of Bayesian Evidential Learning. (b) Detailed workflow of the prediction-focused approach. Steps 1 and 2 are already performed in the pre-field data acquisition step of BEL.

experimental data. The objective is to assess the range of uncertainty in the model responses (both in data and prediction) resulting from the uncertainty in the prior.

2.1.2. Global Sensitivity Analysis and Data-Worth Assessment

In this step, BEL uses a global sensitivity considering all parameters simultaneously to identify sensitive parameters to the prediction, i.e., the evolution of the temperature distribution with time in the pumping well. BEL being a strategy rather than a specific technique, it can be applied with any global sensitivity method such as Sobol' (e.g., Oladyshkin et al., 2012).

In this application, we use the DGSA (distance-based global sensitivity analysis) method to perform the global sensitivity analysis for its ability to handle the time distribution of studied variables, various types of parameters (continuous, discrete, spatial uncertainty), and the limited number of models required to compute the sensitivity and interactions. DGSA calculates the distance between the variables within the prior distribution of the studied variables to derive the sensitivity. For more details on DGSA, we refer to the supporting information Text S1 and to Park et al. (2016) and Fenwick et al. (2014).

An experimental set-up is proposed to collect information on the prediction (e.g., heat storage or heat tracing experiment) and a global sensitivity analysis on the data prior distribution is used to assess the information content of new experimental data regarding the prediction. This requires simulating the proposed

experiment for each sampled realization of the prior. The sensitivity analysis identifies if the set-up is likely to generate useful information to reduce the uncertainty on the prediction. If it is not the case, another data proposal has to be made (Figure 1a).

Note that performing global sensitivity analysis does not require the collection of any field data; instead, it requires the definition of a prior model. Once an adapted experiment has been identified, we can proceed to data acquisition (Figure 1a).

2.2. BEL Postfield Data Acquisition

2.2.1. Prior Falsification

Bayesian methods require that the posterior solution is part of the span of the prior (e.g., Hou & Rubin, 2005). It is important to verify that the acquired data are able to be predicted by the current prior. If it is not the case, there is a risk for the subsequent prediction to be completely erroneous (Chalmers, 2013). If the prior and the data are not consistent, i.e., the observed data cannot be reproduced by the prior, the prior is falsified and revising the prior model is necessary (see Figure 1a).

In BEL, we thus propose to verify that the prior is able to simulate the observed data before any attempt to predict the posterior (Hermans et al., 2015a; Park et al., 2013; Scheidt et al., 2015a). This step, called prior falsification, is done by a direct comparison of field data with its prior distribution (already computed for the sensitivity analysis) both in the physical space and in a reduced-dimension space. Note that prior model falsification does not require matching data (see Scheidt et al, 2015a, 2018). Falsification refers to the idea that priors cannot be proven correct; they can only be proven incorrect (Popper, 1959).

2.2.2. Prediction-Focused Approach

Once the prior is not falsified, posterior prediction can be made. To compute the posterior distribution of the prediction, BEL uses a prediction-focused approach (PFA, Figure 1b). PFA takes advantage of the prior simulations ran in the prefield data acquisition step (sensitivity analysis) to derive a direct relationship between data and prediction variables using machine learning. Therefore, PFA requires only *2 x n* forward runs (*n* for the prediction and *n* for the data) to compute the posterior distribution, which drastically reduces the time to compute the posterior distribution of the prediction. The method differs from standard Bayesian methods developed in hydrogeology (e.g., Fu & Gómez-Hernández, 2009; Linde et al., 2015; Rubin et al., 2010) because its likelihood function is formulated in a reduced-dimension space based on target variables (the prediction) and not on model parameters. In this section, we present a brief summary of the prediction-focused approach used in this paper. As for global sensitivity, BEL is not limited to this specific approach. For a more detailed description, we refer to Hermans et al. (2016b) and Satija and Caers (2015).

Our application of PFA can be divided in 6 steps (Figure 1b). Steps 1 and 2 correspond to the definition and sampling of the prior and the forward simulation of the prediction $\boldsymbol{h}$ and the data $\boldsymbol{d}$ for each realization. Those steps actually correspond to the prefield data acquisition of BEL.

Both data and forecast variables are high-dimensional. To reduce the size of the problem, we independently apply principal component analysis (PCA) on both $\boldsymbol{d}$ and $\boldsymbol{h}$ and retain only their first p and q components, respectively. This provides two reduced sets $\boldsymbol{d}^{\boldsymbol{f}}$ and $\boldsymbol{h}^{\boldsymbol{f}}$ (Step 3). Then, we linearize the relationship between $\boldsymbol{d}^{\boldsymbol{f}}$ and $\boldsymbol{h}^{\boldsymbol{f}}$ into a set of independent linear relationships between $\boldsymbol{d}^{\boldsymbol{c}}$ and $\boldsymbol{h}^{\boldsymbol{c}}$(second dimension reduction) using canonical correlation analysis (CCA, Step 4). This is the learning step of the methodology. If CCA fails to generate a useful relationship (i.e., no clear linear relationship exists between data and prediction variables), then PFA will not produce a realistic solution. This might happen because the data is not informative for the prediction or because a more complex relationship is present, in which case other methods should be applied (Scheidt et al., 2015b).

At this stage, we formulate the posterior distribution of the forecast in the reduced dimension space $\boldsymbol{h}^{\boldsymbol{c}}$ given the field observed data $\boldsymbol{d}^{\boldsymbol{c}}_{\boldsymbol{obs}}$ in reduced dimensions using Bayes' rule:

$$f_H\left(\boldsymbol{h}^{\boldsymbol{c}}|\boldsymbol{d}^{\boldsymbol{c}}_{\boldsymbol{obs}}\right)=kf_H(\boldsymbol{h}^{\boldsymbol{c}})L\left(\boldsymbol{h}^{\boldsymbol{c}}|\boldsymbol{d}^{\boldsymbol{c}}_{\boldsymbol{obs}}\right) \tag{2}$$

where $f_H(\boldsymbol{h}^{\boldsymbol{c}})$ is the prior distribution, $L\left(\boldsymbol{h}^{\boldsymbol{c}}|\boldsymbol{d}^{\boldsymbol{c}}_{\boldsymbol{obs}}\right)$ is a likelihood function and k is a proportionality constant. If we ensure by histogram transformation that $\boldsymbol{h}^{\boldsymbol{c}}$ is multivariate Gaussian, we can use a Gaussian regression with a multivariate Gaussian likelihood to solve the problem analytically with mean and covariance equal to (Tarantola, 2005):

$$\tilde{\boldsymbol{h}}_{\boldsymbol{c}} = \left(\boldsymbol{G}^{T}\boldsymbol{C}_{\boldsymbol{c}}^{\boldsymbol{d}} - 1\boldsymbol{G} + \boldsymbol{C}_{\boldsymbol{Hc}}^{-1}\right)^{-1}\left(\boldsymbol{G}^{T}\boldsymbol{C}_{\boldsymbol{c}}^{\boldsymbol{d}} - 1\boldsymbol{d}_{\boldsymbol{obs}}^{\boldsymbol{c}} + \boldsymbol{C}_{\boldsymbol{Hc}}^{-1}\bar{\boldsymbol{h}}^{\boldsymbol{c}}\right) \quad (3)$$

$$\tilde{\boldsymbol{C}}_{\boldsymbol{H}}^{\boldsymbol{c}} = \boldsymbol{G}^{T}\boldsymbol{C}_{\boldsymbol{c}}^{\boldsymbol{d}-1}\boldsymbol{A} + \boldsymbol{C}_{\boldsymbol{Hc}}^{-1} \quad (4)$$

where $\boldsymbol{G}$ is the linear forward operator modeling the relationship between $\boldsymbol{d}^{\boldsymbol{c}}$ and $\boldsymbol{h}^{\boldsymbol{c}}$, $\boldsymbol{C}_{\boldsymbol{c}}^{\boldsymbol{d}}$ is the observational error covariance matrix, $\bar{\boldsymbol{h}}^{\boldsymbol{c}}$ and $\boldsymbol{C}_{\boldsymbol{Hc}}$ are the mean and the covariance matrix of the sampled distribution, and $\boldsymbol{A}$ is the matrix mapping the transformation of $\boldsymbol{d}^{\boldsymbol{f}}$ to $\boldsymbol{d}^{\boldsymbol{c}}$ (Step 4). A linear regression model is used to account for residual error in the linear regression between $\boldsymbol{d}^{\boldsymbol{c}}$ and $\boldsymbol{h}^{\boldsymbol{c}}$ (Satija & Caers, 2015). The derivation of $\boldsymbol{C}_{\boldsymbol{c}}^{\boldsymbol{d}}$ for field data according to Hermans et al. (2016b) is described in supporting information Text S3. It is now straightforward to generate samples in the reduced dimension space. By back transformation of CCA and PCA, it is possible to get the posterior distribution in the original space (Step 6).

Note that other techniques of dimension reduction and sampling can be used in PFA depending on the specificity of the study. For example, if the process generates a non-linear relationship or if the Gaussian histogram transformation alters the linearity, it is possible to use kernel density estimation to estimate the posterior distribution.

3. Predicting Thermal Energy Storage Efficiency in a Real Aquifer

3.1. Prediction Scenario

To assess the heat storage potential of the studied aquifer, we define a heat storage and recovery cycle scenario. During 30 days, heated water is continuously injected in a well at the rate of 3 m^3/h. The temperature of the injected water is 10°C above the background temperature. After 30 days, water is extracted from the well at the same rate of 3 m^3/h. The change in temperature of the pumped water compared to the initial value in the aquifer (i.e., before injection) is recorded for another 30 days. The thermal energy recovery can be estimated using the temperature of the extracted water and the flowrate. The flowrate being fixed, the heat storage capacity is here directly related to the decrease of temperature with time during the pumping phase. This constitutes our prediction (Figure 1a).

This simplified scenario enables to assess the storage potential of the aquifer without considering changing boundary conditions with time and the interaction with the building. Longer simulations would be necessary for the design of the ATES system itself and the flowrate should be adapted according to the energy requirement of the building.

3.2. Study Site

The study site is located in the alluvial aquifer of the Meuse River in Hermalle-sous-Argenteau, Belgium. We build the conceptual model of the aquifer and the prior, based on previous studies carried out on the study site and similar sites in the alluvial aquifer of the Meuse River, including geological logs, pumping and tracing experiments, geophysical surveys and groundwater models (Brouyère, 2001; Dassargues, 1997; Derouane & Dassargues, 1998; Hermans et al., 2015a, 2015b, Klepikova et al., 2016b; Wildemeersch et al., 2014). The water level is located at 3.2 m depth. Based on borehole logs, the saturated deposits can be divided into two distinct layers. The upper layer, between 3 and about 7 m depth, is composed of gravel in a sandy matrix. The bottom layer is composed of coarse clean gravel. The bedrock composed of low permeability carboniferous shale and sandstones lies at 10 m depth and constitutes the basement of the alluvial aquifer (Figure 2b).

We build a hydrogeological model of the subsurface using the control-volume finite element code HydroGeoSphere (Therrien et al., 2010). The discretization of the model is based on geological information obtained from boreholes (Wildemeersch et al., 2014). The saturated part of the subsurface is modeled using 14 layers 0.5 m thick down to the carboniferous bedrock. The bottom 6 layers correspond to clean gravel, the top 8 layers to sandy gravel. The grid around the wells is discretized with elements 0.025 m wide. The size of the elements is progressively increased by a factor of 1.15 (Klepikova et al., 2016b). The total size of the model is 60 m in the direction of flow, 40 m perpendicularly and 7 m vertically (Figure 2a).

The model is oriented such that its main axis corresponds to the natural direction of flow which was identified in previous studies (Brouyère 2001). Therefore, no-flow boundary conditions were assigned to boundaries

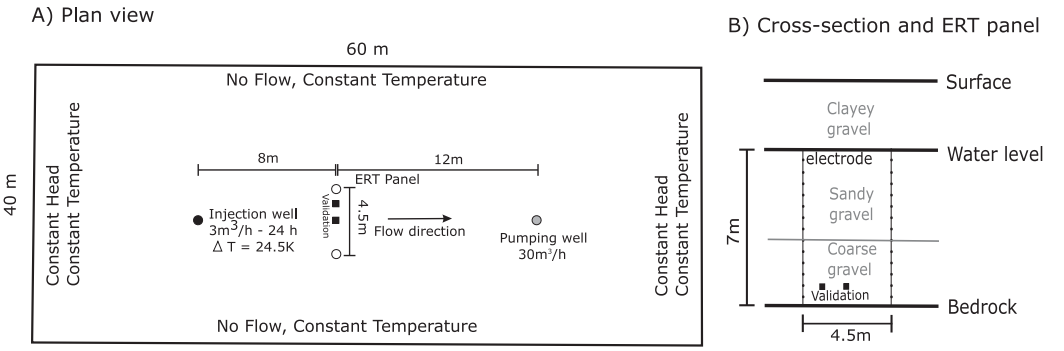

Figure 2. (a) Plan view of the experimental set-up and the model for the heat tracing experiment (boundaries are not to scale). (b) Vertical cross section and electrode layout in the ERT panel.

parallel to this direction. A no-flow boundary condition was also assigned to the bedrock. Between the up and down-gradient boundaries, a natural gradient exists and is imposed through constant head boundaries.

Heat transport is simulated by analogy to solute transport, homogeneous initial temperature conditions, corresponding to the average temperature encountered in the aquifer, are used. Those initial conditions are also used as fixed temperature boundary conditions during the whole simulation.

3.3. Proposed Experiment

The new experiment considered here (Figure 1a) to reduce the uncertainty on the prediction is a forced gradient heat tracing experiment monitored with cross-borehole electrical resistivity tomography (Figure 2). Heat tracing is an interesting candidate because it is influenced by both hydraulic and thermal properties of the aquifer. In addition, in forced gradient conditions, heat tracing enables to investigate a larger volume of subsurface than short-term heat storage experiments of the same duration. The forced gradient is in the direction of the natural one so that it speeds up groundwater flow without affecting too much the flow direction. Geophysical methods enable to collect spatially distributed data, informing on the spatial heterogeneity of the aquifer (Hermans et al., 2014).

In the present study, we consider collecting two different data sets. In the first case, only the first 30h of the experiment, corresponding to the increasing part of the temperature breakthrough curve in the ERT panel, are considered. We will refer to it as the 1 day experiment. In the second case, the tailing part of the breakthrough curve is also considered. It will be referred to as the 5 day experiment. The difference between the two cases illustrates a choice between two tests, one being shorter and less expensive but also likely less informative than the other.

3.4. Description of the Prior and Global Sensitivity Analysis

We consider here three types of uncertainty: hydrogeological properties, hydrogeological conditions, and the model of spatial continuity describing the spatial heterogeneity. Then we analyze the sensitivity of the prediction to these parameters.

3.4.1. Hydrogeological Properties and Boundary Conditions Uncertainties

The main hydrogeological unit responsible for flow and transport in the aquifer is the coarse clean gravel layer lying above the bedrock. To represent realistically our prior knowledge of the hydraulic conductivity distribution (K) and its uncertainty, we consider 2 different parameters: the mean of the logarithmic distribution of K and its variance. Based on previous studies, we consider values ranging from -4 to -1 for the mean of the logarithm distribution. The variance ranges from 0.05 (relatively homogeneous deposits) to 1.5 (heterogeneous deposits).

The porosity value is uncertain, but considered constant within the field. The porosity ranges between 5 and 11% in previous studies. However, we consider in this study a larger range of variation with values up to 40%, which can be encountered in soft sediments. The reason for this is the influence of the porosity on thermal properties (heat capacity, thermal conductivity) of the porous medium. The parameters for heat transport are calculated through volumetric average between the solid and aqueous phases

$$\kappa_b = (1-\theta)\kappa_s + \theta\kappa_w \quad (5)$$

where κ_b is a bulk thermal parameter and θ is the porosity; the subscripts s and w represent the solid and the water phase, respectively. The uncertainty on the porosity indirectly induces uncertainty on the thermal parameters of the aquifer.

Given the importance of the natural gradient on heat storage in aquifers (Possemiers et al., 2015), the natural gradient was also considered as an uncertain parameter. Gradient ranging from 0 to 0.167% were considered (Brouyère, 2001). The direction of the gradient is fixed.

3.4.2. Model of Spatial Continuity

The spatial heterogeneity of K within the coarse gravel layer is described by a model of spatial continuity using a spherical variogram. The spatial continuity on the site is largely unknown. The description of existing destructive boreholes is not precise enough to estimate natural variations within the layers.

We first consider the variogram range (i.e., the correlation length) in the direction of flow to lie between 1 and 10 m. The range in the perpendicular direction is related to the main range through an anisotropy ratio ranging from 0.5 to 10. The orientation of the main ranges can deviate from the direction of flow from $-45°$ to 45°.

The hydraulic conductivity parameter distribution and the model of spatial continuity are used to generate independent realizations of the hydraulic conductivity field using sequential Gaussian simulations (Goovaerts, 1997).

3.4.3. Global Sensitivity Analysis

In absence of information, all parameter prior distributions were modeled with uniform distributions, using the principle of indifference (Jaynes, 2003). All other parameters were considered fixed. A summary of the parameters and their range of variations are provided in Table 1. We generate a representative set of 500 realizations of the prior model parameters described in the previous sections, by means of simple Monte Carlo simulations. DGSA and later PFA are using dimension reduction and averaging techniques that make them robust even for a limited number of samples and relatively large number of parameters. Same ranges of number of models were used by other authors (Fenwick et al., 2014; Park et al., 2016). We further discuss the choice of the number of models in the dedicated section.

For each model, we simulate the heat storage experiment and generate the curve representing the temperature at the well during the pumping phase. This constitutes the prior distribution of the prediction (Figure 3a). To identify the most sensitive parameters (step 1, Figure 1a), we use the prior distribution of the temperature at the well during the whole pumping phase as the input for DGSA (Figure 3b and supporting information Text S1). The parameter exhibiting the most impact on the prediction is the mean value of the hydraulic conductivity. The variance of the hydraulic conductivity distribution and the gradient are also sensitive parameters. The anisotropy and the range influence the prediction to a lesser extent. The influence of the porosity and the orientation on the prediction is limited. Globally, high hydraulic conductivity and high gradient tend to reduce the heat storage capacity of the aquifer (low temperature) due to higher groundwater fluxes.

Table 1
Parameters Used for the Heat Flow and Transport Simulations

Parameters	Fixed/variable	Value
Mean of $\log_{10}$ K (m/s)	Variable	U[-4 -1]
Variance $\log_{10}$ K (m/s)	Variable	U[0.05 1.5]
Range (m)	Variable	U[1 10]
Anisotropy ratio	Variable	U[0.5 10]
Orientation	Variable	U[$-\pi/4$ $-\pi/4$]
Porosity	Variable	U[0.05 0.40]
Gradient (%)	Variable	U[0 0.167]
$\log_{10}$ K (m/s) – upper layer	Fixed	10^{-5}
Longitudinal dispersivity (m)	Fixed	1
Transverse dispersivity (m)	Fixed	0.1
Solid thermal conductivity (W/mK)	Fixed	3
Water thermal conductivity (W/mK)	Fixed	0.59
Solid specific heat capacity (J/kgK)	Fixed	1,000
Water specific heat capacity (J/kgK)	Fixed	4,189

Note. U indicates that the value was randomly sampled from a uniform distribution

To compare with our stochastic approach, we also provide the prediction for a hydrogeological model where the spatially distributed hydraulic conductivity is calibrated through the pilot-point method using temperature data during a heat tracing experiment (Klepikova et al., 2016b; see below). The prediction from this model indicates that the hydraulic conductivity of the site is probably not adequate for heat storage because the recovered energy is low (ΔT is rapidly equal to 0, Figure 3a). However, this model was calibrated under fixed boundary conditions with a gradient at 0.06% that have a strong influence on the thermal energy recovery.

3.5. Prefield Data Worth Assessment: An ERT Heat Tracing Experiment

To reduce uncertainty on the prediction, we consider the collection of a new data set: a heat tracing experiment monitored by ERT (step 2,

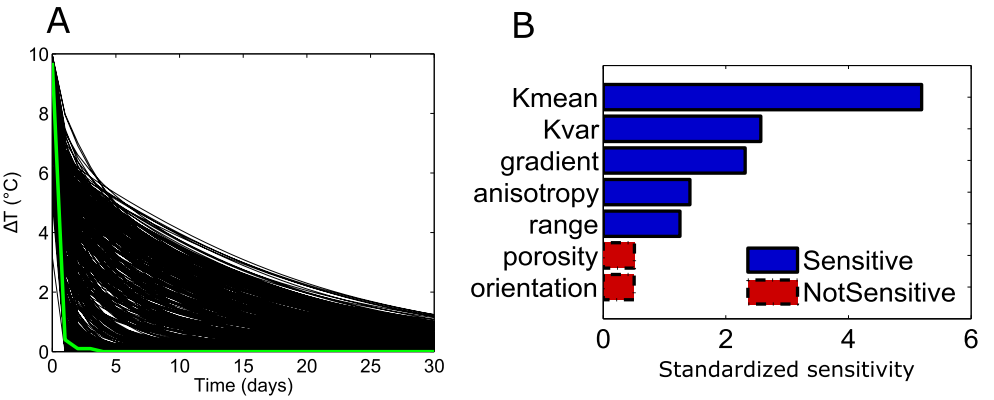

Figure 3. (a) Prior distribution of the prediction (temperature during heat storage). Each black curve represents the prediction for one model of the prior. The green curve indicates the prediction for the calibrated hydraulic conductivity distribution. (b) Global sensitivity analysis of the parameters considered in the prior for the prediction response.

Figure 1a). For each of the 500 models of the prior set, we simulate the geophysical data set collection. Then, we apply DGSA on the simulated data sets. If the data response is sensitive to the same parameters as the prediction, then conditions are favorable to reduce the uncertainty on the prediction with that data set.

The heat tracing experiment is simulated for the 500 models of the prior set. The temperature distribution at each time-step in the cross-borehole ERT panel is extracted and transformed into resistivity variations using a calibrated petrophysical relationship (Hermans et al., 2014, 2015b). In general, a 3D forward ERT model should be used to simulate the change in resistance related to the variation of resistivity induced by the change in temperature. However, in this specific study, we used the 2.5D forward model CRMod (Kemna, 2000) to reduce the memory requirement and the computational cost of geophysical simulations. Working with 2.5D simulations can introduce 3D effects in the data sets because variations of resistivity perpendicular to the panel are ignored. In this case, 3D effects can be related to the initial resistivity distribution and the presence of the heat plume. Indeed, in practice, the plume is detected before it reaches the panel because of the out-of-plane sensitivity (e.g., Vandenborght et al., 2005). The former effects will be reduced by the use of data difference, but not the later. However, in this specific case, previous works have shown that 2.5D modeling was able to reproduce the breakthrough curve in an acceptable manner and 3-D effects were assumed negligible, what was further validated by direct observations (Hermans et al., 2015b, 2016a). The data set for each subsurface model is composed of the change in the measured electrical resistance for 410 different electrode configurations at 6 and 13 different time-steps respectively.

For the 1 day experiment (Figure 4a), the data are sensitive to the same parameters as the heat storage except for the variance of the hydraulic conductivity distribution. This is likely due to the fact that the first time-steps mainly provide information on the arrival time of the tracer, related to the mean hydraulic conductivity. The heterogeneity of the distribution will mostly influence the tailing part of the curve. This can

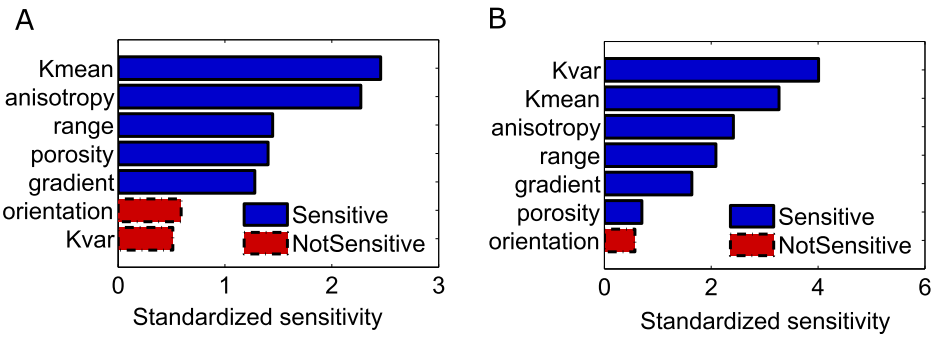

Figure 4. Global sensitivity analysis of the parameters considered in the prior for the data response (ERT data) for the (a) 1 day and (b) 5 day experiment.

be seen in the sensitivity analysis for the 5 day experiment (Figure 4b), mostly sensitive to the variance of the hydraulic conductivity. Both experiments are sensitive to the natural gradient, but its influence is smaller than for heat storage, because the tracing experiment is carried out in forced conditions.

The sensitivity analysis indicates that the proposed experiments are informative for heat storage predictions because they are globally sensitive to the same parameters. At this point, one could consider alternative field experiments. Indeed, the prefield data acquisition step is field-measurement independent. It is only based on the expected variations of the prediction and the proposed data set based on our prior knowledge of the site.

3.6. Prior Falsification Using Field Data

After data acquisition, we use the entire set of changes in resistances as well as a reduced dimension set to falsify the prior distribution (step 3, Figure 1a). In addition, two piezometers within the ERT panel, located at 1 and 2.25 m from the left ERT borehole (Figure 2b), were equipped with temperature loggers (Diver®) measuring constantly the temperature within the piezometers. We also falsify the prior with those direct, independent, temperature measurements.

3.6.1. Raw ERT Measurements

The field ERT data set is composed of 410 electrode configurations. It was obtained after removal of obvious outliers and noisy measurements. An analysis of the noise on the field data set using time-lapse reciprocals (Lesparre et al., 2017) is described in supporting information Text S3 and Figure S3.

The average change in resistance for all electrode configurations is consistent with the prior as shown by Figure 5a. Indeed, the observed curve (red) is within the distribution of curves from the prior (black curves). This ensures that globally, the amplitude of the change can be represented by the prior and that the temporal behavior is consistent. If the maximum in change in resistance was reached at the 10^{th} time step for all models of the prior set but at the 4^{th} time step for the field, it would clearly indicate a falsification of the prior.

We also use each individual electrode configuration for prior falsification. For most quadrupoles, the change in resistance time series is well caught by the prior (Figure 5b). However, for a few of them, the data are at the edge of the prior distribution, or even slightly above the maximum observed in the prior, while the temporal behavior is similar (Figure 5c). The observed difference is not larger than between the extreme cases in Figures 5a and 5b and therefore not sufficient to falsify the prior. This observation is related to the high sensitivity of cross-borehole ERT to resistivity variations close to the electrodes.

3.6.2. ERT Measurements in Reduced Dimension Space

ERT measurements are sensitive to temperature changes, but each measured resistance is integrated over a large volume which depends on the actual electrical resistivity distribution in the subsurface. Measurements with close electrode configurations are relatively similar. A similar temporal behavior is also expected,

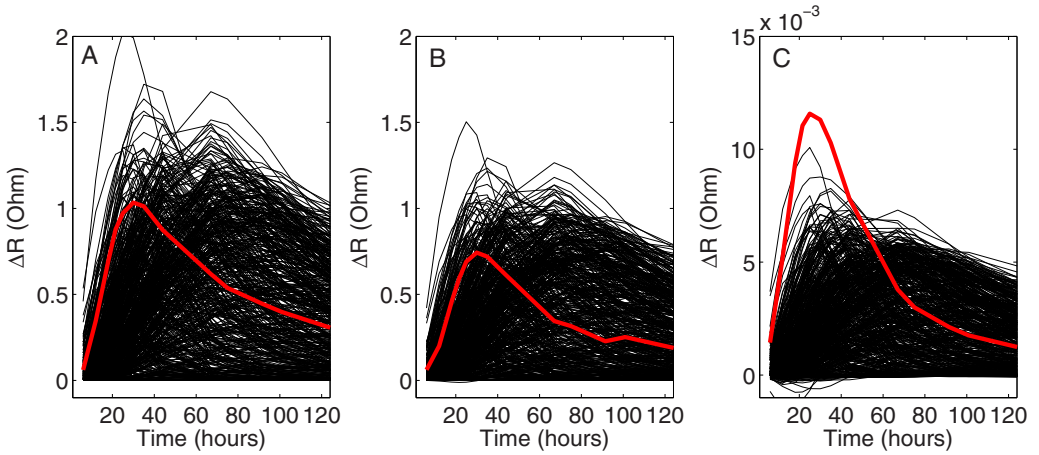

Figure 5. Prior falsification using ERT data. Each black line corresponds to the data of one of the prior models. The red line corresponds to the field data. (a) Average change in resistance for all data configurations. (b) Individual data consistent with the prior. (c) Individual data slightly outside the range of the prior.

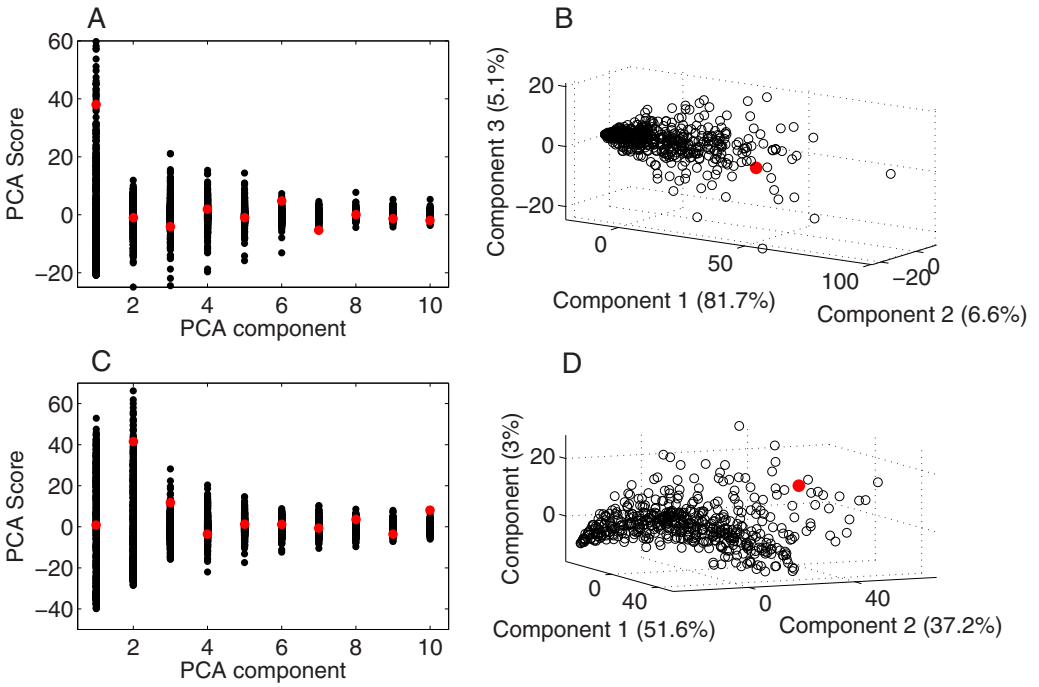

Figure 6. Prior falsification in reduced dimension space for the 1 day experiment: (a) distribution of the scores for each principal component, (b) distribution of the scores in the space composed of the three first components; and the 5 day experiment: (c) distribution of the scores for each principal component, (d) distribution of the scores in the space composed of the 3 first components. The red point corresponds to the field data set.

although small differences might occur (Figure 5). Time-lapse ERT data sets thus have a high level of redundancy. Here, we propose to reduce the dimensionality of the data set using principal component analysis (PCA) and falsify the prior in the reduced dimension space (Figure 6).

For the 1 day experiment, the first principal component explains 81% of the variance, the 3 first components represent 92.3%, and more than 99% of the variance in the data is explained by the 10 first principal components. The same behavior is observed for the 5 day experiment, the first 3 and 10 components represent 93.4% and 98.7% of the variance, respectively.

For both experiments, the observed data lie within the prior distribution (Figures 6a and 6c) for the first 10 components. However, the location of the field data set in the PCA space for the first three dimensions shows that our field set lies at the edge of the distribution (Figures 6b and 6d). Although our field data are part of the prior, it is located in a region poorly sampled. This renders prediction-focused approaches, and stochastic inversions in general, more challenging than in a densely populated zone (Hermans et al., 2016b).

3.6.3. Direct Measurements of Temperature

We use a similar approach for prior falsification with the independent validation data (raw data and in reduced space). In contrast to ERT measurements, direct temperature measurements are not subject to any integration over a large volume. The measured temperature variation is largely within the prior, both for the raw data (Figure 7a) and in the reduced dimension space (Figure 7b). From the three sets used for prior falsification, we are confident that our prior is adequate and able to reproduce the observed data.

3.7. Heat Storage Stochastic Prediction and Uncertainty Quantification

Since the forced heat tracing experiment monitored by ERT is informative and prior-consistent, we will now use PFA to predict the temperature at the well during the heat storage experiment (step 4, Figure 1a). We show only final results, intermediate steps and data noise analysis are presented in supporting information Text S2, Text S3, Figures S1 and S3.

For the prediction of heat storage with the 1 day heat tracing experiment, we use the first 10 dimensions in the data and the first 2 dimensions in the prediction. They represent more than 99% of the variance. The

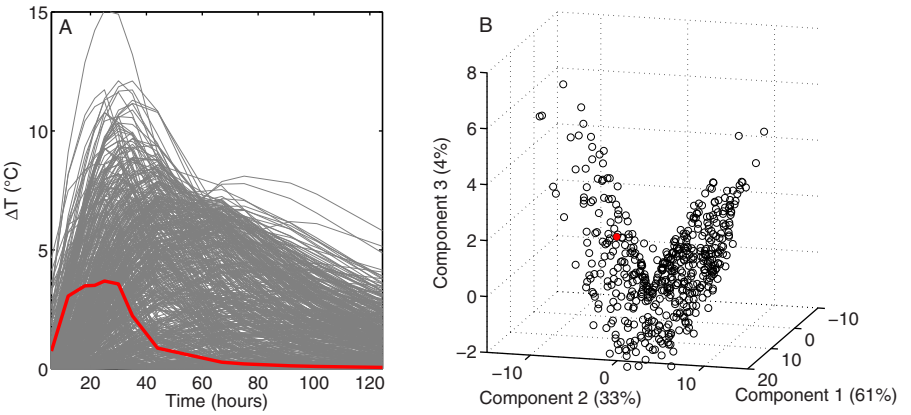

Figure 7. Prior falsification using direct temperature measurements. (a) Temperature change over time, the red line corresponds to the field data, grey lines to the prior distribution. (b) Distribution of the PCA scores, the red point correspond to the field data set.

correlation in the CCA space is not very strong as reflected in the respective scatter plots (supporting information Figures S1a and S1b): 0.75 and 0.40. The calculated posterior distribution (Figure 8a) of the change of temperature at the well indicates that the capacity for heat storage on the field site is relatively low. Most samples of the posterior have a rapid decrease in temperature once the pumping phase starts. Globally, little thermal energy is recovered. The uncertainty in the prediction is relatively low compared to the prior distribution of temperature at the well (Figure 8b). We see that the predictions for the observed data lie within the prior but are at the lower end of the distribution. In the posterior, the percentiles 10 and 50% are almost superimposed and correspond to the quantile 10% of the prior distribution. This signifies that about 50% of the samples of the posterior have very low thermal energy recovery, at a level corresponding to only 10% of the prior models.

For the 5 day experiment, increase correlation coefficients in CCA space are observed (0.89 and 0.72, see supporting information Figures S1c and S1d) indicating that the 5 day experiment contains more useful information for the prediction of heat storage. The low heat storage capacity of the aquifer is confirmed by

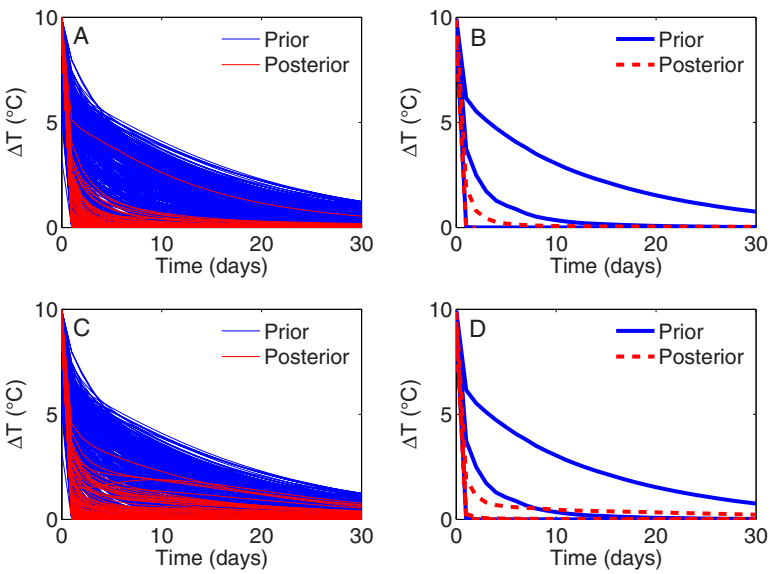

Figure 8. (a) Prior and posterior distributions of the prediction for the 1 day experiment. (b) 10%, 50% and 90% percentiles of the prior and posterior distribution of the prediction. (c) Prior and posterior distributions of the prediction for the 5 day experiment. (d) 10%, 50% and 90% percentiles of the prior and posterior distribution of the prediction. The 10% and 50% are almost superimposed given the large number of predictions with low energy recovery.

the corresponding posterior distribution (Figure 8c). The associated uncertainty is slightly increased as shown by the 10, 50 and 90% quantiles (Figure 8d). The position of the 90% percentile suggests that the probability of the aquifer to have a higher heat storage capacity is slightly increased considering the 5 day experiment. However, the difference is relatively small. The 5 day experiment results are more reliable but are obtained at the cost of additional data acquisition. This illustrates that PFA can be used within the context of experimental design (Hermans, 2017). The results are also coherent with the deterministic solution (Figure 3a, green curve).

3.8. Spatio-Temporal Distribution of Temperature Prediction

A similar approach as for heat storage prediction is used in this section. Since the data are similar, the prior remains valid. The major difference is that the prediction is now a spatio-temporal distribution of temperature in the saturated part of the aquifer (2D + time = 3D) and not a simple curve (1D). The prediction is the temperature distribution observed in the ERT panel during the 1 day heat tracing experiment using the cross-borehole ERT data. It can be validated using direct measurement of the aquifer temperature. Intermediate results (CCA) are again presented in supporting information Text S2 and Figure S2.

3.8.1. Posterior Prediction

The dimensionality of the prediction is much larger and complex than in the previous case. The first 8 dimensions were kept after PCA, corresponding to 89% of the variance. Nevertheless, highly correlated relationships are obtained between data and prediction, although the correlation coefficient decreases as the scattering around the linear relationships increases (supporting information Figure S2).

Three different samples of the posterior distribution at three different time-steps illustrate the variability of the posterior distribution (Figure 9). They all show a heat plume limited to the bottom part of the aquifer and divided in 2 different parts with a lower temperature in the middle of the section. This observation is coherent with results previously obtained with classical inversion methods and direct measurements made during the experiment (Hermans et al., 2015b, 2016a). The confinement of the plume to the bottom part of the aquifer is due to the preferential flow in the coarse gravel layer located above the bed rock.

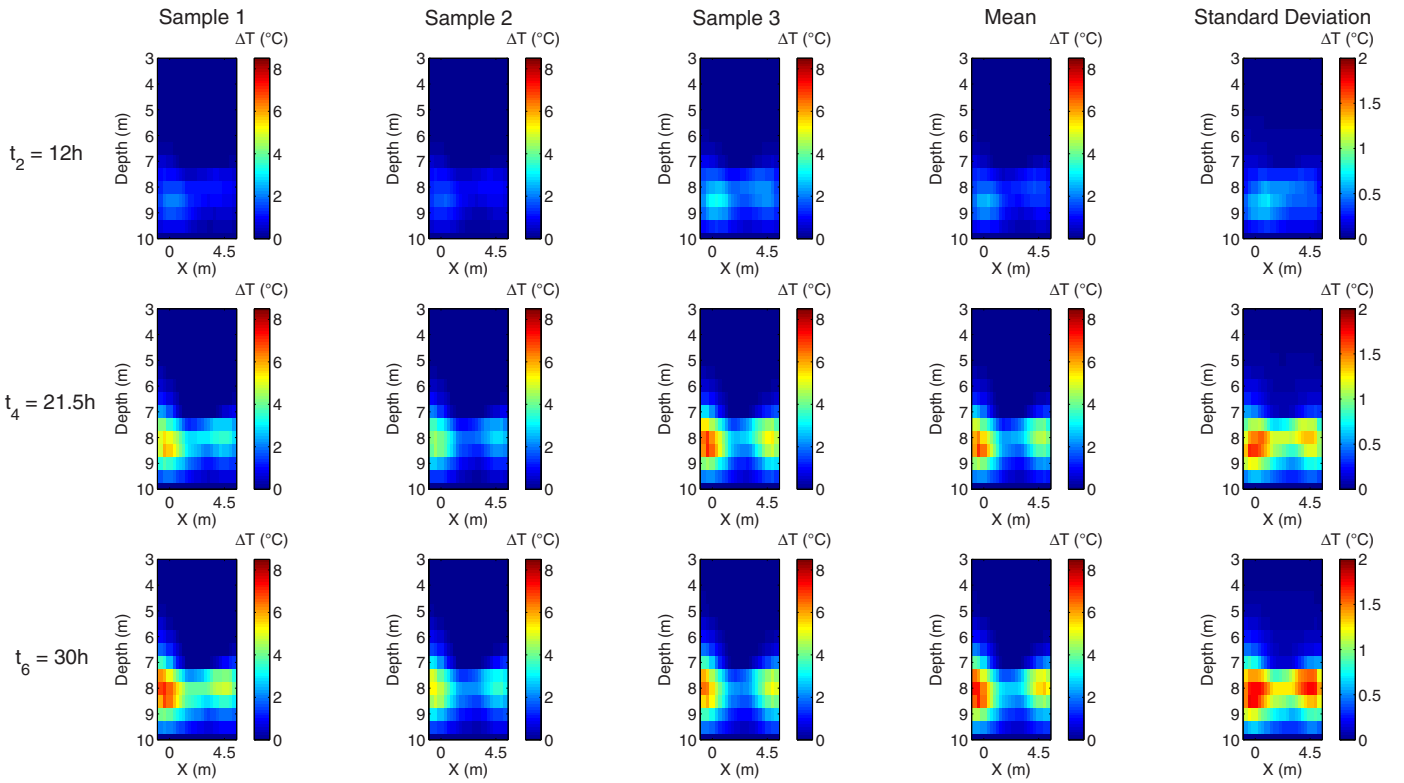

Figure 9. Posterior distribution of the temperature in the ERT panel during the tracing experiment from BEL at three different time-steps. Columns 1 to 3 show three samples of the posterior distribution. Columns 4 is the mean of the posterior distribution and column 5 is the standard deviation.

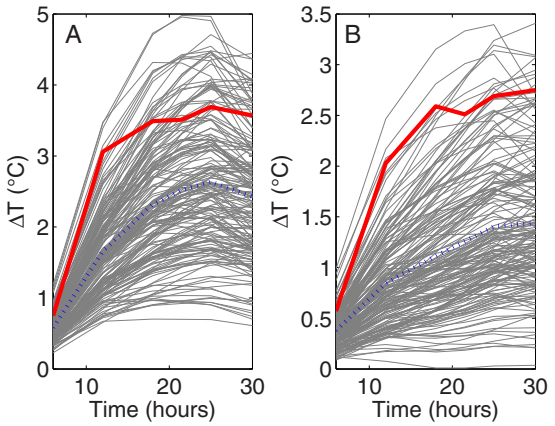

Figure 10. Comparison of the posterior distribution (gray) and its mean (blue) with direct measurements (red) at the two validation locations.

Most uncertainty is related to the level of temperature change and the small-scale temperature distribution. Those spatial variations are highlighted in the standard deviation distribution (Figure 9). The temporal behavior varies only slightly within the samples.

3.8.2. Validation With Direct Measurements

The direct measurements in the piezometers (Figure 2b) offer independent and direct measurements to validate the proposed predictions at the locations of the two sensors. The posterior distribution of temperature changes at those two positions (Figure 10) shows that the temporal behavior of the tracer is very well caught by ERT data. The reduction of uncertainty compared to the prior is large (Figure 7) but the uncertainty on the temperature variations remains relatively high.

A change of 1°C is responsible for a change in electrical conductivity of about 2%. In deterministic ERT inversions, small changes are often disregarded or interpreted only qualitatively (e.g., Chrétien et al., 2014; Doetsch et al., 2012; Hermans et al., 2012; Robert et al., 2012). As an example, Hermans et al. (2015b) estimated the limit of detection of ERT in this experiment at about 1.5°C. The level of uncertainty derived with PFA thus seems relatively coherent with the limitation of the method itself.

The uncertainty of the temperature distribution obtained with PFA within the ERT panel is an illustration of the non-uniqueness of the solution for geophysical inverse problems. It is therefore expected to observe high uncertainty at the location of the piezometers which lie in zones of relatively low resolution. Here, BEL allows a direct estimation of the non-uniqueness (and therefore uncertainty) by deriving the full posterior distribution.

4. Discussion

Although we considered a wide prior, conceptual assumptions made in the numerical model may introduce uncertainty. For instance, the bedrock was modeled as an impermeable layer and the upper sandy gravel layer was modeled as a homogeneous, less permeable layer. The latter is confirmed by direct measurements, but the 5 day tracing data show that a late tracer arrival, with low amplitude is observed in this layer (Wildemeersch et al., 2014). This behavior is not well reproduced by our prior because we did not consider uncertainty within this layer. The effect on the heat storage prediction is limited because it is not significantly influenced by this parameter. However, it would impact predictions for the last time-steps of the heat tracing experiment.

The physics of the problem itself was simplified since the density-dependence of the flow was neglected to speed up hydrogeological forward modeling. Klepikova et al. (2016b) have shown that density effects mostly affect the temperature distribution at the proximity of the injection well for the heat tracing experiment. Given the lower temperature of injection during heat storage, the effect on the prediction is also limited.

Some thermal transport parameters and the petrophysical relationship between temperature and resistivity changes were considered as known. If the relationship was calibrated in the laboratory (Hermans et al., 2015b), small variations related to spatial uncertainty might occur, due to the presence of clayey sediments for example. Uncertainty related to petrophysical parameters could directly be included in BEL, similarly to other parameters. The relationship also supposes the initial water conductivity to be constant.

The application of BEL to a field case highlights the necessity to avoid prior falsification, a concern common to all Bayesian methods. Otherwise, the resulting posterior distribution might be misleading with an unrealistic associated uncertainty. In this specific case, the field data lie at the edge of the data prior distribution. This is clearly a challenging situation for complex forecast since most models of the prior, although globally consistent with the data, are not really representative of the particular data set collected on the field. This is likely related to the spatial heterogeneity of the aquifer.

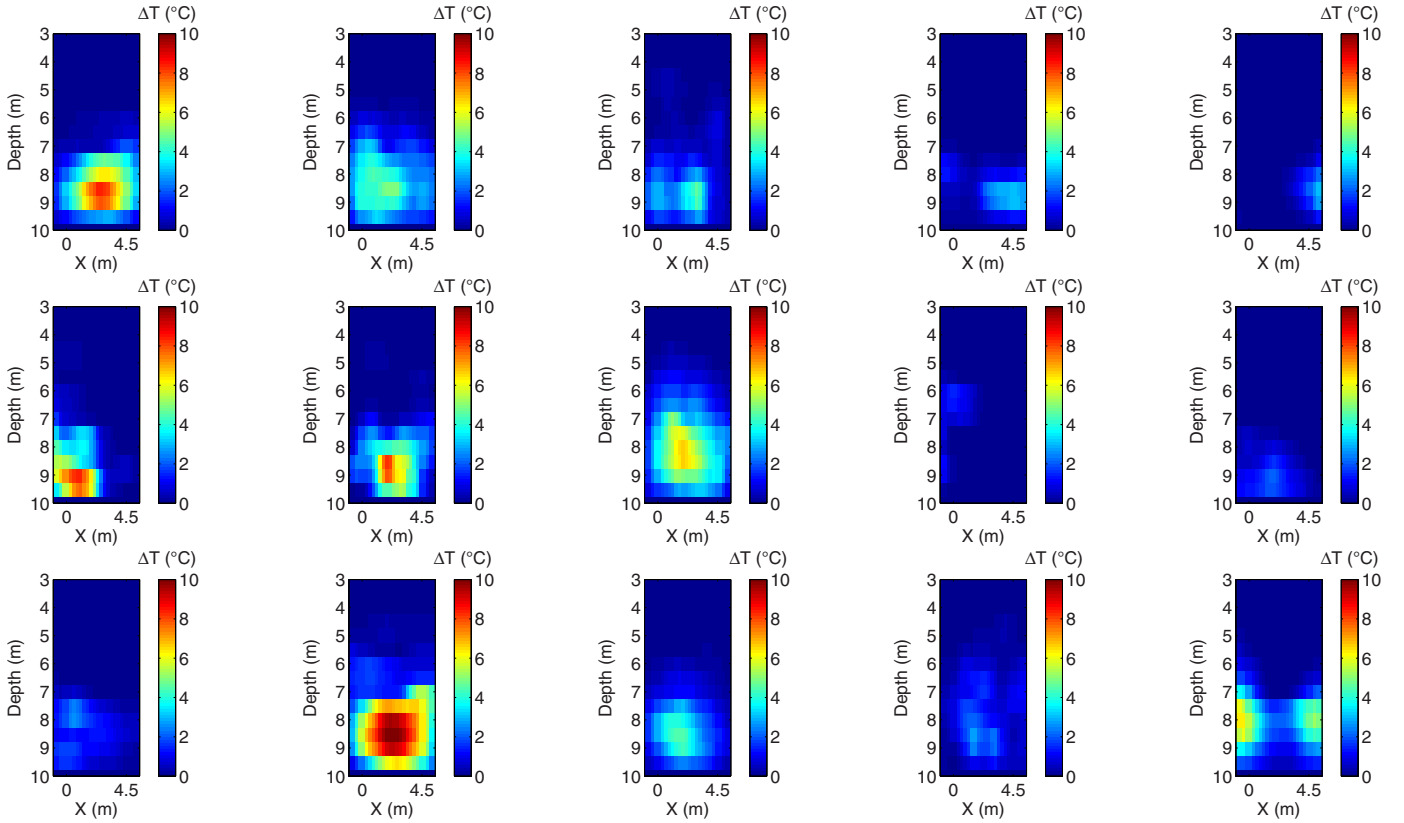

Figure 11. 14 samples of temperature distribution after 30 h during the heat tracing experiment from the prior set of models. The image at the bottom-right is a sample of the posterior distribution for the observed data, close to the mean of the posterior distribution. This illustrates that the field situation is relatively particular with two distinct plumes.

Indeed, the spatial patterns of temperature in the ERT panel during the heat tracing experiment, divided in two separated plumes, indicate a relatively complex flow pattern induced by the heterogeneity within the coarse gravel layer (Klepikova et al., 2016b). This kind of temperature distribution requiring preferential flow paths is not observed in most models of the prior set (Figure 11). Many models from the prior set show a single plume or plumes with several maximum, close to each other. This signifies that it is difficult to accurately represent the field data because only a few spatial models within the prior can represent similar temperature distributions. This can subsequently explain why the field data lie at the edge of the prior distribution (Figure 6).

If an accurate prediction is difficult to obtain, one solution is to update the prior model in order to preferentially sample models in the area of the data. Such prior falsification approaches have been successfully applied in different contexts (Hermans et al., 2015a; Park et al., 2013; Scheidt et al., 2015a, 2018).

The above considerations are not specific to our approach but remain valid for all deterministic and stochastic methods. BEL has the double advantage to (1) indicating relatively easily the incompatibility between the prior and the data and (2) offering a simple and unique framework to integrate any type of uncertainty, would it be related to conceptual decision, parameter distribution or spatial uncertainty. In some cases, model parameters are not independent, but are correlated to each other (e.g., porosity and hydraulic conductivity). If described in the prior, such dependency between parameter can be handled by BEL without modification of the framework.

One component of BEL is the choice of the number of models necessary to adequately sample the prior. It is difficult to give a rule of thumb and a sensitivity analysis on this effect is out of the scope of the paper. It is important to note that BEL focuses on the prediction and not on the model, so that it is the complexity of the prediction that should drive the number of models. In most studies, a range of models between 100 and 1,000 should be sufficient. In cases where less prior information is available, a wider prior is necessary

and therefore more Monte Carlo simulations should be used to adequately characterize the data-prediction relationship. Similarly, if more uncertain parameters are considered or if the spatial extent of the model is larger, the amount of Monte Carlo simulation should be increased in order to properly sample the data and prediction spaces. The required number of models might depend on the observed data. Areas of the data space densely sampled are easier to predict. If the observed data are somehow extreme, a denser sampling of the prior might help to accurately predict the posterior (see also Scheidt et al., 2018). In our example, the validation prediction is complex (2D + time temperature distribution) and 250 models were sufficient to make a prediction close to the average, but we increased the number to 500, making the derived relationship more robust, because our data set lies at the edge of the prior distribution.

A good approach is to start with a limited number of models (e.g., 100–200) for the sensitivity analysis and the prior falsification. Since DGSA provides a confidence interval around the sensitivity, the number of model can be increased to narrow this interval, if needed. The number can be further increased for PFA to derive a robust relationship. Hermans et al. (2016b) introduced two straightforward ways to validate the choice of reduced dimensions and number of models. First, one can calculate the prediction corresponding to some of the models of the prior. If various predictions are correctly estimated, then the chosen prior is likely sufficient to characterize the data-prediction relationship. Second, the later process can be extended to all the models of the prior. In that case, integrating the respective posterior distribution over the parameter space should give a distribution similar to the prior.

5. Conclusion

This paper proposes to use a novel framework, Bayesian Evidential Learning, to estimate the uncertainty of prediction for subsurface models. We demonstrate the method within the context of the prediction of heat storage capacity of an alluvial aquifer.

BEL is divided in two stages: prefield and postfield data acquisition. Before data acquisition, we show that the prior uncertainty of heat storage is large. We identify sensitive parameters and design an informative heat tracing experiment to reduce the uncertainty of heat storage. After data acquisition, we first verify the consistency of the prior with the field data (prior falsification). Then we predict the heat storage capacity of the aquifer. The final results reveal a low ability for the alluvial aquifer to store heat, related to a combination of high hydraulic conductivity and relatively high natural gradient. Our methodology enables to estimate the uncertainty related to the prediction offering a basis for fully-informed decision making and risk analysis. As a validation, we also predict the spatio-temporal distribution of the temperature during the field heat tracing experiment.

The proposed methodology has the advantage over other stochastic methodologies to require only a limited amount of forward simulations. In this case, only 1,000 hydrogeological model forward runs were necessary to estimate the full posterior distribution of the prediction. Each run requiring about 5 minutes on a standard desktop computer, the method does not require advanced computing facility. DGSA and PFA are not computationally time-consuming. Moreover, forward simulations can be fully parallelized. Therefore, the computational cost is limited compared to the ten to hundred thousands of runs that would be required by a McMC method.

BEL allows accounting for all kind of uncertainty (conceptual, structural or spatial) within a unique framework. Therefore, our method should enable to generalize the use of stochastic analysis techniques by researchers, engineers and practitioners for real field applications.

Acknowledgments
This work was partly performed by Thomas Hermans while being the recipient of a fellowship from the Belgian American Educational Foundation for postdoctoral studies at Stanford University. We also thank Wallonia-Brussels International and the Vocatio Foundation for the financial support of T. Hermans during his postdoctoral research stay at Stanford University. Data used in this paper are stored in the H+ database, site of Hermalle-sous-Argenteau at http://hplus.ore.fr/en/enigma/data-hermalle. Codes for DGSA and PFA are freely available at https://github.com/SCRFpublic.

References

Arato, A., Boaga, J., Comina, C., De Seta, M., Di Sipio, E., Galgaro, A., et al. (2015). Geophysical monitoring for shallow geothermal applications––Two Italian case histories. *First Break*, *33*, 75–79.

Bakr, M., van Oostrom, N., & Sommer, W. (2013). Efficiency of and interference among multiple aquifer thermal energy storage systems: A Dutch case study. *Renewable Energy*, *60*, 53–62. https://doi.org/10.1016/j.renene.2013.04.004

Bayer, P., Rybach, L., Blum, P., & Brauchler, R. (2013). Review on life cycle environmental effects of geothermal power generation. *Renewable and Sustainable Energy Reviews*, *26*, 446–463. https://doi.org/10.1016/j.rser.2013.05.039

Bloemendal, M., Olsthoorn, T., & Boons, F. (2014). How to achieve optimal and sustainable use of the subsurface for aquifer thermal energy storage. *Energy Policy*, *66*, 104–114. https://doi.org/10.1016/j.enpol.2013.11.034

Blum, P., Campillo, G., & Kölbel, T. (2011). Techno-economic and spatial analysis of vertical ground source heat pump systems in Germany. *Energy*, *36*, 3002–3011. https://doi.org/10.1016/j.energy.2011.02.044

Bridger, D. W., & Allen, D. M. (2010). Heat transport simulations in a heterogeneous aquifer used for aquifer thermal energy storage (ATES). *Canadian Geotechnical Journal*, *47*, 96–115. https://doi.org/10.1139/T09-078

Bridger, D. W., & Allen, D. M. (2014). Influence of geologic layering on heat transport and storage in an aquifer thermal energy storage system. *Hydrogeology Journal*, *22*, 233–250. https://doi.org/10.1007/s10040-013-1049-1

Brouyère, S. (2001). *Etude et modélisation du transport et du piégeage des solutés en milieu souterrain variablement saturé (study and modelling of transport and retardation of solutes in variably saturated media)* (PhD thesis). Liege, Belgium: University of Liege.

Chalmers, A. (2013). *What is this thing called science*? (4th ed.). Indianapolis, IN: Hackett Publishing.

Chen, M., Tompson, A. F. B., Mellors, R. J., & Abdalla, O. (2015). An efficient optimization of well placement and control for a geothermal prospect under geological uncertainty. *Applied Energy*, *137*, 352–363. https://doi.org/10.1016/j.apenergy.2014.10.036

Chrétien, M., Lataste, J. F., Fabre, R., & Denis, A. (2014). Electrical resistivity tomography to understand clay behavior during seasonal water content variations. *Engineering Geology*, *169*, 112–123. https://doi.org/10.1016/j.enggeo.2013.11.019

Dassargues, A. (1997). Modeling baseflow from an alluvial aquifer using hydraulic-conductivity data obtained from a derived relation with apparent electrical resistivity. *Hydrogeology Journal*, *5*, 97–108.

Derouane, J., & Dassargues, A. (1998). Delineation of groundwater protection zones based on tracer tests and transport modeling in alluvial sediments. *Environmental Geology*, *36*, 27–36.

Doetsch, J., Linde, N., Vogt, T., Binley, A., & Green, A. G. (2012). Imaging and quantifying salt-tracer transport in a riparian groundwater system by means of 3D ERT monitoring. *Geophysics*, *77*, B207–B218. https://doi.org/10.1190/geo2012-0046.1

Fenwick, D., Scheidt, C., & Caers, J. (2014). Quantifying asymmetric parameter interactions in sensitivity analysis: Application to reservoir modeling. *Mathematical Geosciences*, *46*, 493–511. https://doi.org/10.1007/s11004-014-9530-5

Ferguson, G. (2007). Heterogeneity and thermal modeling of ground water. *Ground Water*, *45*, 485–490. https://doi.org/10.1111/j.1745-6584.2007.00323.x

Fu, J., & Gómez-Hernández, J. J. (2009). A blocking Markov Chain Monte Carlo method for inverse stochastic hydrogeological modeling. *Mathematical Geosciences*, *41*, 105–128. https://doi.org/10.1007/s11004-008-9206-0

Gómez-Hernández, J. J., Sahuquillo, A., & Capilla, J. E. (1997). Stochastic simulation of transmissivity fields conditional to both transmissivity and piezometric data–-1. Theory. *Journal of Hydrology*, *203*, 162–174.

Goovaerts, P. (1997). *Geostatistics for natural resources evaluation* (Applied Geostatistics Series). New York: Oxford University Press.

Hermans, T., Kemna, A., & Nguyen, F. (2016a). Covariance-constrained difference inversion of time-lapse electrical resistivity tomography data. *Geophysics*, *81*, E311–E322. https://doi.org/10.1190/geo2015-0491.1

Hermans, T., Nguyen, F., & Caers, J. (2015a). Uncertainty in training image-based inversion of hydraulic head data constrained to ERT data: Workflow and case study. *Water Resources Research*, *51*, 5332–5352. https://doi.org/10.1002/2014WR016460

Hermans, T., Nguyen, F., Robert, T., & Revil, A. (2014). Geophysical methods for monitoring temperature changes in shallow low enthalpy geothermal systems. *Energies*, *7*, 5083–5118. https://doi.org/10.3390/en7085083

Hermans, T., Oware, E. K., & Caers, J. (2016b). Direct prediction of spatially and temporally varying physical properties from time-lapse electrical resistance data. *Water Resources Research*, *52*, 7262–7283. https://doi.org/10.1002/2016WR019126

Hermans, T., Vandenbohede, A., Lebbe, L., & Nguyen, F. (2012). A shallow geothermal experiment in a sandy aquifer monitored using electric resistivity tomography. *Geophysics*, *77*, B11–B21.

Hermans, T., Wildemeersch, S., Jamin, P., Orban, P., Brouyère, S., Dassargues, A., & Nguyen, F. (2015b). Quantitative temperature monitoring of a heat tracing experiment using cross-borehole ERT. *Geothermics*, *53*, 14–26. https://doi.org/10.1016/j.geothermics.2014.03.013

Hermans, T. (2017). Prediction-focused approaches: An opportunity for hydrology. *Groundwater*, *55*(5), 683–687.

Hou, Z., & Rubin, Y. (2005). On minimum relative entropy concepts and prior compatibility issues in vadose zone inverse and forward modeling. *Water Resources Research*, *41*, W12425. https://doi.org/10.1029/2005WR004082

Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge, UK: Cambridge University Press.

Kemna, A. (2000). *Tomographic inversion of complex resistivity-theory and application*. Bochum, Germany: Ruhr-Universität.

Kim, J., Lee, Y., Yoon, W. S., Jeon, J. S., Koo, M. H., & Keehm, Y. (2010). Numerical modeling of aquifer thermal energy storage system. *Energy*, *35*, 4955–4965.

Klepikova, M., Wildemeersch, S., Hermans, T., Jamin, P., Orban, P., Nguyen, F., et al.(2016b). Heat tracer test in an alluvial aquifer: Field experiment and inverse modelling. *Journal of Hydrology*, *540*, 812–823. https://doi.org/10.1016/j.jhydrol.2016.06.066

Klepikova, M. V., Le Borgne, T., Bour, O., Dentz, M., Hochreutener, R., & Lavenant, N. (2016a). Heat as a tracer for understanding transport processes in fractured media: Theory and field assessment from multiscale thermal push-pull tracer tests. *Water Resources Research*, *52*, 5442–5457. https://doi.org/10.1002/2016WR018789

Kuo, C., & Liao, H. (2012). The feasibility of using circulating groundwater as renewable energy sources for air-conditioning in Taipei basin. *Renewable Energy*, *39*, 175–182. https://doi.org/10.1016/j.renene.2011.07.046

Lee, K. S. (2010). A review on concepts, applications, and models of aquifer thermal energy storage systems. *Energies*, *3*, 1320–1334.

Lesparre, N., Nguyen, F., Kemna, A., Robert, T., Hermans, T., Daoudi, M., & Flores-Orozco, A. (2017). A new approach for time-lapse data weighting in electrical resistivity tomography. *Geophysics*, *82*(6), E325–E333.

Linde, N., Renard, P., Mukerji, T., & Caers, J. (2015). Geological realism in hydrogeological and geophysical inverse modeling: A review. *Advances in Water Resources*, *86*, 86–101. https://doi.org/10.1016/j.advwatres.2015.09.019

Lo Russo, S., & Civita, M. V. (2009). Open-loop groundwater heat pumps development for large buildings: A case study. *Geothermics*, *38*, 335–345. https://doi.org/10.1016/j.geothermics.2008.12.009

Macfarlane, A., Foerster, A., Merriam, D., Schroetter, J., & Healey, J. (2002). Monitoring artificially stimulated fluid movement in the Cretaceous Dakota aquifer, western Kansas. *Hydrogeology Journal*, *10*, 662–673. https://doi.org/10.1007/s10040-002-0223-7

Nam, Y., & Ooka, R. (2010). Numerical simulation of ground heat and water transfer for groundwater heat pump system based on real-scale experiment. *Energy and Buildings*, *42*, 69–75. https://doi.org/10.1016/j.enbuild.2009.07.012

Oladyshkin, S., de Barros, F. P. J., & Nowak, W. (2012). Global sensitivity analysis: A flexible and efficient framework with an example from stochastic hydrogeology. *Advances in Water Resources*, *37*, 10–22.

Oliver, D. S., Cunha, L. B., & Reynolds, A. C. (1997). Markov chain Monte Carlo methods for conditioning a permeability field to pressure data. *Mathematical Geology*, *29*, 61–91.

Palmer, C. D., Blowes, D. W., Frind, E. O., & Molson, J. W. (1992). Thermal energy storage in an unconfined aquifer 1. Field injection experiment. *Water Resources Research*, *28*, 2845–2856.

Park, B.-H., Bae, G.-O., & Lee, K.-K. (2015). Importance of thermal dispersivity in designing groundwater heat pump (GWHP) system: Field and numerical study. *Renewable Energy*, *83*, 270–279. https://doi.org/10.1016/j.renene.2015.04.036
Park, H., Scheidt, C., Fenwick, D., Boucher, A., & Caers, J. (2013). History matching and uncertainty quantification of facies models with multiple geological interpretations. *Computational Geosciences*, *17*, 609–621. https://doi.org/10.1007/s10596-013-9343-5
Park, J., Yang, G., Satija, A., Scheidt, C., & Caers, J. (2016). DGSA: A Matlab toolbox for distance-based generalized sensitivity analysis of geoscientific computer experiments. *Computers & Geosciences*, *97*, 15–29. https://doi.org/10.1016/j.cageo.2016.08.021
Popper, K. (1959). *The logic of scientific discovery*. London, UK: Hutchinson & Co.
Possemiers, M., Huysmans, M., & Batelaan, O. (2015). Application of multiple-point geostatistics to simulate the effect of small-scale aquifer heterogeneity on the efficiency of aquifer thermal energy storage. *Hydrogeology Journal*, *23*, 971–981. https://doi.org/10.1007/s10040-015-1244-3
Robert, T., Caterina, D., Deceuster, J., Kaufmann, O., & Nguyen, F. (2012). A salt tracer test monitored with surface ERT to detect preferential flow and transport paths in fractured/karstified limestones. *Geophysics*, *77*, B55–B67.
Rubin, Y., Chen, X., Murakami, H., & Hahn, M. (2010). A Bayesian approach for inverse modeling, data assimilation, and conditional simulation of spatial random fields. *Water Resources Research*, *46*, W10523. https://doi.org/10.1029/2009WR008799
Saner, D., Juraske, R., Kübert, M., Blum, P., Hellweg, S., & Bayer, P. (2010). Is it only CO2 that matters? A life cycle perspective on shallow geothermal systems. *Renewable and Sustainable Energy Reviews*, *14*, 1798–1813. https://doi.org/10.1016/j.rser.2010.04.002
Satija, A., & Caers, J. (2015). Direct forecasting of subsurface flow response from non-linear dynamic data by linear least-squares in canonical functional principal component space. *Advances in Water Resources*, *77*, 69–81. https://doi.org/10.1016/j.advwatres.2015.01.002
Scheidt, C., Jeong, C., Mukerji, T., & Caers, J. (2015a). Probabilistic falsification of prior geologic uncertainty with seismic amplitude data: Application to a turbidite reservoir case. *Geophysics*, *80*, M89–M100. https://doi.org/10.1190/geo2015-0084.1
Scheidt, C., Li, L., & Caers, J. (2018). *Quantifying uncertainty in subsurface sysems* (288 pp.). New York, NY: American Geophysical Union, Wiley.
Scheidt, C., Renard, P., & Caers, J. (2015b). Prediction-focused subsurface modeling: Investigating the need for accuracy in flow-based inverse modeling. *Mathematical Geosciences*, *47*, 173–191. https://doi.org/10.1007/s11004-014-9521-6
Sommer, W., Valstar, J., van Gaans, P., Grotenhuis, T., & Rijnaarts, H. (2013). The impact of aquifer heterogeneity on the performance of aquifer thermal energy storage. *Water Resources Research*, *49*, 8128–8138. https://doi.org/10.1002/2013WR013677
Sommer, W. T., Doornenbal, P. J., Drijver, B. C., van Gaans, P. F. M., Leusbrock, I., Grotenhuis, J. T. C., & Rijnaarts, H. H. M. (2014). Thermal performance and heat transport in aquifer thermal energy storage. *Hydrogeology Journal*, *22*, 263–279. https://doi.org/10.1007/s10040-013-1066-0
Tarantola, A. (2005). *Inverse problem theory and methods for model parameter estimation*. Philadelphia, PA: Society for Industrial and Applied Mathematics.
Therrien, R., McLaren, R., Sudicky, E., & Panday, S. (2010). *HydroGeoSphere: A three-dimensional numerical model describing fully-integrated subsurface and surface flow and solute transport*. Waterloo, ON: Groundwater Simulation Group.
Vandenbohede, A., Hermans, T., Nguyen, F., & Lebbe, L. (2011). Shallow heat injection and storage experiment: Heat transport simulation and sensitivity analysis. *Journal of Hydrology*, *409*, 262–272.
Vandenbohede, A., Louwyck, A., & Lebbe, L. (2009). Conservative solute versus heat transport in porous media during push-pull tests. *Transport in Porous Media*, *76*, 265–287. https://doi.org/10.1007/s11242-008-9246-4
Vandenborght, J., Kemna, A., Hardelauf, H., & Vereecken, H. (2005). Potential of electrical resistivity tomography to infer aquifer transport characteristics from tracer studies: A synthetic case study. *Water Resources Research*, *41*, W06013. https://doi.org/10.1029/2004WR003774
Vanhoudt, D., Desmedt, J., Van Bael, J., Robeyn, N., & Hoes, H. (2011). An aquifer thermal storage system in a Belgian hospital: Long-term experimental evaluation of energy and cost savings. *Energy and Buildings*, *43*, 3657–3665.
Vrugt, J. A., ter Braak, C. J. F., Diks, C. G. H., & Schoups, G. (2013). Hydrologic data assimilation using particle Markov chain Monte Carlo simulation: Theory, concepts and applications. *Advances in Water Resources*, *51*, 457–478. https://doi.org/10.1016/j.advwatres.2012.04.002
Wagner, V., Li, T., Bayer, P., Leven, C., Dietrich, P., & Blum, P. (2014). Thermal tracer testing in a sedimentary aquifer: Field experiment (Lauswiesen, Germany) and numerical simulation. *Hydrogeology Journal*, *22*, 175–187. https://doi.org/10.1007/s10040-013-1059-z
Wildemeersch, S., Jamin, P., Orban, P., Hermans, T., Klepikova, M., Nguyen, F., et al. (2014). Coupling heat and chemical tracer experiments for estimating heat transfer parameters in shallow alluvial aquifers. *Journal of Contaminant Hydrology*, *169*, 90–99. https://doi.org/10.1016/j.jconhyd.2014.08.001
Yapparova, A., Matthäi, S., & Driesner, T. (2014). Realistic simulation of an aquifer thermal energy storage: Effects of injection temperature, well placement and groundwater flow. *Energy*, *76*, 1011–1018. https://doi.org/10.1016/j.energy.2014.09.018
Yoon, H., Hart, D. B., & McKenna, S. A. (2013). Parameter estimation and predictive uncertainty in stochastic inverse modeling of groundwater flow: Comparing null-space Monte Carlo and multiple starting point methods. *Water Resources Research*, *49*, 536–553. https://doi.org/10.1002/wrcr.20064

Erratum

The originally published version of this article misstated the affiliation of Jef Caers as "Department of Geology, Ghent University, Ghent, Belgium." The error has been corrected, and this may be considered the official version of record.