Manuscript draft accepted for publication in *Sciene of the Total Environement* (July 2020)

Multi-fidelity approach to Bayesian parameter estimation in subsurface heat and fluid transport models

Kathrin Menberg[1], Asal Bidarmaghz[2], Alastair Gregory[3], Ruchi Choudhary[3,4], Mark Girolami[3,4]

[1] Institute of Applied Geosciences, Karlsruhe Institute of Technology, Kaiserstraße 12, 76131 Karlsruhe, Germany, Email: menberg@kit.edu

[2] School of Civil and Environmental Engineering, University of New South Wales, Sydney, Australia, Email: a.bidarmaghz@unsw.edu.au

[3] The Alan Turing Institute, Data-Centric Engineering, London, United Kingdom, Email: agregory@turing.ac.uk

[4] Department of Engineering, University of Cambridge, Trumpington Street CB2 1PZ, United Kingdom, Email: rc488@cam.ac.uk, mag92@eng.cam.ac.uk

*Corresponding author: Kathrin Menberg, Email: menberg@kit.edu

Abstract

The increased use of the urban subsurface for competing purposes, such as anthropogenic infrastructures and geothermal energy applications, leads to an urgent need for large-scale sophisticated modelling approaches for coupled mass and heat transfer. However, such models are subject to large uncertainties in model parameters, the physical model itself and in available measured data, which is often rare. Thus, the robustness and reliability of the computer model and its outcomes largely depend on successful parameter estimation and model calibration, which are hampered by the computational burden of large-scale coupled models.

To tackle this problem, we develop a novel Bayesian approach for parameter estimation, which allows us to account for different sources of uncertainty, is capable of dealing with sparse field data and makes optimal use of the output data from expensive numerical model runs. This is achieved by combining

output data from different models that represent the same physical problem, but at different levels of fidelity, e.g. reflected by different spatial resolution. By applying this new approach to a 1D analytical heat transfer model and a large-scale semi-3D numerical model while using synthetic data, we show that the accuracy and precision of parameter estimation by this multi-fidelity framework by far exceeds the standard single-fidelity results. The consideration of different error terms in the Bayesian framework also allows assessment of the model bias and the discrepancy between the different fidelity levels. These are emulated by Gaussian Process models, which facilitate re-iteration of the parameter estimation without additional model runs.

Keywords

Large-scale hydro-thermal modelling; parameter estimation; Bayesian inference; numerical modelling; subsurface temperature

Nomenclature

Abbreviations

GP	Gaussian Process model
HMC	Hamiltonian Monte Carlo
KOH	Kennedy & O'Hagan framework
MF	Multi-Fidelity
RBKC	Royal Borough of Kensington and Chelsea

Variables

m	measured field/synthetic data
n	model output data
p	number of state variables
q	number of calibration parameters
t	time [s]
T	temperature [°C]
x	state variable
y	model output
z	combined data set for parameter estimation
X	geographical X-ccordinate in the model domain
Y	geographical Y-ccordinate in the model domain
Z	depth layer of the model

β	smoothness hyper-parameter of GP models
δ	model bias
ε	random error
ζ	non-observable physical process
η	emulator term
θ	calibration parameters
κ	thermal diffusivity [m²/s]
λ	precision hyper-parameter of GP models
μ	model discrepancy
Σ	covariance function of GP model
τ	frequency

Subscripts

a	annual variation
c	computed data/model output
f	field/measured data
h	high-fidelity
l	low-fidelity

1 Introduction

Due to the rapid rate of urbanisation, the shallow subsurface of dense cities is exploited for various purposes such as transport, additional residential/commercial spaces, storage, and industrial processes. Recent studies have clearly demonstrated that anthropogenic heat flow from such underground structures influences the subsurface temperatures, especially in the proximity of urban aquifers (Attard et al., 2016; Bidarmaghz et al., 2019a; Epting and Huggenberger, 2013; Menberg et al., 2013). Tracking these temperature variations is crucially important for long term resilience of ground resources (such as water, energy, etc.), as well as for the energy efficiency of underground structures and geothermal systems. At the same time, modelling of spatial variations of temperatures in the subsurface also poses immense challenges, not only due to the large scale of these models, but also due to the significant role of local and often unknown thermal-hydrological effects and subsurface heterogeneities (Bayer et al., 2019). Various studies have thus tested different types of analytical and numerical geothermal models of varying levels of complexity in terms of spatio-temporal resolution, boundary conditions, type of heat sources, etc. (see Bayer et al. (2019) for a review).

While analytical or simple numerical models have been found to lack in accuracy to replicate the local thermal effects, detailed 3D numerical models are often computationally too expensive to allow iterative evaluations (Bayer et al., 2019). Yet, these are needed to perform some form of parameter estimation and model calibration, and to quantify the uncertainty in the inputs using inverse modelling techniques. In combination with sparse measurement data this renders any meaningful inference from calibrating the model with subsurface data difficult.

Unlike geothermal modelling, large-scale modelling of the subsurface has a more sustained record in the field of groundwater modelling, particularly with respect to parameter estimation. As there are a number of similarities between the two applications, it makes sense to review the state-of-the-art in the calibration and parameter estimation of groundwater models. Indeed, a common feature amongst groundwater flow and subsurface heat and fluid transport model is the problem of ill-posedness, or underdetermination owing to a large number of model inputs and few observations (Moore and Doherty, 2006). Several tools for automatic parameter estimation and model calibration are used extensively in the groundwater modelling community: for example, PEST (Doherty, 2010), PEST++ (Welter et al., 2015), UCODE (Poeter and Hill, 1998; Poeter and Hill, 1999) and UCODE_2014 (Lu et al., 2014). These tools apply different regularisation techniques to reduce the problem of high parameterisation and to prevent over-fitting. They are most widely utilised to tune parameters that obtain a simple ‘best fit’ between model outputs and data (Voss, 2011a; Voss, 2011b). More recent studies have emphasised the importance of accounting for uncertainty in a more comprehensive manner by explicitly including different error terms into the inference procedure (Rajabi et al., 2018).

In the context of solving inverse problems under uncertainty, and with sparse data, Bayesian approaches provide a suitable framework. This has lead to an increasing interest in Markov Chain Monte Carlo (MCMC)-based Bayesian techniques in hydro(geo)logy (Rajabi et al., 2018). Bayesian approaches offer a different perspective on uncertainty than classical (i.e. frequentist) statistical methods by interpreting probability as a reasonable expectation representing a state of knowledge, and allow to update this knowledge by combining prior beliefs with measured data (Gelman et al., 2014). Thus, Bayesian

inference can be applied to obtain posterior probability distributions of model parameters based on the defined prior probability distributions and the likelihood of the match between model outputs and the measured data. The approximation of the complex likelihood function, however, requires iterative evaluation of the model using MCMC methods, which is often computationally prohibitve for large-scale numerical subsurface models.

In order to improve the MCMC performance in the presence of a high-dimensional parameter space (i.e. when dealing with a large number of uncertain parameters), Cui et al. (2011) introduced an adaptive delayed-acceptance Metropolis-Hastings (MH) algorithm (ADAMH) and applied it to a large-scale deep geothermal reservoir model with unknown spatial permeability variation. Their algorithm increases the computational efficiency of estimating the Bayesian posterior distribution of model parameters by using a coarse model (3,335 elements) that approximates the fine numerical model (26,005 elements), accounting for the error between the two models. However, as the two models are coupled within the sampling process, they require the same parameterisation and an upscaling scheme for the model outputs (Cui et al., 2011).

This approach also makes the simultaneous use of surrogate models or emulators infeasible, which are a convenient mean to reduce the computational costs for optimisation problems in hydrological modeling (Asher et al., 2015; Razavi et al., 2012). Razavi et al. (2012) distinguish between two families of surrogate models: response surface models (or model emulators) and lower-fidelity models. Approaches for response surface models used in hydro(geo)logical modeling include polynomials, neural networks, Gaussian process models, and radial basis functions amongst others, which are all data-driven approximations of the physical model (Asher et al., 2015). Lower-fidelity surrogates on the other hand are less-detailed, physical representations of the more sophisticated, original model (Razavi et al., 2012).

Often models can be built at different levels of fidelity (e.g. spatial resolutions), which can be joined into a multi-fidelity framework that builds on a few runs from an expensive model and a larger number of fast runs from a more simple approximation (Fernández-Godino et al., 2016). Such multi-fidelity

approaches are particularly convenient for optimization of complex models (Forrester et al., 2007; Robinson et al., 2008), and have been applied to a wide field of optimisation problems in engineering and natural sciences in combination with different surrogate models (Durantin et al., 2017; Sun et al., 2010). In the field of groundwater modelling, multi-fidelity approaches have received some attention in recent years (Asher et al., 2015; Zhang et al., 2018; Zheng et al., 2019), in particular schemes that apply multi-scale finite element methods (Efendiev et al., 2013) and multi-scale finite volume method (Jenny et al., 2003). However, as dicussed by Asher et al. (2015) most multi-fidelty approaches applied in the field of groundwater modelling so far, are intrusive methods that are implemented specifically for certain software codes (Panday et al., 2013). In the context of uncertainty quantification, multi-fidelity Bayesian frameworks were also developed and employed on environmental and hydrodynamical models (Goh et al., 2013; Kennedy and O'Hagan, 2000).

Gaussian Process (GP) emulators are particularly suited to be used in conjunction with multi-fidelity frameworks (Zaytsev and Burnaev, 2017). In the field of large-scale groundwater modeling Cui et al. (2018) employed a GP emulator to approximate a large-scale groundwater model of a river catchment (> 200,000 elements). By doing so they were able to apply Approximate Bayesian Computation (ABC) to calibrate the 38 uncertain parameters in their model against 900 groundwater head measurements. Originally introduced for diagnostic model evaluation, ABC does not require the explicit formulation of a likelihood function, which makes it particularly suitable for problems that have large sets of measured data available (Vrugt and Sadegh, 2013). Instead, the use of an acceptance criterion based on the distance between the observed and simulated data, makes inference straightforward and efficient, albeit at the cost of not explicitly considering different sources of uncertainty (Vrugt et al., 2009). Indeed, informal Bayesian approaches provide similar results in terms of the total predictive uncertainty (Beven and Binley, 1992; Beven and Binley, 2014), but require sufficiently large quantities of data.

Untangling the effect of the different error terms stemming from data, parameter, and model uncertainty is, however, prerequisite for improving the underlying numerical models and to better interpret model predictions. Following the conceptual framework from Kennedy and O'Hagan (2001), Xu and Valocchi

(2015) and Xu et al. (2017b) account for different sources of uncertainty, i.e. structural model error, parameter and input data uncertainty in groundwater flow models. By decomposing the predictive variance of the model outputs, they showed that the structural model error can be the primary source of uncertainty in predicted hydraulic heads (Xu et al., 2017a). Thus, considering different sources of uncertainty is not only crucial in order to get robust model predictions, but analysis of these errors also improves understanding of model deficiencies. Indeed, formal Bayesian approaches (e.g. the DREAM framework (Vrugt, 2016; Vrugt et al., 2008)) have been shown to perform better for parameter estimation of hydrological models (Vrugt et al., 2009).

In the field of geothermal modelling, formal Bayesian approaches for systematic uncertainty quantification are still scarce, and the adoption of approaches from other related fields, such as hydrology, is very limited (Heße et al., 2019). There are no applications that leverage recent developments in inverse modelling, and good quality field observations are scarce. This paper addresses this gap with the objective of tackling a challenge that is equally applicable to both the geothermal and groundwater modelling communities: performing parameter estimation (or calibration) of computationally expensive, large-scale numerical models, under different sources of uncertainties (including the uncertainty from structural errors in the numerical model and random measurement errors), and with sparse and limited amount of observations. We present a methodology and the first parameter estimation and model calibration of an urban-scale geothermal numerical model. The methodology is based on using multi-fidelity numerical models with a formal Bayesian implementation, which employs an efficient Hamiltonian Monte Carlo algorithm to sample from the posterior distribution. In addition, the framework can account for random (i.e. data) and structural (i.e. model) uncertainty by following the conceptual calibration framework introduced by Kennedy and O'Hagan (2001). We apply Gaussian Process models to emulate the model bias function and model outputs, which stem from different fidelities of numerical models for the same physical problem. While structural uncertainty can stem from many sources (conceptualisation, spatial heterogeneity etc.), we here focus

on spatial discretization, and modify the spatio-temporal resolution of our models to create representations with different fidelities.

The methodology is first explained and tested on an analytical model for 1D heat transfer in the subsurface. We use the analytical solution to generate synthetic data for the test. Through this example, we demonstrate the ability of our proposed method to identify the true parameter values under different settings. We then apply the method to calibrate a semi-3D large-scale numerical model of coupled thermal and hydraulic transport in the subsurface (Bidarmaghz et al., 2020), where the main quantity of interest is spatial variations of long term ground temperatures resulting from the combined influence of anthropogenic heat sources, geological, and hydrogeological make-up of the ground at different depths.

2 Bayesian parameter estimation

2.1 Bayesian framework with single fidelity level

The Bayesian parameter estimation framework is based on Bayes' paradigm, which relates the probability p of an event (or a specific parameter value, θ) given evidence (or data, y), $p(\theta|y)$, to the probability of the event, $p(\theta)$, and the likelihood $p(y|\theta)$ (Gelman et al., 2014):

$$p(\theta|y) \propto p(\theta) \times p(y|\theta) \tag{1}$$

Using this relation allows one to combine prior belief (i.e. expert judgement) about an event and evidence about this event (i.e. measured data), to update the prior knowledge and to quantify it in the form of posterior probabilistic distributions. Kennedy and O'Hagan (2001) adopted this principle and formulated a Bayesian framework (hereafter KOH framework) for parameter estimation and model calibration, which considers multiple sources of errors in the data and the model itself. The KOH framework relates field observations, y_f, to one set of computer simulation outputs, y_c, over a range of state (or forcing) variables x (Kennedy and O'Hagan, 2001):

$$y_f(x) = \zeta(x) + \varepsilon = y_c(x, \theta) + \delta(x) + \varepsilon = \eta(x, \theta) + \delta(x) + \varepsilon \tag{2}$$

Here ζ represents the true, non-observable physical process, ε is the random measurement errors corresponding to the field observations, and $\delta(x)$ is the structural discrepancy between the model and the true process. Instead of iteratively evaluating computationally demanding computer models, most studies approximate these models by an emulator $\eta(x, \theta)$, depending on a set of model parameters, θ, of interest. In line with previous studies that applied the KOH framework (Chong and Menberg, 2018; Goh et al., 2013; Higdon et al., 2004), we use Gaussian Processes (GP) to emulate the simulation model $\eta(x, \theta)$ and the model bias function $\delta(x)$. Both emulators are built simultaneously based in the same training data. All GP models in this study are assigned a zero mean function, while the covariance functions for the emulator, Σ_η, and the model discrepancy, Σ_δ, are specified as suggested by Higdon et al. (2004).

$$\Sigma_{\eta(i,j)} = \frac{1}{\lambda_\eta}\exp\left[-\sum_{k=1}^{p}\beta_{\eta,k}\left(x_{i,k}-x_{j,k}\right)^2-\sum_{k\prime=1}^{q}\beta_{\eta,p+k}\left(\theta_{i,k\prime}-\theta_{j,k\prime}\right)^2\right] \quad (3)$$

$$\Sigma_{\delta(i,j)} = \frac{1}{\lambda_\delta}\exp\left[-\sum_{k=1}^{p}\beta_{\delta,k}\left(x_{i,k}-x_{j,k}\right)^2\right] \quad (4)$$

This formulation introduces several unknown hyper-parameters, which need to be estimated in the inversion alongside the model parameters (θ). The precision hyper-parameters λ_η and λ_δ determine the magnitude of the covariance function, and thus the variation in the output explained by the model emulator and the model bias, while the correlation hyper-parameters β_η and β_δ determine the smoothness of the emulator and model bias function in dimensions of x and θ. Here, p represents the number of state variables (x), and q the number of calibration parameters (θ). The random measurement error, ε, is independent of x and θ, and represented by the covariance Σ_ε. All hyper-parameters are uncertain and assigned prior distributions following the suggestions in previous studies (Guillas et al., 2009; Higdon et al., 2004). For details about the selection of prior distributions of hyper-parameters the reader is referred to the study by Chong and Menberg (2018).

The covariance function of the combined data set, z, used for parameter estimation, which contains both field observations and computer model outputs, can be written as (Higdon et al., 2004):

$$\Sigma_z = \Sigma_\eta + \begin{pmatrix} \Sigma_\delta + \Sigma_\varepsilon & 0 \\ 0 & 0 \end{pmatrix} \quad (5)$$

2.2 Multi-fidelity Bayesian framework

For expensive computer models, generating a few hundred model outputs, y_c, required to build the GP emulator of sufficient quality, might be infeasible. To overcome this problem, Goh et al. (2013) suggested a hierarchical modification of the KOH framework that uses outputs from computer models with different physical accuracies and computational expenses to solve inverse problems. This approach combines a few parametric model outputs from a physically accurate, but expensive, high-fidelity computer code, η_h, with a larger number of evaluations from a less expensive and less accurate model, η_l. These two simulators can share the same set of unknown model parameters, θ, or individual sets for low-fidelity, θ_l, and high-fidelity, θ_h. Following the nomenclature of the original KOH framework (Kennedy and O'Hagan, 2001), the discrepancy between the measured or synthetic data, y_f, and η_h is termed model bias, δ. The systematic difference between the two computer model outputs, η_h and η_l, is referred to as model mismatch, μ, which describes the relationship between the low- and high-fidelity model. Accordingly, the mathematical concept from eq.(1) can be re-written as (Goh et al., 2013):

$$y_f(x) = \eta_h(x, \theta_h, \theta_l) + \delta(x) + \varepsilon = \eta_l(x, \theta_h, \theta_l) + \mu(x, \theta_h, \theta_l) + \delta(x) + \varepsilon \quad (6)$$

This represents a hierarchical workflow, where the low-fidelity model (or emulator) is first calibrated against the high-fidelity under consideration of the mismatch between the two. This step requires outputs from forward runs of both the low (n_l) and the high-fidelity model (n_h), which also results in the dependency of η_l on the parameters of the high-fidelity model θ_h (eq. 6) (Fig. 1). In a second step this updated model is calibrated against the field data, y_f, while accounting for the bias in the high-fidelity model and a random error inherent to field observations.

In line with the KOH framework described above, Gaussian Process models for the approximation of the different terms (η_l, μ, δ) in eq. (6) are used. As the low-fidelity emulator, η_l, and the model mismatch, μ, depend on both the calibration parameters and the state variables, their covariances, $\Sigma_{\eta l}$ and Σ_μ, can

be formulated by adapting eq. (3) to new hyper-parameters, i.e. $\lambda_{\eta l}$, $\beta_{\eta l}$ and λ_{μ}, β_{μ}, respectively, and considering separate sets of model parameters, θ_l and θ_h, if needed (eqs. 7 & 8).

$$\Sigma_{\eta_l(i,j)} = \frac{1}{\lambda_{\eta_l}} \exp\left[-\sum_{k=1}^{p} \beta_{\eta_l,k} \left(x_{i,k} - x_{j,k}\right)^2 - \sum_{k\prime=1}^{q_h} \beta_{\eta_l,p+k\prime} \left(\theta_{h_{i,k\prime}} - \theta_{h_{j,k\prime}}\right)^2 - \sum_{k\prime\prime=1}^{q_l} \beta_{\eta_l,p+q_h+k\prime\prime} \left(\theta_{l_{ik\prime\prime}} - \theta_{l_{jk\prime\prime}}\right)^2 \right] \quad (7)$$

$$\Sigma_{\mu(i,j)} = \frac{1}{\lambda_{\mu}} \exp\left[-\sum_{k=1}^{p} \beta_{\mu,k} \left(x_{i,k} - x_{j,k}\right)^2 - \sum_{k\prime=1}^{q_h} \beta_{\mu,p+k\prime} \left(\theta_{h_{i,k\prime}} - \theta_{h_{j,k\prime}}\right)^2 - \sum_{k\prime\prime=1}^{q_l} \beta_{\mu,p+q_h+k\prime\prime} \left(\theta_{l_{ik\prime\prime}} - \theta_{l_{jk\prime\prime}}\right)^2 \right] \quad (8)$$

As for eqs. (3) and (4), the uknown hyper-parameters in eqs. (7) and (8) will be sampled and estimated alongside the thermal and hydraulic model parameters, which are specified in more detail for each model below. For the model bias, δ, in the multi-fidelity approach eq. (4) is adopted as it stands. Accordingly, the overall covariance function of the combined data set, z, which here contains outputs from the low- and high-fidelity computer models and the measured data, can be written as (Goh et al., 2013):

$$\Sigma_Z = \Sigma_{\eta l} + \begin{pmatrix} \Sigma_{\mu} & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} \Sigma_{\delta} + \Sigma_{\varepsilon} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad (9)$$

Thus, eq. (9) represents the mathematical implementation of the conceptual framework in eq. (6), specifically for the use with Gaussian Process emulators for the approximation of the low-fidelity model, model mismatch and model bias. Figure 1 shows a comparison of the mathematical concepts of the single and multi-fidelity approaches based on the different model terms used for the calculation of the covariances, as well as the field observations and model output data considered in the individual components.

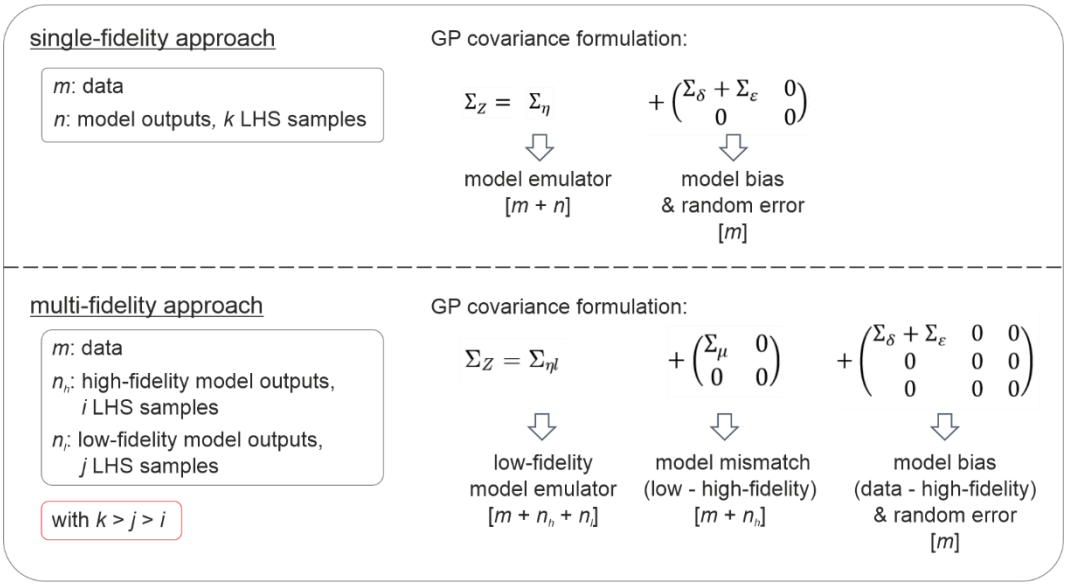

Figure 1: Conceptual overview on the data sets and the components of the Gaussian Process (GP) covariance formulation of the single-fidelity approach (Kennedy & O'Hagan framework) and the multi-fidelity approach for Bayesian parameter estimation.

2.3 Implementation of the Bayesian frameworks

In order to obtain the likelihood and approximate the posterior distributions of the unknown parameters, repeated model evaluations with iterative sample draws for x and θ are required to generate the data sets n, n_h and n_l. We implemented the KOH and the multi-fidelity framework using the STAN programming language (mc-stan.org), which provides a software platform for solving inverse problems using Bayesian statistics (Carpenter et al., 2017). It employs an efficient Hamiltonian Monte Carlo (HMC) algorithm for sampling from the posterior distributions (Betancourt, 2016; Hoffman and Gelman, 2014). We run HMC with 500 samples on four independent sampling chains for each parameter estimation shown in this study (the first 250 runs are discarded as the burn-in phase). In order to check the convergence of the HMC, we employ the $\hat{R}$ criterion which compares the inner and inter-chain variance of the posterior samples (Gelman and Rubin, 1992).

A detailed step-by-step guidance through the implementation process of the Bayesian calibration framework from Kennedy and O'Hagan (2001) in STAN is provided by Chong and Menberg (2018).

The multi-fidelity framework is set up in the same manner by adding the formulations for the additional covariance terms introduced in eqs. (7) and (8) and Figure 1 to the STAN code.

3 Demonstrative analytical heat transfer model

3.1 Model description and setup

The newly adapted multi-fidelity approach for parameter estimation of subsurface models is first tested on a 1D analytical heat transport model. This small-scale application allows a clear assessment of different setups with respect to the prior distributions of the unknown (hyper-)parameters. The analytical model represents a one-dimensional solution for temperature, T (°C), variation in the subsurface over time, t (s), and depth, Z (m). It is calculated based on the thermal diffusivity, κ (m²/s), of the ground and seasonal variation of the surface temperature, T_a (°C) (eq. (10)) (Grathwohl, 2012):

$$T(Z,t) = T_a + A\, exp\left[\frac{Z}{(\kappa * \tau * \pi^{-1})^{0.5}}\right] * sin\left[\frac{2\,\pi\,t}{\tau} - \frac{Z}{(\kappa * \tau * \pi^{-1})^{0.5}}\right] \quad (10)$$

where A (°C) is the annual amplitude of the surface temperature and τ (s) is the frequency of the seasonal variations. Accordingly, the average temperature over depth at a certain time is the quantity of interest, y, while different times of the year are used as state variable values for x. T_a and A are set to 12 °C and 15 °C, respectively, reflecting typical climatic conditions in central Europe. Thermal diffusivity of the ground is the unknown model parameter to be estimated. In contrast to the examples given in Goh et al. (2013) the high- and low-fidelity simulators in our study share the same set of parameters to be estimated, θ, as well as the same state variables, x, so the mathematical formulations stated above can be simplified.

The analytical model is expressed at multiple levels of fidelity by varying the resolution in depth, which leads to significantly different outputs of the average temperature at the considered time steps. Accordingly, the low-fidelity model outputs are obtained for a lower depth resolution (2m) than the high-fidelity model outputs (0.05m) (Figure 2).

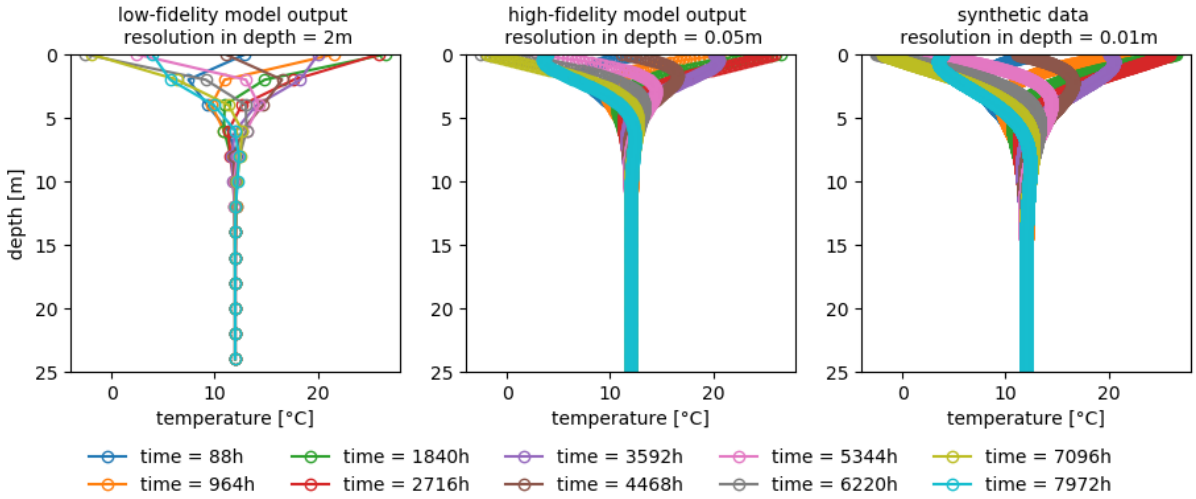

Figure 2: Model outputs of the 1D analytical heat transfer model at different levels of fidelity, i.e. different resolutions in depth for the ten time steps used as state variables.

These data sets are calibrated against synthetic data, generated from the model run with a very fine resolution in depth, and a thermal diffusivity of 7.0 * 10^{-7}m²/s, which is a typical value for saturated sand and gravel deposits (Stauffer et al., 2013). Knowing the true value of the parameter to be estimated allows us to verify that the Bayesian framework is able to quantify the "unknown" parameter correctly. Since synthetic data is used in this exercise and no random observational error is assigned to y_f, the covariance Σ_ε is not included in the mathematical framework (see eqs. (5) and (9)).

As 10 points in time are considered and the very fine model is evaluated once at each time step, the size of data set m = 10. The high-fidelity output data is supposed to stem from a computationally expensive model, so 20 Latin Hypercube samples (LHS) are chosen here (representing 20 iterative model evaluations) each evaluated at the 10 values of x, which leads to the size of data n_h = 200. Based on the study by Goh et al. (2013), a higher number of 50 LHS is used for the low-fidelity output data so that the size of n_l = 500. These larger LHS samples sizes in the multi-fidelity approach, as compared to the samples sizes typically used in the single-fidelity KOH framework, are related to the higher number of unknown, non-physical hyper-parameter to be estimated here. Prior probability distributions need to be defined for all unknown quantities in the KOH as well as the multi-fidelity framework. Regarding the

unknown model parameter thermal diffusivity, we test different normal distributions around the true value to evaluate the replicability of the estimation results. The prior distributions for the different GP hyper-parameters, λ and β, are set according to recommendations given in previous studies (Table 1) (Chong and Menberg, 2018; Guillas et al., 2009; Heo et al., 2012; Higdon et al., 2004).

Table 1: List of uncertain (hyper-) parameters in the single and multi-fidelity parameter estimation frameworks.

parameter	parameter description	prior probability distribution
θ	model (calibration) parameters	Normal (μ, σ)
λ_η	precision parameter for model emulator	Gamma (10, 1)
$\lambda_{\eta l}$	precision parameter for low-fidelity model emulator	Gamma (10, 1)
λ_δ	precision parameter for model bias	Gamma (10, 0.03)
λ_μ	precision parameter for model mismatch	Gamma (10, 0.03)
β_η	correlation strength parameter for model emulator	Beta (1, 0.5)
$\beta_{\eta l}$	correlation strength parameter for low-fidelity model emulator	Beta (1, 0.5)
β_δ	correlation strength parameter for model bias	Beta (1, 0.7)
β_μ	correlation strength parameter for model mismatch	Beta (1, 0.7)

3.2 Parameter estimation results

We compare the thermal diffusivity value obtained from the single-fidelity KOH framework (using the synthetic data and either the low- *or* high-fidelity model) against the multi-fidelity approach (using the synthetic data and both high- *and* low-fidelity models) given different prior information on the thermal diffusivity. All parameter estimation results show $\hat{R}$ values smaller than 1.1, indicating good convergence. Figure 3 shows the prior and posterior probability density functions of the thermal diffusivity in relation to the defined value in the analytical model used to obtain the synthetic data. For the higher prior probability distribution (Figure 3a) both KOH and the multi-fidelity method estimate posterior values around the true values, although the posterior from the KOH with the high fidelity model and the multi-fidelity framework perform slightly better. For the second prior distribution, however, the parameter inference with only the high- or low-fidelity models suggest the thermal diffusivity to be smaller than the true value, while the multi-fidelity approach correctly estimates higher

diffusivity values. Here, the combination of highly accurate information from the high-fidelity model and the additional information of the low-fidelty model steming from a large number of iterative runs with a dense sampling of the parameter space, is crucial for a reliable estimation of the true parameter value.

The different error functions in the multi-fidelity framework also allow a comparison of the magnitude of the model bias (i.e. discrepancy between the synthetic data and high-fidelity model output) and the model discrepancy (i.e. discrepancy between high- and low-fidelity model). The posterior distribution of the model bias precision hyper-parameter, λ_δ, has a higher mean value of 336, than the posterior of the model discrepancy hyper-parameter, λ_μ, with 36 (Figure S1). According to the covariance formulation (eq. (4)) this indicates that the model bias is smaller than the model discrepancy. This difference of about one order of magnitude is in good agreement with the difference in resolution along depth in Figure 2.

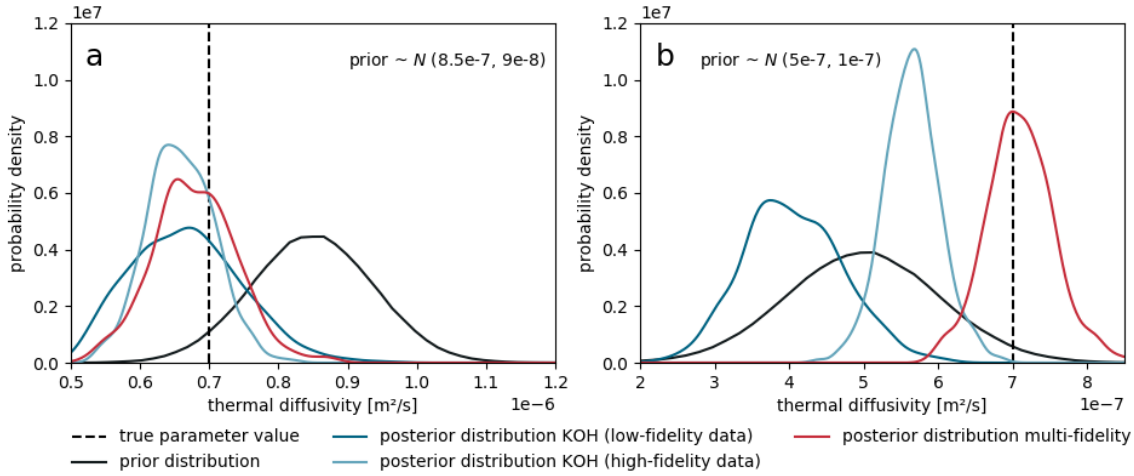

Figure 3: Posterior distribution of thermal diffusivity from the standard KOH framework using either the low *or* high-fidelity model (blue lines) versus the posterior distribution obtained from the multi-fidelity approach (MF) (red line) using both model outputs for two different prior distributions (a & b). Dashed line shows the true parameter value.

4 Hydro-thermal model of urban subsurface

4.1 Model description

Sustainable utilisation of geothermal energy in urban areas requires reliable and detailed knowledge of subsurface temperature distribution. Due to the scarcity of ground temperature measurements, numerical simulations are often utilised to estimate the ground temperature variations in urban subsurface. A key challenge in modelling urban ground temperatures at large scales are the model size and computational expense. The thermal interactions between the ground, surface, underground structures, and groundwater variations make the full 3D hydro-thermal modelling of urban subsurface an invalid option. Considering the large lateral scale of urban districts (many square kilometres) compared to the vertical scale (~ 100m) in such problems, Bidarmaghz et al. (2020) developed and validated a semi-3D modelling methodology, in which the 3D volume of the urban subsurface is numerically divided into several 2D horizontal planes. Conductive and convective heat transfer and fluid flow are considered at horizontal planes. Each set of equations (with variables of temperature, pressure, and velocity) refers to one specific plane – at a specific depth – taking into account the geology, hydrogeology and underground built environment characteristics at the relevant depth. Heat transfer in Z direction (depth) is evaluated by coupling the temperature distribution at each depth (T_n) to the temperature distribution of a shallower depth (T_{n-1}) and a deeper depth (T_{n+1}) by accounting for the selected vertical distance between the planes and effective thermal conductivity of the porous ground. This model is developed and used to analyse the impact of heat rejection from 13,000 residential basements in the Royal Borough of Kensington and Chelsea (RBKC), London. The basements are single level with a floor area of 50m^2, reflecting the basement of a typical two-bedroom terrace house and an average ceiling height of 3m. The consequent shallow subsurface temperature disturbance is investigated by coupling and solving the heat transfer and fluid flow equations in a porous medium (ground) with groundwater flow using COMSOL Multiphysics, which is a general purpose finite element solver. Depending on the depth that each 2D plane represents in the semi-3D model, its geological distribution and groundwater regime would vary following the thickness of the permeable gravel layer overlaying the London Clay Formation. To account for these

variations, the correspondent geological units and hydraulic head differences are assigned to the relevant 2D planes from surface the bottom of the permeable layer.

Similar to the analytical model presented in Section 3, three levels of fidelity are created for this semi-3D numerical model. This is done by varying the model resolution by depth, i.e. the number of horizontal planes representing the total 3D volume. The resulting so-called *fine*, *intermediate* and *coarse* models each comprises a certain number of layers in Z direction (depth) directly correlated to the resolution of the model. A maximum depth of 50m is considered in all models to: 1) capture all the geological variations down to the base of the London Clay Formation, 2) set the model base boundaries deep enough to avoid numerical edge effects and forced temperature constraints.

The finest, computationally most expensive model (Figure 4a) is used to produce synthetic data for the calibration of the coarse and the intermediate model. The fine model consists of 50 horizontal planes with a uniform 1m plane interval. A successful parameter estimation also requires computationally robust and efficient models for the high- and low-fidelity representations of the problem without significantly compromising on the level of details and accuracy. Therefore, the model resolution is reduced for the intermediate model (high-fidelity model) by considering a non-uniform plane interval distribution by depth. The high-fidelity model consists of 16 planes (Figure 4b), with plane intervals selected such that the model captures all the changes from surface to 50m depth including geological, hydrogeological and subsurface built environment variations. For example, in RBKC, the depth of the permeable River Terrace Deposits varies between 2m to 10m overlaying the London Clay Formation, and residential basements have a depth of 3m. The intermediate model, therefore, consists of plane intervals of 1-2m for the first 10m below the ground surface, and a uniform 4m plane interval from 10m to 50m depth. For the model with the lowest level of fidelity (coarse model), the number of planes is reduced to 10 with non-uniform plane intervals varying between 1m, 2m and 10m (Figure 4c). For the first 10m where the majority of changes occur, the coarse model has a slightly lower resolution than the intermediate model. Deeper than 10m, where the subsurface variations are minor, the model resolution is significantly reduced. Model predictions of time and depth-dependent ground temperature at 8

selected locations (shown in Figure 4d) are calibrated against synthetically generated observations (fine model outputs).

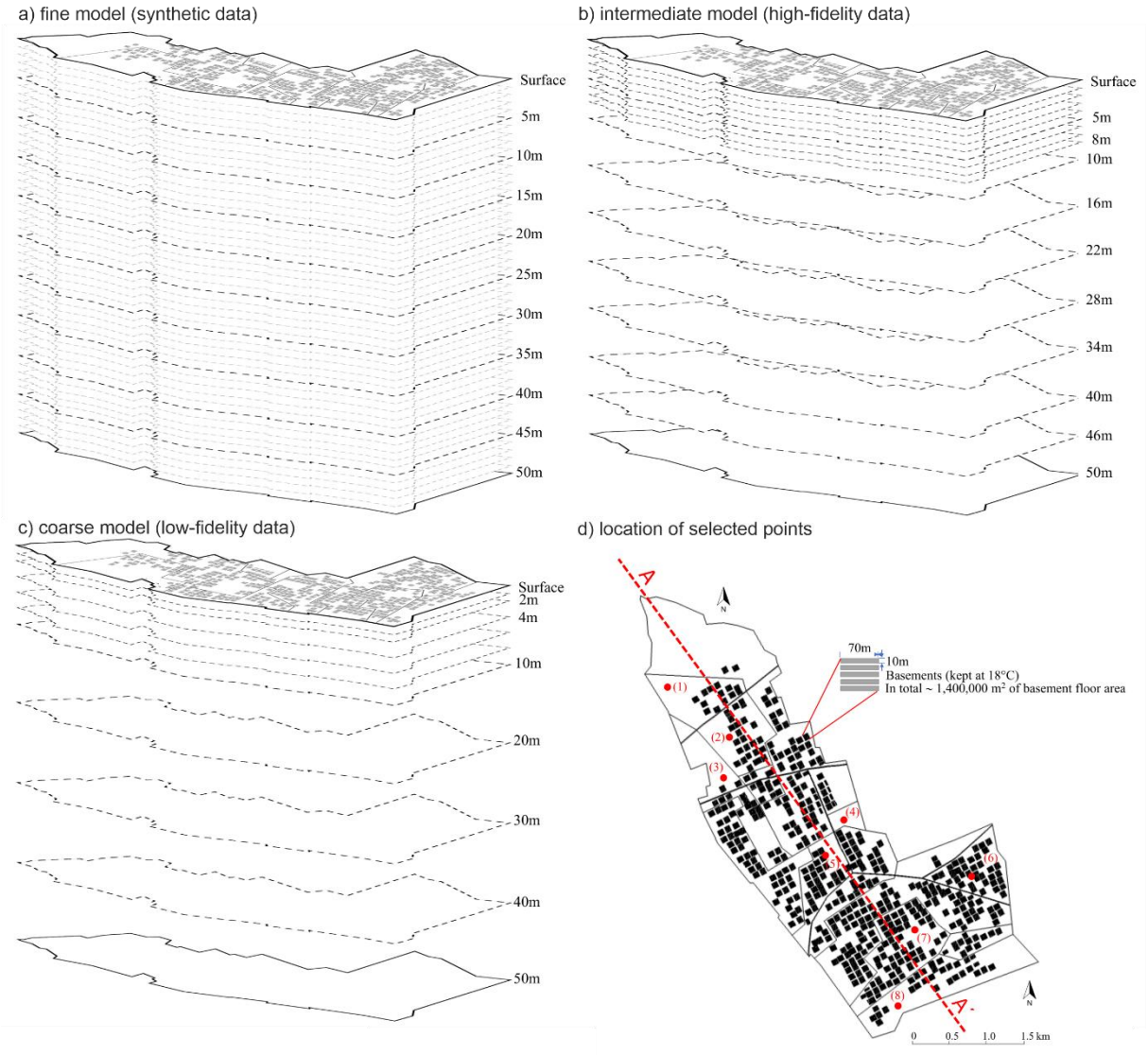

Figure 4: Semi-3D model setup for a) fine model used to generate synthetic data, b) intermediate model reflecting high-fidelity output data and c) coarse model reflecting low-fidelity output data, and d) location of the eight selected points used in the parameter estimation and section line A-A`.

4.2 Uncertainty and Sensitivity Analysis

In order to reduce the computational demand of the parameter estimation process, a sensitivity analysis was conducted on the heat and mass transfer finite element model of RBKC using the Morris method with factorial sampling for parameter screening (Menberg et al., 2016; Morris, 1991). In this process the parameter space is transformed into a unit-length hypercube, where repeated sampling sequences (so-called trajectories, *t*) start at a randomly chosen point and in which one parameter at the time is varied by a pre-defined value. The variations in the model output (so-calles elementary effects, *EE*) due to these changes in parameter values from the different trajectories can be statistically evaluated for each parameter. Here, we evaluate the absolute mean value μ^* as an indication of the magnitude of influence (Campolongo et al., 2007) (eq. 11), and standard deviation σ as a measure for the spread in the model output due to changes in a specific parameter *i* (Morris, 1991) (eq. 12):

$$\mu^*_i = 0.5 \sum_{t=1}^{r} |EE_{it}| \tag{11}$$

$$\sigma = \sqrt{\frac{1}{(r-1)} \sum_{t=1}^{r} (EE_{it} - \mu_i)^2} \tag{12}$$

The sensitivity analysis evaluated the modelled ground temperature sensitivity to the 10 potentially uncertain and important parameters of the finite element model (details of the model set up in (Bidarmaghz and Narsilio, 2018; Bidarmaghz et al., 2017)). The uncertain parameters in the model and the range of their values are listed in Table 2. The range of initial temperature values is based on measured groundwater temperatures in the London area, which are mostly between 10°C and 15.5°C at 60m depth, with a trend of lower temperatures in the shallow subsurface (Headon et al., 2009). Table 2 also shows the result of the sensitivity analysis, as modified mean μ^* and standard deviation σ, which were obtained with 110 runs of the fine model (10 parameters and a chosen trajectory number of t=10). We use the parameter estimation process to infer the uncertainty in the four most influential parameters: 1) ground initial temperature, 2) thickness of gravel, 3) surface cover type and 4) density of basements.

Table 2: Uncertain model parameters and sensitivity analysis results from Morris method.

rank	parameter	unit	true value	minimum	maximum	mean (μ^*)	standard deviation (σ)
1	ground initial temperature[a,b]	°C	12.5	9	15	4.45	4.06
2	thickness of gravel[c,d]	m	8	2	10	3.02	3.64
3	surface cover type[e]	-	0.8	0.1	0.9	2.56	2.99
4	density of basements	-	1	1	2	2.22	2.39
5	heat source temperature[f]	°C	18	18	22	1.18	1.28
6	depth of heat source	m	3	3	18	0.88	1.21
7	ground hydraulic conductivity[g]	m/s	5.6e-4	3.3e-6	3.3e-3	0.62	0.77
8	thermal conductivity (gravel)[a]	W/(m.K)	2.5	0.77	2.6	0.37	0.53
9	hydraulic head difference[c,h]	m	10	4	12	0.40	0.39
10	thermal conductivity (clay)[a]	W/(m.K)	1.7	1.5	2.45	0.18	0.27

[a] Busby et al. (2009) [b] Price et al. (2018) [c] Bidarmaghz et al. (2019b) [d] BGS (2017)
[e] Baggs (1983) [f] Heo et al. (2012) [g] Bricker and Bloomfield (2014)
[h] Mansour et al. (2018)

4.3 Setup of the inverse problem

As aforementioned, 8 locations in RBKC were selected to investigate the time and depth-dependent ground temperature variations as the quantity of interest. The synthetic data (m) mimics field measurements of ground temperatures at these 8 different locations (combinations of X and Y values), and at each location, for 10 points in depth (Z). Thus, there are $m = 80$ total observations, which follows the rationale that field data measurements are typically scarce and limited in depth. These are generated by running the fine model with the 'true' parameter values listed in Table 2 and shown in Figure 5.

The 3D spatial coordinates (X, Y, and Z) are used as state variables (x) in the Bayesian framework, as the temperature is expected to vary spatially. Indeed, as shown in Figure 5, the temperatures as well as the difference between the models vary from point to point within RBKC. Locations that have larger depths of permeable ground and therefore, larger ground temperature disturbance, show significant output dependency to the subsurface resolution of the model (such as point 7 in Figure 4d with 10m of River Terrace Deposits and the associated temperature profile shown in Figure 5a). In contrast, at points where the permeable material is shallow, the impact of model resolution on temperature outputs

decreases, resulting in smaller deviations between synthetic data, and high- and low-fidelity outputs (points 6, 5 and 1 in Figure 4d with 5m, 2m and 0m of permeable ground respectively).

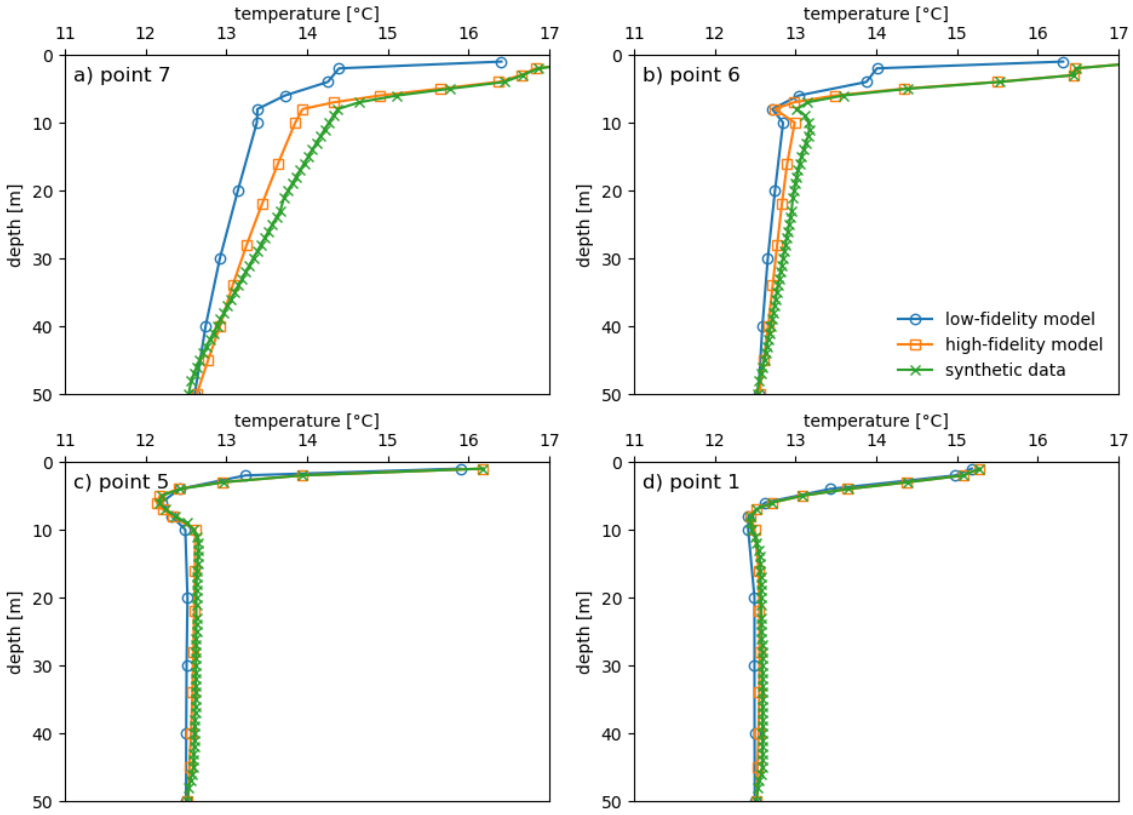

Figure 5: Model outputs of the heat and mass transfer processes obtained with the true values from Table 2 after 25 years at different points in RBKC (refer to Figure 4d) with different depths of permeable River Terrace Deposits: a) 10m, b) 5m, c) 2m and d) 0m.

As aforementioned, the intermediate model represents a high-fidelity model of heat and fluid transport in RKBC. As for the analytical problem, it is run with 10 LHS samples of parameter combinations. Each model execution is evaluated at the same 8 values of X and Y, but only for 3 points in depth (1m, 10m, 50m). Hence, we have n_h = 8 x 3 x 10 = 240 high-fidelity model outputs. The low-fidelity, i.e. the coarse model, is run with 50 LHS samples of parameter combinations, and each model execution is evaluated at only 2 points in depth (1m, 10m) at the 8 locations. Thus, we have n_l = 8 x 2 x 50 = 800 low-fidelity model outputs. Accordingly, the dimension of the overall data matrix, z, equals 1120*1120 (including the m = 80 total observations), which is still suitable for inversion. These are shown in Figure 6.

The selection of LHS sample size follows the rationale that the expensive numerical model is executed fewer times to minimize computational burden, yet the few samples provide accurate information at a few crucial points in the parameter space. The cheaper, low-fidelity model is evaluated for a larger range of combinations of model parameter values, providing information that is less accurate but covers the parameter space well. The prior distributions for the GP hyper-parameters are set identically to the analytical problem (Table 1). The four model parameters to be estimated are assigned normal distributions with mean values set arbitrarily higher and lower than their true values, and standard deviations that cover the true values (Table 3).

Table 3: Specified prior distributions for the calibration parameters of the RBKC model.

parameter	**distribution**
ground initial temperature	Normal (10.9, 1.2)
thickness of gravel	Normal (6.0, 1.6)
surface cover type	Normal (0.5, 0.16)
density of basements	Normal (1.5, 0.2)

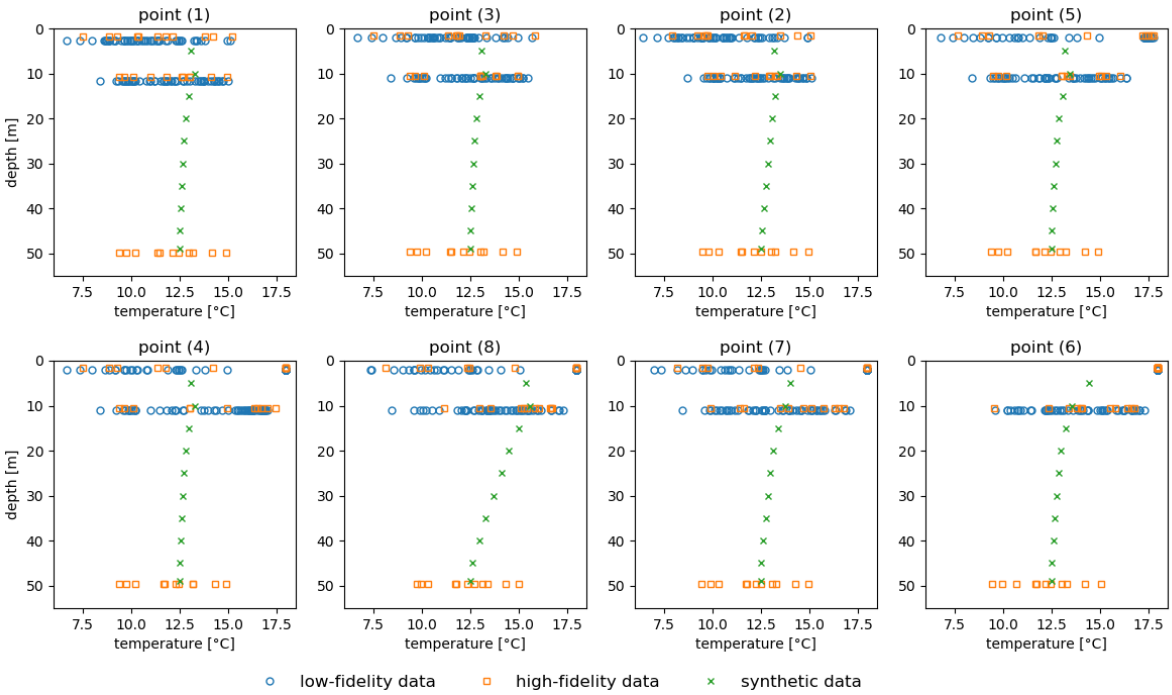

Figure 6: Input data for the single and multi-fidelity parameter estimation at each location. The depth refers to the corresponding layer of the semi 3D numerical model picked for temperature evaluation.

4.4 Parameter estimation results

The single (KOH) and multi-fidelity parameter estimation frameworks are applied using the data sets shown in Figure 6 to infer the posterior probability distributions of the most influential parameters of the numerical model. $\hat{R}$ values for all unknown (hyper-)parameters were below 1.1 indicating good convergence of the HMC chains.

The results from the two parameter estimation frameworks in Figure 7 reveal a significant difference. The KOH framework using only the low-fidelity model yields rather wide posterior distributions, which for the first two parameters show shifts away from the true parameter values. Although the KOH framework correctly identifies a large model bias function (i.e. between the low-fidelity model and the synthetic data), the information contained in the sparse amounts of synthetic data is insufficient to learn about the model parameters. The exercise with only the high-fidelity model significantly improves the parameter estimation for the gravel thickness and surface cover. This is because the high-fidelity model is composed of more layers in the shallow subsurface (up to 10 m depth). Hence it contains temperature outputs that contain information relevant to these two parameters (Figure 4b & c). Neither the low-fidelity nor the high-fidelity model parameter estimation result in correct inference of the other two parameters: initial ground temperature and density of basements. On the other hand, the mode values of the posterior distributions of all four parameters obtained from the multi-fidelity framework agree very well with their true parameter values. Also, the narrow ranges of uncertainty around the mode values reflect a high degree of confidence about these values, which leads to overall excellent results for this parameter estimation exercise (Figure 7). Additional parameter estimation exercises with varying prior distributions for the hyper-parameters of the KOH and the multi-fidelity approach show that the median values of resulting posteriors of the four model parameters are very consistent (Figure S2, S3 & S4), which higlights the robustness of the parameter estimation results.

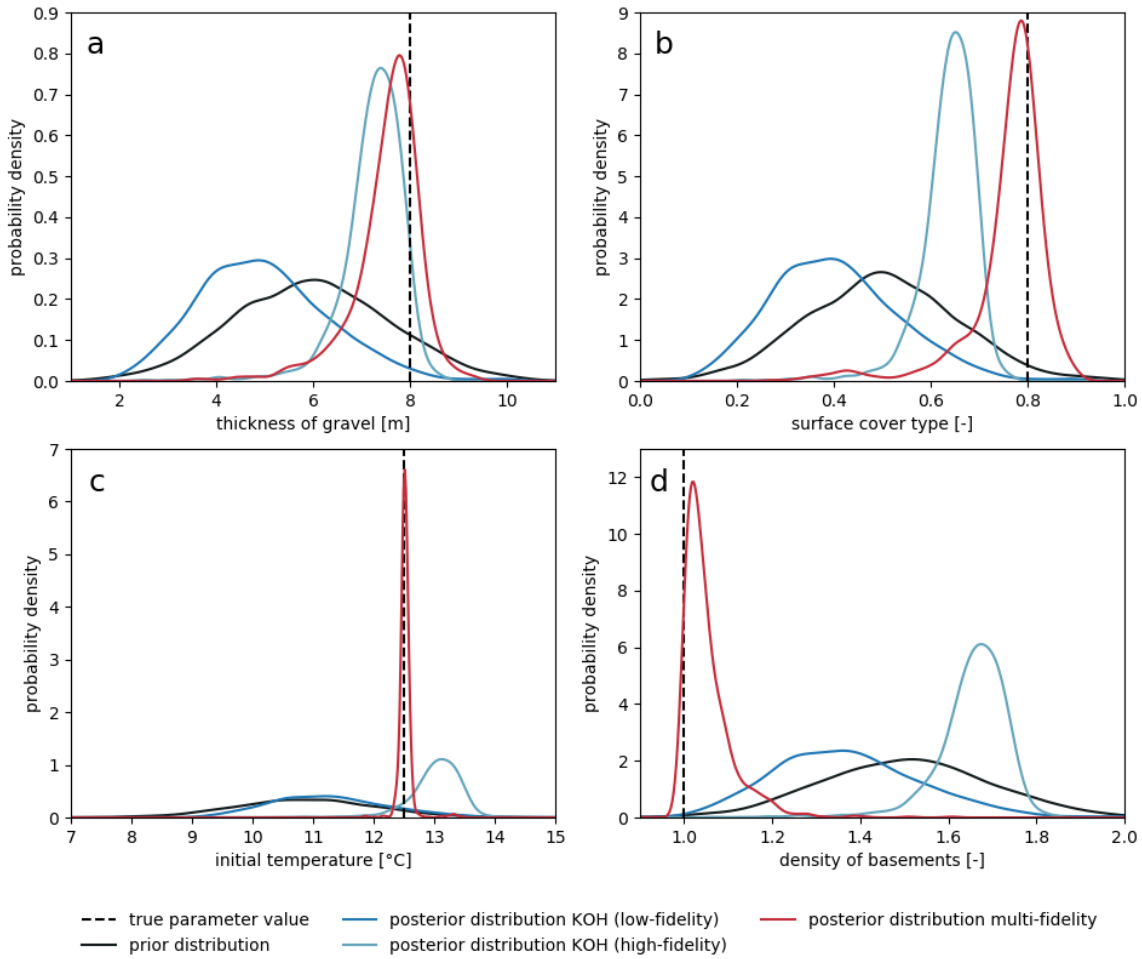

Figure 7: Parameter estimation results for the four influential parameters (a-d) of the numerical model shown as assumed prior probability distribution, posterior probability distributions using one set of model outputs (KOH) each, and both data sets at different levels of model fidelity (MF).

To investigate the contribution of the proposed parameter estimation framework to the accuracy of the subsurface hydro-thermal model, the mode values of the inferred parameters (posteriors) are fed back into the high-fidelity numerical model. Results show that the deviation of ground temperatures from the synthetic data are significantly reduced compared to the deviation resulting from using the prior estimates of model parameters. These are displayed in Figure 8 which shows the temperature discrepancy resulting from the mode values of the prior and posterior estimates along section A-A' of the modelled area. More importantly, Figure 8 shows that there are distinct differences in the magnitude of discrepancy reduction across the different areas of the modelling domain, both horizontally and in

the exemplary depth layers. In areas where the ground is dominantly consisting of London Clay (e.g. 0-4500m), the ground temperatures obtained from the model with the estimated values show relatively smaller deviation than the model with the prior values for all depths. Conversely, in the southern part of the district (ca. 4700-6700m), where shallow ground consists of sand and gravel, both prior and posterior model results show smaller deviation in shallow depths, which becomes larger by depth. These observations are attributed to different heat transfer mechanisms occurring in different parts of the studied area. Within the London Clay Formation, heat transfer is dominated by conduction, in which the ground temperature distribution is largely impacted by the initial thermal state. In the southern parts of the area, however, heat transfer at shallow depth is mainly by convection, which reduces the effect of initial ground temperature. Therefore, deviations for both prior and posterior models are relatively small at shallow depths and become larger where the ground consists of London Clay at greater depths. When interpreting these findings one has to bear in mind, that the model results are compared against synthetic data obtained from a very accurate model, and not to real data. Thus, the results are a reflection of increased accuracy of the heat and mass transport modelling and their influence on subsurface temperature elevations. These, in turn impact the robustness of planned geothermal systems, in terms of efficient and sustainable use, as well as cost-efficient system installation.

It has to be noted that due to the use of synthetic data from the fine model, the training data contains no noise. Using real data would require additional random error terms in the Gaussian Process emulation (see eqs. (6) and (9)), and that would increase the number of unknowns to be estimated. Also, previous studies that applied the KOH framework to other types of numerical models observed a potential confounding of inference between the model bias (in our synthetic example the error between the finest and the intermediate model) and the random (measurement) error, which leads to larger confidence intervals of the estimated parameters (Li et al., 2016; Menberg et al., 2017; Menberg et al., 2018). On the other hand, the comparably intuitive interpretation of the GP hyper-parameters of the error terms as precision and smoothness of the error functions over x and θ also allows a detailed assessment on where improvement of the model or more informative data is needed.

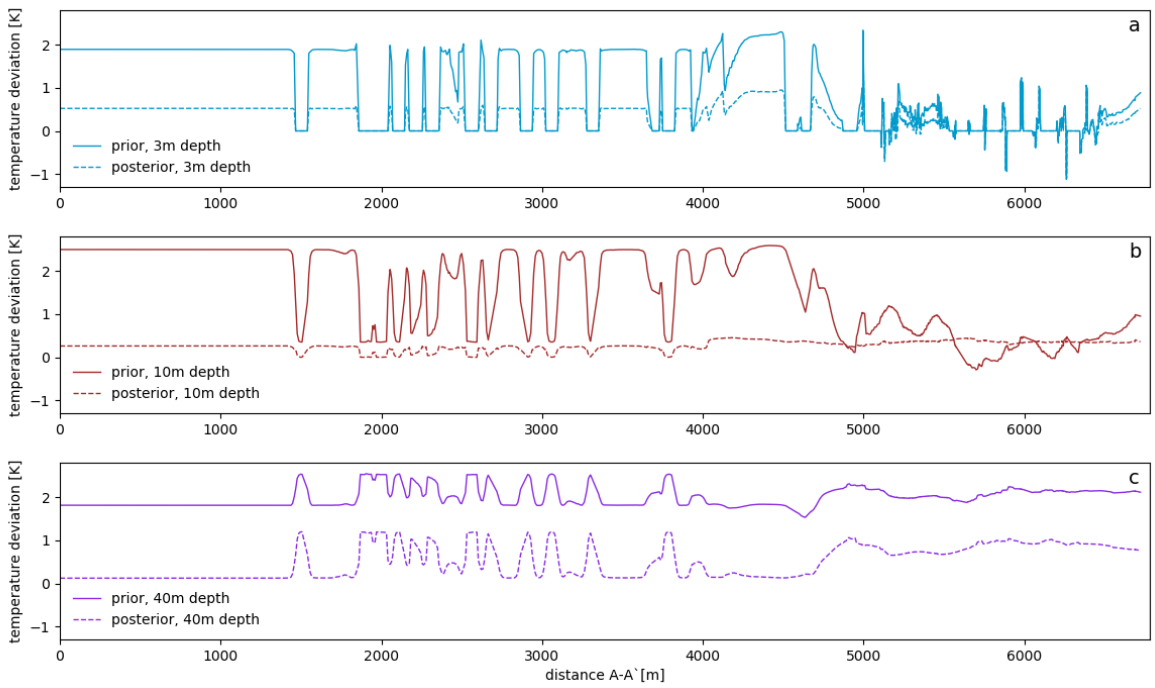

Figure 8: Temperature deviation between the synthetic model output (true parameter values) and the model output using the prior and posterior mode values in the high-fidelity model, respectively, along section line A-A' (see Figure 4d) at a) 3m, b) 10m and c) 40m depth.

Apart from the run time of the numerical models, the computational costs of matrix inversion during the parameter estimation represent a significant factor. Obviously, the multi-fidelity framework with a matrix size of z_{MF}=1120*1120, has higher computational expenses than a single model framework (z_{SF}=880*880). While the increase in precision and accuracy of the estimated parameter is significantly worthwhile, it is advised to use a programming language that is efficient at matrix calculations, such as Cholesky decomposition. It also has to be mentioned, that compared to traditional parameter estimation approaches, which aim at minimizing solely the error between model output and observations, our Bayesian multi-fidelity approach is likely to be more time-consuming, due to the large sets of model output needed for the inference of the hyper-parameters. Yet, it provides potentially interesting information about the different error terms.

The multi-fidelity Bayesian framework presented in this study is also specifically designed to work with Gaussian Process emulators. This means that once the training data for these is generated, there is no

need to run the expensive numerical models again. This allows testing different parameter estimation setups (e.g. in terms of prior distributions, data sub-sets, etc.) more efficiently than in frameworks, which calibrate directly against outputs from the numerical codes. In our study, we were thus able to use different combinations of *XY*-planes at different depths *Z* from the high- and low-fidelity model outputs for building the GP models. Parameter estimation results from those different GP models revealed a strong sensitivity to the information contained at different depths of the numerical model. In particular, a good estimation of the initial ground temperature value requires information from the deep layers, as the temperature change over time in this depth is minimal, and thus the absolute temperature after 25 years of simulation time is very sensitive to the initial state. This example shows that multiple re-iterations of parameter estimation with the emulators under different input data settings can provide valuable knowledge and enhance our understanding of subsurface heat and mass transport processes.

5 Conclusion

This paper presents a novel Bayesian parameter estimation approach using outputs from models at different levels of fidelity, reflecting different spatial resolutions and accordingly different computational burdens. Our framework combines information from a few parametric model outputs from a physically accurate, but expensive, high-fidelity computer model, with a larger number of evaluations from a less expensive and less accurate low-fidelity model. This enables us to include accurate information about the model output at sparse points in the parameter space, as well as dense samples across the entire parameter space, albeit with a lower physical accuracy.

We first apply the multi-fidelity approach to a simple 1D analytical heat transfer model, and secondly on a semi-3D coupled mass and heat transport numerical model, and estimate the unknown model parameters. By using synthetic data generated with known parameter values, we were able to test the reliability of the new method, as well as the improved performance over the standard single-fidelity approach, under different framework settings. Overall, the results from the analytical and numerical model show that combining 50 runs of the low resolution model with data from only 10 runs of a higher

resolution model significantly improved the posterior distribution results, both in terms of agreement with the true parameter values and confidence interval around this value.

In addition to estimating unknown model parameters, the Bayesian formulation allows examination of error terms, such as model bias and the discrepancy between high- and low-fidelity model, and thus the loss in accuracy when decreasing the model resolution. As we use synthetically generated data in our studies, the resulting errors terms showed realistic magnitudes. However, it is well known that the inference of error terms is likely to be more limited when large random (measurement) noise is present in the data. Also, the novel framework hugely benefits from the use of Gaussian Process models for emulating numerical model outputs, as well as error functions, which allows re-estimation (and thus re-calibration) under different settings without additional runs from the expensive numerical model. In particular, for large-scale, coupled heat and mass transport models this represents a significant advantage over manual or existing parameter estimation frameworks.

While the use of synthetic data allows assessment of the performance of the single and multi-fidelity approach, it also simplifies the inverse problems and reduces the number of unknown parameters to be estimated in the Bayesian framework. The next steps for further testing of the method are therefore employing real data from field measurements and adding statistical formulations for model prediction based on the inferred posterior distributions of the estimated parameters. This will also facilitate a comprehensive comparison with other calibration frameworks, in terms of accuracy and predictive power. Also, many other forms of structural uncertainty exist, besides spatial discretization that is employed here. It might be worthwhile for future studies to look into spatial heterogeneity, as a major contributor to overall uncertainty in many subsurface flow and heat models.

Acknowledgements

This work was supported by AI for Science and Government (ASG), UKRI's Strategic Priorities Fund awarded to the Alan Turing Institute, UK and the Lloyd's Register Foundation programme on Data-centric Engineering.

References

Asher MJ, Croke BFW, Jakeman AJ, Peeters LJM. A review of surrogate models and their application to groundwater modeling. Water Resources Research 2015; 51: 5957-5973.

Attard G, Winiarski T, Rossier Y, Eisenlohr L. Impact of underground structures on the flow of urban groundwater. Hydrogeology Journal 2016; 24: 5-19.

Baggs SA. Remote prediction of ground temperature in Australian soils and mapping its distribution. Solar Energy 1983; 30: 351-366.

Bayer P, Attard G, Blum P, Menberg K. The geothermal potential of cities. Renewable and Sustainable Energy Reviews 2019; 106: 17-30.

Betancourt M. A Conceptual Introduction to Hamiltonian Monte Carlo, 2016, pp. 60.

Beven K, Binley A. The future of distributed models: model calibration and uncertainty prediction. Hydrological processes 1992; 6: 279-298.

Beven K, Binley A. GLUE: 20 years on. Hydrological processes 2014; 28: 5897-5918.

BGS. Ground Thermal and Hydraulic Property Data. In: Survey BG, editor, 2017.

Bidarmaghz A, Choudhary R, Soga K, Kessler H, Terrington L, Thorpe S. Influence of Geology and Hydrogeology on Heat Rejection from Residential Basements in Urban Areas. Tunnelling and Underground Space Technology 2019a; 92.

Bidarmaghz A, Choudhary R, Soga K, Kessler H, Terrington RL, Thorpe S. Large-scale urban underground hydro-thermal modelling–A case study of the Royal Borough of Kensington and Chelsea, London. Science of The Total Environment 2019b: 134955.

Bidarmaghz A, Choudhary R, Soga K, Terrington L, Kessler H, Thorpe S. Large-scale urban underground hydro-thermal modelling – A case study of the Royal Borough of Kensington and Chelsea, London Science of The Total Environment 2020; 700.

Bidarmaghz A, Narsilio G. Heat exchange mechanisms in energy tunnel systems. Geomechanics for Energy and the Environment 2018.

Bidarmaghz A, Narsilio G, Buhmann P, Moormann C, Westrich B. Thermal Interaction Between Tunnel Ground Heat Exchangers and Borehole Heat Exchangers. Geomechanics for Energy and the Environment 2017; 10: 29-41.

Bricker S, Bloomfield J. Controls on the basin-scale distribution of hydraulic conductivity of superficial deposits: a case study from the Thames Basin, UK. Quarterly Journal of Engineering Geology and Hydrogeology 2014; 47: 223-236.

Busby J, Lewis M, Reeves H, Lawley R. Initial geological considerations before installing ground source heat pump systems. Quarterly Journal of Engineering Geology and Hydrogeology 2009; 42: 295-306.

Campolongo F, Cariboni J, Saltelli A. An effective screening design for sensitivity analysis of large models. Environmental Modelling & Software 2007; 22: 1509-1518.

Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, et al. Stan: A probabilistic programming language. Journal of statistical software 2017; 76.

Chong A, Menberg K. Guidelines for the Bayesian calibration of building energy models. Energy and Buildings 2018; 174: 527-547.

Cui T, Fox C, O'Sullivan M. Bayesian calibration of a large-scale geothermal reservoir model by a new adaptive delayed acceptance Metropolis Hastings algorithm. Water Resources Research 2011; 47.

Cui T, Peeters L, Pagendam D, Pickett T, Jin H, Crosbie RS, et al. Emulator-enabled approximate Bayesian computation (ABC) and uncertainty analysis for computationally expensive groundwater models. Journal of Hydrology 2018; 564: 191-207.

Doherty J. PEST Model-Independent Parameter Estimation User Manual: 5th Edition. Watermark Numerical Computing 2010.

Durantin C, Rouxel J, Désidéri J-A, Glière A. Multifidelity surrogate modeling based on radial basis functions. Structural and Multidisciplinary Optimization 2017; 56: 1061-1075.

Efendiev Y, Galvis J, Hou TY. Generalized multiscale finite element methods (GMsFEM). Journal of Computational Physics 2013; 251: 116-135.

Epting J, Huggenberger P. Unraveling the heat island effect observed in urban groundwater bodies–Definition of a potential natural state. Journal of hydrology 2013; 501: 193-204.

Fernández-Godino MG, Park C, Kim N-H, Haftka RT. Review of multi-fidelity models. arXiv preprint arXiv:1609.07196 2016.

Forrester AI, Sóbester A, Keane AJ. Multi-fidelity optimization via surrogate modelling. Proceedings of the royal society a: mathematical, physical and engineering sciences 2007; 463: 3251-3269.

Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian data analysis. Vol 2: Chapman & Hall/CRC Boca Raton, FL, USA, 2014.

Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. Statistical science 1992: 457-472.

Goh J, Bingham D, Holloway JP, Grosskopf MJ, Kuranz CC, Rutter E. Prediction and computer model calibration using outputs from multifidelity simulators. Technometrics 2013; 55: 501-512.

Grathwohl P. Diffusion in natural porous media: contaminant transport, sorption/desorption and dissolution kinetics. Vol 1: Springer Science & Business Media, 2012.

Guillas S, Rougier J, Maute A, Richmond A, Linkletter C. Bayesian calibration of the Thermosphere-Ionosphere Electrodynamics General Circulation Model (TIE-GCM). Geoscientific Model Development 2009; 2: 137-144.

Headon J, Banks D, Waters A, Robinson V. Regional distribution of ground temperature in the Chalk aquifer of London, UK. Quarterly Journal of Engineering Geology and Hydrogeology 2009; 42: 313-323.

Heo Y, Choudhary R, Augenbroe G. Calibration of building energy models for retrofit analysis under uncertainty. Energy and Buildings 2012; 47: 550-560.

Heße F, Comunian A, Attinger S. What we talk about when we talk about uncertainty. Toward a unified, data-driven framework for uncertainty characterization in hydrogeology. Frontiers in Earth Science 2019; 7: 118.

Higdon D, Kennedy M, Cavendish JC, Cafeo JA, Ryne RD. Combining field data and computer simulations for calibration and prediction. SIAM Journal on Scientific Computing 2004; 26: 448-466.

Hoffman MD, Gelman A. The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. Journal of Machine Learning Research 2014; 15: 1593-1623.

Jenny P, Lee S, Tchelepi HA. Multi-scale finite-volume method for elliptic problems in subsurface flow simulation. Journal of Computational Physics 2003; 187: 47-67.

Kennedy MC, O'Hagan A. Predicting the output from a complex computer code when fast approximations are available. Biometrika 2000; 87: 1-13.

Kennedy MC, O'Hagan A. Bayesian calibration of computer models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 2001; 63: 425-464.

Li Q, Augenbroe G, Brown J. Assessment of linear emulators in lightweight Bayesian calibration of dynamic building energy models for parameter estimation and performance prediction. Energy and Buildings 2016; 124: 194-202.

Lu D, Ye M, Hill MC, Poeter EP, Curtis GP. A computer program for uncertainty analysis integrating regression and Bayesian methods. Environmental Modelling & Software 2014; 60: 45-56.

Mansour M, Wang L, Whiteman M, Hughes A. Estimation of spatially distributed groundwater potential recharge for the United Kingdom. Quarterly Journal of Engineering Geology and Hydrogeology 2018; 51: 247-263.

Menberg K, Blum P, Schaffitel A, Bayer P. Long-term evolution of anthropogenic heat fluxes into a subsurface urban heat island. Environmental science & technology 2013; 47: 9747-9755.

Menberg K, Heo Y, Choudhary R. Efficiency and Reliability of Bayesian Calibration of Energy Supply System Models. IBPSA Building Simulation Conference, San Fransisco, 2017.

Menberg K, Heo Y, Choudhary R. Influence of error terms in Bayesian calibration of energy system models. Journal of Building Performance Simulation 2018: 1-15.

Menberg K, Heo Y, Choudhary RJE, Buildings. Sensitivity analysis methods for building energy models: Comparing computational costs and extractable information. 2016; 133: 433-445.

Moore C, Doherty J. The cost of uniqueness in groundwater model calibration. Advances in Water Resources 2006; 29: 605-623.

Morris MDJT. Factorial sampling plans for preliminary computational experiments. 1991; 33: 161-174.

Panday S, Langevin C, Niswonger R, Ibaraki M, Hughes J. MODFLOWUSG: An Unstructured Grid Version of MODFLOW for Simulating Groundwater Flow and Tightly Coupled Processes Using a Control Volume Finite-Difference Formulation. US Geological Survey, Reston, Virginia 2013.

Poeter EP, Hill MC. Documentation of UCODE, a computer code for universal inverse modeling: DIANE Publishing, 1998.

Poeter EP, Hill MC. UCODE, a computer code for universal inverse modeling. Computers & Geosciences 1999; 25: 457-462.

Price SJ, Terrington RL, Busby J, Bricker S, Berry T. 3D ground-use optimisation for sustainable urban development planning: A case-study from Earls Court, London, UK. Tunnelling and Underground Space Technology 2018; 81: 144-164.

Rajabi MM, Ataie-Ashtiani B, Simmons CT. Model-data interaction in groundwater studies: Review of methods, applications and future directions. Journal of hydrology 2018; 567: 457-477.

Razavi S, Tolson BA, Burn DH. Review of surrogate modeling in water resources. Water Resources Research 2012; 48.

Robinson T, Eldred M, Willcox K, Haimes R. Surrogate-based optimization using multifidelity models with variable parameterization and corrected space mapping. Aiaa Journal 2008; 46: 2814-2822.

Stauffer F, Bayer P, Blum P, Giraldo NM, Kinzelbach W. Thermal use of shallow groundwater: CRC Press, 2013.

Sun G, Li G, Stone M, Li Q. A two-stage multi-fidelity optimization procedure for honeycomb-type cellular materials. Computational Materials Science 2010; 49: 500-511.

Voss CI. Editor's message: Groundwater modeling fantasies—part 1, adrift in the details. Hydrogeology Journal 2011a; 19: 1281-1284.

Voss CI. Editor's message: Groundwater modeling fantasies—part 2, down to earth. Hydrogeology Journal 2011b; 19: 1455-1458.

Vrugt JA. Markov chain Monte Carlo simulation using the DREAM software package: Theory, concepts, and MATLAB implementation. Environmental Modelling & Software 2016; 75: 273-316.

Vrugt JA, Sadegh M. Toward diagnostic model calibration and evaluation: Approximate Bayesian computation. Water Resources Research 2013; 49: 4335-4345.

Vrugt JA, Ter Braak CJ, Clark MP, Hyman JM, Robinson BA. Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation. Water Resources Research 2008; 44.

Vrugt JA, ter Braak CJF, Gupta HV, Robinson BA. Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling? Stochastic Environmental Research and Risk Assessment 2009; 23: 1011-1026.

Welter DE, White JT, Hunt RJ, Doherty JE. Approaches in highly parameterized inversion—PEST++ Version 3, a Parameter ESTimation and uncertainty analysis software suite optimized for large environmental models. US Geological Survey, 2015.

Xu T, Valocchi AJ. A Bayesian approach to improved calibration and prediction of groundwater models with structural error. Water Resources Research 2015; 51: 9290-9311.

Xu T, Valocchi AJ, Ye M, Liang F. Quantifying model structural error: Efficient Bayesian calibration of a regional groundwater flow model using surrogates and a data-driven error model. Water Resources Research 2017a; 53: 4084-4105.

Xu T, Valocchi AJ, Ye M, Liang F, Lin Y-F. Bayesian calibration of groundwater models with input data uncertainty. Water Resources Research 2017b; 53: 3224-3245.

Zaytsev A, Burnaev E. Large scale variable fidelity surrogate modeling. Annals of Mathematics and Artificial Intelligence 2017; 81: 167-186.

Zhang J, Man J, Lin G, Wu L, Zeng L. Inverse modeling of hydrologic systems with adaptive multifidelity Markov chain Monte Carlo simulations. Water Resources Research 2018; 54: 4867-4886.

Zheng Q, Zhang J, Xu W, Wu L, Zeng L. Adaptive multifidelity data assimilation for nonlinear subsurface flow problems. Water Resources Research 2019; 55: 203-217.