Data-Centric Engineering (2022), 3: e32
doi:10.1017/dce.2022.32

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

Bayesian parameter inference for shallow subsurface modeling using field data and impacts on geothermal planning

Monika J. Kreitmair[1,*], Nikolas Makasis[1], Kathrin Menberg[2], Asal Bidarmaghz[3], Gareth J. Farr[4,5], David P. Boon[4] and Ruchi Choudhary[1,6]

[1]Department of Engineering, University of Cambridge, Cambridge CB2 1PZ, United Kingdom
[2]Institute of Applied Geosciences, Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany
[3]School of Civil and Environmental Engineering, University of New South Wales, Sydney, Australia
[4]British Geological Survey, Cardiff University, Cardiff CF10 3AT, United Kingdom
[5]The Coal Authority, Mansfield, Nottinghamshire NG18 4RG, United Kingdom
[6]Data-centric Engineering, Alan Turing Institute, British Library, London NW1 2DB, United Kingdom
*Corresponding author. E-mail: mk2040@cam.ac.uk

Received: 10 December 2021; **Revised:** 25 August 2022; **Accepted:** 25 September 2022

Keywords: Bayesian calibration; finite element methods; parameter inference; shallow geothermal energy; uncertainty quantification

Abstract

Understanding the subsurface is crucial in building a sustainable future, particularly for urban centers. Importantly, the thermal effects that anthropogenic infrastructure, such as buildings, tunnels, and ground heat exchangers, can have on this shared resource need to be well understood to avoid issues, such as overheating the ground, and to identify opportunities, such as extracting and utilizing excess heat. However, obtaining data for the subsurface can be costly, typically requiring the drilling of boreholes. Bayesian statistical methodologies can be used towards overcoming this, by inferring information about the ground by combining field data and numerical modeling, while quantifying associated uncertainties. This work utilizes data obtained in the city of Cardiff, UK, to evaluate the applicability of a Bayesian calibration (using GP surrogates) approach to measured data and associated challenges (previously not tested) and to obtain insights on the subsurface of the area. The importance of the data set size is analyzed, showing that more data are required in realistic (field data), compared to controlled conditions (numerically-generated data), highlighting the importance of identifying data points that contain the most information. Heterogeneity of the ground (i.e., input parameters), which can be particularly prominent in large-scale subsurface domains, is also investigated, showing that the calibration methodology can still yield reasonably accurate results under heterogeneous conditions. Finally, the impact of considering uncertainty in subsurface properties is demonstrated in an existing shallow geothermal system in the area, showing a higher than utilized ground capacity, and the potential for a larger scale system given sufficient demand.

Impact statement

This paper demonstrates the use of Bayesian techniques in engineering applications for which data is scarce, specifically applications within the subsurface. The combined use of field data, numerical modeling, and Bayesian calibration to infer accurately important (and often difficult to quantify) parameters at the district scale are showcased. The applicability of this methodology under heterogeneous conditions is demonstrated, as

is that different measurement locations can contribute different amounts of information to the inference of parameter values, understanding of which can reduce costs incurred in site investigations. The significance of calibration and accounting for uncertainty is highlighted in the design of an existing shallow geothermal system, indicating higher ground capacity than utilized and thus potential for a larger scale system.

1. Introduction

As cities grow in size and density, shortage of space above ground drives the increasing development of underground infrastructures. Residential and commercial spaces, transport systems, industrial processes, and energy applications all compete for the use of the shallow subsurface. Structures built into the ground and the associated heat flux from these are known to impact ground temperatures, with anomalous temperatures propagating particularly far through groundwater flow (Ferguson and Woodbury, 2004; Bidarmaghz et al., 2020). Such temperature anomalies can have knock-on effects, impacting cooling and heating requirements to maintain underground spaces at comfortable levels (particularly due to groundwater not transporting away as much heat (Blum et al., 2021)), groundwater quality, and the functioning of underground biospheres (Benz et al., 2015; Attard et al., 2016; Bayer et al., 2019; Krcmar et al., 2020).

Effective and equitable utilization of the shallow sub-surface as a space and energy resource requires numerical thermo-hydraulic modeling of the ground at city-scale, taking into account anthropogenic influences and how they interact with one another (Meng et al., 2019; Epting et al., 2020; Bidarmaghz et al., 2021). Large-scale numerical models have proven sufficiently reliable in capturing anthropogenic influences on the shallow subsurface. However, such models are computationally expensive and depend on measured data for calibration and parameter inference to yield reliable results. Inevitably, uncertainties arise within the modeling process, stemming from various sources such as noise in field measurements, simplifications adopted to make computational models tractable, and parametric uncertainty. Accounting for such uncertainties is important for the reliability, reproducibility, and interpretability of model outputs (Volodina and Challenor, 2021). Bayesian frameworks allow for the quantification of uncertainty and its incorporation into the calibration process, as well as accounting for prior beliefs about the system that is modeled (Kaipio and Somersalo, 2007).

While Bayesian calibration methods are used in several fields, for example, engineering mechanics (Rappel et al., 2020; Pepi et al., 2020), natural hazards engineering (Zheng et al., 2021), building energy modeling (Hou et al., 2021), and ecological and environmental modeling (Speich et al., 2021), they are not widely utilized in the context of large-scale subsurface models for shallow geothermal applications, and their performance, when applied to field data, has not been tested fully. Case studies utilizing such methods on groundwater models and deep geothermal reservoirs exist (Cui et al., 2019; Omagbon et al., 2021; Scott et al., 2022), however, the complexities present in the shallow subsurface, such as thermal influence from the surface as well as from anthropogenic infrastructure and activity, as well as the contrast in scale and magnitude make the problem context significantly different. Moreover, compared to synthetic data (i.e., data generated numerically or under controlled conditions) calibration on field data poses a number of non-trivial challenges: (a) observations are expensive to obtain and hence scarce, (b) while the spread of observations should capture meaningful variations for robust inference, these are often not easily tractable (even when data is available), leading to a lower content of information compared to synthetic data, and (c) observation noise and measurement errors introduce additional elements of uncertainty to the calibration process.

The scarcity of measurements is of particular relevance to the calibration of the subsurface models due to the cost and labor intensity of observations within the ground, requiring the drilling of boreholes and the deployment of measurement equipment (Nicholson et al., 2020). Hence data is often sparse relative to the complexity of the system. Bayesian techniques are particularly useful as they include probabilistic distributions for prior knowledge and measurements, thereby accounting explicitly for the lack of knowledge about the system being modeled (Omlin and Reichert, 1999). For example, Cui et al.

https://doi.org/10.1017/dce.2022.32 Published online by Cambridge University Press

(2011) estimate the posterior distribution for parameters used in a deep geothermal reservoir model using a two-stage Markov chain Monte Carlo (MCMC) sampling scheme, where the two stages consist of running the MCMC samples through a surrogate model, and subsequently running samples producing acceptable results through the original model. A possible issue with such an approach, however, is that it often requires a large number of forward runs of the original model which can be computationally expensive. An alternative methodology is proposed by Menberg et al. (2020), based on the work by Kennedy and O'Hagan (2001) and Goh et al. (2013). It employs models of two different levels of fidelity (i.e., mesh resolutions) to generate a large set of numerical data to train Gaussian Processes (GP) emulators for use in the calibration. The use of multi-fidelity approaches has also been suggested in a hierarchical framework by Maclaren et al. (2020), that incorporates posterior-informed approximation errors and has been tested with deep geothermal applications. Adopting GP emulators is also used successfully by Cui et al., (2018) in regional groundwater modeling as well as by Rajabi and Ketabchi (2017) in coastal groundwater management applications. A limitation of the training of GP models is the required run-time complexity of $\mathcal{O}(N^3)$, where N is the sum of the number of field observations and the number of numerical outputs (Chong et al., 2017). That is to say, large data sets can make calibration infeasible if the process is prohibitively costly. Therefore, it becomes crucial to assess how much individual data-points contribute to parameter inference. It calls for a process by which points that provide the most information can be identified and curated as a manageable but high-value data set. For subsurface models covering large areas, this in turn requires the identification of critical locations where measurements can be improved and uncertainty in parameter posteriors reduced (Nicholson et al., 2020).

A further compounding factor, limiting the ability of parameter inference for spatially large models, is the issue of non-identifiability for parameters that vary spatially in large regions where measured quantities in the ground tend to be highly local due to this heterogeneity. Perego et al. (2016) consider the impact of heterogeneity in soil profile on GSHP efficiency for a system installed within the town of Alessandria, Italy. A (deterministic) 3D heat transport modeling suite, GeoSIAM (Integrated System for GeoModelling Analyses), was used to model the system and validated against reference cases. The authors find that the assumption of homogeneity and averaged thermal properties can underestimate the thermal impact on the surrounding soil, leading to greater thermal imbalances. This issue also may be addressed through Bayesian methods in that the resulting inferred parameter values are distributions, rather than single-valued estimates (Omlin and Reichert, 1999). However, it is useful to understand the degree to which the distributions inferred from data represent the spatial variations of parameters across a given region. Characterization of heterogeneity and its impact on inferred parameters using Bayesian methods is a growing topic of interest in hydrogeology (Zhou et al., 2014; Linde et al., 2017). For example, Painter et al. (2007) demonstrate the ability of Bayesian methods to out-perform deterministic approaches in inferring a hydraulic conductivity field for a numerical model of the highly heterogeneous Edwards aquifer, Texas, USA. The authors compare the performance of simple interpolation and of numerical upscaling combined with cokriging against a revision to the hydraulic conductivity field using Bayesian updating, finding that the latter greatly improved the forward groundwater model, reducing the mean residual error by a factor of 10.

To address the challenges highlighted above, we apply a multi-fidelity Bayesian calibration approach using field data measured across locations in the city center of Cardiff, UK. The calibration methodology is also applied with synthetic, numerically generated data. The synthetic data enables us to have confidence in the setup of the calibration process and allows control of the conditions under which data for calibration is produced. Additionally, comparison between posteriors inferred from field data and those from synthetic data permits exploration of the additional complexities contained within field data and provides a measure for the increased uncertainty arising from field data. Our first objective is to investigate and identify a subset of high-value data points, thus addressing the issues of data scarcity and information gain. We do so by performing the calibration process with subsets of the available data and analyzing how the posteriors of the different parameters are affected by the size of the data subsets and by the physical phenomena experienced by the measurement points included in the subset. Our second objective is to test the ability of the calibration methodology to infer reasonable distributions for spatially

https://doi.org/10.1017/dce.2022.32 Published online by Cambridge University Press

heterogeneous parameters. We do so by numerically creating synthetic scenarios where the aquifer hydraulic conductivity is varied and examining the parameters inferred using these synthetic data sets. Finally, we demonstrate the importance of model calibration and uncertainty quantification by propagating the uncertainties in the inferred parameters to the design of an existing open-loop Ground Source Heat Pump system within the modeled area. In sum, the novelty of the paper lies in demonstrating data-efficient calibration of large-scale ground thermal models with field data and the application of the information gained on geothermal applications, thus contributing to the on-going efforts towards utilizing our environment's resources in efficient manners.

2. Methodology

2.1. Bayesian multi-fidelity framework formulation

We employ a multi-fidelity approach developed by Menberg et al. (2020), which extends the Kennedy and O'Hagen "single-fidelity" Bayesian framework (Kennedy and O'Hagan, 2001) (hereafter KOH) by including outputs from numerical models at both high and low levels of fidelity for the calibration. The advantage of this method is that the low-fidelity model may be run at a lower computational cost than the high-fidelity one, albeit with lower accuracy, and thereby more frequently, providing a greater number of numerical data points for calibration. The multi-fidelity framework relates field observations y_f to simulation outputs y_c (ultimately re-expressed in terms of the low-fidelity model output term η_l and a term representing the mismatch between output from the high- and the low-fidelity models, μ) and a discrepancy function δ, for a given set of state variables x, as

$$\begin{aligned} y_f(x) &= \zeta(x) + \varepsilon \\ &= y_c(x, \theta) + \delta(x) + \varepsilon \\ &= \eta_l(x, \theta_h, \theta_l) + \mu(x, \theta_h, \theta_l) + \delta(x) + \varepsilon, \end{aligned} \tag{1}$$

where θ_h and θ_l are the sets of unknown model parameters of the high-fidelity and the low-fidelity models, respectively. $\zeta(x)$ represents the true, non-observable, physical process and ε the random measurement error. GP models are used for the approximation of the terms in equation (1), namely (η_l, μ, δ), with covariance matrices adapted from the KOH approach accordingly, shown in Supplementary Appendix A.

The multi-fidelity framework is implemented using the STAN programming language, employing a Hamiltonian Monte Carlo (HMC) algorithm to sample the posterior distributions. Four independent sampling chains are run during a calibration process, with 2000 samples each, the first 1000 being discarded as part of the burn-in phase. The outcome of a calibration consists of posterior distributions of the model parameters θ, as well as the hyper-parameters, which define the range of prior uncertainty of the parameters. While some methods for Bayesian inference, such as the Bayesian approximation error (BAE) approach (e.g., Maclaren et al. (2020)), use empirical estimates for approximation of the posterior distributions, our approach follows the concept of the KOH framework. Here, the posteriors distributions of all unknown (hyper-) parameters are inferred simultaneously by HMC sampling. However, it should be noted that due to limitations in the number of samples our posterior are approximations of the true distributions, and indeed several studies advise caution when interpreting the posteriors of error terms from the KOH framework quantitatively (Li et al., 2016; Menberg et al., 2019). In fact, even with an infinite number of samples, the use of GP emulators would result in an approximation of the true posterior. The role of the hyper-parameters in the GP models and the calibration process is further explained in Supplementary Appendix 6. Further information on Bayesian inference can be found in works such as Gelman et al. (2013), Chong and Menberg (2018), Choi et al. (2018), and Menberg et al. (2019).

2.2. Description of the domain and field data

The city center of Cardiff, UK, has been monitored extensively by the British Geological Survey (BGS) and the Cardiff Harbour Authority over 5–20 years, for geological and hydrological data (Williams, 2008; Heathcote et al., 1997). The data include soil profiles and geological characteristics of the area (Kendall

https://doi.org/10.1017/dce.2022.32 Published online by Cambridge University Press

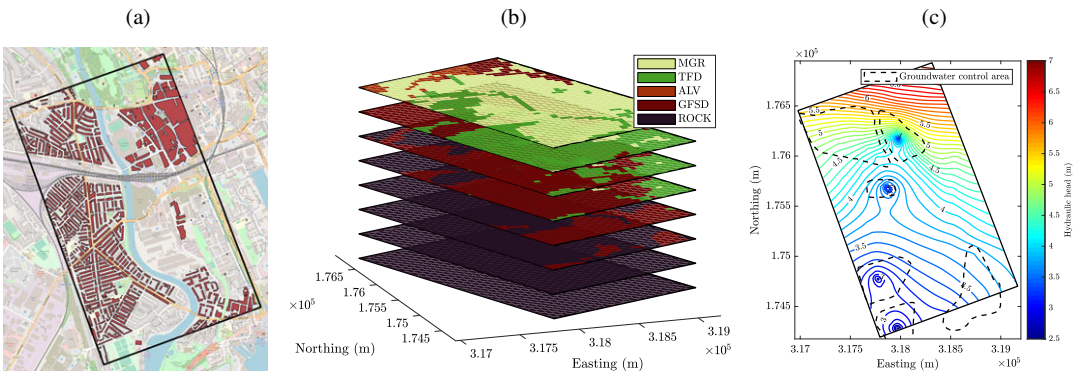

Figure 1. *Outlines of buildings with basements (a), geological distribution (b), and hydraulic head distribution (c) within the study domain. Groundwater levels are pumped within dashed regions (copyright BGS, UKRI). © Crown copyright and database rights 2021 Ordnance Survey [100021290 EUL]. Use of this data is subject to terms and conditions.*

et al., 2020; Patton et al., 2020), hydrological information such as groundwater level and hydraulic head measurements, as well as temperature time series measurements from a number of monitoring boreholes within the area (Farr et al., 2019,0).

The domain considered in this study consists of a rectangle of approximately 3.5 km^2 in area, in the southern part of the city, illustrated in Figure 1a where footprints of buildings likely to be with basements are shown. The geological layers down to a depth of 45 m below ground level are shown in Figure 1b, and relevant properties of the materials are listed in Table 1. Below a depth of approximately $z = -25$ m, the subsurface consists almost exclusively of Mercia Mudstone bedrock (ROCK), which is considered to be the base of the glaciofluvial sand and gravel deposits (GFSD) that constitute the aquifer, while the shallower layers consist of made ground (MGR), tidal flat deposits (TFD), and alluvium (ALV). An annual mean hydraulic head distribution of the area, shown in Figure 1c, is generated from a detailed hydrogeological model developed by the BGS and includes pumped groundwater control areas, mitigating rise in groundwater levels (Williams, 2008). The average groundwater level in this domain is found to be 4 m below the ground surface.

The temperature monitoring boreholes at 24 locations scattered throughout the domain, at depths ranging from 1.5 to 12.5 m below the ground surface are shown in Figure 2a (see Table 2 in Patton et al. (2020) for co-ordinates of sensors). A linearly increasing sinusoidal function is fitted to each of the measured temperature time series. The sinusoidal variation observed in the temperatures is a consequence of the seasonal temperature fluctuation at the ground surface, and so the oscillation period is set to 365 days. Key features, such as the mean temperature T_{mean}, amplitude T_{amp}, annual temperature shift T_{inc}, and phase ϕ of the temperature oscillations, are extracted. This provides a suitable framework for data management, allowing the use of a large data set in a lower dimensional space by removing the temporal dimension, and enabling efficient utilization of multiple points in the calibration process. Figure 2b,c illustrates the fitting applied to data measured at two of the boreholes.

2.3. Numerical modeling

The domain within Cardiff is modeled using finite element methods and is implemented within the software suite COMSOL Multiphysics (COMSOL ® v. 5.6, 2020). A semi-3D numerical modeling approach (Bidarmaghz et al., 2019, 2020; Makasis et al., 2021) is utilized to model the subsurface response to thermal and hydraulic phenomena. This approach entails modeling a collection of horizontal planes using the governing equations for conductive and convective heat transfer and for fluid flow through porous media in 2D, and thermally coupling neighboring planes via conduction. A key aspect of

https://doi.org/10.1017/dce.2022.32 Published online by Cambridge University Press

Table 1. *Thermal and hydraulic properties of geological materials present in the modeled domain and concrete material used for heat sources (SimonHydrotechnica, 1993; Howard et al., 2008; Dalla Santa et al., 2020; Hobbs et al., 2002; Parkes et al., 2020; Boon et al., 2021). Thermal diffusivity was calculated according to* $\alpha = \lambda / (\rho C_p)$. *Where appropriate, top values in a row represent partially saturated conditions and bottom values fully saturated conditions.*

Geology/material	Average depth range (m)	Thermal conductivity, λ (W/(m K))	Density, ρ (Mg/m^3)	Specific heat capacity, C_p (kJ/(kg K))	Effective porosity, ε (−)	Hydraulic conductivity, k_h (m/s)	Thermal diffusivity, α (m^2/s)
Made ground, MGR	0–2.3	1.00[a]	1.80	1.27	0.35	2.31×10^{-5}	4.37×10^{-7}
		2.00[a]					8.75×10^{-7}
Alluvium, ALV	0.8–2.7	1.40	1.67	1.18	0.35	1.00×10^{-5}	7.10×10^{-7}
		2.40					1.21×10^{-6}
Tidal flat deposits, TFD	2.0–6.7	1.20	1.67	1.18	0.2	1.00×10^{-8}	6.09×10^{-7}
		1.50					7.61×10^{-7}
Glacio-fluvial sediment deposits, GFSD	5.0–10.9	0.50[a]	2.00	1.75	0.2	2.5×10^{-3} [a]	1.43×10^{-7}
		1.80[a]					5.14×10^{-7}
Bedrock (Mercia Mudstone), ROCK	10.9–45.0	1.10[a]	2.01	0.80	0.25	1.00×10^{-7}	6.84×10^{-7}
		1.80[a,b]	2.10				1.07×10^{-6}

[a]Values further explored in sensitivity analysis and uncertainty quantification, Sections 2.4 and 3.

[b]This value is a conservative as recent studies by Boon et al. (2021) suggest that the thermal conductivity for fully-saturated dolomitic mudstone could be as high as 2.7 W/(m K).

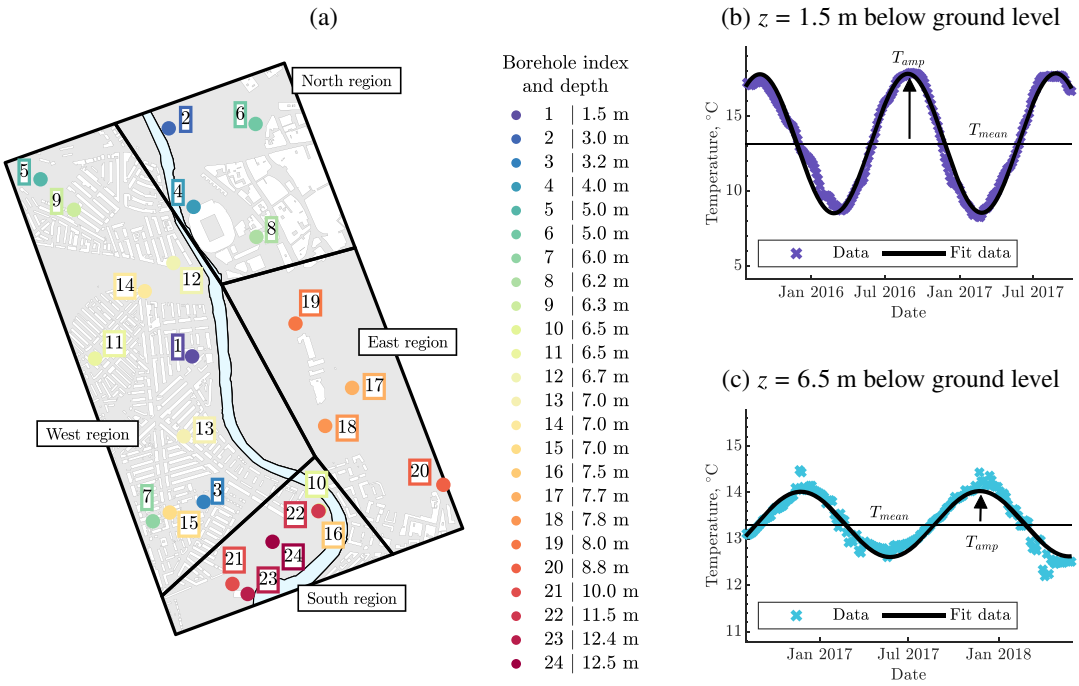

Figure 2. *Measurement locations in domain with color denoting depth of sensor (left panel) and examples of data fitting (right two panels). © Crown copyright and database rights 2021 Ordnance Survey [100021290 EUL]. Use of this data is subject to terms and conditions.*

Table 2. *Ranking of uncertain parameters considered in the model according to mean temperature sensitivity.*

Rank	Parameter	Symbol	Unit	Standard value		Minimum		Maximum	
1	Far-field ground temperature	T_{ground}	°C	12.9		11.7		14.7	
2	Basement temperature	T_{room}	°C	18.0		13.0		22.0	
3	Basement percentage	B_{perc}	%	60		0		100	
4	Aquifer (GFSD) hydraulic conductivity	$k_{h,\text{GFSD}}$	m/s	2.5×10^{-3}		1.0×10^{-4}		1.0×10^{-2}	
				Unsat.	Sat.	Unsat.	Sat.	Unsat.	Sat.
5	Thermal conductivity of made ground	λ_{MGR}	W/(m K)	1.0	2.0	0.3	0.9	1.2	4.0
6	Thermal conductivity of the bedrock	λ_{ROCK}	W/(m K)	1.1	1.8	0.8	1.6	2.0	3.6
7	Thermal conductivity of aquifer	λ_{GFSD}	W/(m K)	0.5	1.8	0.4	1.6	0.5	2.5
8	Vegetation coefficient	k_v	n/a	1.2		0.9		1.4	

the modeling is the inclusion of heated basements to assess their anthropogenic influence on the thermal state of the ground, as well as the incorporation of the river Taff in Cardiff, modeled using turbulent flow governing equations. Detailed information on the numerical modeling methodology, including governing equations, initial and boundary conditions, and a labeled schematic of the model set-up can be found in Supplementary Appendix B. The methodology has been validated against more complex detailed 3D

https://doi.org/10.1017/dce.2022.32 Published online by Cambridge University Press

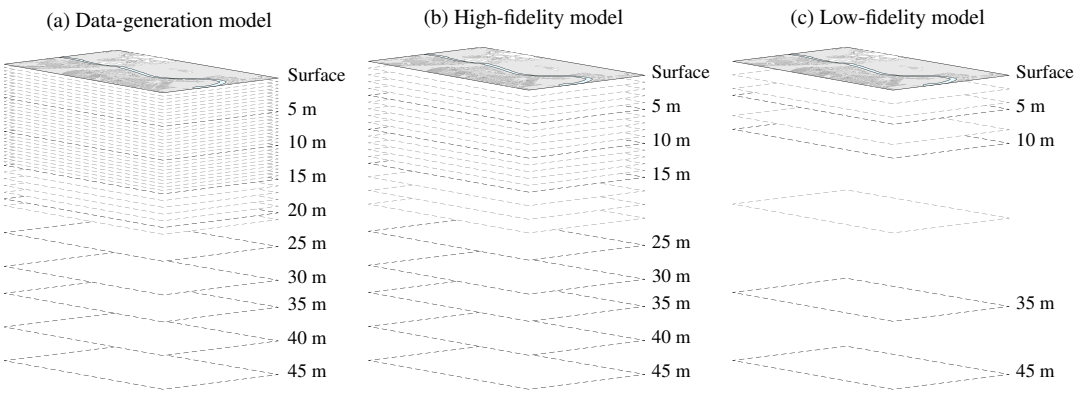

Figure 3. *Schematic of the three models with different levels of fidelity (number of modeling planes). The 2D planes are connected to their nearest neighbor by convective heat flux transfer and temperature boundary conditions are applied at the top- and bottom-most planes.*

numerical simulations which have, in turn, been validated against both experimental and numerical data (Bidarmaghz and Narsilio, 2018; Bidarmaghz et al., 2020; Makasis et al., 2020).

For the purposes of the Bayesian calibration utilizing the multi-fidelity approach defined in Section 2.1, two numerical models with different levels of fidelity are required. In addition to these, a model with very high fidelity is created to generate output data of controllable scenarios (i.e., both input and output values are known), hereafter called the data-generating model and its output synthetic data. No noise was added to the synthetically-generated training data. These allow the systematic exploration of the heterogeneity across the domain on parameter inference, to understand at what degree this methodology is applicable under conditions with heterogeneity (which is expected in large-scale modeling of the ground). The schematics for these three models are shown in Figure 3. The number of planes implemented in the models defines the fidelity of the models and determines computational costs: the data-generating model with the highest resolution (41 layers) requires about 15 hr to run; the high-fidelity model (23 layers) requires approximately 5 hr to run; and the low-fidelity model (8 layers) requires less than 30 min for one computational run. All three models implement a higher layer density near the surface, with planes spaced more closely together (i.e., smaller dz). This is to better resolve the topmost layers of the domain, where more pronounced geological variation and the presence of heat sources produce a more complex convective flow pattern, compared to deeper underground where the soil conditions are more homogeneous, there is a lack of heat sources, and the presence of the mostly impermeable bedrock diminishes convection. Each plane is meshed with 38,415 triangular elements, irrespective of the model fidelity. Simulations were run on high-performance computers with two 2.1 GHz processors (24 cores) and 192 GB of RAM operating at 2400 MHz.

2.4. Parameter screening and calibration set-up

The computational cost of the calibration process (both in terms of the cost taken to perform necessary model runs as well as the time taken for the calibration process itself) necessitates parameter screening to identify parameters with the greatest impact on subsurface temperatures. Limiting the calibration to only the most influential parameters also addresses the possibility of over-fitting (Chong and Menberg, 2018). To this end, the Morris Method (Morris, 1991; Menberg et al., 2016; Chong and Menberg, 2018) is used to identify the model parameters with the greatest influence on the model output of interest, that is, the mean temperature T_{mean} at a given location. The Morris Method is a global sensitivity analysis whereby the input parameter space is non-dimensionalized onto a unit hyper-cube H^k (for a set of k parameters), and each dimension i is divided into a regular grid of spacing Δ_i with a total of p levels along each dimension. A series of t trajectories is defined within H^k, each beginning at a randomly chosen point in the parameter

https://doi.org/10.1017/dce.2022.32 Published online by Cambridge University Press

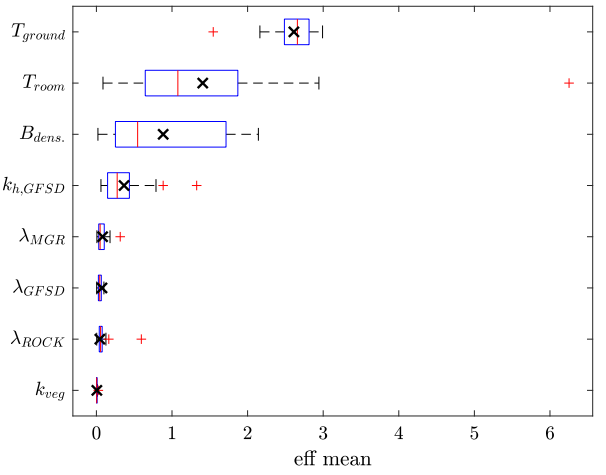

Figure 4. *Results of the Morris method sensitivity analysis of parameter impact on mean temperature, shown as box plots of the effective mean across the 24 measurement locations and indicating the average value as a black cross. The higher the effective mean, the more sensitive the model output is to changes in this parameter, giving it a higher ranking.*

space with the following points differing from the previous one by changing a single parameter by one increment at a time for a total sequence of $k+1$. For a set of $t=10$ trajectories with $k=8$ parameters (see below), then, this yields a total of 90 model runs. The results from these runs are used to determine the effective mean for each parameter, a measure of the magnitude of influence of a parameter X on model output Y, which is given by

$$\mu_i^* = \frac{1}{2}\sum_{t=1}^{r}\left|\frac{Y(X+e_i\Delta_i)-Y(X))}{\Delta_i}\right|. \tag{2}$$

This can then be used to rank parameters according to the sensitivity of model output on parameter variation.

The parameters considered for screening are listed in Table 2 along with their plausible ranges, and the ranking of the parameters according to their effective mean values. The ranking is determined by taking the average of the effective mean calculated across the 24 measurement locations considered in the domain. Results from this analysis are illustrated in Figure 4, showing the effective mean of each parameter as well as the variation in this metric across the 24 locations. The most influential parameter is seen to be the far-field ground temperature T_{ground}, which is unsurprising as it affects most directly the mean temperature at a location, itself being the mean temperature of the ground in absence of any external temperature fluctuations. The sensitivity of the model output to T_{ground} is fairly consistent across the 24 measurement locations in the domain, with the percentiles of the box-plot nestled closely around the average value.

The next two most influential parameters are the temperature to which the basements are heated T_{room} and the percentage of basements in the domain that are heated B_{perc}. This percentage is varied in the sensitivity analysis by "activating" (i.e., applying the temperature boundary condition of $T=T_{\text{room}}$) to varying fractions of the basements shown in Figure 1. The parameters exhibit a signification variation in sensitivity across the measurement points, confirming that heat fluxes from any individual basement have a highly localized effect on ground temperatures. Despite being explored independently in the sensitivity study, the two parameters are closely linked in terms of how they impact the mean temperature of the surrounding ground. Because the basement percentage is a parameter with a more localized impact on mean temperature than the basement temperature, T_{room} is chosen as the parameter for calibration, while

https://doi.org/10.1017/dce.2022.32 Published online by Cambridge University Press

keeping the percentage of basements active constant at 60%, following initial investigations where both B_{perc} and T_{room} were varied.

The third-highest ranking parameter is the hydraulic conductivity of the glacio-fluvial sediment deposits, that is, the aquifer. The hydraulic conductivity affects the groundwater flow velocity and thereby how far groundwater heated from nearby basements flows before returning to the ambient ground (i.e., far-field) temperature, hence indirectly impacting the mean temperature measured in the vicinity of basements. Slower groundwater flow disperses less of the heat generated by basements, leading to the accumulation of heat beneath heated infrastructures in the ground.

The three parameters to be estimated are therefore the far-field ground temperature, the basement percentage, and the aquifer hydraulic conductivity. Being a Bayesian approach, the calibration requires prior estimates of the model parameters (and hyper-parameters). For the three model parameters θ, the priors span the entire plausible range of the parameter, defined through literature review and local expertise, and follow a normal distribution of $\mathcal{N}(\mu = 0.5, \sigma = 0.2)$. (The priors of the hyper-parameters follow Menberg et al. (2020) and are given in Supplementary Table 4.) Latin Hypercube Sampling (LHS) is used to determine the combinations of parameter inputs for the runs of the numerical model, and the model outputs are used to train the statistical emulator. The high-fidelity model is sampled 20 times and the low-fidelity model, with its lower computational cost, is sampled 50 times. It should be noted that, for the data-generating model, defined values are used for the three uncertain parameters. Namely a far-field ground temperature of $T_{\text{ground}} = 13.3$ °C, hydraulic conductivity of the aquifer of $k_{h,\text{GFSD}} = 2 \times 10^{-3}$ m/s, and a temperature of the anthropogenic heat sources of $T_{\text{room}} = 15$ °C.

Mean temperature values are shown in Figure 5 for the synthetic data (black crosses) alongside the field data (black asterisks). Also shown is the output from the 20 high-fidelity (colored pluses) and 50 low-fidelity model runs (colored circles), based on LHS covering the expected parameter range. Both the synthetic and the field data fall well within the range of outputs from the high- and low-fidelity models, providing confidence in the chosen range of prior parameter estimates.

3. Calibration Results

The posterior parameter distributions inferred from calibration on both the synthetic and the field data are shown in Figure 6. The $\widehat{R}$ criterion (Gelman and Rubin, 1992), which compares the inner and inter-chain variance of the posterior samples is applied for convergence diagnostics and approaches unity (i.e., < 1.01 (Vehtari et al., 2021)) within the 1000 samples (after burn-in) of the HMC algorithm for all parameters

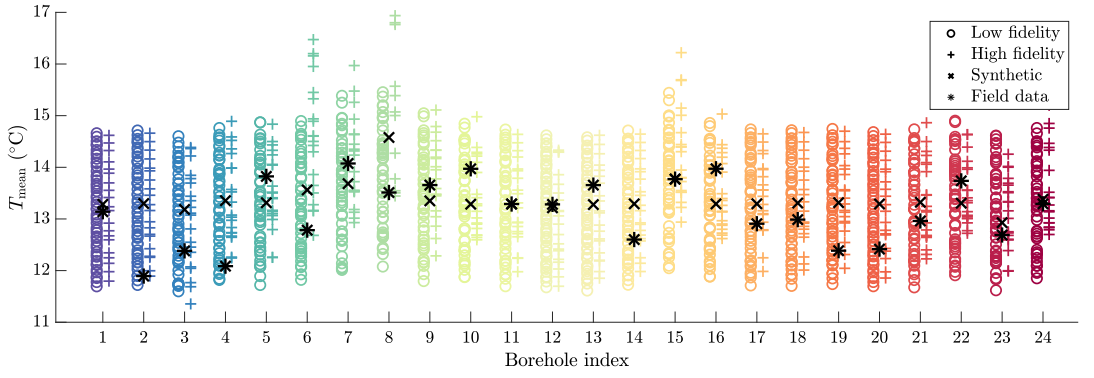

Figure 5. *Mean temperature data determined from fitting the temperature output from the data-generating model (black crosses) and the field data temperature time-series (black asterisks) across the 24 measurement locations. Also shown is the output from the 20 LHS samples run using the high-fidelity model (colored pluses) and the 50 LHS samples using the low-fidelity model. The numerical model runs capture the extent of the measured and synthetic data well.*

https://doi.org/10.1017/dce.2022.32 Published online by Cambridge University Press

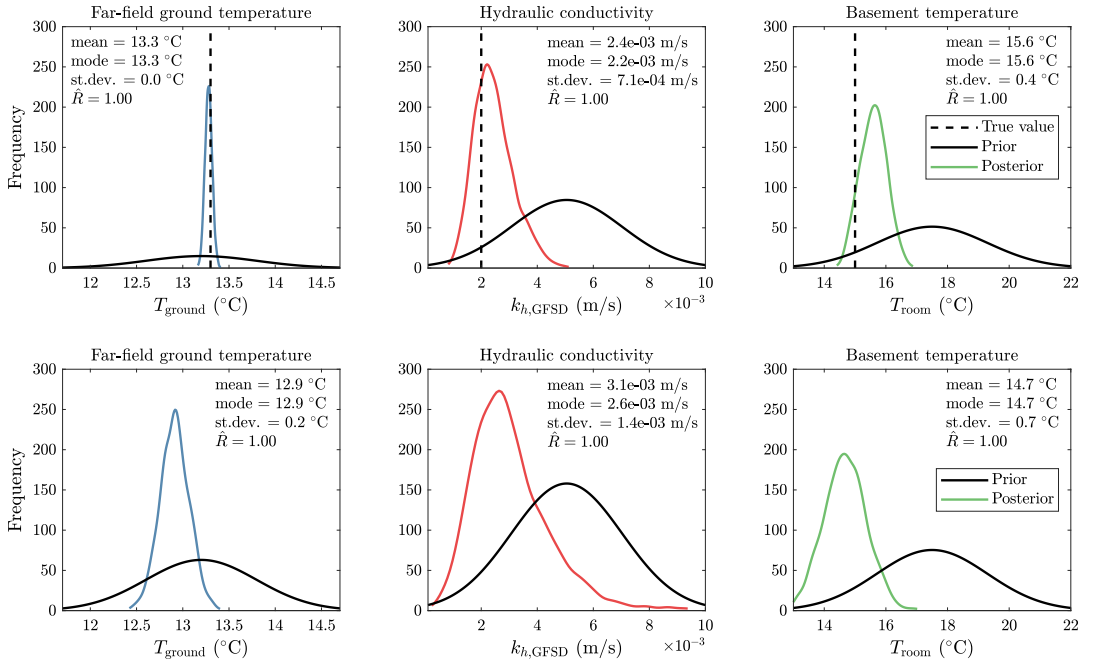

Figure 6. *Parameter posterior and prior distributions for parameter values from calibration using numerically-generated (top row) and field (bottom row) temperature data sets. The posteriors for the synthetic data set show good narrowing around the input value (indicated by the vertical dashed line) and the posteriors inferred from the field measurements exhibit a reduction in uncertainty.*

shown. The Effective Sample Size (ESS) is a further metric related to convergence and is computed to determine sample independence, included in Supplementary Appendix C. For the synthetic data (top row) values used as parameter input in the model, that is, the values that are to be inferred, are indicated by a dashed vertical line. The calibration performs very well for the far-field ground temperature (left-most panel), with both the mean and mode values agreeing with the input of T_{ground}=13.3°C to within two decimal places, and with a very narrow distribution around this value reflecting the high degree of confidence, mirroring the observations from the sensitivity analysis of the impact of far-field ground temperature on local soil temperature. The hydraulic conductivity and basement temperature (central and right-most panels) are also inferred well, with the posteriors exhibiting narrower uncertainty compared to the prior distributions, albeit with less precision and accuracy than for the far-field ground temperature. The aquifer hydraulic conductivity posterior overestimates the input value by about 0.4×10^{-3} m/s, while the posterior for the basement temperature by 0.6 °C. These results provide further confidence in the methodology as well as indicating the amount of information the 24 monitoring boreholes can provide for the three calibration parameters.

For calibration using field data, shown in the bottom row of Figure 6, the resulting mean and modal values for T_{ground} agree well with deep borehole measurements taken in the region (12.9 °C for the far-field ground temperature (Farr et al., 2017)), providing further confidence in the results. The posterior for the basement temperature shows lower values than expected, indicating that basement structures in the city center of Cardiff are only about two degrees higher than the temperature of the ground. The hydraulic conductivity posteriors indicate a value of about 2.9×10^{-3} m/s, which is higher than found in the literature or measured from previous work in the area, but still within the plausible range for the type of material in the area. The observed differences in uncertainty between the two rows of inferred parameter posteriors (i.e., standard deviation values) are to be expected as calibration on data from field observations is subject to noise, model error, and further complexities not present in the numerically-generated data

https://doi.org/10.1017/dce.2022.32 Published online by Cambridge University Press

calibration. The resulting precision hyper-parameter distributions for the field data calibration are included in Appendix C.

3.1 Impact of subset size

The influence of data set volume on parameter inference is examined by incrementally reducing the size of data subset from the 24 locations. In addition to the size of the data set used for calibration, the amount of information contained in the data from the points included in the data set determines the ability of the process to infer the parameter values. To determine which points in the data set provide the most information, calibration would need to be performed on all possible combinations of subsets of a given size. This is constrained, however, by the prohibitively high computational costs of the calibrations, particularly for large subset sizes as the matrix for inversion grows with the number of data points, and is outwith the scope of this work. Subsets are therefore selected systematically, whereby measurement points are selected so as to maximize the variety of points in the subset, in terms of the spatial coordinates (x, y, z) and to cover the largest amount of volume possible, thereby maximizing mutual information spatially. This is based on the assumption that points that are in close proximity to one another are more likely to experience similar conditions and not provide complementary information. For this, the measurement locations are clustered spatially to find points that are closest together, as illustrated in the dendrogram in Figure 7a. Pair-wise comparison is performed on points nearest to one another, retaining only the most dissimilar points in the subset, decrementing from the maximum of 24 points. This includes approximately maintaining the ratio between the number of data points within and outwith the aquifer, noted with blue and orange colored labels respectively, at each step of the reduction. For example, the subset containing 16 points is shown in Figure 7b, resulting from having removed eight points, eliminating points with indices (10,14,15,18,24,9,22,7), of which five are located in the aquifer, and three outside. It is worth noting that this approach of determining successive measurement locations is a naive approach to maximize mutual information. The purpose of this exercise is to demonstrate the impact of using different data locations on the inferred posterior distributions rather than optimizing the locations of

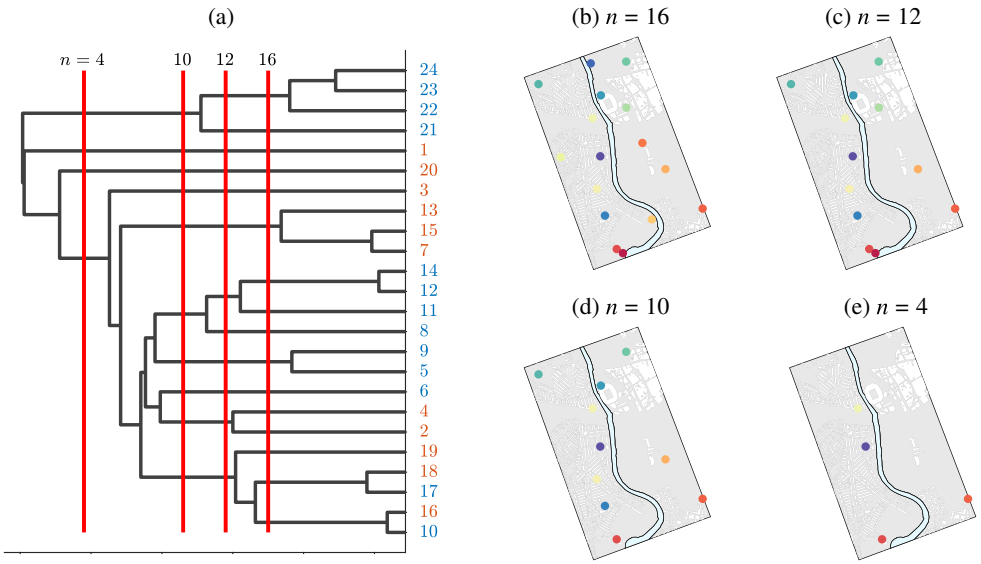

Figure 7. *Dendrogram indicating spatial proximity of measurement points. Borehole indices colored blue indicate points located within the aquifer, and orange colored indicate points outwith the aquifer. The indices of the points removed upon each reduction in subset side, in order are 10, 14, 15, 18, 24, 9, 22, 7, 16, 11, 19, 2, 8, 23, 6, 5, 4, 17, 13, 3.*

https://doi.org/10.1017/dce.2022.32 Published online by Cambridge University Press

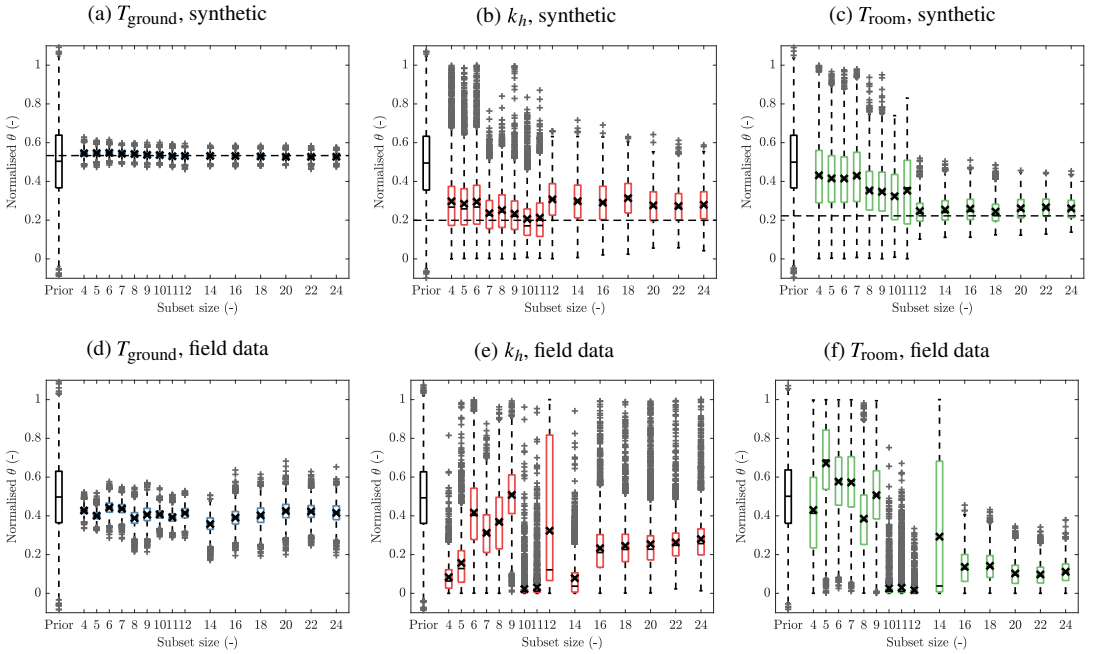

Figure 8. *Box plots showing the progression of parameter calibration with increasing subset size for the three calibration parameters (left to right) for the single-valued synthetic data (top row) and the field data collected from boreholes (bottom row). The mean distribution values are marked with "x" for each case. The posteriors exhibit a narrowing around the input value with increasing data set size for the synthetic data and more clearly defined distribution behavior for the field-data.*

measurements for this example of a (non-linear) Bayesian inverse problem for which work exists (e.g., Huan and Marzouk, 2013; Alexanderian et al., 2016; Alexanderian, 2021).

Results of calibrations performed on the selected subsets are shown in Figure 8. The plots show the change of the posterior distributions for each of the θ parameters with a change in subset size. This illustration choice enables understanding at what size the data set becomes sufficient for the posterior distribution found from calibration on the subset to become equivalent to that of the calibration performed on the entire set. The top row (panels a, b, c) shows results from application to synthetic data, with the "true" value used as input in the data-generating model shown as black dashed lines. While the posterior distribution for T_{ground} remains mostly unchanged, there exist notable changes in the other two parameter posteriors as the subset size increases. For example, at a size of 12 points the distributions for the parameters narrow, indicating a significant contribution of information form the 12th data point. This point, labeled "8" in Figure 2a, is both within the aquifer and close to basements, likely making it a good source of information for the calibration algorithm. For calibration on synthetic data, as few as 10 data points (shown in Figure 7d) can be used to predict the combined three parameters reasonably confidently. However, the uncertainty is reduced even further with more than 12 data points, especially for T_{room}. On the other hand, it is also interesting to note that the value for k_h is better predicted with a subset size of 10 rather than any other size (despite the high number of outliers). This suggests that using more data points could, in some cases, reduce the predictive accuracy for certain parameters and thus data points should be carefully selected.

The bottom row of Figure 8 (panels d, e, f) shows results for calibration on field data. Compared with results from synthetically created scenarios, these results are less well-behaved, with more uncertainty persisting throughout the different set sizes. The inferred ground temperature T_{ground} exhibits a greater variance in the posteriors and the mean value of individual calibrations varies more compared to the

https://doi.org/10.1017/dce.2022.32 Published online by Cambridge University Press

posterior calibrated on the synthetic data, reinforcing that real data stem from systems with greater complexity, heterogeneity, and/or may be affected by measurement noise and external factors and heat sources not accounted for within the numerical modeling. The posteriors for k_h and T_{room} exhibit inconclusive outcomes for the smaller subsets, with unchanged distributions (same as the prior) or bi-modality where the algorithm identifies multiple possible solutions. At a data set size of 16 (points shown in Figure 7b) the posterior distributions appear to start converging onto a more defined value, narrowing and moving away from the extremal ends of the calibration range. Therefore, even though there is a reduction in uncertainty with increasing subset size, the 16 data points are likely sufficient to predict reasonable values for the parameters. The irregular pattern with which the posterior distributions change with increasing subset size also suggests that different measurement locations/data points contain different levels of information and therefore differently influence the parameter posterior distributions.

The difference in behavior of the posteriors with increasing subset size is a result of the level of uncertainty and variation within the data used for calibration. The synthetic data is generated using a model with a single input parameter and without measurement noise such that there is less uncertainty than for parameter values inferred from the field data. As a result, additional data points from the synthetic data predominantly act to narrow the posterior distribution to a single value. In contrast, further information from additional data points from the field data can in fact contradict information from other data points, causing a significant change in the posterior upon updating using the new evidence.

To illustrate this further a small-scale study is conducted, for which calibrations are performed on 20 unique data subsets, consisting of 12 measurement locations. The borehole indices of the points contained in each subset are given in Supplementary Table 5 (in Supplementary Appendix D), and results, in terms of the posteriors inferred from calibrating the model on the data from the given subset of the measured field data, are shown in Figure 9. The significant variation in the parameter value distributions inferred from the different data sets indicates that the measurement locations provide different levels of information or even different information altogether. This is illustrated by the fact that the variation in the inferred far-field ground temperature, Figure 9a, is significantly less than for the other two parameters which are likely to be subject to much more local variations. The influence of measurement locations and volume of data sets on parameter inference is explored more in the subsequent section.

In terms of obtaining data for a site investigation, the findings of this section showcase the importance of identifying the additional information contained in different data points, as this could significantly reduce costs for obtaining data, for example, due to drilling unnecessary boreholes. It should be noted that this exercise is performed after the measurements were taken. The value of this *ex post* exploration of measurement locations is to illustrate the impact that data set size and measurement point distribution can have on parameter inference and the importance of identifying appropriate measurement locations which are often subject to constraints, particularly in densely built-up areas

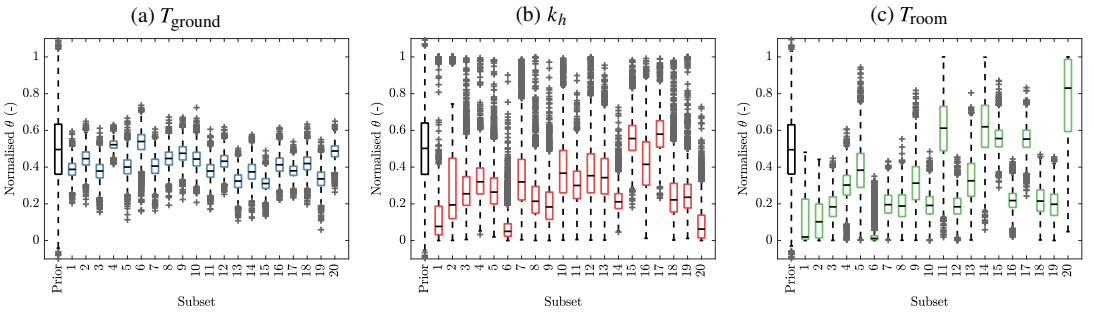

Figure 9. *Posteriors inferred from 20 unique field-data subsets consisting of 12 measurement locations. The variation in posterior distribution indicates that the different data points provide different information to the calibration.*

https://doi.org/10.1017/dce.2022.32 Published online by Cambridge University Press

where drilling locations are limited due to existing infrastructure and due to legal issues such as land-ownership, to minimize both cost as well as uncertainty.

3.2. Heterogeneity

The previous validation uses single values as parameter inputs in the data-generating model, assuming the homogeneous distribution of parameter values throughout the domain. However, that is rarely the case for many parameters in reality, where natural variability creates spatial heterogeneity in domain properties. In particular, hydro-geological properties of the subsurface, such as the hydraulic conductivity, are known to vary greatly (Rehfeldt et al., 1992). To explore the ability of the calibration methodology to infer model parameters in scenarios where heterogeneity is present, two test scenarios with spatially varying aquifer hydraulic conductivity values are considered, designated as "noisy" and "regional". The data used in these calibrations is generated using the very high-fidelity data-generating model, with spatially varying hydraulic conductivity values.

In the "noisy" scenario, hydraulic conductivity values change randomly within the aquifer following a normal distribution with mean $\mu_{k_h} = 7 \times 10^{-3}$ m/s and standard deviation $\sigma_{k_h} = 5 \times 10^{-4}$ m/s, while the far-field and basement temperatures were kept the same as for the single value case considered above. This represents a scenario where the aquifer permeability is subject to random heterogeneity as a result of variation in the compaction of the gravel and sand deposits making up the aquifer. This implementation of the hydraulic conductivity neglects spatial correlation, as is often incorporated in hydro-geological applications by means of a random field formulation. However, the purpose of our approach here is not to infer the spatial distribution of posterior hydraulic conductivity values, in favor of a simultaneous calibration of all parameters of interest, hence spatial correlation is not considered. It would be expected that the noise in the hydraulic conductivity would make it more difficult for the calibration to infer "correct" parameter values, giving rise to broader distributions and greater uncertainty in values. Results from the calibration using output data from this scenario are shown in Figure 10, with the calibration algorithm having achieved convergence for all three parameters, as indicated by the $\widehat{R}$ values. As for the single-valued case considered above, the far-field ground temperature and basement temperature are well predicted, with accurate and precise posterior distributions and mean and modal values very close to the ones used as data-generating model input. The basement temperature posterior is significantly more accurate compared to that of the single value calibration, suggesting that higher values for the ground hydraulic conductivity may reduce the uncertainty in T_{room} as this allows anthropogenically heated groundwater to propagate further, providing more information about heat sources to a greater number of measurement points. The hydraulic conductivity (central panel) is seen to be inferred reasonably well, despite the input values in the domain following a distribution. The posterior reduces in uncertainty compared to the prior, although not to a degree as to fully match the normal input distribution,

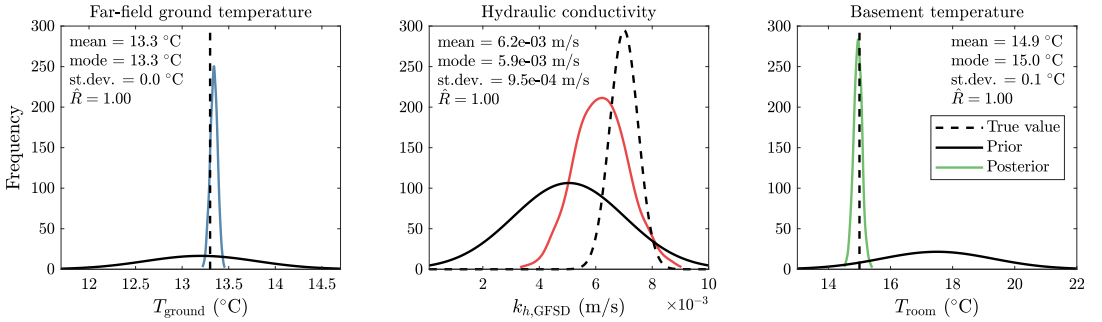

Figure 10. *Prior and posterior distributions for calibration performed on the "noisy" hydraulic conductivity output data set. The single-valued input parameters are well inferred and the posterior for the hydraulic conductivity approaches the input distribution.*

https://doi.org/10.1017/dce.2022.32 Published online by Cambridge University Press

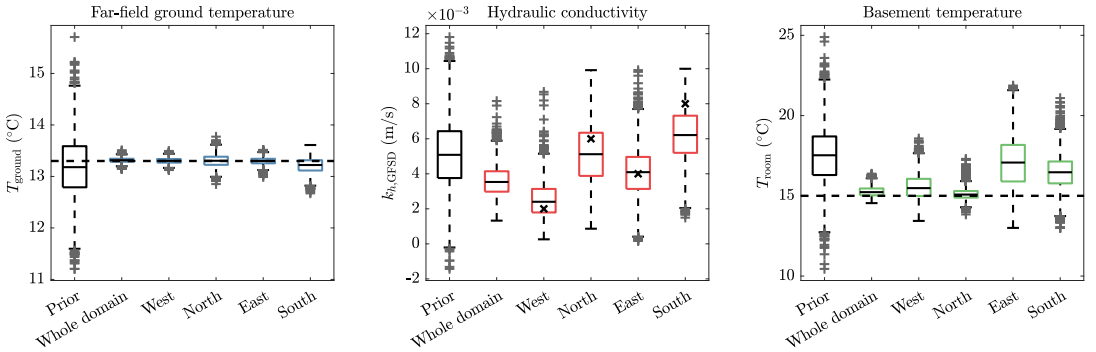

Figure 11. *Posteriors for regionally varying hydraulic conductivity synthetic data set, shown for calibration using regional data subsets. The dashed lines in the left and right-most panels indicate the "true" parameter value, used as an input in the data-generating model, and the black crosses in the central panel indicate the hydraulic conductivity values used in each of the four regions of the domain. The posteriors for the hydraulic conductivity approximate well the regional input values.*

underestimating the "true" mean value by about 0.8×10^{-3} m/s. Overall "noisy" heterogeneity, where the value of the parameter varies based on a (sufficiently narrow) distribution, may be considered to be captured relatively well by the calibration methodology.

In the "regional" scenario, the model domain is divided into four regions (shown in Figure 2a), each with a different hydraulic conductivity assigned to it. These regions are defined using natural and anthropogenic features of the domain: the western residential side of the river, densely populated with basements (10 data points), the northern commercial part (4 points), the eastern side with lower basement density (4 points), and the southern part which is close to the river outflow (6 points). The parameter posteriors inferred from calibration applied to data from this scenario are shown in Figure 11, with the "true" parameter value shown as either a line (if the plotted parameter is applied homogeneously at the domain-level) or as black markers (if the parameter is applied regionally). Results are shown for calibration performed on the entire data set, labeled "Whole domain", and for regional subsets, with data points within each of the four regions of different hydraulic conductivity labeled by the cardinal direction of the region. The calibration methodology continues to perform well for the far-field ground temperature (left-most panel) as this is a parameter with a common value across all locations and with a strong direct impact on the mean temperature. The basement temperature (right-most panel) is inferred well for the whole domain and the west and north regions, with data points close to basements, but less well for the regions with measurement locations further away from basements, illustrating the localized nature of the parameter. Moreover, the fact that the north and east regions contain as few as 4 data points could affect the information the calibration algorithm has to meaningfully infer parameter values. The hydraulic conductivity values are inferred well regionally, as may be seen in the central panel, where the calibrations performed using data points from the individual regions cluster the interquartile range around the input value that is to be inferred, with only the south region underestimating the true value. All of the calibrations converge, with $\widehat{R}$ values of 1.0, including the one for the whole domain (that is using the data points from all four regions within one calibration to obtain a single "representative" distribution for k_h), suggesting that, at the domain level the, posterior distribution could be used as a representative, capturing regional heterogeneity. To test this, the mean parameter values inferred from calibration on data points from the whole domain are used as input to a simulation of the numerical model, and the mean temperature values found at the measurement locations are compared with the data from the regional data-generating model. Results from this exercise are shown in Figure 12, showing the good agreement between the synthetic data used for the calibration ($T_{\text{mean, regional}}$, x-axis) and the output from the model using the mean parameter values inferred from the calibration ($T_{\text{mean, calibrated}}$, y-axis), with changes in mean temperature of $|\Delta T| = |T_{\text{mean, calibrated}} - T_{\text{mean, regional}}| \leq 0.2°\text{C}$. This suggests that, in cases where parameters vary

https://doi.org/10.1017/dce.2022.32 Published online by Cambridge University Press

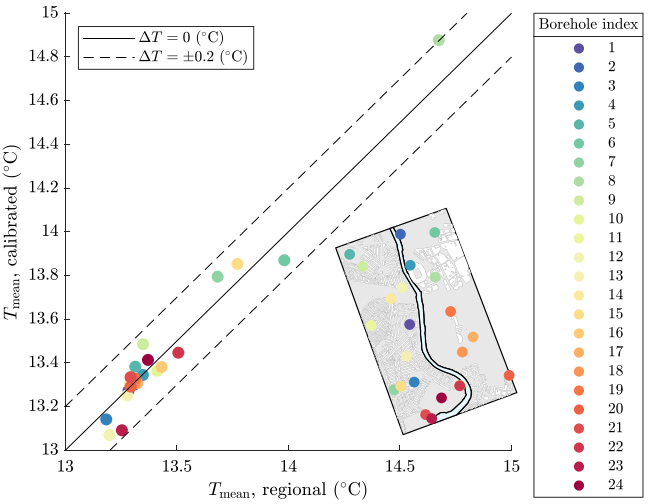

Figure 12. *Comparison between mean temperatures outputted at the borehole locations for the regional (x-axis) and the calibrated (y-axis) data. The results obtained from the calibrated model are in good agreement (i.e., within 0.2 °C) of the numerically generated "field" measurements.*

regionally within a domain, calibration with sufficient data points could be able to infer a representative domain value for those parameters, which could be used in urban planning and design.

4. Impact of Uncertainty on Geothermal Planning

The inherent variation within the subsurface and the difficulty and costs associated with obtaining high-resolution data typically result in uncertainty in its properties. This can impact numerous applications relating to the underground and can lead to over-designing as a measure of accounting for this uncertainty. One such application is the utilization of shallow geothermal energy, which uses ground-source heat pump systems (GSHPs) to efficiently provide renewable thermal energy for heating and cooling purposes by rejecting heat to or absorbing heat from the ground using ground heat exchangers (GHEs) (Mustafa Omer, 2008; Bayer et al., 2019). To assess the potential impact of parameter uncertainty on GSHP design, an existing open-loop GSHP system within the studied area of the city center of Cardiff is examined as an example case. This system has been installed at the Grangetown Nursery School, in the southwestern region of the modeled domain, and is monitored by the BGS (Boon et al., 2019). It includes two ground-source heat pumps that can provide up to 11 kW each, noting that, at the time of designing, few data from the site and aquifer were available and thus assumptions were made in the design. The hydraulic gradient of the local area is estimated to be 0.002 and the two wells (one abstraction and one injection) have a diameter of 0.2 m, a screen length of about 9 m, screened across the full thickness of the gravel aquifer, and are constructed at a distance of 20 m apart. Boon et al. (2019) note the close (non-ideal) well spacing resulted from a lack of space for drilling on site during the installation work, a situation not uncommon in highly built-up urban environments. Typical open-loop GHE design processes for feasibility studies are followed to determine the potential geothermal power that this system can provide for different values of the hydraulic conductivity of the aquifer, using the calibration results from Section 3. More details on the project specifications may be found in Boon et al. (2019) and detailed information on GSHP systems and open-loop design processes followed within this work can be found in Supplementary Appendix E.

This analysis focuses on the uncertainty in the hydraulic conductivity value of the aquifer, and uses various values for this parameter as found in the literature (listed in Table 3), as well as values obtained from the calibration presented above to explore the impact of uncertainty on the design of a typical open-loop shallow geothermal system. Results of the analysis are presented in Figure 13, showing the

https://doi.org/10.1017/dce.2022.32 Published online by Cambridge University Press

Table 3. *Hydraulic conductivity (k_h) values used in the analysis.*

Label	k_h (m/s)	Description	Reference
Lit 1	6.3×10^{-4}	Average of range reported for sand and gravel mix	Kruseman and De Ridder (2000)
Lit 2	1.0×10^{-2}	Typical value for clean sands and gravel deposits	Hobbs et al. (2002)
Pump test	5.8×10^{-4}	Obtained via pumping tests in the Cardiff area	Heathcote et al. (2003)
Preliminary	2.9×10^{-4}	Used in preliminary modeling of this site	Boon et al. (2019)

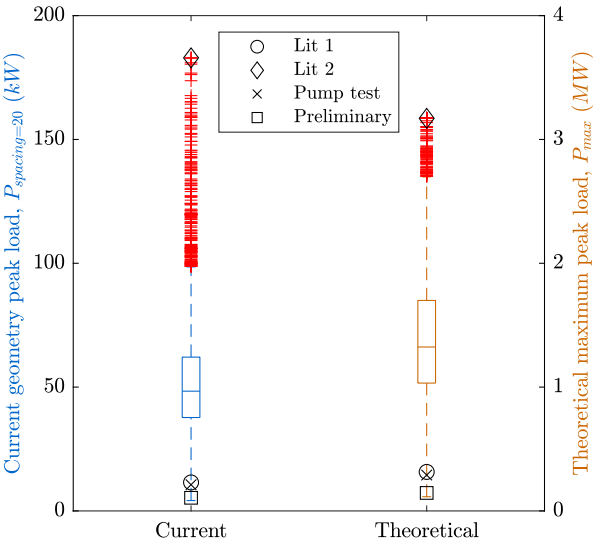

Figure 13. *Comparison of geothermal energy potential for different values of k_h. Results are shown in terms of the peak load for the current well spacing (left axis) and hypothetical maximum when the well spacing is such that no thermal interference happens between the wells (right axis).*

significant impact that the value of aquifer hydraulic conductivity can have on system design and expected geothermal potential. The figure shows the peak load distribution, calculated using the calibrated hydraulic conductivity distribution from Section 3 and the process outlined in Supplementary Appendix E, as box plots for both the current geometry and the theoretical maximum load that can be provided (which does not constrain the spacing of the wells). The peak loads obtained using hydraulic conductivity values found in the literature are also shown. For the current system geometry and a well spacing of 20 m, the estimated achievable peak power spans a wide range, from 5 to 183 kW, with the box plot interquartile range (25th, and 75th percentiles) narrowing this to between 38 kW and 62 kW. Three of the four literature hydraulic conductivity values result in peak loads on the lower range of the distribution. The fourth value, labeled "Lit 2", results in an extreme peak load outside the calibrated range. This value is likely less suitable for this scenario, given current knowledge of the domain, but is included for completeness. Results for the theoretical maximum peak load for the area, where well spacing constraints are ignored, exhibit a range of peak loads between 0.15 MW and 3.17 MW and box plot interquartile range between 1.05 MW and 1.70 MW, relatively wider than the results obtained from considering the current

https://doi.org/10.1017/dce.2022.32 Published online by Cambridge University Press

geometry. Similarly, theoretical peak load values determined using literature values for the hydraulic conductivity are on the lower end of the calibrated range, save the one labeled "Lit 2". Given the heterogeneity in the soil structure that can exist between aquifers of similar geological designation (as well as within a single aquifer) and the resulting uncertainties in the hydraulic properties, the results suggest that lack of detailed information from the site can result in a range of different GSHP designs of varying suitability for a given scenario. This can lead to under-utilizing the available resources or, perhaps more concerningly, attempting to use more resources than available, thereby risking damage to the GSHP system or the natural environment. In this case, results suggest that the open-loop system installed in the city of Cardiff is under-utilized, since the peak load of 22 kW currently installed (from the heat pumps) is at the lower end of the calculated range and it is highly likely that a higher load could be provided. Given the uncertainty, a conservative approach could, for example, be to use the lower percentile calibration value of 38 kW as peak load, which is still close to double the current design peak load. It is worth noting that the current system was designed to fulfill the heating and cooling demands of the school, which are less than the ground capacity. However, a clearer understanding of the subsurface capacity could have potentially enabled a larger scale system, where the geothermal energy is shared with nearby buildings, assuming sufficient demand. Quantifying the uncertainty in important design parameters can, therefore, contribute towards better utilizing natural resources, even within conservative approaches, resulting in relatively lower risk designs with higher thermal yields.

The discrepancy in geothermal potential between the two sets of results presented in Figure 13 highlights one of the potential drawbacks of open-loop systems, namely the physical constraints associated with the distance between wells. To fully avoid thermal interference, this distance typically needs to be large, often unreasonably so for urban locations such as London (Banks, 2009; Fry, 2009; Abesser, 2010). Figure 14 shows the optimal well spacing calculated for a given value of ground hydraulic conductivity, as explained in Supplementary Appendix E, noting that the advised minimum spacing according to the open-loop code of practice for the UK by CIBSE is 100 m (while numerical modeling is recommended) (Wincott et al., 2019). The well spacing values obtained using calibrated parameters peak at around 250 m with 25th and 75th percentiles at around 230 and 290 m. Using the hydraulic conductivity values from literature results in optimal spacing values outside that range, a very low value of around

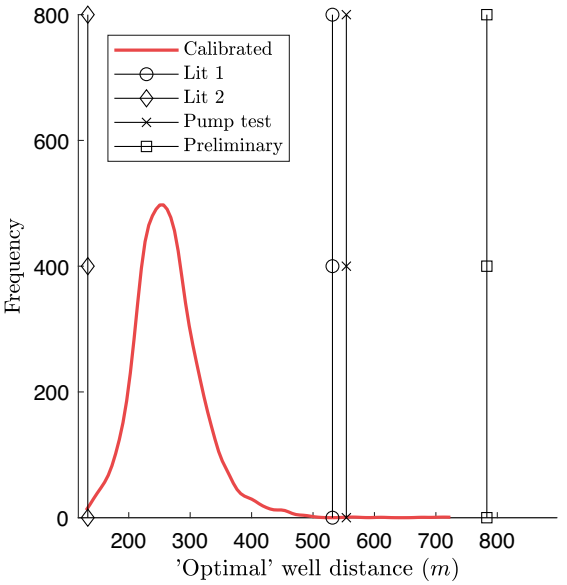

Figure 14. *Distance between abstraction and injection wells in order to avoid thermal interference in the open-loop system, for the different values of k_h used in this analysis (Banks, 2009).*

https://doi.org/10.1017/dce.2022.32 Published online by Cambridge University Press

130 m for "Lit 2", and values higher than 550 m for "Lit 1", "Pump test", and "Preliminary". In the case of attempting to design a GSHP system without thermal interference, the first value would likely result in under-designing the system, leading to lower whole-system efficiency and, in the worst-case scenario, system failure, while the other three would likely result in very costly over-designed systems, or, more realistically, would not be constructed due to the estimated costs and/or low thermal potential. It is apparent that uncertainty in important design parameters, such as hydraulic conductivity, can result in significantly different GSHP system designs, and costs, and could heavily influence whether a design is deemed a financially feasible solution. Moreover, the magnitudes of the distances calculated suggest that it would be difficult to utilize the maximum open-loop geothermal potential in most urban situations for private land owners and indicate the need for planning and investigations to be undertaken at a larger (city-)scale, with the potential of sharing geothermal resources across multiple buildings and land owners.

5. Conclusion

In this work, uncertainty in underground thermal modeling parameters is investigated, identifying the impacts this uncertainty can have on our understanding of how anthropogenic heat fluxes affect the shallow subsurface and how to utilize the ground as a resource. To this end, the city center of Cardiff is modeled using a semi-3D finite element approach to solve the governing equations of heat transfer and fluid flow. A Bayesian statistical framework is adopted, combining field observations and data from simulations at two levels of fidelity, to calibrate the numerical model and infer values for the most influential parameters within the model. Namely, these are the far-field ground temperature, T_{ground}, the hydraulic conductivity of the aquifer, k_h, and the temperature of the anthropogenic heat sources considered in the domain (i.e., heated basements), T_{room}. The calibration methodology is applied to numerically generated (synthetic) and to field data at a set of measurement locations, both to further validate the methodology as well as to highlight the additional unknowns and challenges faced when working with real data. Results for both sets of calibrations show convergence, with the synthetic data case resulting in inferred posterior distributions with mean and modal values that match closely the model inputs, and the field data case producing distributions with reduced uncertainty compared to the prior used and reasonable values for the explored parameters. Specifically, the field data calibration posteriors show a far-field ground temperature T_{ground} of 12.9 °C, agreeing with deep borehole measurements from the area, a k_h of about 3.1×10^{-3} m/s, somewhat higher than previous investigations for the area and local soil type, and a T_{room} of 14.7 °C, indicating basements in the area are not significantly heated.

A notable limitation of parameter inference methodologies, such as the one presented in this work, is the lack of data availability, as not many locations have the quality and quantity of data that is available for the study area chosen in this work. The size of data set required to yield satisfactory calibration results is investigated, for both synthetic and field data, by methodically decreasing the number of data points used for calibration. Results show that with synthetic data calibration can be done to a high degree of accuracy using as few as 10 data points, while with field data at least 16 points become necessary due to the complexity, unknowns, and noise associated with realistic conditions. The results highlight the importance of identifying information contained within data measured at different locations for the purposes of parameter estimation, which in the context of a site investigation could help mitigate costs from drilling boreholes in locations that provide little additional information.

The impact of heterogeneity is also investigated, as this can be an important aspect of the subsurface. This work focuses on heterogeneity in k_h, particularly (a) heterogeneity throughout the material, represented as noise in the value of k_h, and (b) regional heterogeneity, implemented by subdividing the area into four regions with different local k_h values. Noise in the parameter value is shown to not negatively impact the calibration process, with very reasonable posterior distributions found for the three investigated parameters (inferred mean within 0.1 °C of true values for T_{ground} and T_{room} and within 10% for k_h). The introduction of regional heterogeneity also resulted in reasonably accurate posterior distributions, and importantly, a fully converged calibration using the entire data set (i.e., the combination of data from all four sub-regions) suggests that "equivalent" parameter values inferred using data from the

https://doi.org/10.1017/dce.2022.32 Published online by Cambridge University Press

entire domain can be used as a model value for the model of the whole domain, giving rise to a difference of less than 0.2 °C between mean temperatures determined from the calibrated model and the calibration data. Considering the heterogeneity of realistic parameters, this result provides confidence in using the calibration methodology to predict parameter values, including area-representative values that could be used in urban design and planning.

Finally, the impact of the identified uncertainty on engineering applications is assessed by considering the effect of different parameter values on an installed open-loop GSHP system present within the modeled area. The shallow geothermal potential of this system is computed using theoretical methods for a range of different values for k_h, found in the above investigations and in the literature. The results indicate that uncertainty in k_h can affect significantly the design and energy potential of such a system, with capacity (peak load) values in this case ranging greatly from 5 to 183 kW. The plausible (interquartile) range using the inferred distribution for k_h is between 38 kW and 62 kW, much higher than the system's current design load, indicating that if this information had been available during the design stage and there was equivalent demand from nearby buildings, the system could have been utilized even further, providing more geothermal energy for heating and cooling. Investigating the maximum geothermal potential of this system without accounting for its specific configuration shows much higher load values, for which the representative well spacing would vary from 133 to 783 m depending on the value of k_h used, illustrating the importance of reducing uncertainty in sensitive parameters early on in the design phase of such projects and thus the importance of subsurface monitoring and characterization of aquifer properties at appropriate scales.

Acknowledgments. The authors are grateful to Ricky Terrington (BGS), Johanna Scheidegger (BGS), and Ashley Patton (BGS) for their valuable input to the data acquisition for this work and to Cardiff Harbour Authority for the provision of supporting data and access to boreholes. Gareth Farr and David Boon publish with the permission of the executive director, BGS, UKRI.

Data Availability Statement. The data that support the findings of this study are available from the British Geological Survey. Restrictions apply to the availability of these data, which were used under license for this study. Data are available from the British Geological Survey via https://www.bgs.ac.uk/about-bgs/contact-us/ with the permission of the British Geological Survey.

Author Contributions. Conceptualisation: M.J.K.; N.M.; R.C.; K.M.; A.B. Methodology: M.J.K; N.M.; K.M.; A.B.; R.C. Data curation: G.J.F.; D.B.P. Data visualisation: M.J.K.; N.M. Writing original draft: M.J.K.; N.M. All authors approved the final submitted draft.

Funding Statement. This work was supported by CMMI-EPSRC: Modeling and Monitoring of Urban Underground Climate Change (EP/T019425/1), by the Centre for Smart Infrastructure & Construction (EP/N021614/1) and the Centre for Digital Build Britain at the University of Cambridge, as well as AI for Science and Government (ASG), UKRI's Strategic Priorities Fund awarded to the Alan Turing Institute, UK (EP/T001569/1). The financial support for Kathrin Menberg via the Margarete von Wrangell program of the Ministry of Science, Research and the Arts (MWK) of the State of Baden-Württemberg is gratefully acknowledged. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests. The authors declare no competing interests exist.

References

Abesser C (2010). *Open-loop ground source heat pumps and the groundwater systems: A literature review of current applications, regulations and problems. British Geological Survey Open Report.* OR/10/045. 31pp.

Alexanderian A (2021) Optimal experimental design for infinite-dimensional Bayesian inverse problems governed by PDEs: A review. *Inverse Problems 37*(4). 31pp.

Alexanderian A, **Petra N**, **Stadler G and Ghattas O** (2016) A fast and scalable method for a optimal design of experiments for infinite-dimensional Bayesian nonlinear inverse problems. *SIAM Journal on Scientific Computing 38*(1), A243–A272.

Ampofo F, **Maidment GG and Missenden JF** (2006) Review of groundwater cooling systems in London. *Applied Thermal Engineering 26*(17–18), 2055–2062.

Attard G, **Rossier Y**, **Winiarski T and Eisenlohr L** (2016) Deterministic modeling of the impact of underground structures on urban groundwater temperature. *Science of the Total Environment 572*, 986–994.

Baggs SA (1983) Remote prediction of ground temperature in Australian soils and mapping its distribution. *Solar Energy 30*(4), 351–366.

Banks D (2009) Thermogeological assessment of open-loop well-doublet schemes: A review and synthesis of analytical approaches. *Hydrogeology Journal 17*(5), 1149–1155.

Bayer P, **Attard G**, **Blum P and Menberg K** (2019) The geothermal potential of cities. *Renewable and Sustainable Energy Reviews 106*, 17–30.

Beardsmore GR and Cull JP (2001) Heat generation. In *Crustal Heat Flow*. Cambridge: Cambridge University Press, pp. 23–44.

Benz SA, **Bayer P**, **Menberg K**, **Jung S, and Blum P** (2015) Spatial resolution of anthropogenic heat fluxes into urban aquifers. *Science of the Total Environment* 524–525, 427–439.

Bidarmaghz A, **Choudhary R**, **Narsilio G and Soga K** (2021) Impacts of underground climate change on urban geothermal potential: Lessons learnt from a case study in London. *Science of the Total Environment 778*. 31pp.

Bidarmaghz A, **Choudhary R**, **Soga K**, **Kessler H**, **Terrington RL and Thorpe S** (2019) Influence of geology and hydrogeology on heat rejection from residential basements in urban areas. *Tunnelling and Underground Space Technology 92*, 103068.

Bidarmaghz A, **Choudhary R**, **Soga K**, **Terrington RL**, **Kessler H and Thorpe S** (2020) Large-scale urban underground hydrothermal modelling – A case study of the Royal Borough of Kensington and Chelsea, London. *Science of the Total Environment 700*. 31pp.

Bidarmaghz A and Narsilio G (2018) Heat exchange mechanisms in energy tunnel systems. *Geomechanics for Energy and the Environment 16*, 83–95.

Blum P, **Menberg K**, **Koch F**, **Benz SA**, **Tissen C**, **Hemmerle H and Bayer P** (2021) Is thermal use of groundwater a pollution? *Journal of Contaminant Hydrology 239*, 103791.

Boon DP, **Farr GJ**, **Abesser C**, **Patton AM**, **James DR**, **Schofield DI and Tucker DG** (2019) Groundwater heat pump feasibility in shallow urban aquifers: Experience from Cardiff, UK. *Science of the Total Environment 697*. 31pp.

Boon DP, **Farr GJ and Hough E** (2021) *Thermal properties of Triassic Sherwood (Bunter) Sandstone Group and Mercia Mudstone Group (Keuper Marl) lithologies. In 2nd Geoscience & Engineering in Energy Transition Conference.*

Böttcher F, **Casasso A**, **Götzl G and Zosseder K** (2019) TAP – Thermal aquifer potential: A quantitative method to assess the spatial potential for the thermal use of groundwater. *Renewable Energy 142*, 85–95.

Choi W, **Menberg K**, **Kikumoto H**, **Heo Y**, **Choudhary R and Ooka R** (2018) Bayesian inference of structural error in inverse models of thermal response tests. *Applied Energy 228*, 1473–1485.

Chong A, **Lam KP**, **Pozzi M and Yang J** (2017) Bayesian calibration of building energy models with large datasets. *Energy and Buildings 154*, 343–355.

Chong A and Menberg K (2018) Guidelines for the Bayesian calibration of building energy models. *Energy and Buildings 174*, 527–547.

COMSOL Multiphysics (R) v. 5.6 (2020) www.comsol.com. COMSOL AB, Stockholm, Sweden.

Cui T, **Fox C and O'Sullivan MJ** (2011) Bayesian calibration of a large-scale geothermal reservoir model by a new adaptive delayed acceptance Metropolis Hastings algorithm. *Water Resources Research 47*(10). 31pp.

Cui T, **Fox C and O'Sullivan MJ** (2019) A posteriori stochastic correction of reduced models in delayed-acceptance mcmc, with application to multiphase subsurface inverse problems. *International Journal for Numerical Methods in Engineering 118*(10), 578–605.

Cui T, **Peeters L**, **Pagendam D**, **Pickett T**, **Jin H**, **Crosbie RS**, **Raiber M**, **Rassam DW and Gilfedder M** (2018) Emulator-enabled approximate bayesian computation (abc) and uncertainty analysis for computationally expensive groundwater models. *Journal of Hydrology 564*, 191–207.

Dalla Santa G, **Galgaro A**, **Sassi R**, **Cultrera M**, **Scotton P**, **Mueller J**, **Bertermann D**, **Mendrinos D**, **Pasquali R**, **Perego R**, **Pera S**, **Di Sipio E**, **Cassiani G**, **De Carli M and Bernardi A** (2020) An updated ground thermal properties database for GSHP applications. *Geothermics 85*, 101758.

Epting J, **Böttcher F**, **Mueller MH**, **García-Gil A**, **Zosseder K and Huggenberger P** (2020) City-scale solutions for the energy use of shallow urban subsurface resources – Bridging the gap between theoretical and technical potentials. *Renewable Energy 147*, 751–763.

Fabreti LG and Höhna S (2022) Convergence assessment for bayesian phylogenetic analysis using mcmc simulation. *Methods in Ecology and Evolution 13*(1), 77–90.

Farr G, **Patton AM**, **Boon D**, **James D**, **Coppell L and James L** (2019) Cardiff Urban Geo Observatory, Groundwater Temperature Data 2014-2018. British Geological Survey. (Dataset). https://doi.org/10.5285/bf150dd6-7b28-49ca-b66f-8b543a33a5c0

Farr GJ, **Patton AM**, **Boon DP**, **James DR**, **Williams B and Schofield DI** (2017) Mapping shallow urban groundwater temperatures, a case study from Cardiff, UK. *Quarterly Journal of Engineering Geology and Hydrogeology 50*(2), 187–198.

Ferguson G and Woodbury AD (2004) Subsurface heat flow in an urban environment. *Journal of Geophysical Research: Solid Earth 109*(B2), 1–9.

Fry VA (2009) Lessons from London: Regulation of open-loop ground source heat pumps in Central London. *Quarterly Journal of Engineering Geology and Hydrogeology 42*(3), 325–334.

García-Gil A, **Vázquez-Suñe E**, **Alcaraz MM**, **Juan AS**, **Sánchez-Navarro JÁ**, **Montlleó M**, **Rodríguez G and Lao J** (2015) GIS-supported mapping of low-temperature geothermal potential taking groundwater flow into account. *Renewable Energy 77*, 268–278.

Gelman A, **Carlin JB**, **Stern HS**, **Dunson DB**, **Vehtari A and Rubin DB** (2013) *Bayesian Data Analysis*, 3rd Edn. CRC Press.

Gelman A and Rubin DB (1992) Inference from iterative simulation using multiple sequences. *Statistical Science 7*(4), 457–472.

https://doi.org/10.1017/dce.2022.32 Published online by Cambridge University Press

Goetzl G, **Steiner C**, **Hofmann K**, **Riedel P and Görz** (2017) Resource mapping of open loop systems. (Report Delierable D.T2.2.2 Synopsis of geothermal mapping methods). Interreg CENTRAL EUROPE GeoPlasma-CE.

Goh J, **Bingham D**, **Holloway JP**, **Grosskopf MJ**, **Kuranz CC and Rutter E** (2013) Prediction and computer model calibration using outputs from multifidelity simulators. *Technometrics 55*(4), 501–512.

Guillas S, **Rougier J**, **Maute a**, **Richmond AD and Linkletter CD** (2009) Bayesian calibration of the thermosphere-ionosphere electrodynamics general circulation model (TIE-GCM). *Geoscientific Model Development 2*(2), 137–144.

Heathcote J, **Lewis R and Sutton J** (2003) Groundwater modelling for the Cardiff Bay barrage, UK – Prediction, implementation of engineering works and validation of modelling. *Quarterly Journal of Engineering Geology and Hydrogeology 36*, 159–172.

Heathcote JA, **Lewis RT**, **Russell DI and Soley RW** (1997) Cardiff Bay barrage: Investigating groundwater control in a tidal aquifer. *Quarterly Journal of Engineering Geology 30*(1), 63–77.

Hobbs PRN, **Hallam JR**, **Forster A**, **Entwisle DC**, **Jones LD**, **Cripps AC**, **Northmore KJ**, **Self SJ and Meakin JL** (2002) Engineering geology of British rocks and soils Mudstones of the Mercia Mudstone Group. British Geological Survey Research Report, 106 pp.

Hou D, **Hassan IG and Wang L** (2021) Review on building energy model calibration by Bayesian inference. *Renewable and Sustainable Energy Reviews 143*, 110930.

Howard AS, **Warrington G**, **Ambrose K and Rees JG** (2008) *A formational framework for the Mercia Mudstone Group (Triassic) of England and Wales National Geoscience Framework Programme. British Geological Survey.*

Huan X and Marzouk YM (2013) Simulation-based optimal Bayesian experimental design for nonlinear systems. *Journal of Computational Physics 232*(1), 288–317.

Huan X and Marzouk YM (2014) Gradient-based stochastic optimization methods in BAYESIAN experimental design. *International Journal for Uncertainty Quantification 4*(6), 479–510.

Kaipio J and Somersalo E (2007) Statistical inverse problems: Discretization, model reduction and inverse crimes. *Journal of Computational and Applied Mathematics 198*(2), 493–504.

Kendall RS, **Williams LR**, **Patton AM and Thorpe S** (2020) Metadata report for the Cardiff superficial deposits 3D geological model. This item has been internally reviewed, but not externally peer-reviewed.

Kennedy MC and O'Hagan A (2001) Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63*(3), 425–464.

Krcmar D, **Flakova R**, **Ondrejkova I**, **Hodasova K**, **Rusnakova D**, **Zenisova Z and Zatlakovic M** (2020) Assessing the impact of a heated basement on groundwater temperatures in Bratislava, Slovakia. *Groundwater 58*(3), 406–412.

Kruseman GP and De Ridder NA (2000) *Analysis and Evaluation of Pumping Test Data*. 2nd Edn. The Netherlands: ILRI Publication, 47.

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

Linde N, **Ginsbourger D**, **Irving J**, **Nobile F and Doucet A** (2017) On uncertainty quantification in hydrogeology and hydrogeophysics. *Advances in Water Resources 110*, 166–181.

Loveridge F, **McCartney JS**, **Narsilio GA and Sanchez M** (2020) Energy geostructures: A review of analysis approaches, in situ testing and model scale experiments. *Geomechanics for Energy and the Environment*, *22*, 100173.

Maclaren OJ, **Nicholson R**, **Bjarkason EK**, **O'Sullivan J and O'Sullivan M** (2020) Incorporating posterior-informed approximation errors into a hierarchical framework to facilitate out-of-the-box MCMC sampling for geothermal inverse problems and uncertainty quantification. *Water Resources Research 56*.

Makasis N (2019) *Further Understanding Ground-Source Heat Pump System Design Using Finite Element Methods and Machine Learning Techniques.* PhD Thesis, The University of Melbourne.

Makasis N, **Kreitmair M**, **Bidarmaghz A**, **Farr G**, **Scheidegger J and Choudhary R** (2021) Impact of simplifications on numerical modelling of the shallow subsurface at city-scale and implications for shallow geothermal potential. *Science of the Total Environment 791*, 148236.

Makasis N, **Narsilio GA**, **Bidarmaghz A**, **Johnston IW and Zhong Y** (2020) The importance of boundary conditions on the modelling of energy retaining walls. *Computers and Geotechnics 120*, 103399.

Menberg K, **Bidarmaghz A**, **Gregory A**, **Choudhary R and Girolami M** (2020) Multi-fidelity approach to Bayesian parameter estimation in subsurface heat and fluid transport models. *Science of the Total Environment 745*, 140846.

Menberg K, **Heo Y and Choudhary R** (2016) Sensitivity analysis methods for building energy models: Comparing computational costs and extractable information. *Energy and Buildings 133*, 433–445.

Menberg K, **Heo Y and Choudhary R** (2019) Influence of error terms in Bayesian calibration of energy system models. *Journal of Building Performance Simulation 12*(1), 82–96.

Meng B, **Vienken T**, **Kolditz O and Shao H** (2019) Evaluating the thermal impacts and sustainability of intensive shallow geothermal utilization on a neighborhood scale: Lessons learned from a case study. *Energy Conversion and Management 199*, 111913.

Milnes E and Perrochet P (2013) Assessing the impact of thermal feedback and recycling in open-loop groundwater heat pump (GWHP) systems: A complementary design tool. *Hydrogeology Journal 21*(2), 505–514.

Morris MD (1991) Factorial sampling plans for preliminary computational experiments. *Technometrics 33*(2), 161–174.

Mustafa Omer A (2008) Ground-source heat pumps systems and applications. *Renewable and Sustainable Energy Reviews 12*(2), 344–371.

https://doi.org/10.1017/dce.2022.32 Published online by Cambridge University Press

NCAS British Atmospheric Data Centre (2020) Met office integrated data archive system (midas) land and marine surface stations data (1853-current).

Nicholson R, **Alferink H**, **Paton-simpson E**, **Gravatt M**, **Guzman S**, **Popineau J**, **Sullivan JPO**, **Sullivan MJO and Maclaren OJ** (2020) An introduction to optimal data collection for geophysical model calibration problems. In *Proceedings 42nd New Zealand Geothermal Workshop*, November, Waitangi, New Zealand.

Omagbon J, **Doherty J**, **Yeh A**, **Colina R**, **O'Sullivan J**, **McDowell J**, **Nicholson R**, **Maclaren OJ and O'Sullivan M** (2021) Case studies of predictive uncertainty quantification for geothermal models. *Geothermics 97*, 102263.

Omlin M and Reichert P (1999) A comparison of techniques for the estimation of model prediction uncertainty. *Ecological Modelling 115*(1), 45–59.

Painter SL, **Woodbury AD and Jiang Y** (2007) Transmissivity estimation for highly heterogeneous aquifers: Comparison of three methods applied to the Edwards aquifer, Texas, USA. *Hydrogeology Journal 15*(2), 315–331.

Parkes D, **Busby J**, **Kemp SJ**, **Petitclerc E and Mounteney I** (2020) The thermal properties of the Mercia Mudstone Group. *Quarterly Journal of Engineering Geology and Hydrogeology 54*(1). pp10

Patton AM, **Farr G**, **Boon DP**, **James DR**, **Williams B**, **James L**, **Kendall R**, **Thorpe S**, **Harcombe G**, **Schofield DI**, **Holden A and White D** (2020) Establishing an urban geo-observatory to support sustainable development of shallow subsurface heat recovery and storage. *Quarterly Journal of Engineering Geology and Hydrogeology 53*(1), 49–61.

Pepi C, **Gioffrè M and Grigoriu M** (2020) Bayesian inference for parameters estimation using experimental data. *Probabilistic Engineering Mechanics 60*, 103025.

Perego R, **Guandalini R**, **Fumagalli L**, **Aghib FS**, **De Biase L and Bonomi T** (2016) Sustainability evaluation of a medium scale GSHP system in a layered alluvial setting using 3D modeling suite. *Geothermics 59*, 14–26.

Popiel CO and Wojtkowiak J (2013) Temperature distributions of ground in the urban region of Poznan City. *Experimental Thermal and Fluid Science 51*, 135–148.

Pujol M, **Ricard LP and Bolton G** (2015) 20 years of exploitation of the Yarragadee aquifer in the Perth Basin of Western Australia for direct-use of geothermal heat. *Geothermics 57*(2015), 39–55.

Rajabi MM and Ketabchi H (2017) Uncertainty-based simulation-optimization using gaussian process emulation: Application to coastal groundwater management. *Journal of Hydrology 555*, 518–534.

Rappel H, **Beex LA**, **Hale JS**, **Noels L and Bordas SP** (2020) A tutorial on Bayesian inference to identify material parameters in solid mechanics. *Archives of Computational Methods in Engineering 27*(2), 361–385.

Rehfeldt KR, **Boggs JM and Gelhar LW** (1992) Field study of dispersion in a heterogeneous aquifer: 3. Geostatistical analysis of hydraulic conductivity. *Water Resources Research 28*(12), 3309–3324.

Rode A, **Liesch T and Goldscheider N** (2015) Open-loop geothermal heating by combined extraction-injection one-well systems: A feasibility study. *Geothermics 56*, 110–118.

Rupprecht D, **Steiner C**, **Heiermann M and Riedel P** (2017) Summary of National Legal Requirements, Current Policies and Regulations of Shallow Geothermal Use. Technical Report, GeoPLASMA-CE, Austria.

Scott SW,**O'Sullivan JP**, **Maclaren OJ**, **Nicholson R**, **Covell C**, **Newson J and Gujónsdóttirð MS** (2022) Bayesian calibration of a natural state geothermal reservoir model, Krafla, North Iceland. *Water Resources Research 58*(2), e2021WR031254.

SimonHydrotechnica (1993) Cardiff dewatering pilot study: Final report, Vol. 1. Technical Report, SimonHydrotechnica.

Sorensen T and Vasishth S (2015) Bayesian linear mixed models using Stan: A tutorial for psychologists, linguists and cognitive scientists. *Preprint*, arXiv:1506.06201.

Speich M, **Dormann CF and Hartig F** (2021) Sequential Monte-Carlo algorithms for Bayesian model calibration – A review and method comparison. *Ecological Modelling 455*, 0–3.

UK Environment Agency (2014) Guidance notes on registration of your ground source heating and cooling system as exempt from the need for an environmental permit.

Ungemach P (2003) Reinjection of cooled geothermal brines into sandstone reservoirs. *Geothermics 32*(4), 743–761.

Vehtari A, **Gelman A**, **Simpson D**, **Carpenter B and Bürkner P-C** (2021) Rank-normalization, folding and localization: An improved R^ for assessing convergence of MCMC (with discussion). *Bayesian Analysis 16*(2), 667–718.

Volodina V and Challenor P (2021) The importance of uncertainty quantification in model reproducibility. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 379*(2197), pp16.

Williams B (2008) Cardiff Bay barrage: Management of groundwater issues. *Proceedings of the Institution of Civil Engineers: Water Management 161*(6), 313–321.

Wincott N, **Billings J**, **CIBSE, HPA and GSHPA** (2019) *CP3 Open-loop groundwater source heat pumps: Code of Practice for the UK (2019).* CIBSE.

Zheng Y, **Xie Y and Long X** (2021) A comprehensive review of Bayesian statistics in natural hazards engineering. *Natural Hazards 108*(1), 63–91.

Zhou H, **Gómez-Hernández JJ and Li L** (2014) Inverse methods in hydrogeology: Evolution and recent trends. *Advances in Water Resources 63*, 22–37.

https://doi.org/10.1017/dce.2022.32 Published online by Cambridge University Press

A. Gaussian process covariance formulation

Gaussian Process models are used for the approximation of the terms in Equation 1, namely (η_l, μ, δ), with covariance matrices adapted from the KOH approach accordingly, that is,

$$\Sigma_{\eta_l(i,\,j)} = \frac{1}{\lambda_{\eta_l}} \exp\left[-\sum_{k=1}^{p} \beta_{\eta_l,k} \left(x_{i,k} - x_{j,k}\right)^2 - \sum_{k'=1}^{q_h} \beta_{\eta_l,p+k'} \left(\theta_{hi,k'} - \theta_{hj,k'}\right)^2 - \sum_{k''=1}^{q_l} \beta_{\eta_l,p+q_h+k''} \left(\theta_{li,k''} - \theta_{lj,k''}\right)^2 \right], \tag{3}$$

$$\Sigma_{\mu(i,\,j)} = \frac{1}{\lambda_{\mu}} \exp\left[-\sum_{k=1}^{p} \beta_{\mu,k} \left(x_{i,k} - x_{j,k}\right)^2 - \sum_{k'=1}^{q_h} \beta_{\mu,p+k'} \left(\theta_{hi,k'} - \theta_{hj,k'}\right)^2 - \sum_{k''=1}^{q_l} \beta_{\mu,p+q_h+k''} \left(\theta_{li,k''} - \theta_{lj,k''}\right)^2 \right], \tag{4}$$

$$\Sigma_{\delta(i,\,j)} = \frac{1}{\lambda_{\delta}} \exp\left[-\sum_{k=1}^{p} \beta_{\delta,k} \left(x_{i,k} - x_{j,k}\right)^2 \right], \tag{5}$$

where p is the number of state variables and q_h and q_l are the number of parameters of the high- and the low-fidelity models, respectively. Equations 3-5 introduce unknown hyper-parameters, namely the precision hyper-parameters (λ_{η_l}, λ_{μ}, λ_{δ}) and the correlation hyper-parameters ($\beta_{\eta_l}, \beta_{\mu}, \beta_{\delta}$) which are sampled along with the model parameters θ. The precision hyper-parameters are a measure for the magnitude of the covariance functions and thus to what extent the variation in output may be explained by the associated term. For example, $\lambda_{\eta l}$ accounts for the covariance in the field and model outputs that is captured by the low-fidelity model emulator term in Equation (1). A large value for a given λ (and, correspondingly, a small value for λ^{-1}) indicates that the related term absorbs only a small portion of the variance in the model output, i.e. explains a small amount of the variation in y_f (Choi et al., 2018). The correlation hyper-parameters are an indication for the smoothness of the covariance functions. The priors for the hyper-parameters are given in Table 4, is chosen in accordance with work done in previous studies (Menberg et al., 2020; Chong and Menberg, 2018; Guillas et al., 2009).

The posterior distribution depends on the unknown calibration parameters, θ, the GP hyper-parameters, β, and the GP precision hyper-parameters, λ, and is estimated using HMC, which is suitable for high-dimensional distributions. The likelihood function is given by

$$\mathcal{L}\left(z|\theta, \beta_{\eta_l}, \lambda_{\eta_l}, \beta_{\mu}, \lambda_{\mu}, \beta_{\delta}, \lambda_{\delta}, \lambda_{\varepsilon}\right) \propto |\Sigma_Z|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} (z-\mu)^T \Sigma_Z^{-1} (z-\mu) \right\}, \tag{6}$$

where z is a single vector containing the combined observations and outputs of low- and high-fidelity simulations and

$$\Sigma_Z = \Sigma_{\eta_l} + \begin{pmatrix} \Sigma_{\mu} & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} \Sigma_{\delta} + \Sigma_{\varepsilon} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{7}$$

Table 4. *Uncertain hyper-parameters in the multi-fidelity framework and prior probability distributions.*

Parameter	Description	Prior pdf
$\lambda_{\eta l}$	Precision parameter, low-fidelity model emulator	Gamma(10, 1)
λ_{μ}	Precision parameter, model mismatch	Gamma(10,0.03)
λ_{δ}	Precision parameter, model bias	Gamma(10,0.03)
β_{η_l}	Correlation strength parameter, low-fidelity model emulator	Beta(10,0.5)
β_{μ}	Correlation strength parameter, model mismatch	Beta(10,0.7)
β_{δ}	Correlation strength parameter, model bias	Beta(10,0.7)

https://doi.org/10.1017/dce.2022.32 Published online by Cambridge University Press

B. Numerical modelling details

This work utilises a validated semi-3D numerical modelling approach (Bidarmaghz et al., 2020; Makasis et al., 2021), to model the subsurface response to thermal and hydraulic phenomena. A collection of *x*-*y* planes is modelled using the governing equations for conductive and convective heat transfer, and fluid flow through porous media, i.e.,

$$\mathrm{d}z\left(\rho C_p\right)_{\mathrm{eff}}\frac{\partial T_g}{\partial t}+\mathrm{d}z\rho_f C_{p,f}\mathbf{v}_f\nabla T_g+\nabla\cdot\mathbf{q}=q_{0,\mathrm{up}}+q_{0,\mathrm{down}}+Q_{\mathrm{applied}}, \tag{8}$$

$$\mathbf{v}_f=-\frac{K}{\mu_f}\left(\nabla p_f-\rho_f\mathbf{g}\nabla Z\right), \tag{9}$$

$$\nabla\cdot\rho_f\left[-\frac{K}{\mu_f}\left(\nabla p_f-\rho_f\mathbf{g}\nabla Z\right)\right]=0, \tag{10}$$

where $\mathrm{d}z$ is the relative width of the plane (or the distance between planes) (m), ρ_{eff} is the effective density (kg/m^3), $C_{p,\mathrm{eff}}$ the effective specific heat capacity (J/(kg K)), t is time (s), ρ_f is the fluid (groundwater) density (kg/m^3), $C_{p,f}$ is the specific heat capacity of the fluid (J/(kg K)), $\mathbf{v_f}$ is the Darcy velocity of the fluid (m/s), $\mathbf{q}$ is the heat flux (W/m^2), the permeability K (m^2) of the material is related to the hydraulic conductivity k_h (m/s) by $K\mu_f=k_h\left(\rho_f\mathbf{g}\right)$, μ is the dynamic viscosity of water (Pa · s), p_f is the pressure of water (Pa), ∇Z is the total head gradient (m) and Q_{applied} represents an external heat source (W), such as a ground heat exchangers. Each plane is thermally coupled to its nearest vertical neighbours via conductive heat transfer, which is the dominant mode of heat transfer in the z-direction. The upwards and downwards out-of-plane conductive heat fluxes ($q_{0,\mathrm{up}}$ and $q_{0,\mathrm{down}}$, respectively) from neighbouring planes are given by

$$q_{0,\mathrm{up}}=\lambda_{\mathrm{eff}}\left(T_{n-1}-T_n\right)/\mathrm{d}z, \tag{11}$$

$$q_{0,\mathrm{down}}=\lambda_{\mathrm{eff}}\left(T_{n+1}-T_n\right)/\mathrm{d}z, \tag{12}$$

where λ_{eff} is the effective thermal conductivity (W/(m K)), and T_n represents the temperature at the n^{th} plane. Further details on the modelling can be found in the cited literature.

The dynamics of the River Taff flowing through the domain are modelled using incompressible turbulent single-phase flow physics and coupled to the heat transfer physics in terms of temperature, pressure and velocity, as explained in detail in (Makasis et al., 2021). The Reynolds-averaged Navier-Stokes (RANS) equations are implemented for conservation of momentum, and the continuity equation for conservation of mass. The governing equations for turbulent flow are solved for the pressure and velocity vectors of the fluid flow within the river which are then coupled to the heat transfer equations to calculate the transfer of heat and thus distribution of temperature within the domain.

A series of boundary and initial conditions applied to the model are shown in Figure 15. At the uppermost plane, a seasonally varying temperature is applied, utilising as a semi-empirical function in time and depth (Beardsmore and Cull, 2001; Baggs, 1983) fitted to local conditions, i.e.

$$T_g(z,t)=T_{0,g}-1.07k_vT_{\mathrm{amp}}e^{-\varepsilon z}\cos\left[\omega(t-t_0)+\varepsilon z\right], \tag{13}$$

where $T_{0,g}$ is the mean annual ground temperature (°C), assumed as 12.9°C (Farr et al., 2017), T_{amp} the seasonal heating cycle amplitude (°C), assumed as 6.5°C (NCAS British Atmospheric Data Centre, 2020), $\omega=2\pi/P$ is the angular frequency of the heating cycle (rad) with period $P=365$ days, $\varepsilon=\sqrt{\pi/(P\alpha)}$, α is the thermal diffusivity of the ground (m^2/s), k_v is the vegetation coefficient (defined spatially based on the surface cover conditions, adopting a value of 0.9 for suburban and 1.0 for urban terrain (Popiel and Wojtkowiak, 2013)), and t_0 is the day of coldest temperature after January 1st, found to be 26 days (NCAS British Atmospheric Data Centre, 2020). At the bottom-most plane it is assumed that any influence from the surface will have dissipated and a constant temperature of $T=T_{0,g}$ is applied. For each plane, thermal symmetry is applied at all domain boundaries, and the time independent hydraulic head distribution in Figure 1 is used to define the groundwater flow conditions. In the shallow layers, where appropriate, the temperature and velocity of the river flowing into the domain are also defined.

C. Further calibration results

The Effective Sample Size (ESS) is an important metric for MCMC approaches, showing the correlation within parameter samples and identifying the equivalent number of effectively independent samples drawn from the Markov Chains (Fabreti and Höhna, 2022). For the calibrations on the complete dataset of 24 points, using 1000 samples for each of the four chains (after burn-in), the ESS for the three calibration parameters were computed to be 3791, 2198, and 2140, respectively, when using synthetic data, and 4102, 1866, and 2024 when using real data. The values computed, which are relatively high, support the calibration configuration and provide confidence in its findings.

Convergence is further indicated by plotting of the four chains of the HMC sampling algorithm. These are shown in the top row of Figure 16 for the calibration on synthetically generated data, and in the bottom row for calibration on the field data. A visual inspection of these suggests adequate convergence of the calibrations, as the trace plots resemble a "fat, hairy caterpillar" that does not bend (Sorensen and Vasishth, 2015).

https://doi.org/10.1017/dce.2022.32 Published online by Cambridge University Press

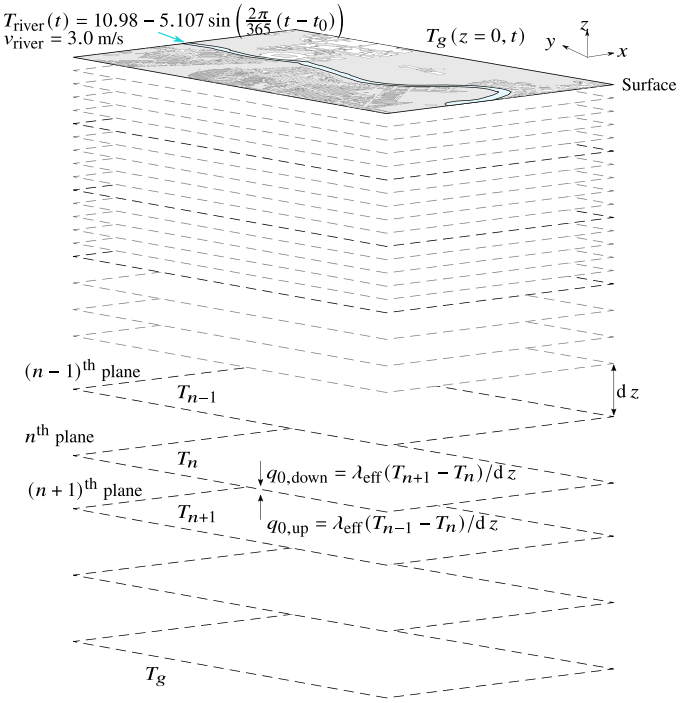

Figure 15. *Schematic of the modelling approach, showing the collection of 2D planes interconnected by heat flux transfer and the temperature boundary conditions at the top and bottom planes.*

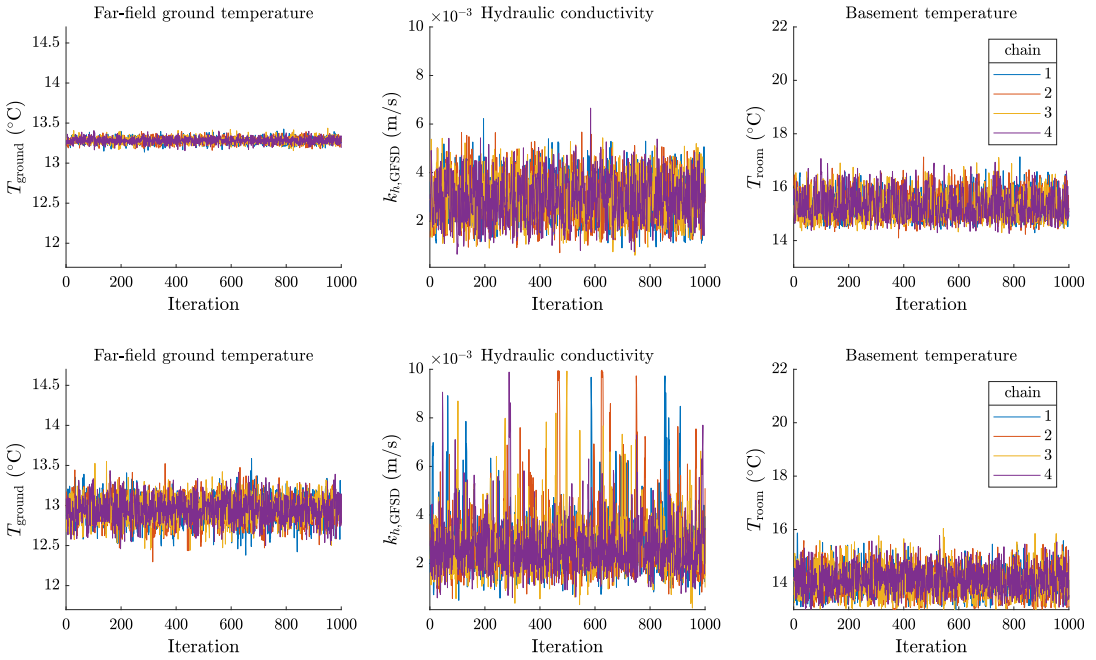

Figure 16. *Trace plots for the calibrations using both the synthetically generated (top row) and 'real' field data (bottom row).*

https://doi.org/10.1017/dce.2022.32 Published online by Cambridge University Press

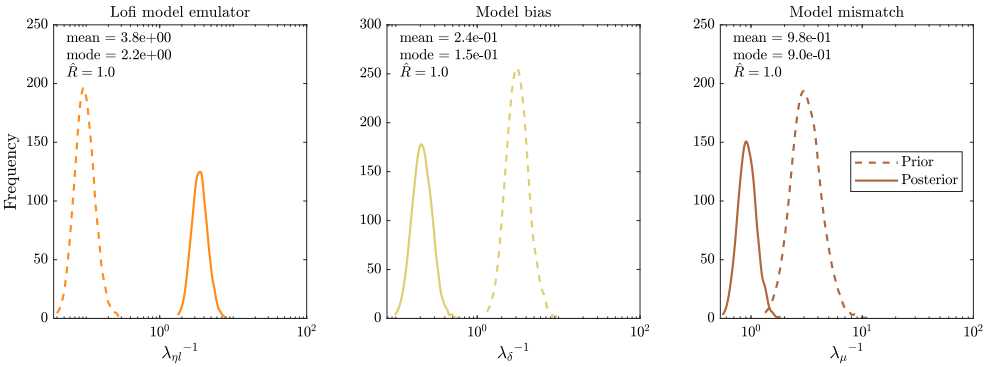

Figure 17. *Hyper-parameter posteriors determined from calibrating on field-data using the multi-fidelity framework.*

D. Data subset study

The indices contained in the data subsets used to infer the posteriors shown in Figure 9 are given in Table 5. Subsets are unique and contain 12 distinct measurement locations.

E. Open-loop shallow geothermal energy systems

GSHP systems can provide renewable geothermal energy for heating and cooling purposes. These systems can be installed in various ways, most commonly in close-loop configurations which circulate water in a closed pipe loop, implemented in GHEs such as purpose-built boreholes, trenches or energy geo-structures (Makasis, 2019; Makasis et al., 2020; Loveridge et al., 2020). Another economically viable approach, used in this work to demonstrate the impact of uncertainty on the use of the subsurface as a resource, is using open-loop GSHP systems (Banks, 2009; Milnes and Perrochet, 2013). These systems work by abstracting groundwater from a subsurface aquifer, passing it through a water-to-water GSHP to either extract or reject heat from/to it, and then re-inject the resulting cooled or heated water back to the aquifer. The abstraction and injection processes are commonly implemented using two wells (Banks, 2009), one for each process, however a single well combining both can also be used (Rode et al., 2015) and in rare cases only an extraction well may be required (Ampofo et al., 2006). In addition to posterior distributions for the calibration parameters, the calibration yields a posteriors for the hyper-parameters. The precision hyper-parameter prior and posterior distributions are shown in Figure 17.

In designing an open-loop shallow geothermal system, it is important to quantify the thermal capacity, defined as (Ungemach, 2003; Pujol et al., 2015)

$$P = \eta c_w \rho_w Q \Delta T, \tag{14}$$

where P is the thermal load/capacity of the well (W), η the efficiency of the heat pump, assumed as 0.95 (Pujol et al., 2015), c_w the specific heat capacity of groundwater, assumed as 4180 J/(kg K), ρ_w the density of groundwater, assumed as 998 kg/m^3, Q the pumping flow rate (m^3/s), and ΔT the difference in the water temperature between abstraction and injection (K). The thermal capacity is a product of the hydraulic productivity (the pumping flow rate) and thermal productivity (given by the water temperature difference between abstraction and injection points) (Goetzl and Steiner, 2017). Due to the re-injection of water to the underground and its interaction with groundwater resources, these values have strong dependencies on local restrictions and legislation and different guidelines can be found in different countries (Goetzl and Steiner, 2017; UK Environment Agency, 2014). In addition, the design parameters depend on properties of the local subsurface and project specifications, such as thermal demand and associated costs.

The thermal productivity aspect of the design is typically restricted, such that groundwater resources are not significantly affected. In the UK, it is required for ΔT to be less than 10 °C and the maximum temperature of the water injected into the ground to not exceed 25 °C (UK Environment Agency, 2014), while in Austria more conservative values of a maximum ΔT of 6 °C, a minimum water temperature of 5 °C and a maximum of 20 °C are adopted (Rupprecht et al., 2017). In this scenario, a ΔT of 8 °C is adopted, taking the temperature of the water from around 13 °C to about 5 °C, close to the limit of UK-based guidelines. The hydraulic productivity, or suitable pumping rate, depends on the hydraulic properties of the aquifer, such as aquifer thickness and hydraulic conductivity. Different equations exist in the literature to calculate this value, many consisting of adaptations of Thiem's theorem (for example (García-Gil et al., 2015)). Bottcher et al. (2019) summarise these and propose an approach that determines the flow rate to so as to avoid (i) depletion of the aquifer, (ii) thermal interference between injection and abstraction wells, and (iii) groundwater flooding at the injection well. This methodology is subsequently adopted herein to calculate the hydraulic productivity

https://doi.org/10.1017/dce.2022.32 Published online by Cambridge University Press

Table 5. *Subset information for small-scale study*

Subset	Borehole indices in subset											
1	1	2	4	5	6	8	12	14	18	19	23	24
2	5	6	7	8	9	11	12	14	15	17	19	20
3	1	2	6	7	8	11	12	16	18	19	20	23
4	1	10	11	12	13	15	16	17	18	21	23	24
5	2	4	5	6	7	9	12	17	19	21	22	24
6	4	6	7	8	10	11	12	13	15	16	17	23
7	2	5	6	8	11	13	15	16	18	20	21	22
8	2	8	9	10	11	12	14	15	17	21	22	24
9	2	5	6	9	11	12	13	14	15	16	17	24
10	1	2	3	5	7	8	12	17	19	21	22	24
11	1	4	10	12	13	14	16	18	19	20	22	23
12	5	8	9	13	14	15	17	19	20	21	22	23
13	3	4	6	11	13	14	18	20	21	22	23	24
14	4	6	7	9	11	12	16	17	18	21	22	23
15	2	3	7	9	11	15	16	17	18	20	21	23
16	1	2	5	8	9	14	15	16	18	19	22	24
17	3	5	7	10	11	14	15	17	18	20	21	24
18	1	2	5	6	8	9	12	15	16	19	20	24
19	2	3	6	7	8	12	14	16	17	18	19	23
20	1	5	6	7	12	13	16	17	18	20	21	22

and thus the peak thermal load the geothermal system can provide, additionally implementing an upper limit of 100 L/s due to technical limitations as suggested by (García-Gil et al., 2015). The formulation of the hydraulic productivity is summarised below, while for a detailed explanation readers can refer to the original paper (Böttcher et al., 2019):

$$Q = \min\left(Q_{\text{drawdown}}, Q_{\text{breakthrough}}, Q_{\text{injection}}\right), \tag{15}$$

$$Q_{\text{drawdown}} = 0.195Kb^2, \tag{16}$$

$$Q_{\text{breakthrough}} = \frac{\pi}{1.96} v_D b x_w, \tag{17}$$

$$Q_{\text{injection}} = (z_{\max} - z)Kb^{0.798}e^{29.9i}, \tag{18}$$

where K is the hydraulic conductivity of the aquifer (m/s), b the saturated aquifer thickness (m), in this case 9 m, v_D the Darcy velocity (m/s), x_w the distance between abstraction and injection wells (m), $z_{\max}$ the maximum allowed groundwater level (m), z the natural groundwater level (m), and i the hydraulic gradient (-), in this case 0,002.

In an open-loop system, feedback from the injection well might affect the operation of the abstraction well and create thermal interference. An ideal design places the two wells such that no thermal interference occurs, even though, this can be arguably difficult. Banks presents an equation calculating the minimum distance for the two wells and argues that, over time, thermal interference may be inevitable (Banks, 2009):

$$L > \frac{2Q}{T\pi i}, \tag{19}$$

where L is the well spacing (m), Q the abstraction flow rate (m^3/s), T the aquifer transmissivity (m^2/s) and i the hydraulic gradient (-).

Cite this article: Kreitmair MJ, Makasis N, Menberg K, Bidarmaghz A, Farr GJ, Boon DP and Choudhary R (2022). Bayesian parameter inference for shallow subsurface modeling using field data and impacts on geothermal planning. *Data-Centric Engineering*, 3, e32. doi:10.1017/dce.2022.32