

This is the preprint of the contribution published as:

Petrova, E., **Selzer, P.**, Kranz, S., Zeilfelder, S., Hebig, K.H., Machida, I., Marui, A., Blöcher, G., Scheytt, T. (2025):

Surrogate-model-based calibration of effective transport parameters from push-pull tests in the Horonobe aquifer (Japan)

Geothermics **133** , art. 103449

The publisher's version is available at:

<https://doi.org/10.1016/j.geothermics.2025.103449>

Surrogate-Model-Based Calibration of Effective Transport Parameters from Push-Pull Tests in the Horonobe Aquifer (Japan)

Elena Petrova¹, Philipp Selzer², Stefan Kranz¹, Sarah Zeilfelder³, Klaus H. Hebig⁴, Isao Machida⁵, Atsunao Marui⁵, Guido Blöcher^{1,6}, Traugott Scheytt⁷

¹ GFZ Helmholtz Centre for Geosciences, Telegrafenberg, 14473 Potsdam, Germany

² Department of Environmental Informatics, Helmholtz Centre for Environmental Research – UFZ, Permoserstraße 15, 04318 Leipzig, Germany

³ Department of Integrative Environmental Protection (II), Senate Department for Urban Mobility, Transport, Climate Action and the Environment, Brückenstraße 6, 10179 Berlin, Germany

⁴ BASE Federal Office for the Safety of Nuclear Waste Management, Department A – Supervision, Wegelystraße 8, 10623 Berlin, Germany

⁵ National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan

⁶ Technical University Berlin, Department of Engineering Geology, Berlin, Germany

⁷ TU Bergakademie Freiberg, Gustav-Zeuner-Straße 12, 09599 Freiberg, Germany

Corresponding author: elena.petrova@gfz.de

Author statement

E.P.: Writing - Original Draft, Writing Review & Editing, Conceptualization, Investigation, Validation, Visualization, Methodology, Software,;

P.S.: Methodology, Software, Conceptualization, Writing - Review & Editing;

S.K.: Methodology, Conceptualization, Project administration, Writing - Review & Editing;

S.Z.: Field Tests, Conceptualization, Data Curation;

K. H.-S.: Field Tests, Data Curation, Conceptualization;

I.M.: Field Tests, Conceptualization, Writing- Review & Editing;

A.M.: Field Tests, Conceptualization, Writing- Review & Editing;
GB: Supervision, Writing- Review & Editing, Methodology, Visualization, Project
administration, Funding acquisition, Conceptualization;
T. S.: Supervision, Writing- Review & Editing, Methodology, Conceptualization;
All authors contributed to the article and approved the submitted version.

Conflict of interest: None.

Abstract

Aquifer characterization is essential for optimizing Aquifer Thermal Energy Storage (ATES) systems. Single well tests, also known as push-pull tests, are a common method to identify effective solute and heat transport parameters of the aquifer, which are crucial for the design and assurance of long-term performance of ATES systems. Tracer breakthrough curves from push-pull tests are commonly used to calibrate analytical or numerical models of heat and solute transport in order to infer effective transport parameters like dispersivity of heat and solutes, retardation factors, and porosity. The main bottleneck of such multiparametric calibration is the non-uniqueness of the inverse problem solution which requires ensemble-based optimization to address the parametric uncertainty. In addition, the field measurements can only be performed up to a certain confidence as well, which introduces additional uncertainty to the calibration results. To account for both sources of uncertainty while targeting computationally affordable simulation, we have developed a surrogate model-based optimization framework for stochastic parameter optimisation. The surrogate model uses Gaussian process regression (GPR) to train and predict the objective function based on up to six aquifer and tracer properties. For training and fast model evaluation, we implemented a stable 1D radial finite difference representation of the advection-dispersion equation for sorbing compounds including measured input time-series as transient boundary condition and wellbore storage to accurately model push-pull tests. The surrogate model is used to calibrate this model and to propose plausible parameter combinations. The optimisation framework was applied to push-pull experiments using uranine, iodide, lithium, and heat as tracers in a sandy aquifer in Horonobe (Hokkaido, Japan).

The samples drawn from the posterior distribution resulting from the GPR-based optimisation show an overall good fit to the field observations. Based on the posterior parameter distribution, it was possible to shrink the uncertainty intervals of the solute and heat dispersivity and porosity. The outcome suggests low sensitivity to the solute retardation factors. However, the study also reveals that slight sorption may be acting in the Horonobe aquifer for some of the solute tracers commonly assumed to be conservative. Moreover, the study shows that exact porosity measurements may reveal the presence of sorption and thus improve the understanding of the tracers' behaviour. We demonstrate the benefits of using multiple tracers and high-resolution measurements to improve calibration accuracy under measurement uncertainty. The demonstrated approach offers a computationally efficient framework for addressing parametric uncertainty in push-pull test analysis, improving the design and optimization of ATEs systems.

Key words: ATEs, Push-Pull test, Simulation-based inference, Modelling, Solute and heat transport

1. Introduction

Aquifer thermal energy storage (ATEs) has been gaining growing popularity recently as an alternative to fossil fuel-based heating and cooling (De Schepper et al., 2020; Stemmler et al., 2023). Being an efficient energy storage option making use of the subsurface, ATEs enables large energy savings and massive reduction of CO₂ emissions (Vanhoudt et al., 2011; Bloemendal and Olsthoorn, 2018; Beernink et al., 2022). The efficiency of geothermal power generation as well as aquifer thermal energy storage systems is highly affected by the subsurface heat storage and transport properties (Doughty et al., 1982; Tang and Rijnaarts, 2023). Therefore, one of the essential steps in ATEs design is an accurate characterisation of the target aquifer (Blöcher et al., 2024). This is typically done by hydraulic field tests, core logging, and laboratory experiments (Müller and Regenspurg, 2014; Wagner et al., 2014; Park et al., 2015).

Among hydraulic field tests for aquifer characterization for an optimal design of ATEs systems, the push-pull test (PPT) using tracers is a well-established and commonly applied method to approximate the effective transport parameters of an aquifer (Schroth and Istok, 2005; Vandenbohede et al., 2009; Vandenbohede et al., 2011; Park et al., 2015; Stettler, et al., 2022). PPT results have been widely used to infer solute and heat transport parameters as well

as aquifer properties such as dispersivity, retardation factors, porosities, attenuation factor (Blöcher et al., 2024), groundwater velocity, and heat capacities (Leap and Kaplan, 1988; Kim et al., 2019; Ueckert et al., 2020; Kruisdijk and Breukelen, 2021; Johnson et al., 2023). To approximate the necessary parameters, multiple analytical (Schroth and Istok, 2005), semi-analytical (Shi et al., 2020; Suk et al., 2023), as well as numerical models (e.g., Vandenbohede et al., 2009) have been employed to account for PPTs of varying complexity including different processes and geological conditions as well as varying boundary conditions.

The majority of studies named above were dedicated to inference of a unique parameter set. However, in subsurface systems data collection is typically limited to sparse information from available boreholes, leading to incomplete information about the spatial distribution of properties. This limited data introduces significant uncertainty, particularly when inferring parameters for spatially distributed and physics-based modelling approaches (Kitanidis, 1998). As the number of unknown parameters increases, the solution may become non-unique, meaning different parameter combinations can yield outcomes of similar quality fitting the data equally well within the given uncertainty. To address these challenges, stochastic methods are employed, as they provide a robust framework to account for uncertainty and explore multiple possible parameter realizations, ultimately improving the reliability and predictive capability of physics-based models (Rasmusson et al., 2014; Cirpka and Valocchi, 2016; Jin et al., 2024). Although both, the simulation of PPTs and stochastic methods have been extensively developed already, only a limited number of studies aimed to apply stochastic methods to the calibration of real word PPTs.

This article presents a computationally efficient framework to stochastically calibrate PPT data under uncertainty using field data. We aim to provide a surrogate model-based workflow for uncertainty-acknowledged stochastic parameter estimation of the PPT data. For training and rapid model evaluation, we developed a fast and stable 1D finite-difference (FD) implementation of the advection-dispersion equation for sorbing tracers based on an analytical radial flow field. This model employs an adaptive explicit time-stepping scheme that ensures numerical stability while minimizing numerical diffusion. Furthermore, it can account for wellbore storage, the injection of a chaser after the push phase, as well as it can handle an input-time-series of concentration or temperature as transient boundary conditions. The workflow was applied to the Horonobe aquifer, which is a sandy aquifer in Hokkaido, Japan

(Hebig et al. 2014, 2015a,b), demonstrating its effectiveness in calibrating real-world PPT data serving for a better design of ATEs systems. The PPT data was modelled stochastically while reducing the number of model runs required with ongoing calibration and maintaining calibration precision. Posterior distributions of thermal and solute dispersivity, porosity, and retardation factors of solutes and heat were analysed with the aim to investigate which parameters can be reliably deduced from the calibration procedure, enhancing our understanding of aquifer characterization through PPTs. This approach addresses the challenges of parameter estimation in complex subsurface systems, offering a computationally efficient method for improving ATEs design and optimization. Calibrated parameter sets indicate that tracers, traditionally assumed to be conservative under most conditions in a sandy aquifer, such as uranine, iodide, or lithium may experience in fact slight sorption. Exact measurements of porosity may reveal, if this is actually the case in reality.

2. Site description and push-pull tests

The field experiments were performed at the Hamasato test site, which is the part of the municipality of Horonobe, at the north-western coast of Hokkaido, Japan (Matsumoto et al., 2020). Horonobe is situated in a coastal sedimentary basin composed primarily of loosely compacted sandstones, siltstones, and mudstones (Ikawa et al., 2014). The well field is located approximately 250 meters from the shoreline of the Sea of Japan, with the site elevation being around 5 meters above sea level (Hebig et al., 2016), moreover, it is screening the Sarabetsu Formation. This formation consists of unconsolidated Quaternary alluvial deposits, characterized by interbedded layers of coarse sand, fine gravel, and clay lenses (Hebig et al., 2015). The well screens a confined aquifer, primarily composed of sand and gravel, which is located between 90.7 and 99.7 meters below ground level. Although detailed hydraulic gradient and ambient groundwater flow velocity data for the monitoring well are unavailable, groundwater flow is generally from the recharge zone in the Horonobe Anticline, located about 10 km northeast, towards the Sea of Japan. Isotopic analysis and numerical simulations suggest groundwater ages ranging from 8,000 to 18,000 years, indicating an estimated flow velocity of approximately 0.56–1.25 meters per year (Matsumoto et al., 2020). Thus, we can safely neglect background flow for the analysis of the PPTs as the flow velocities induced by pumping in the near-field around the well are several orders of magnitude higher.

156 2.1. Experiments

157 A total of seven PPTs were performed in the year 2012 in the well DD-2 having a diameter of
158 5.08 cm (Figure 1), fully penetrating the aquifer with the thickness of 9.0 m. The experimental
159 setup is discussed in details in (Hebig, 2015; Hebig et al., 2016, 2016; Matsumoto et al.,
160 2020). To ensure the targeted investigations of the aquifer, the well was sealed with a packer.
161 The general setup of the PPTs included ~920 L of the injected test solution with an average
162 pumping rate of 5 – 10 L/min. Immediately after the push phase 120 L of chaser solution was
163 injected with the same pumping rate to replace the well volume (95 L). The following
164 extraction of a 10 times higher volume of 10367 to 10404 L ensured tracer recovery. The tests
165 took place with a descent recovery time of at least 12 hours in between to ensure the recovery
166 of the natural aquifer conditions by the time of the next experiment. Mass recovery for the
167 tests varied from 65 to 85 % depending on the setup. The identical tests “PPT1” and “PPT2”
168 were used for the current study to identify aquifer parameters.

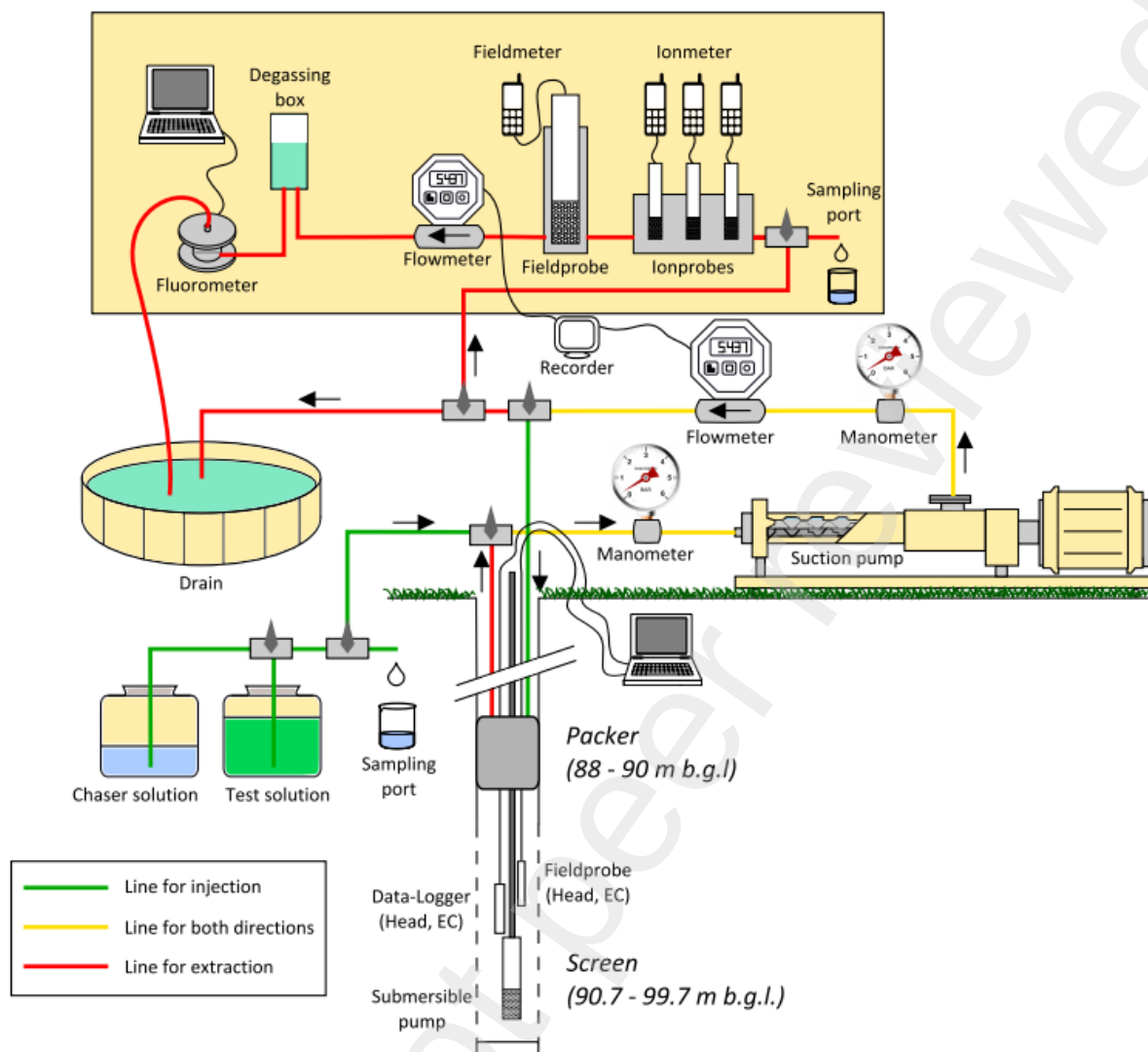


Figure 1: Experimental setup for the PPTs in the well DD-2 (after Zeilfelder et al., 2015).

2.2. Tracers

In the push-pull tracer tests, uranine, iodide, lithium, and temperature were simultaneously injected in the aquifer. Heat transport is known to exhibit retardation behaviour similar to sorption (Section 3). While iodide and uranine are generally regarded as conservative tracers (Adams and Davis, 1991), studies have noted weak sorptive behaviour with retardation factors ranging from 1.0 to 2.0 (Breuer, 2016). Lithium, as an alkali metal, typically exhibits minimal to no retardation in sandy aquifers due to its weak sorption onto the according aquifer material, represented predominantly by quartz sand with low organic carbon content. If sorbed, the retardation factor for lithium is close to that of a conservative tracer, typically ranging from 1.0 to 1.6, attributed to its small ionic size and limited interaction with the

180 mineral matrix (Fuentes et al., 1989; Mojid and Vereecken, 2005). In our field tests
 181 considered, the breakthrough curves of lithium and uranine exhibit very similar close to
 182 conservative behaviour while the iodide breakthrough may be retarded in comparison to
 183 cations (Figure 2, top) in porous media mainly composited of quartz. Based on this, we tested
 184 two calibration scenarios: (i) conservative behaviour for all solutes (ii) conservative behaviour
 185 for lithium and uranine and weakly sorbing behaviour of iodide; (iii) weakly sorbing
 186 behaviour for all solutes (Table 2). Given the relatively short duration of the experiments (6.5
 187 hours for injection and 35 hours for extraction), we assume linear sorption behaviour to be
 188 valid over the timescale of the tests. The pore diffusion coefficient for the solutes was set to
 189 $1 \cdot 10^{-9} [\text{m}^2 \text{s}^{-1}]$ while the thermal diffusion coefficient was set to $1 \cdot 10^{-6} [\text{m}^2 \text{s}^{-1}]$.

190 Table 1. Parameter ranges for the model calibration.

Parameter		Unit	Lower bound	Upper bound
Porosity	n	-	0.01	0.3
Solute dispersivity	α_l	m	0.001	0.2
Thermal dispersivity	α_T	m	0.001	0.8
Iodide retardation factor	R_I	-	1	2
Uranine retardation factor	R_U	-	1	2
Heat retardation factor	R_T	-	computed from porosity (Eq. 8)	
Lithium retardation factor	R_{Li}	-	1	2

191

192 2.3. Approximate confidence intervals

193 PPT 1 and PPT 2 were conducted under identical experimental conditions to facilitate
 194 an estimation of measurement accuracy and to evaluate approximate confidence intervals for
 195 the observed data. The measurements and their difference for each tracer are illustrated in
 196 Figure 2, where the bottom panel displays the deviation between values recorded during PPT 1
 197 and PPT 2 at each time step, and the top panel presents the corresponding breakthrough
 198 curves. The maximum observed deviations were around 5% for temperature, 4% for uranine,
 199 8% for iodide, and 2% for lithium.

200 To account for a conservative calibration scenario, the maximum value of a duplicate error
 201 was assigned as an approximate confidence interval for each component. These confidence
 202 intervals were then incorporated into a calibration strategy to normalize the root mean square
 203 error (RMSE) as an objective function.

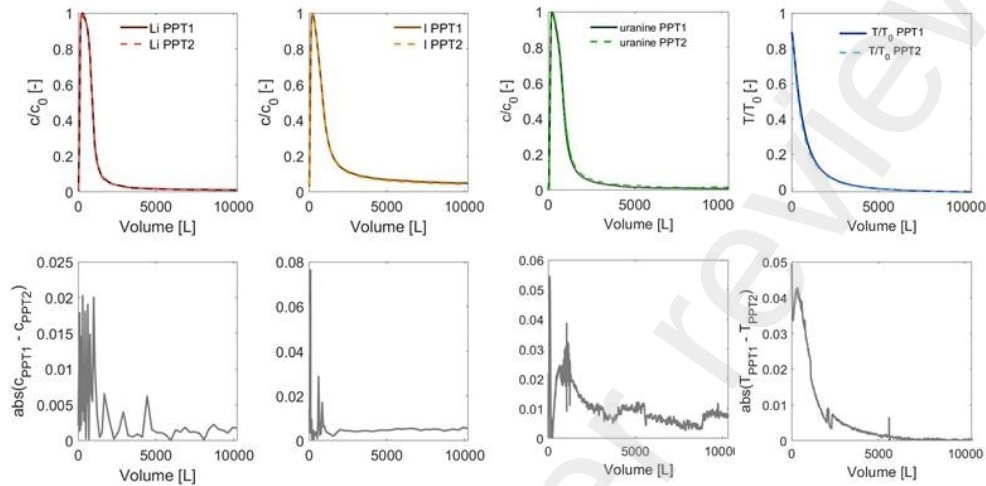


Figure 2: Top: breakthrough curves of lithium, iodide, uranine, and temperature (from left to right) during the extraction phase of duplicate PPTs performed in Hokkaido, Japan; bottom: absolute concentration difference between two duplicate PPTs. Concentrations are normalized in order to have the same maximum value considering mass recovery.

204 3. Process-based model of the push-pull test

205 Push-pull tests (PPTs) are single-well tests and as such an established and broadly
 206 applied technique for in situ characterization of effective transport parameters of the near-well
 207 area. The advantages of a PPT compared to multi-well tests include: easier operation, lower
 208 cost, and greater efficiency. However, the aquifer volume, which is sampled, is typically
 209 smaller with such a single-well test than with tracer tests involving multiple injection,
 210 extraction and observation wells. During the push phase, water is injected in a screened well,
 211 then one borehole volume of tracer-free water is injected as chaser, and during the pull phase,
 212 water is extracted from the borehole (Figure 3). The concentration/temperature is being
 213 monitored throughout the whole experiment. This relatively basic setting is complemented by
 214 using different kinds of tracers – conservative, decaying, sorptive, or heat. Based on the tracer
 215 properties, their corresponding effective transport parameters are evaluated. Analytical
 216 solutions are widely utilized for inferring parameters of relatively simple test setups (constant

input, linear behaviour), whereas numerical modelling is usually used for more sophisticated setups such as non-linear reaction rates, heterogeneous media, or transient tracer inputs. Conservative tracers allow the estimation of longitudinal dispersivity and porosity or groundwater flow velocity, while the comparison with a sorbing tracer yields a solute-specific retardation coefficient, and reactive tracers reveal properties like reaction rates, aquifer reactive capacities, or, by inverse modelling, Monod- and Michaelis-Menten coefficients etc. (Gelhar, L.W, Collins, 1971; Snodgrass and Kitanidis, 1998; Schroth and Istok, 2005; Boisson et al., 2013; Kruisdijk and Breukelen, 2021).

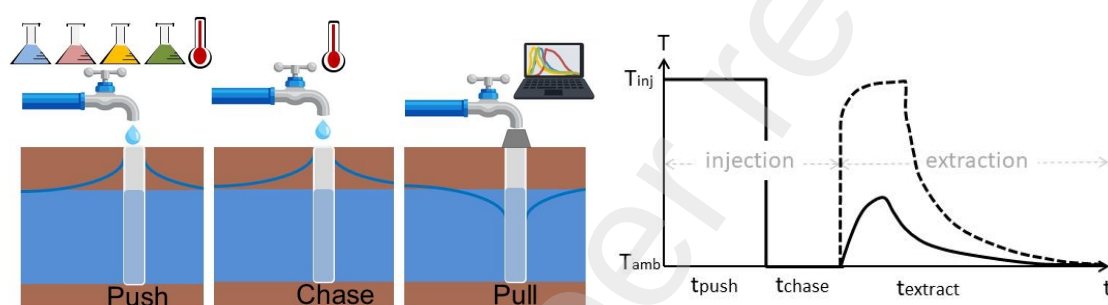


Figure 3. Conceptual representation of a multi-tracer PPT with push, chase and pull phases (left), and corresponding sketch of a typical breakthrough curve of a single tracer (heat via temperature as primary variable in this case) observed the borehole (right); dashed black curve corresponds to almost conservative tracer whereas solid line is a retarded tracer.

3.1. Model equations

The physical model describes PPTs in a fully screened well in a confined aquifer, where flow is solely induced by pumping and the ambient hydraulic gradient is neglected. As such, we aim to reduce the dimensionality of our transport problem via considering the advection-dispersion equation (ADE) for a sorbing solute in cylindrical coordinates. This is a classical assumption together with the one assuming spherical coordinates. We neglect all fluxes in the vertical, which was our Cartesian axes, if considered, and assume the same concentration over depth. Moreover, we neglect transverse dispersion and end up with a radially symmetric equation derived from cylindrical coordinates from which we start our further consideration (compare e.g. Moench et al., 1989; Leiji et al., 1991; Hwang, 2021):

$$n_e R \frac{\partial c}{\partial t} + q_r \frac{\partial c}{\partial r} - \frac{1}{r} \frac{\partial}{\partial r} \left(n_e D_r r \frac{\partial c}{\partial r} \right) = 0$$

239 (1)

240 where $n_e = n[-]$ is effective porosity, $R [-]$ is the retardation factor, $c [N L^{-1}]$, $t [T]$ is
 241 time, $q_r [L T^{-1}]$ is the specific discharge in radial direction, $r [L]$ is the radial coordinate in
 242 space, and D_l is the longitudinal dispersion coefficient as simplification of the dispersion
 243 tensor:

$$244 \quad D_l = \alpha_l |v_r| + D_p$$

245 (2)

246 where $v_r [L T^{-1}]$ is the seepage velocity as $v_r = q_r/n_e$, and $D_p [L^2 T^{-1}]$ is the pore diffusion
 247 coefficient including an estimation of tortuosity, and $\alpha_l = \alpha$ is the longitudinal dispersivity. In the
 248 following, we assume that the effective porosity is constant throughout the physical domain,
 249 thus we divide by the porosity to get rid of the additional pre-factor in front of the temporal
 250 derivative and operate with the seepage velocity in the advective term. Moreover, we neglect
 251 the spatial variability of dispersion, thus $(\partial D_l / \partial r) (\partial c / \partial r) = 0$. Finally, we observe the
 252 emergence of the term $(D_l / r) (\partial c / \partial r)$, which acts as correction term to advection due to
 253 cylindrical coordinates. However, together with our other assumptions and the assumption
 254 that α_l is a constant factor independent of seepage velocity, we observe that in a numerical
 255 model very close to well advective flow may get effectively reverted due to this factor for
 256 high dispersivity, which is an unphysical behaviour. Several authors emphasized that
 257 Scheidegger's model of dispersion (Equation 2) may be wrong for high dispersivity close to a
 258 pumping well (Parker and van Genuchten, 1984; Dagan and Bresler, 1985; Moench et al.,
 259 1989). PPTs are acting on relatively small spatial scales, thus the near-field of the well is
 260 important to us, requiring an adequate numerical resolution in space. To avoid unphysical
 261 behaviour close to a well, we thus neglect the correction factor inducing spurious flow
 262 reversal and employ the classical radial advection-dispersion equation (Wang and Crampon,
 263 1995; Schroth et al., 2001; Chen, 2010; Huang et al., 2010; Kang et al., 2015):

$$264 \quad R \frac{\partial c}{\partial t} + v \frac{\partial c}{\partial r} - D_l \frac{\partial^2 c}{\partial r^2} = 0$$

265 (3)

266 We assume radial flow induced by the well over the area of a cylinder, which implies
 267 divergent flow in the radial direction, such that the seepage velocity becomes:

$$v_r = \frac{Q_w}{2\pi r h_a n_e}$$

(4)

where Q_w [$L^3 T^{-1}$] is the volumetric pumping rate of the well, where different signs lead to flow reversal in radial direction, and h_a [L] is the thickness of the aquifer. With $A = (Q_w/2\pi h_a n_e)$, and dividing by R we end up with our model equation for a solute:

$$\frac{\partial c}{\partial t} + \frac{A}{rR} \frac{\partial c}{\partial r} - \left(\frac{A\alpha_l}{rR} + \frac{D_p}{R} \right) \frac{\partial^2 c}{\partial r^2} = 0$$

(5)

For retardation, we solely consider linear sorption, which may be justified by the low retardation coefficients considered in our study for the according tracers and by the high induced flow velocities as well as considering a single continuum porous medium only. Thus, the retardation factor for solutes is:

$$R = 1 + \frac{(1 - n_e)}{n_e} \rho_s K_d = 1 + \frac{\rho_b K_d}{n_e}$$

(6)

where ρ_s [$M L^{-3}$] is the solid density of the grains, ρ_b [$M L^{-3}$] is the dry bulk density of the porous medium, and K_d is the linear sorption coefficient in equilibrium. Please note, that in our model application, we consider not c but solely normalized concentration c/c_0 [-], where c_0 is the input concentration in the well.

For heat transport, we rewrite Equation (5) considering temperature T [$^\circ$]:

$$\frac{\partial T}{\partial t} + \frac{A}{rR_T} \frac{\partial T}{\partial r} - \left(\frac{A\alpha_T}{rR_T} + \frac{D_T}{R_T} \right) \frac{\partial^2 T}{\partial r^2} = 0$$

(7)

where α_T [L] is the dispersion length for temperature, D_T is the diffusion coefficient for temperature, and R_T [-] is the retardation factor for temperature, which is defined as follows:

$$R_T = \frac{n_e p_w C_{p,w} + (1 - n_e) \rho_s C_{p,s}}{n_e p_w C_{p,w}}$$

(8)

where ρ_w [M L⁻³] is the density of water, $C_{p,w}$ [L² T⁻² Θ⁻¹] is the specific heat capacity of water, and $C_{p,s}$ [L² T⁻² Θ⁻¹] is the specific heat capacity of the grains, which equal the solid matrix in the considered porous medium.

3.2. Numerical implementation

For solving the governing equations (5) and (7) we combine a fast and stable finite-difference approximation with an explicit Euler scheme on a regular grid. This has the advantage that we do not need to assembly storage and mobility matrices and can thus just push an initial condition forward in time without the need to solve a linear equation system. The initial normalized concentrations in solute transport are zero in the domain while for temperature the initial temperature is background temperature. In the following, all our considerations are for a quantity $a \in \{c, T\}$ which is a solute concentration, c , or temperature, T .

We define an effective seepage velocity v_{eff} , which is valid for either concentration or temperature:

$$v_{eff} = \begin{cases} \frac{A}{rR}, & \text{if } a = c \\ \frac{A}{rR_T}, & \text{if } a = T \end{cases}$$

(9)

Note that that in our case the velocities are either from left to right (having a positive sign) or from right to left (having a negative sign). The exact value for v_{eff} depends on r . For advection, we use an upstream finite-difference scheme, where the effective seepage velocity for the flux between the nodes i and $i - 1$, where numbering is from left to right, considering i is:

$$v_{eff,i}^{up} = \begin{cases} v_{eff,i-1}, & \text{for } v_{eff} > 0 \\ v_{eff,i}, & \text{for } v_{eff} < 0 \end{cases}$$

(10)

where $v_{eff,i}$ is the effective seepage velocity at node i , $v_{eff,i-1}$ is the effective seepage velocity at node $i-1$, $v_{i \rightarrow i-1}$ is the seepage velocity at node i towards the node $i - 1$, and $v_{i-1 \rightarrow i}$ is the seepage velocity at node $i - 1$ towards the node i . Note that that in our case the velocities are either from left to right (having a positive sign) or from right to left (having a negative sign). Note that seepage velocity gets reverted in orientation for the pull phase compared to the push

phase, node orientation thus needs to be accordingly. For a Courant number of unity the upstream finite-difference scheme is shock-capturing for advection, for a larger Courant number the scheme is instable, and for a lower Courant number numerical diffusion is introduced, which can be minimized by grid refinement. The approximation of the advective flux together with an explicit Euler discretization related to a node i reads:

$$J_{adv} = \begin{cases} \frac{v_{eff,i}^{up} \Delta t}{\Delta x} (a_{i-1}(t) - a_i(t)), & \text{for } v_{eff,i}^{up} > 0 \\ \frac{v_{eff,i}^{up} \Delta t}{\Delta x} (a_i(t) - a_{i-1}(t)), & \text{for } v_{eff,i}^{up} < 0 \end{cases} \quad (11)$$

where Δx is the spatial increment and Δt is the global time-step size. For the approximation of dispersion, we employ central finite differences. First, we define an effective dispersion coefficient:

$$D_{eff} = \begin{cases} \frac{A\alpha_l}{rR} + \frac{D_p}{R}, & \text{if } a = c \\ \frac{A\alpha_T}{rR_T} + \frac{D_T}{R_T}, & \text{if } a = T \end{cases} \quad (12)$$

Thus, a central differentiation scheme for the dispersive flux considering node i reads:

$$J_{disp} = \frac{D_{eff,i} \Delta t}{\Delta x} (a_{i-1}(t) - 2a_i(t) + a_{i+1}(t)) \quad (13)$$

Finally, our overall scheme for an internal node i reads for a quantity $a \in \{c, T\}$, which can be either concentration or temperature:

$$a(t + \Delta t) = a(t) + J_{adv} + J_{disp} \quad (14)$$

Moreover, we employ free inflow and outflow boundary conditions. For solute transport, we consider a constant normalized inflow concentration of unity and account for well-bore storage, whereas for heat, we consider an input time-series for temperature. In such a scheme it is not necessary to explicitly account for well-bore storage for temperature propagation. To render the scheme to be fast and stable, while not overly introducing spurious diffusion, our scheme computes an optimal global time step on a fixed grid size, Δx , as such:

$$\Delta t = \min_{\forall i \in N} \left(\frac{\Delta x}{v_{eff,i}}, \frac{\Delta x^2}{3D_{eff,i}} \right) f$$

where N is the set of nodes, the first entry in the vector is the optimal timestep size in an explicit Euler scheme for advection based on the Courant number, which needs to be unity at maximum, whereas the second is the optimal time-step size due to dispersion based on the condition that the Neumann number should be less or equal $1/3$ for a stable approximation of diffusion/dispersion. For stability reasons the smaller value of the two values needs to be taken such that at the end the smallest time step for all nodes considering advection and dispersion is chosen to be the global time step for the next iteration. Expression (15) is eventually multiplied with a dimensionless safety factor f , slightly smaller than unity for enhanced stability and to account for numerical precision. The chosen global time step is the smallest time step for advection and dispersion over all nodes. We implemented the described methodology in Matlab 2024a.

3.3. Surrogate model formulation

Parameterization of the advection-dispersion equation (ADE) to fit the observed experimental breakthrough-curve data is typically done by optimizing an objective function such as RMSE which reflects on a match between the model and the observed data. For breakthrough-curve data obtained from natural geological media, this may involve high-dimensional parametric spaces, non-convex objective functions, and inherent uncertainties as well as unresolved heterogeneities, which potentially make deterministic methods stuck in local optima or fail to capture the variability of the solution space (Cirpka and Valocchi, 2016; Fiori et al., 2016). Overcoming these limitations is possible by stochastic modelling (Kitanidis, 1998), which can effectively handle multimodal functions and may include an uncertainty estimate. However, such stochastic optimization techniques are often limited by the high computational costs associated with ensemble simulations of physics-based numerical models. To minimize the time required for the optimization process, a surrogate model capturing the relationship between aquifer parameters and the objective function was applied. Surrogate modeling emerged as a powerful approach to mitigate computational speed limitations by employing an approximation of the model to be fitted, which is the surrogate model. This enables efficient stochastic parameter estimation and uncertainty quantification (Erdal et al., 2020; Allgeier, 2023; Rohmer et al., 2023; Ershadi et al., 2024). Among the different surrogate modelling approaches available, Gaussian process regression (GPR) was selected since it preserves the continuous and smoothed-out nature of concentrations fields in

natural porous media and can include a combination of several multiparametric ADEs (Degen et al., 2023). The principles and mathematical formulation of GPR originates from the kriging principles, applied to multidimensional parameter spaces (Allgeier and Cirpka, 2023; Rohmer et al., 2023).

The continuous input parameters varying within site-specific ranges (Table 1) are solute and heat dispersivity treated separately, porosity as an aquifer property, as well as individual solute and heat retardation factors for each component. The response for each tracer is the sum of squared errors normalized by confidence intervals residuals and number of points considered:

$$RMSE(\mathbf{p}) = \sqrt{\frac{\sum_{i=1}^N (y_{mod,i} - y_{obs,i})^2 / CI}{N}}, \quad (18)$$

where $\mathbf{p}$ is the set of parameters to be optimized in the solute or heat transport model, $y_{mod,i}$ and $y_{obs,i}$ refer to the modelled and the observed value at the discrete points i in time considering the breakthrough curves, where the modelled values depend i.a. on $\mathbf{p}$, N is the total number of simulated/measured values, and CI is an approximated confidence interval. The differences between observed and modelled values within the confidence interval are considered equally acceptable due to the resolution limits of the measurements.

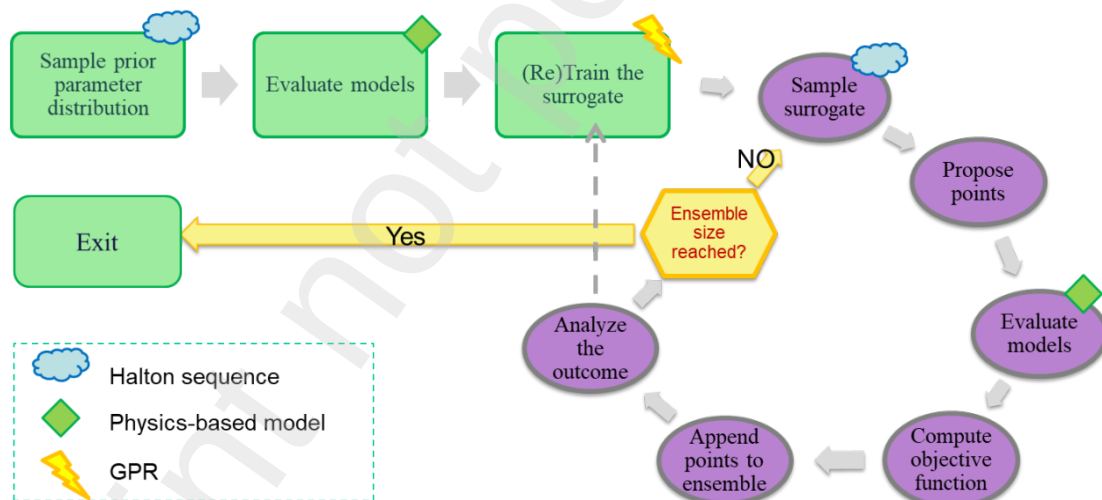
The best-estimate calibration aims for all tracers to minimize the objective function f_{obj} , which is the deviation between measured and modelled values considering the maximum RMSE:

$$f_{obj} = \max((RMSE(\mathbf{p}))_{j=1,\dots,n}), \quad (19)$$

where n is the number of considered solutes plus temperature. For calibration, only favourable unique parameter combinations are sampled in order to obtain an ensemble of a defined size. Once trained, the GPR is employed to propose potentially plausible parameter combinations by computing the probability of satisfying the condition (Equation 19).

Possible parameter combinations are drawn using a Halton sequence using an adaptive sampling approach to ensure the local optimization and global accuracy of the surrogate model. The infill criteria for the new samples focus on a combination of exploitation of high-probability regions (indicated by the GPR model) with exploration of regions where the model demonstrates high uncertainty. At every iteration 20 initial candidate parameter

407 combinations are drawn from the Latin hypercube sampling (Figure 4). To balance between
 408 exploitation and exploration, we proportionally added a small amount of exploration
 409 (searching for new regions) and large amount of exploitation (refining predictions in
 410 promising areas). The selected candidate parameter combinations are then optimized using the
 411 prediction capacity of the surrogate model. The interior-point algorithm, which is suitable for
 412 high-dimensional constrained problems (Byrd et al., 2000), is employed for optimization of
 413 GPR by minimizing the predicted RMSE value. In addition, a secondary small set of five
 414 candidate parameter combinations is generated based on sampling a large (10000 realizations)
 415 sample of surrogate models and predicting their likelihood of being optimal (Erdal et al.,
 416 2020). The cumulative distribution function (CDF) was employed for sampling to identify
 417 points in the input space where the predictive error is expected to be close to the lowest
 418 RMSE. From the set of 10000 GPR realizations and corresponding CDF values, five points
 419 with the highest probability were selected to prioritize the points which are most likely to
 420 improve the surrogate model and reduce the uncertainty further.



421

Figure 4. Flowchart for the surrogate modelling process for optimization of advective-dispersive transport parameters of a PPT; the purple circles refer to the main adaptive sampling loop.

422 The dynamically improving surrogate modelling strategy is outlined in the flowchart
 423 (Figure 2) and is similar to (Allgeier and Cirpka, 2023): (i) Sample a uniform prior of the
 424 Latin hypercube, (ii) train the GPR based on the model response, (iii) Based on the GPR

results, compute the probability of satisfying the condition (Eq.19) in n random locations in the parametric space. Select points with the highest probability; (iv) create additional (small) sample by a 1D random walk in each parametric space for existing favourable samples, (v) for the selected points, run the full model and add new ensemble members to the training set; (vi) periodically retrain the model with the updated training data. Steps 2 – 5 are repeated until the ensemble is filled by successful model runs. To account for measurement uncertainty, all members of the posterior parameter distribution are treated as equally valid, and thus no ranking was performed within the parameter distributions that satisfy Equation 19. The prior was sampled from the normalized parameter space using the Halton quasi-random sequence with reverse-radix scrambling (Mascagni and Chi, 2004).

The Figure 5 (right) shows an exemplary response surface of an example calibration for a single conservative tracer which requires only two parameters to be estimated: dispersivity and porosity based on the analytical solution of Schroth and Istok (2005). Thanks to the simplicity of the response surface (Fig.3, right), the acceptance rate increases drastically already after 100 model runs (Fig. 3, left). As shown in analytical solutions (Schroth and Istok, 2005), the dispersivity inversion from the push pull test is plausible while porosity may leave a broad range of possible realizations. At the same time, the area of the minimum RMSE on the response surface (Fig .3, right) covers up the whole prior range of possible porosity values revealing lower sensitivity of the output to the porosity. This consideration is kept for calibration of the sorbing solutes and heat.

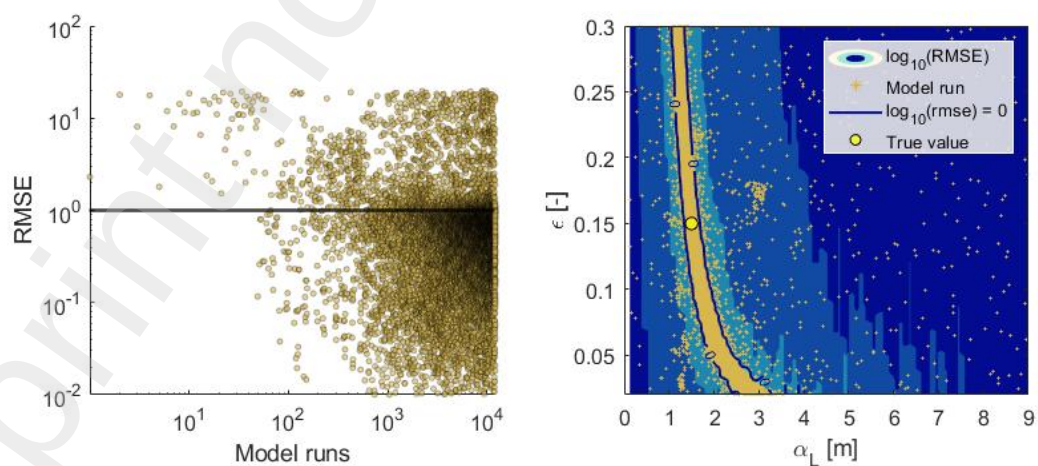


Figure 5 Calibration of conservative tracer data using the analytical approximation of (Schroth and Istok, 2005). Left: Performance progress of the calibration is shown reflecting the RMSE decrease with increase of model runs leading to surrogate model specification. The solid line refers to the acceptance threshold of the normalized RMSE=1, where all model

results below it are considered plausible and added to ensemble. Right: Response surface which was created based on all model runs. The colour code refers to the decadic logarithm of the RMSE with the darker colours (blue) corresponding to higher RMSE values and the lighter colours (yellow) representing smaller RMSE values, the yellow dots denote the model runs and the bold yellow circle denotes the “true” value based on the values used to generate a hypothetically true parameter combination.

The ensemble size was defined based on a stop criterion (Zhu et al., 2023) which in the current application corresponds to 600 plausible samples. To validate the suggestion of Zhu et al. (2023), the ensemble size was gradually increased from 50 to 700 plausible samples to confirm no further change in the posterior distribution.

The surrogate model performance was evaluated using 20% holdout of the data which did not contribute to training by evaluating the coefficient of determination:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_{obs} - y_{pred})^2}{\sum_{i=1}^n (y_{obs} - \bar{y})^2}$$

where y_{obs} and y_{pred} denote the observed and predicted RMSE values, respectively, and $\bar{y}$ denotes the mean of the observed RMSE values.

4. Results

This section outlines performance metrics of the surrogate modelling and numerical calibration results that reveal the sensitivity of the model to dispersivity, porosity, and retardation factors demonstrating the influence of using multiple tracers on calibration quality of effective transport parameters.

4.1. Surrogate model evaluation

The optimization constitutes a three-dimensional optimization space for all solutes being conservative (porosity, longitudinal diversities of solutes and heat are unknown), a four-dimensional space for the case when lithium and uranine are considered conservative tracers (retardation factor of iodide is added to the unknowns), and a six-dimensional space in case all solutes are sorbing (retardation factors of lithium and uranine are added to the unknowns).

The hyperparameters of the GPR were optimized using Bayesian optimization (Snoek, et al., 2012) with a holdout 20% of the total data for validation. Special attention was given to

sampling data points with lower RMSE values, as these regions are crucial for precise optimization in the area of interest. The optimized length scale of 0.87 (not shown) indicates that the model assumes a relatively smooth underlying function which is typical for the ADE (Degen et al., 2023). The relatively high variance of 9.0084 corresponds to the non-normalized RMSE array resulting in substantial variability, which the model is capturing. The noise standard deviation 0.0246 is small compared to the signal variance, meaning that the model is performing well at distinguishing the underlying trend from noise in the data.

The performance improvement of the surrogate model is depicted in Figure 6. The right plot presents the distribution of RMSE values, reflecting that the majority of RMSE values are concentrated between 3 and 5, with a peak probability density of ~ 0.6 . This indicates that, while some variability in performance exists, a significant portion of the predictions achieves low RMSE values, which enables using GPR for proposing the plausible parameter combinations. The parity plot at the end-state of GPR training (Fig. 6, left) illustrates good agreement between the actual RMSE values running the physical model and the predicted values from the GPR model on the holdout set with R^2 of 0.9985. The predictions generally follow the 1:1 line, indicating a reasonable correlation between the actual and predicted RMSE values. Although some deviation is observed suggesting potential limitations in the model's ability to generalize in this range, the trained model is sufficient to propose the potentially plausible parameter combinations leading to smaller RMSE values.

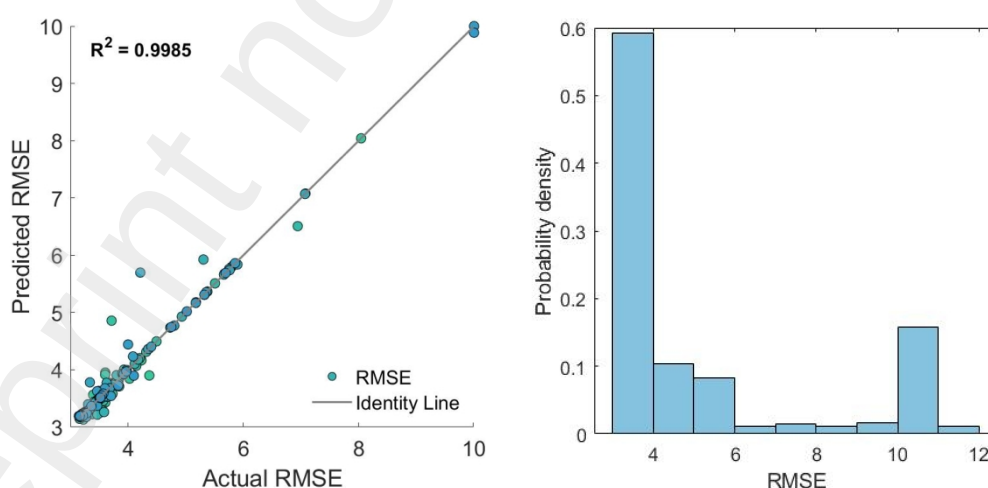


Figure 6 Performance evaluation of the GPR model considering all tracers exhibit sorptive behaviour. Left: parity plot between surrogate model prediction and the holdout dataset of the

physics-based model outcomes; right: distribution of the RMSE based on the physics-based model evaluation.

4.2. Application to Hokkaido site and posterior parameter distributions

Lithium, iodide, and uranine are generally considered to be conservative tracers. However, several studies have shown that they may exhibit slightly sorptive behaviour under some conditions (Breuer, 2016; Fuentes et al., 1989; Mojid and Vereecken, 2005). In the calibration strategy, we studied the importance of accounting for the possible sorption of the tracers for parameter inference. As such, three calibration scenarios were considered: (i) all three tracers, lithium, iodide, and uranine are exhibiting sorptive behaviour; (ii) only iodide is sorbing whereas lithium and uranine are conservative; (iii) all tracers are conservative on the time scale of the experiment.

4.2.1. Lithium, iodide, and uranine are sorptive tracers

This scenario corresponds to a hypothesis of slightly sorptive behaviour of all solutes considered, i.e., lithium, uranine, and iodide, and corresponds to calibration of their retardation factors together with porosity, solute dispersivity, and thermal dispersivity. This case corresponds to the highest amount of six parameters in total for calibration, which define the quality of breakthrough curves (BCs) for each tracer. To calibrate the model, we used the breakthrough curves of iodide, uranine, lithium, and temperature to estimate the normalized RMSE.

The calibrated breakthrough curves demonstrate generally a good fit to the data meeting the critical points - maximum concentration and the half-front for all solutes (Fig 7). The observed breakthrough curves of iodide and uranine expose higher tailing concentrations at the later times of the extraction which could not be captured by the numerical model. The model of lithium displays lower tailing concentrations for the later times with slightly increased concentrations directly after the peak concentration. The higher concentrations measured especially for iodide may correspond to an overall larger amount of injected mass compared to the input assumptions. Also, a higher measurement error as assumed may play a role as well as retention of the component associated with aquifer heterogeneity or the in-fact existence of a dual porosity medium including a micro-porosity e.g. only available for anions, which is not included in the current physical model. Notably, the temperature (Fig. 7, bottom-

right) exhibits the opposite behaviour – the modelled concentration fits very well the higher relative temperatures while exceeding the measured values for later times.

Posterior distributions of parameters are depicted in Figure 8. As expected from the analytical solutions and sensitivity studies (Schroth and Istok, 2005; Maier and Kocabas, 2013; Shi, Wang and Zhan, 2020), the posterior distributions of solute dispersivity and thermal dispersivity narrowed down to a much smaller range than the uniform prior. Posterior parameter distributions of the Horonobe aquifer reflect the effect of high sensitivity to dispersivity leading to substantial uncertainty reduction (Fig. 8). The identified thermal dispersivity values belong to the range between $1 \cdot 10^{-3}$ m and 0.2 m with the mean value around 0.1 m, while the solute dispersivity is, as expected, characterized by lower values less than 0.05 m. Considering the injected radius may vary from 0.5 to 0.9 m depending on the porosity, the dispersion of solutes corresponds to 6-10% of the spread. Porosity is a less sensitive parameter and tends to be within the broad range of 0.05 to around 0.19 in comparison to 0.01 to 0.3 as a prior distribution while being most likely around 0.1 according to the calibration results. The posterior distribution of heat retardation factor as a function of porosity (Equation 8) represents a narrow interval between 2.3 (for porosity 19%) to 7.4 (for porosity 5 %).

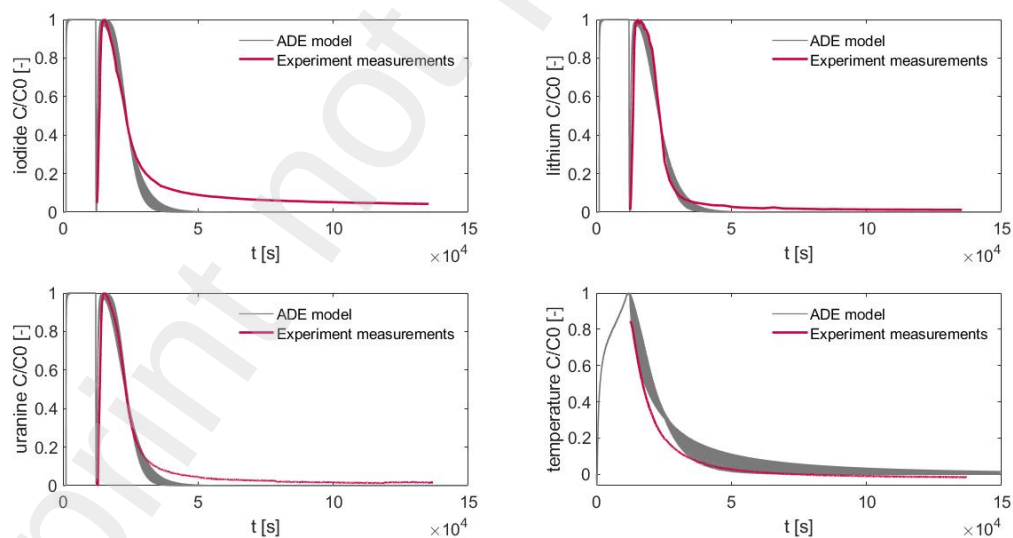


Figure 7 Ensemble optimization results considering iodide, lithium, and uranine to exhibit sorptive behaviour. Breakthrough curves for the solutes and temperature; red lines refers to the concentration measurements during the extraction phase of the PPT, grey lines refer to the fitted modelled data for the extraction period, time equals to 0 corresponds to the start of the push phase.

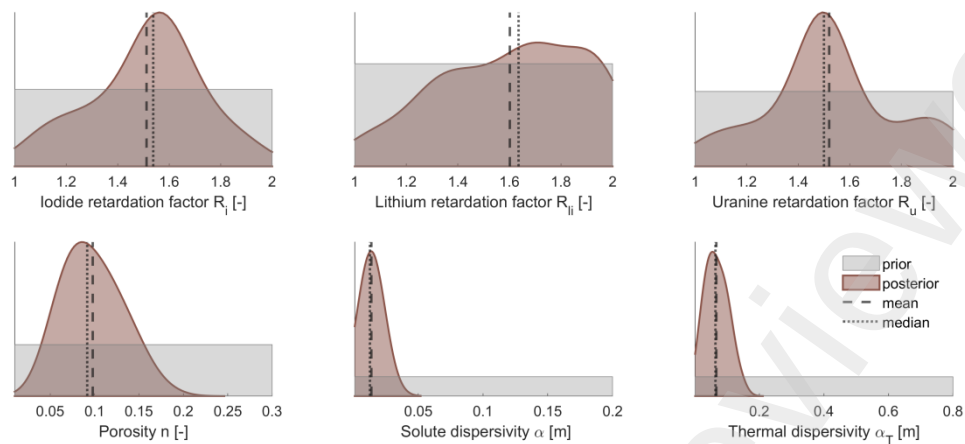


Figure 8 Posterior distributions of calibrated retardation factors of iodide, lithium, uranine, porosity, solute and thermal dispersivity considering lithium, uranine, and iodide to exhibit sorptive behaviour; vertical dashed lines indicate the mean parameter values and the dotted lines indicate the median.

The retardation factors of solutes are not very sensitive to the calibration which highlights the lowest sensitivity of the model to its variation (Schroth and Istok, 2005). However, they tend to be around 1.5 for iodide and uranine and around 1.6 for lithium, still the full prior range between 1.0 and 2.0 remains possible. Thus, the model shows a low sensitivity to retardation factors of all tracers – lithium, iodide, and uranine, making their transport behaviour inconclusive from the breakthrough-curve observations. Although representing an uncertainty in tracers' behaviour, this fact suggests to reduce the number of parameters to solute and thermal dispersivity, and porosity in the following, thus simplifying the calibration process of the latter parameters.

4.2.2. Lithium and Uranine are conservative tracers while Iodide exhibit slightly sorptive behaviour

Lithium and uranine are typically considered conservative in sandy aquifers because lithium due to its small molecule size and positive charge does not interact significantly with quartz, while uranine being neutral at lower pH values and an anion at higher pH values, is not significantly interacting with the often slightly negatively charged sand surfaces (Jada et al., 2006). In contrast, iodide may exhibit slight sorption due to its high polarizability,

allowing weak van der Waals interactions or surface complexation with minerals like iron oxides or organic matter. This difference in behaviour was considered in this calibration scenario. Under the assumption of conservative tracer behaviour of lithium and uranine, the four other parameters were calibrated which are: retardation factor of iodide, aquifer porosity, solute dispersivity, and thermal dispersivity.

The posterior parameter distributions (Figure 9) for dispersivity of solute and heat are generally similar to the previous scenario, however, solute dispersivity shows a narrower range than before with a mean value around 0.03 m. Thermal dispersivity, while overall being very similar, shows a slightly higher likelihood for smaller and larger values compared to the former results. The posterior distribution of porosity is a bit narrower now allowing also for higher porosity values. However, the most likely value is a little bit smaller than 0.1, which is a little bit less compared to the former results. Since smaller porosity values increase seepage velocity, this causes an earlier breakthrough of the tracer. Therefore, the model compensates by reducing the retardation factor a bit to maintain the correct arrival time in the breakthrough curve. The corresponding heat retardation factors vary from 2.0 to 15.6 which corresponds to a broader range and higher heat attenuation.

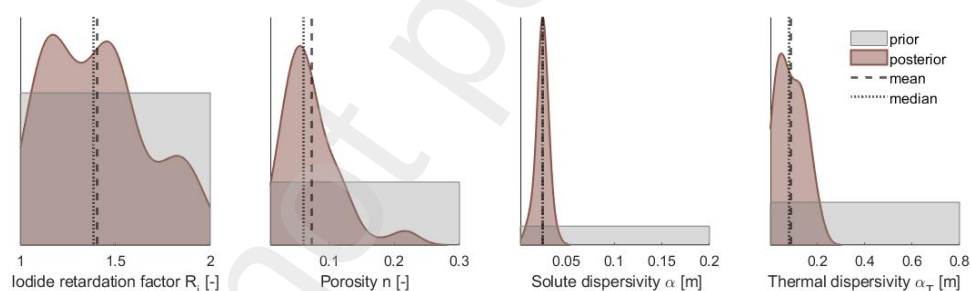


Figure 9 Posterior distributions of calibrated retardation factor of iodide, porosity, solute and thermal dispersivity considering lithium and uranine to be conservative tracers and iodide to exhibit sorptive behaviour; vertical dashed lines indicate the mean parameter values and the dotted lines indicate the median.

4.2.3. Optimization results considering all tracers to be conservative

Due to the relatively small duration of a push pull test and relatively low expected sorption to the quartz minerals of sand, all tracers may have experienced in fact a conservative

behaviour. In this case, only three parameters need to be calibrated – solute and thermal dispersivity as well as porosity.

Posterior distributions of parameters are shown in Figure 10. The histograms indicate the posterior distributions of the calibrated parameters, constrained within the predefined prior ranges. While solute dispersivity shows a very similar pattern compared to the former results, where only iodide was considered potentially sorbing, the distribution of thermal dispersivity shrank with a mean value considerably closer to solute dispersivity and a slight bi-modal shape. The posterior distribution for porosity is similar to the case, where all tracers were conservative allowing for a range of values below 0.2 but having a most likely value of around 0.1 like in the case where all tracers were considered to be sorptive. The value range of heat retardation factors for this scenario is 2.4-5.4. As depicted in Figure 8, the PPT exhibits low sensitivity to the retardation factors of the solute tracers, making it challenging to exactly quantify the physically correct transport parameters based solely on breakthrough curve analysis. Also, a case considering all tracers to be conservative yields reasonable posterior parameter distributions. Otherwise, solute and thermal dispersivities are sensitive parameters to the calibration, whereas porosity shows consistently a similar mean value throughout the calibration scenarios, which is a promising finding. However, the mean values of solute and thermal dispersivities shown in Figure 10 are actually close to each other, which makes sense if dispersivity is considered to be a pure hydrodynamic parameter and thermal and solute diffusion coefficients are chosen correctly.

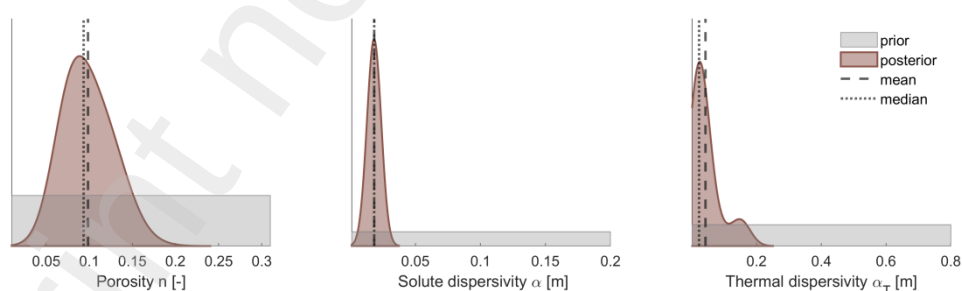


Figure 10 Ensemble optimization results for porosity, solute and thermal dispersivity considering that all tracers exhibit conservative behaviour; vertical dashed lines indicate the mean parameter values and the dotted lines indicate the median.

5. Discussion

This study presents a novel application of GPR for stochastic optimization of a multitracer PPT for inferring effective transport parameters in a sandy aquifer in Horonobe, Japan. A surrogate model-based optimization framework was developed for parameter estimation of PPT data, addressing both parametric uncertainty and approximate measurement confidence intervals. The framework employed GPR to train and predict the objective function based on aquifer and tracer properties, significantly reducing computational costs compared to traditional ensemble-based methods. The GPR training resulted in a good convergence between the modelled and predicted RMSE values confirming the hypothesis that the optimization of a multitracer push pull test can be done by Gaussian processes as the underlying solution is close enough to a Gaussian hypersurface. This provides a benefit in terms of calibration capacity and applicability of the described method.

To evaluate the breakthrough curves of tracers, a 1D ADE solver accounting for the transient tracer input was developed in this study. By assuming homogeneity, the model provided a clear and interpretable baseline for analysing transport behaviour, which can be further refined to account for spatial variability e.g. in porosity or flow velocity coupled with travel time approaches (Selzer and Cirpka, 2020). The solver provides a firm base for understanding fundamental transport dynamics and serves as a computationally efficient tool for simulating advective-dispersive transport in porous media. While its 1D structure simplifies the representation of flow and transport, this approach enables efficient numerical simulations and facilitates sensitivity analyses. To enhance its applicability to complex field conditions, the model could be eventually extended to a dual porosity medium to eventually describe the tailing observed in the breakthrough curves especially for iodide and uranine.

Using three solutes and temperature with potentially different sorption characteristics for calibration of a PPT allowed for determining both, solute and thermal dispersivity, identify the heat retardation factor and slightly narrow down the possible porosity range. The thermal dispersivity varies from $1 \cdot 10^{-3}$ m to 0.2 m with a mean value which is significantly shifted towards lower values for the case that all solutes are considered conservative, which indicates the benefits of performing PPTs with solutes and heat simultaneously for later evaluation. For the cases that all solutes sorb or at least iodides sorbs, solute dispersivity is significantly smaller than thermal dispersivity, while the mean values of both are around the same for the case that no solute is sorbing. Thus, the latter case is consistent with the assumption that dispersivity should be hydrodynamic parameter, if solute and thermal diffusion coefficients

are set about correctly. The heat retardation factors expose some variability from one scenario to another depending on the exact porosity distribution exposing smallest range and minimal value for scenario of all reactive tracers (1.57-2.08), and the highest range and values for the scenario when only iodide exposes sorptive behaviour (1.7 – 4.6). The RMSE is normalized by the approximate confidence interval of the measurements creating a “window” of possible truth values instead of a discrete point, which introduces some spread to the posterior parameter distribution. The more precise the measurements are, the lower will be the uncertainty in the parameter inference.

A GPR trained on RMSE values considering PPT measurements and the ADE model provides a powerful data-driven approach for improving the calibration of a physics-based model of advective-dispersive transport while capturing the parametric uncertainty. By learning from discrepancies between the model results and measurements, the GPR can refine parameter estimations and enhance the interpretability of simulation results. While its performance is influenced by the physical model assumptions delivering the training data, the framework applied in this study offers flexibility also to incorporate more complex transport behaviour including more sophisticated models as dual-porosity modelling approaches. Additionally, while RMSE-based training prioritizes overall error minimization, further refinements can focus on improving localized transport phenomena, such as sharp concentration fronts or early breakthrough times. By integrating additional process knowledge or multi-fidelity approaches, the GPR emulator can be further optimized to represent real-world transport behaviour with greater precision. Overall, the presented approach offers a computationally efficient method for characterizing aquifer properties crucial for ATEs system design and long-term performance prediction via improved effective transport parameter characterization. The presented uncertainty-acknowledged calibration strategy is generally applicable to more complicated models as well with more unknown parameters, considering proportionally increased amount of data or additional knowledge about the aquifer functioning.

6. Conclusions

In this study we developed a parameter optimization methodology based on Gaussian process regression used for improved calibration of a finite-difference based advection-dispersion model of push-pull tests capturing parametric uncertainty. This framework was applied to

analyse push-pull tests performed in a sandy aquifer in Horonobe, Japan, using four tracers (heat, uranine, lithium, and iodide) to infer effective transport parameters. We demonstrated the method's effectiveness in constraining uncertainty intervals for solute and heat dispersivity and to some extent, porosity and retardation factors. The hypothesis of sorbing and non sorbing tracers behaviour was tested by considering calibration scenarios which account for sorption of all tracers, only iodide, and conservative tracer behaviour.

In all calibration scenarios, the solute dispersivity and thermal dispersivity ranges in Horonobe site exhibit almost identical posterior distributions. As such, solute dispersivity constrains from 6% to 10% of the injected radius with the values of up to 0.05m. Thermal dispersivity covers predictably larger range of up to 0.2 m constraining considerable 20% to 40% of the injected radius. Porosity posterior distributions are varying for different scenarios within ranges from 0.04 to 0.19 for all sorptive scenario, 0.02 to 0.28 in iodide only sorbing scenario, and from 0.04 to 0.18 for scenario when all tracers are considered conservative. The calibration scenario analysis of including linear sorption behaviour of the tracers typically assumed to be conservative highlights the value of the conceptual model choice and the bias introduced by the conceptual model selection in interpretation of hydrogeological systems. However, the calibration results for solute retardation factors remain inconclusive with little sensitivity of the model considering solute retardation factors, though assuming all tracers to be conservative lead to similar mean solute and thermal dispersivities, which is consistent with the assumption that dispersivity is a hydrodynamic parameter, if solute and thermal diffusion coefficients are well chosen. For the case where all solutes or only iodide are assumed to undergo slight sorption, solute and thermal dispersivity significantly differ, with thermal dispersivity being larger than solute dispersivity. Still, the inherent biases introduced by choosing a particular conceptual model - whether considering cations as conservative or including sorption - can impact the estimated parameter ranges significantly and, consequently, the interpretation of subsurface processes. The choice to model slight sorption behaviour not only slightly broadens the overall range of acceptable porosity values but also emphasizes the degree to which model selection influences our understanding of key system properties.

The findings of this study suggest that some parameters, such as the dispersivity, exhibit limited variation across model scenarios, being very well inferable. Other parameters such as solute retardation factors are less sensitive to calibration and are closely related to estimated

porosity values due to the physical model resulting in broad porosity distributions and inconclusive results for the retardation factors. One way to further narrow down the uncertainty and to finally estimate if slight sorption may take place for the considered solutes can be the exact measurement of porosity in the lab based on drilling cores narrowing down the broad distribution of possible values of porosity to ideally only one value with only a small uncertainty associated to it, if the medium is assumed to be homogeneous for the aquifer volume the push-pull test samples. Thus, porosity may be a relatively easy to measure proxy to identify whether solute tracers, which are typically assumed to show a conservative non-sorbing transport behaviour, may in fact experience slight sorption even in sandy aquifers or not.

Our study highlights that push-pull test experiments can significantly improve the knowledge about aquifer characteristics like dispersivity, porosity, and heat retardation, which are essential for ATEs design. By identifying flow and transport parameters of the aquifer, push-pull tests enable more accurate long-term predictions of ATEs system functioning. Results however suggest low sensitivity to solute retardation factors, indicating potential limitations in inferring certain parameters from push-pull test data alone and calling for more exact measurements, which can be performed on cores, like exact porosity measurements. By using different types of tracers (conservative, sorptive, heat), push-pull tests allow for a comprehensive evaluation of various aquifer properties relevant to ATEs performance and to improve calibration accuracy capturing the ambiguity inherent in many calibration scenarios and reducing parameter uncertainty.

Acknowledgements

This work was funded the German Federal Ministry of Education and Research (BMBF) project "SpeicherCity", grant number 03G0911.

717 References

- 718 Allgeier, J. and Cirpka, O. A. (2023) ‘Surrogate-Model Assisted Plausibility-Check,
719 Calibration, and Posterior-Distribution Evaluation of Subsurface-Flow Models’, *Water*
720 *Resources Research*, 59(7), p. e2023WR034453. doi: 10.1029/2023WR034453.
- 721 Beernink, S., Bloemendal, M., Kleinlugtenbelt, R. and Hartog, N. (2022) ‘Maximizing the use
722 of aquifer thermal energy storage systems in urban areas : effects on individual system
723 primary energy use and overall GHG emissions’, *Applied Energy*, 311(February), p. 118587.
724 doi: 10.1016/j.apenergy.2022.118587.
- 725 Blöcher, G., Regenspurg, S., Kranz, S., Lipus, M., Pei, L., Norden, B., Reinsch, T.,
726 Henniges, J., Siemon, R., Orenczuk, D., Zeilfelder, S., Scheytt, T. and Saadat, A. (2024)
727 ‘Best practices for characterization of High Temperature-Aquifer Thermal Energy Storage (
728 HT-ATES) potential using well tests in Berlin (Germany) as an example’, *Geothermics*,
729 116(May 2022), p. 102830. doi: 10.1016/j.geothermics.2023.102830.
- 730 Bloemendal, M. and Olsthoorn, T. (2018) ‘Geothermics ATES systems in aquifers with high
731 ambient groundwater flow velocity’, *Geothermics*, 75(January), pp. 81–92. doi:
732 10.1016/j.geothermics.2018.04.005.
- 733 Boisson, A., Anna, P. De, Bour, O., Borgne, T. Le, Aquilina, L., Boisson, A., Anna, P. De,
734 Bour, O., Borgne, T. Le and Labasque, T. (2013) ‘Abstract SC’, *Journal of Contaminant*
735 *Hydrology*. doi: 10.1016/j.jconhyd.2013.02.006.
- 736 Breuer, F. (2016) ‘Investigation of Sorption and Degradation Parameters of Selected
737 Emerging Organic Contaminants under Different Hydrological Conditions’.MS dissertation at
738 Freiburg im Breisgau University
- 739 Byrd, R. H., Gilbert, J. C. and Nocedal, J. (2000) ‘A trust region method based on interior
740 point techniques for nonlinear programming’, *Mathematics Subject Classification*, 185, pp.
741 149–185. doi: 10.1007/s101070000189
- 742 Chen, J. (2010) ‘Analytical model for fully three-dimensional radial dispersion in a finite-
743 thickness aquifer’, *Hydrological Processes*, 24, 945(December 2009), pp. 934–945. doi:
744 10.1002/hyp.7541.
- 745 Cirpka, O. A. and Valocchi, A. J. (2016) ‘Debates—Stochastic subsurface hydrology from

theory to practice: Does stochastic subsurface hydrology help solving practical problems of
contaminant hydrogeology?', *Journal of the American Water Resources Association*, 5(3), pp.
2–2. doi: 10.1111/j.1752-1688.1969.tb04897.x.

Dagan, G. and Bresler, E. (1985) 'Comment on "Flux-Averaged and Volume-Averaged
Concentration in Continuum Approaches to Solute Transport" by J. C. Parker and M. Th. van
Genuchten', *Water Resources Research*, 21(8), pp. 1299–1300. doi:
<https://doi.org/10.1029/WR021i008p01299>.

Degen, D., Voullième, D. C., Buiter, S. Hendricks Franssen, H.-J., Vereecken, H., González-
Nicolás, A., and Wellmann, F. (2023) 'Perspectives of physics-based machine learning
strategies for geoscientific applications governed by partial differential equations',
Geoscientific Model Development, 16, 7375–7409, 2023 doi: 10.5194/gmd-16-7375-2023.

Doughty, C., Hellström, G., Tsang, C. F. and Claesson, J. (1982) 'A dimensionless parameter
approach to the thermal behavior of an aquifer thermal energy storage system', *Water
Resources Research*, 18(3), pp. 571–587. doi: 10.1029/WR018i003p00571.

Erdal, D., Xiao, S., Nowak, W. and Cirpka, O. A. (2020) 'Sampling behavioral model
parameters for ensemble-based sensitivity analysis using Gaussian process emulation and
active subspaces', *Stochastic Environmental Research and Risk Assessment*, 34(11), pp.
1813–1830. doi: 10.1007/s00477-020-01867-0.

Ershadi, A., Finkel, M., Liu, B., Cirpka, O. A. and Grathwohl, P. (2024) 'Ensemble surrogate
modeling of advective-dispersive transport with intraparticle diffusion model for column-
leaching test', *Journal of Contaminant Hydrology*, 267(June), p. 104423. doi:
[10.1016/j.jconhyd.2024.104423](https://doi.org/10.1016/j.jconhyd.2024.104423).

Fiori, A., Cvetkovic, V., Dagan, G., Attinger, S., Bellin, A., Dietrich, P., Zech, A. and
Teutsch, G. (2016) 'Debates—Stochastic subsurface hydrology from theory to practice: The
relevance of stochastic subsurface hydrology to practical problems of contaminant transport
and remediation. What is characterization and stochastic theory good for?', *Water Resources
Research*, 52(12), pp. 9228–9234. doi: 10.1002/2015WR017525.

Fuentes, H R, Polzer, W. L., Essington, E. H. , Newman, B. D. (1989) Characterization of
Reactive Tracers for C-Wells Field Experiments I : Electrostatic Sorption Mechanism ,
Lithium, Los Alamos National Laboratory, Report LA— 11691-MS.

776 Gelhar, L.W, Collins, M. A. (1971) ‘General Analysis of Longitudinal Dispersion in
777 Nonuniform Flow’, *Water Resources Research*, 7(6). doi: 10.1029/WR007i006p01511.

778 Hebig, K. H., Ito, N., Tecklenburg, J., Machida, I., Marui, A. and Scheytt, T. (2016) ‘Use of
779 single-well push – pull tracer tests to simulate a dynamic saltwater – freshwater interface’,
780 *Environmental Earth Sciences*. doi: 10.1007/s12665-015-4999-x.

781 Hebig, K. H., Zeilfelder, S., Ito, N., Machida, I., Marui, A. and Scheytt, T. J. (2015)
782 ‘Geothermics Study of the effects of the chaser in push – pull tracer tests by using temporal
783 moment analysis’, *Geothermics*, 54, pp. 43–53. doi: 10.1016/j.geothermics.2014.11.004.

784 Hebig-Schubert, K. (2015) Deep groundwater flow systems and their characterization in
785 single-well settings by "push-pull" tracer tests. Doctoral Thesis, TU Berlin. doi:
786 10.14279/depositonce-4272

787 Huang, J., Christ, J. A. and Goltz, M. N. (2010) ‘Analytical solutions for efficient
788 interpretation of single - well push - pull tracer tests’, 46(April), pp. 1–16. doi:
789 10.1029/2008WR007647.

790 Hwang, G. (2021) ‘International Journal of Heat and Mass Transfer A unified approach to
791 two-dimensional linear advection-dispersion equation in cylindrical coordinates on a finite
792 domain’, *International Journal of Heat and Mass Transfer*, 164, p. 120569. doi:
793 10.1016/j.ijheatmasstransfer.2020.120569.

794 Ikawa, R., Machida, I., Koshigai, M., Nishizaki, S., Matui, A. (2014) ‘Coastal aquifer system
795 in late Pleistocene to Holocene deposits at Horonobe in Hokkaido, Japan’, *Hydrogeology*
796 *Journal*, 22, pp. 987–1002. doi: 10.1007/s10040-014-1106-4.

797 Jada, A., Ait Akbour, R. and Douch, J. (2006) ‘Surface charge and adsorption from water
798 onto quartz sand of humic acid’, *Chemosphere*, 64(8), pp. 1287–1295. doi:
799 10.1016/j.chemosphere.2005.12.063.

800 Jin, A., Wang, Q. and Zhan, H. (2024) ‘A novel four phase slug single-well push–pull test
801 with regional flux: forward modeling and parameter estimation’, *Journal of Hydrology*, 630,
802 p. 130705. doi: 10.1016/j.jhydrol.2024.130705.

803 Johnson, R. H., Paradis, C. J., Kent, R. D., Tigar, A. D. and Reimus, P. W. (2023) ‘Single-
804 Well Push–Pull Tracer Test Analyses to Determine Aquifer Reactive Transport Parameters at

805 a Former Uranium Mill Site (Grand Junction, Colorado)', *Minerals*, 13(2), pp. 1–26. doi:
806 10.3390/min13020228.

807 Kang, P. K., Le Borgne, T., Dentz, M., Bour, O. and Juanes, R. (2015) 'Impact of velocity
808 correlation and distribution on transport in fractured media: Field evidence and theoretical
809 model', *Water Resources Research*, 51(2), pp. 940–959. doi: 10.1002/2014WR015799.

810 Kim, H. H., Koh, E. H., Lee, S. S. and Lee, K. K. (2019) 'Biased estimation of groundwater
811 velocity from a push-pull tracer test due to plume density and pumping rate', *Water*
812 (Switzerland), 11(8). doi: 10.3390/w11081558.

813 Kitanidis, P. K. (1998) 'Stochastic approaches to inverse problems', *Computer Methods in*
814 *Water Resources*, 12. doi: 10.2495/CMWR980352.

815 Kruisdijk, E. and Breukelen, B. M. Van (2021) 'Applied Geochemistry Reactive transport
816 modelling of push-pull tests : A versatile approach to quantify aquifer reactivity', *Applied*
817 *Geochemistry*, 131(March), p. 104998. doi: 10.1016/j.apgeochem.2021.104998.

818 Leap, D. I., and P. G. Kaplan (1988), A single-well tracing method for estimating regional
819 advective velocity in a confined aquifer: Theory and preliminary laboratory verification,
820 *Water Resour. Res.*, 24(7), 993–998, doi:10.1029/WR024i007p00993..

821 Leiji, F. J., Skaggs, T. H., van Genuchten, M. T. (1991) 'Analytical solutions for solute
822 transport in three-dimensional semi-infinite porous media', *Water Resources Research*, 27,
823 pp. 2719–2733. doi: 10.1029/91WR01912.

824 Maier, F. and Kocabas, I. (2013) 'Efficient analytical solution for parameter estimation of
825 push shut-in pull experiments in an idealized single fracture system', *Pangea.Stanford.Edu*,
826 (Ipcc 2011). PROCEEDINGS, Thirty-Eighth Workshop on Geothermal Reservoir
827 Engineering Stanford University, Stanford, California, February 11-13, 2013 SGP-TR-198.

828 Mascagni, M. and Chi, H. (2004) 'On the Scrambled Halton Sequence', *Monte Carlo*
829 *Methods and Applications*, 10(3–4), pp. 435–442. doi: doi:10.1515/mcma.2004.10.3-4.435.

830 Matsumoto, S., Machida, I., Hebig, K. H., Zeilfelder, S. and Ito, N. (2020) 'Estimation of very
831 slow groundwater movement using a Single-Well Push- Pull test', *Journal of Hydrology*,
832 591(June), p. 125676. doi: 10.1016/j.jhydrol.2020.125676.

833 Moench, A. F., (1989), Convergent radial dispersion: A Laplace transform solution for aquifer

834 tracer testing, *Water Resources Research*, 25(3), 439–447, doi:10.1029/WR025i003p00439.

835 Mojidi, M. A. and Vereecken, H. (2005) ‘On the physical meaning of retardation factor and
836 velocity of a nonlinearly sorbing solute’, 302, pp. 127–136. doi:
837 10.1016/j.jhydrol.2004.06.041.

838 Müller, D. and Regenspurg, S. (2014) ‘Geochemical characterization of the lower Jurassic
839 Aquifer in Berlin (Germany) for aquifer thermal energy storage applications’, *Energy*
840 *Procedia*, 59, pp. 285–292. doi: 10.1016/j.egypro.2014.10.379

841 Park, B. H., Bae, G. O. and Lee, K. K. (2015) ‘Importance of thermal dispersivity in
842 designing groundwater heat pump (GWHP) system: Field and numerical study’, *Renewable*
843 *Energy*, 83(August 2021), pp. 270–279. doi: 10.1016/j.renene.2015.04.036.

844 Parker, J. C. and van Genuchten, M. T. (1984) ‘Flux-Averaged and Volume-Averaged
845 Concentrations in Continuum Approaches to Solute Transport’, *Water Resources Research*,
846 20(7), pp. 866–872. doi: 10.1029/WR020i007p00866.

847 Rasmusson, K., Rasmusson, M., Fagerlund, F., Bensabat, J., Tsang, Y. and Niemi, A. (2014)
848 ‘Analysis of alternative push–pull-test-designs for determining in situ residual trapping of
849 carbon dioxide’, *International Journal of Greenhouse Gas Control*, 27, pp. 155–168. doi:
850 10.1016/j.ijggc.2014.05.008.

851 Rohmer, J., Armandine Les Landes, A., Loschetter, A. and Maragna, C. (2023) ‘Fast
852 prediction of aquifer thermal energy storage: a multicyclic metamodelling procedure’,
853 *Computational Geosciences*, 27(2), pp. 223–243. doi: 10.1007/s10596-023-10192-8.

854 De Schepper, G., Bolly, P.-Y., Vizzotto, P., Weexsteen, H. and Robert, T. (2020)
855 ‘Investigations into the First Operational Aquifer Thermal Energy Storage System in
856 Wallonia (Belgium): What Can Potentially Be Expected?’, *Geosciences*, 10(1), p. 33. doi:
857 10.3390/geosciences10010033.

858 Schroth, M. H. and Istok, J. D. (2005) ‘Approximate solution for solute transport during
859 spherical-flow push-pull tests’, *Ground Water*, 43(2), pp. 280–284. doi: 10.1111/j.1745-
860 6584.2005.0002.x.

861 Schroth, M. H., Istok, J. D. and Haggerty, R. (2001) ‘In situ evaluation of solute retardation
862 using single-well push ± pull tests’, *Advances in Water Resources*, 24, pp. 105–117, doi:

10.1016/S0309-1708(00)00023-3.

Selzer, P., Cirpka, O.A (2020). ‘Postprocessing of standard finite element velocity fields for accurate particle tracking applied to groundwater flow’. *Computational Geoscience*, 24, 1605–1624. doi:10.1007/s10596-020-09969-y.

Shi, W., Wang, Q. and Zhan, H. (2020) ‘New Simplified Models of Single-Well Push-Pull Tests With Mixing Effect’, *Water Resources Research*, 56(8), pp. 1–11. doi: 10.1029/2019WR026802.

Snodgrass, M. F. and Kitanidis, P. K. (1998) ‘A Method to Infer In Situ Reaction Rates from Push-Pull Experiments’, *Groundwater*, 36(4), pp. 645–650. doi: 10.1111/j.1745-6584.1998.tb02839.x.

Stemmle, R., Lee, H., Blum, P. and Menberg, K. (2023) ‘Residential heating and cooling with Aquifer Thermal Energy Storage (ATES) on city scale’, *EGU General Assembly 2023*, Vienna, Austria, 23–28 Apr 2023, EGU23-1030, <https://doi.org/10.5194/egusphere-egu23-1030>, 2023..

Stettler, M., Dentz, M. and Cirpka, O. A. (2022) ‘Linear Stochastic Analysis of the Partial Reversibility of Ensemble and Effective Dispersion in Heterogeneous Porous Media Water Resources Research’. doi: 10.1029/2022WR033570.

Suk, H., Han, W. S., Chen, J. S. and Yang, M. (2023) ‘Semi-analytical solution of single-well push–pull test under transient flow conditions’, *Journal of Hydrology*, 620(PB), p. 129542. doi: 10.1016/j.jhydrol.2023.129542.

Tang, D. W. S. and Rijnaarts, H. H. M. (2023) ‘Dimensionless Thermal Efficiency Analysis for Aquifer Thermal Energy Storage Water Resources Research’. doi: 10.1029/2023WR035797.

Ueckert, M., Wismeth, C. and Baumann, T. (2020) ‘Crystallization of calcium carbonate in a large-scale push–pull heat storage test in the Upper Jurassic carbonate aquifer’, *Geothermal Energy*, 8(1). doi: 10.1186/s40517-020-0160-5.

Vandenbohede, A., Hermans, T., Nguyen, F. and Lebbe, L. (2011) ‘Shallow heat injection and storage experiment: Heat transport simulation and sensitivity analysis’, *Journal of Hydrology*, 409(1–2), pp. 262–272. doi: 10.1016/j.jhydrol.2011.08.024.

- 892 Vandenbohede, A., Louwyck, A. and Lebbe, L. (2009) ‘Conservative solute versus heat
893 transport in porous media during push-pull tests’, *Transport in Porous Media*, 76(2), pp. 265–
894 287. doi: 10.1007/s11242-008-9246-4.
- 895 Vanhoudt, D., Desmedt, J., Van Bael, J., Robeyn, N. and Hoes, H. (2011) ‘An aquifer thermal
896 storage system in a Belgian hospital: Long-term experimental evaluation of energy and cost
897 savings’, *Energy and Buildings*, 43(12), pp. 3657–3665. doi: 10.1016/j.enbuild.2011.09.040.
- 898 Wagner, V., Bayer, P., Bisch, G., Kübert, M. and Blum, P. (2014) ‘Hydraulic characterization
899 of aquifers by thermal response testing: Validation by large-scale tank and field experiments’,
900 *Water Resources Research*, 50(1), pp. 71–85. doi: 10.1002/2013WR013939.
- 901 Wang, H. Q. and Crampon, N. (1995) ‘Method for interpreting tracer experiments in radial
902 flow using modified analytical solutions’, *Journal of Hydrology*, 165(1), pp. 11–31. doi:
903 10.1016/0022-1694(94)02588-3.
- 904 Zeilfelder, S., Hebig, KH, Ito, N., Machida, I., Scheytt, T. ‘The single well “ push-pull ”
905 tracer method : A systematic approach for setup optimization’, 17, p. 9573. - EGU General
906 Assembly Conference Abstracts, 2015
- 907 Zhu, J., Yang, M. and Ren, Z. J. (2023) ‘Machine Learning in Environmental Research :
908 Common Pitfalls and Best Practices’ *Environ. Sci. Technol.* 2023, 57, 17671–17689. doi:
909 10.1021/acs.est.3c00026.

910