This is the preprint of the contribution published as:

Buchwald, J., Chaudhry, A.A., **Yoshioka, K.**, **Kolditz, O.**, **Attinger, S.**, Nagel, T. (2020):
DoE-based history matching for probabilistic uncertainty quantification of thermo-hydro-mechanical processes around heat sources in clay rocks
Int. J. Rock Mech. Min. Sci. **134** , art. 104481

The publisher's version is available at:

http://dx.doi.org/10.1016/j.ijrmms.2020.104481

DoE-based history matching for probabilistic uncertainty quantification of thermo-hydro-mechanical processes around heat sources in clay rocks

J. Buchwald[a,b,*], A. A. Chaudhry[b], K. Yoshioka[a], O. Kolditz[a,c], S. Attinger[a], T. Nagel[b]

[a] *Helmholtz Centre for Environmental Research – UFZ, Permoserstraße 15, 04318 Leipzig, Germany*
[b] *Chair of Soil Mechanics and Foundation Engineering, Geotechnical Institute, Technische Universität Bergakademie Freiberg, Freiberg, Germany*
[c] *Applied Environmental Systems Analysis, Technische Universität Dresden, Germany*

Abstract

In the context of geotechnical and geological barriers, a thorough analysis of uncertainty and sensitivity is a crucial aspect of any physics-based performance assessment. While experimental data are scarce in actual waste repositories, large-scale experiments in underground research laboratories (URLs) provide such data that can be used to not only qualify THMC process models but also uncertainty assessment methodologies. In this paper, we adopt a Design of Experiments (DoE)-based history matching workflow – an approach popular in the oil and gas industry – and scrutinize its applicability for multiphysical analyses of nuclear waste disposal-related processes using synthetic experimental data. Based on an analytical solution of a coupled thermo-hydro-mechanical (THM) problem of a heat source embedded in a fluid-saturated porous medium mimicking a disposal cell in an argillaceous host formation, we discuss the adaptability of the workflow as a way to address parameter and model uncertainties for barrier integrity assessment. We thereby put particular focus on the relative importance of providing defined input parameter distributions for quantities generally afflicted with epistemic uncertainty and the constraints imposed by experimental (URL) or monitoring (repository) data. We found that once constraining data is available, the particular a priori distribution plays only a minor role for the outcome, such that we can conclude that the often unknown distributions can be substituted by uniform priors under such conditions. However, detailed knowledge of parameter distributions can increase the efficiency of the workflow significantly. We conclude that the presented workflow is particularly suitable for performing uncertainty quantification and sensitivity analysis for geotechnical applications where monitoring or other experimental data are available, as it allows us to deal with models of great complexity, epistemic uncertainty and it incorporates canonically to use of measured data in order

to reduce uncertainty.

Keywords: Design of Experiments, history matching, thermo-hydro-mechanical, radioactive waste, geological repository, uncertainty analysis, uncertainty quantification, sensitivity analysis, OpenGeoSys

2000 MSC: 62K20

2000 MSC: 49Q12

2000 MSC: 60-08

2000 MSC: 60G15

1. Introduction

One of the objectives in the design of nuclear waste repositories is to provide solutions that are robust and simple in the sense that their future evolution can be subjected to predictive analyses based on the current physical understanding of the involved processes. Nevertheless, in conjunction with the large spatial and temporal scales relevant to system evolution, any predictive analysis of thermo-hydro-mechanical processes around radioactive waste repositories retains a considerable degree of uncertainty. Safety regulations, therefore, require implementors to address this uncertainty at different levels.

The focus of this contribution is uncertainty quantification for the use in performance assessment of radioactive waste repositories and their components, which remains a thoroughly challenging task. Commonly, one distinguishes between uncertainty analysis/quantification and sensitivity analysis. The former is related to the determination of the overall uncertainty of a model system in terms of its input, whereas the latter refers to evaluating the relative contribution of each input (cf. e.g., [1]) to this uncertainty. Both aspects are naturally linked and addressed in this work. Dependent on the host rock in which nuclear waste is to be emplaced, different thermo-hydro-mechanical-chemical (THMC) processes need to be analyzed that might influence the transport of radionuclides in the geological disposal system [2, 3, 4]. In a repository for high-level waste, some of the most significant effects during the post-closure phase are triggered by the waste packages' decay heat, causing major changes in the physical properties of the host-rock in the near-field and driving the system away from its former equilibrium state.

*Corresponding author

Many other disturbances to the natural state are caused by the excavation itself and dominate the early stages of a repository evolution. The excavation-induced changes of the HM boundary conditions and chemical equilibria in the near-field also represent initial conditions for the post-closure phase. Such effects are neglected here. A detailed overview and description of processes that play an important role can be found elsewhere [5, 3, 4].

J. C. Helton [6] provided a general overview of different uncertainty and sensitivity analysis techniques applicable to a broad set of problems. Most of the outlined techniques remain pertinent, while some new methods emerged (e.g., SSFEM [7, 8, 9], random set theory [10, 8], or new approaches to address global sensitivity [11, 12]) or existing methods have been developed further since the early nineties (response-surface methods [13, 14] and Monte-Carlo methods [15]). The choice of a particular method depends—aside from conceptual aspects—to a large extent on the specific problem, its complexity, and the available computing power.

Indeed, when performing distribution sampling, the resulting response distributions will heavily depend on the a priori distributions of the input variables. In the context of studying radioactive waste disposal, this might be crucial. Commonly, we distinguish between *aleatory* and *epistemic* uncertainty. Whereas aleatory uncertainty is due to a random process or a stochastic variability of a phenomenon, epistemic uncertainty represents insufficient knowledge about a parameter (cf. e.g., [16]). Here, we are dealing mostly with epistemic uncertainty because most parameters vary intricately in space and time, meaning that it would be, in most cases, an onerous task to precisely measure them. It is a matter of on-going debate, whether it is possible in all cases to adequately describe these kinds of uncertainties in terms of classical probability theory [17, 18, 10, 19].

In our work, by model error, we are referring to structural discrepancies coming from a lack of knowledge of the underlying physics for the problem at hand. This also emphasizes the need to validate a computer model (this includes not only the governing equations and numerics but also the geometry, scales that need to be considered, and also its boundaries). In order to judge the practical significance of a validation, the model's uncertainty needs to be quantified. Therefore, an approach to uncertainty quantification that incorporates experimental data that can be directly compared to model predictions would be very appealing. Aside from parameter uncertainties and model inadequacy, other sources of uncertainties constitute numerical inaccuracies and observation errors that are not tackled by our investigation as they can be subsumed by the others or are of low relevance in most cases.

The software employed for performance assessment and coupled process simulations of geotechnical and geological barriers in repositories in various host rocks has undergone significant development over the past decades both in terms of physical representation and computational efficiency [20, 21, 22, 23, 24]. This development, however, has been largely disconnected from the developments in the UQ [25, 26] community, and very few links exist [27]. Thus, while improvements in hardware and software over the past decade have provided us with modeling tools that enable the modeling of radioactive waste repository sites in increasing detail, a direct Monte-Carlo approach would still pose enormous demands in terms of computing power and time. Response surface, proxy, or surrogate model approaches are one avenue for keeping the analyses tractable.

While in past discussions of parameter and model uncertainties in radioactive waste repositories, the focus has been mainly on transport phenomena [28, 16, 6, 29], a future challenge remains to put the entire coupled system (i.e. THMC processes) with all its relevant uncertainties under scrutiny as only a few stochastic case studies exist to the present day [30, 31, 32].

Data to calibrate and validate models against require a monitoring program to be installed as part of repository construction. For nuclear waste disposal research, it discloses another important application area: model development and validation using large-scale in-situ experiments in underground research laboratories cannot merely be about chasing the 'best match'. Instead, a meaningful model development and validation endeavour needs to take both experimental and modeling uncertainty into account. Global sensitivity analysis (GSA) and uncertainty quantification (UQ) can help not only in obtaining a better physical understanding but also in the identification of meaningful target validation corridors for modelers to aim at.

This article focuses on the applicability of the DoE-based history matching workflow to the multiphysical modeling of radioactive waste repository sites and components. In lieu of validating the approach based on ongoing in-situ experiments and envisioning the implementation of a certain monitoring phase of a future repository, our objective is to calibrate the THM model against continuously monitored in-situ data and subsequently to perform a probabilistic predictive analysis, i.e., a forecast. To pave the way for such analyses, this paper evaluates the workflow based on synthetic experimental data.

2. The underlying multiphysical (THM) problem

The DoE-based history matching workflow is applied to a coupled thermo-hydro-mechanical (THM) model of a heat source embedded in an isotropic fluid-saturated porous medium. Although clay-rock typically exhibits thermal, hydraulic, and mechanical anisotropy, this was ignored for the purposes of this study, which used synthetic experimental data derived from an isotropic model. The same workflow, as described below, can be applied to more general settings without the need of further changes. The model can be formulated in terms of the three primary variables temperature, pore pressure, and displacement in the balance equations of mass, momentum, and energy complemented by the constitutive relationships of the fluid and solid phases as well as their interaction in the context of porous-media mechanics [33, 34]. As the heat source causes an increase in local temperatures, solids and fluids expand, creating pore pressure and effective stress variations. The emerging pressure gradient causes the fluid to flow away from the heat source, resulting thereby in a dissipation of the pore pressure in a thermally induced consolidation process. The corresponding equations of the linear problem can be written in terms of a thermal, hydraulic and mechanical part that are coupled to each other. The thermal part is described in terms of the energy balance equation which reads (for a brief nomenclature, see Tab. 1)

$$m\dot{T} + \rho_{\mathrm{w}} c_{\mathrm{w}} T_{,i} v_i - \left(K T_{,i}\right)_{,i} = q_T \tag{1}$$

where q_T is a heat source per unit volume and, m (volumetric heat capacity), K (heat conductivity) and v_i (Darcy velocity) are given as

$$m = \phi \rho_{\mathrm{w}} c_{\mathrm{w}} + (1 - \phi)\, \rho_{\mathrm{s}} c_{\mathrm{s}} \tag{2}$$

$$K = \phi K_{\mathrm{w}} + (1 - \phi)\, K_{\mathrm{s}} \tag{3}$$

$$v_i = -\frac{k_{\mathrm{s}}}{\mu} \left(p_{,i} - \rho_{\mathrm{w}} g_i\right) \tag{4}$$

The the mass balance equation describes the hydraulic part including couplings and is given by

$$\beta \dot{p} - a_{\mathrm{u}} \dot{T} + \alpha_{\mathrm{B}} \dot{u}_{i,i} + v_{i,i} = q_{\mathrm{H}} \tag{5}$$

where q_H is the source term for the fluid while a_u is given as

$$a_{\mathrm{u}} = \phi a_{\mathrm{w}} + (1 - \phi)\, a_{\mathrm{s}} \tag{6}$$

The mechanical part can be derived from the momentum balance equations and reads

$$\sigma_{ij,j} + \rho g_i = 0 \tag{7}$$

where $\rho = \phi\rho_\mathrm{w} + (1-\phi)\rho_\mathrm{s}$ and σ_{ij} is the total stress which is given by

$$\sigma_{ij} = \sigma'_{ij} - \alpha_\mathrm{B} p \delta_{ij} \tag{8}$$

where δ_{ij} refers to Kronecker delta and σ'_{ij} is the effective stress tensor which is given as

$$\sigma'_{ij} = C_{ijkl}\left(\epsilon_{kl} - \frac{1}{3}a'\Delta T\delta_{kl}\right) \tag{9}$$

Where C_{ijkl} and ϵ_{kl} are the stiffness and strain tensors, respectively. For isotropic case, the above equation can be rewritten as

$$\sigma'_{ij} = 2G\epsilon_{ij} + \lambda\epsilon_{kk}\delta_{ij} - b'\Delta T\delta_{ij} \tag{10}$$

where

$$b' = \left(\lambda + \frac{2G}{3}\right)a'. \tag{11}$$

This specific problem can be solved analytically under few simplifying assumptions. The solution and its comparison with a corresponding numerical model can be found elsewhere [35, 36]. The model can be understood as a simplified version of a single disposal cell filled with radioactive material and emitting decay heat emplaced in an underground repository in a low-permeability host rock such as clay rock. In our workflow, the analytical model (further denoted as *AM*) is used on the one hand for the generation of synthetic experimental data, and on the other for the major part of the system analysis, i.e., for proxy generation (the kriging proxy model is further referred to as *PM*) as well as for history match selection/parameter identification.

In order to demonstrate the role of the forecast, i.e. a predictive analysis, in the workflow, we increased the power of the heat-flux in a step-wise manner, as done in actual heater experiments [37]. As a consequence, the resulting forecast response curves are non-trivial and predict behavior that goes beyond the calibration domain of the model. As the analytical solution is not capable of reflecting this change in heat power, the forecast uses the finite element method instead (the numerical model is further denoted as *NM* in the text). For this purpose, a two-dimensional model domain with axial symmetry representing a half-space of the spherically symmetric problem was

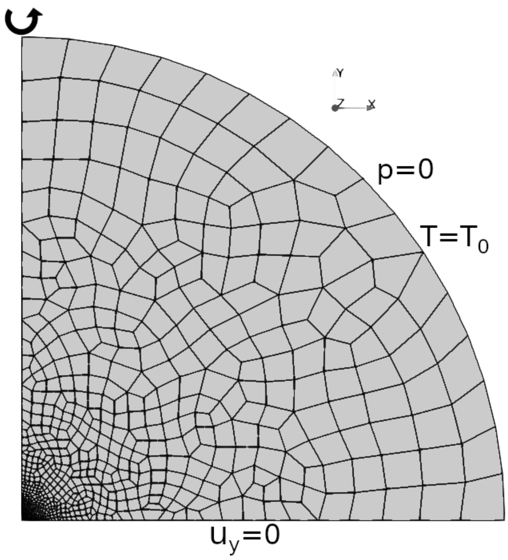


Figure 1: Mesh used for the numerical model with the applied boundary conditions.

created. Alongside the inner symmetry boundaries, the axis-normal displacements are set to zero, while at the outer boundaries the pore pressure is required to vanish[1], and the temperature is fixed to its initial value (293.15 K) along the outer boundaries. The mesh together with a summary of the applied boundary conditions is depicted in Fig. 1.

Further computational details are given in [36]. In Tab. 1, we present a list of parameters along with their lower and upper limits that will be used in the uncertainty and sensitivity analyses. The corresponding probability density functions are displayed in Fig. 1 of the electronic supplementary information (SI).

3. History matching (HM) and uncertainty quantification (UQ)

3.1. General workflow

The applied approach, known as experimental design (DoE)-based history matching, is closely related to the above-mentioned response-surface methods and Monte-Carlo sampling and has been applied to production forecasts in the oil and gas industry [13, 38]. To the best of our knowledge it has not been applied to problems in radioactive waste management. This is likely because usually, there is no historical data to match and most of the analyses are purely predictive. However, underground research laboratories (URLs) provide experimental data sets that allow us to explore

[1]We depart from vanishing initial stresses and pore pressures due to the linearity in the constitutive models used. Stresses and pore pressures thus represent increments rather than absolute values.

Table 1: Used bounds of material parameters for the analytical (AM) and numerical model (NM). More details given in Tab. 1, SI.

Parameter	**symbol**	**low**	**best**	**high**	**unit**
Young's modulus	E	$2.1 \cdot 10^9$	$2.7 \cdot 10^9$	$3.5 \cdot 10^9$	Pa
Poisson's ratio	ν	0.28	0.33	0.38	-
Vol. thermal expansion coefficient of the solid	$a_\mathrm{s} = a'$	$1.5 \cdot 10^{-6}$	$4.2 \cdot 10^{-6}$	$1 \cdot 10^{-5}$	K^{-1}
Vol. thermal expansion coefficient of water	a_w	$1.695 \cdot 10^{-4}$	$3.98 \cdot 10^{-4}$	$5.63 \cdot 10^{-4}$	K^{-1}
Porosity	ϕ	0.14	0.183	0.247	-
Water density	ρ_w	979.4736	991.46	998.767	$\mathrm{kg\,m^{-3}}$
Solid grain density	ρ_s	2700.0	2768.5	2872.0	$\mathrm{kg\,m^{-3}}$
Specific isobaric heat capacity of water	c_w	3941.38	4065.12	4167.71	$\mathrm{J\,kg^{-1}K^{-1}}$
Specific isobaric heat capacity of the solid	c_s	760.0	860.0	960.0	$\mathrm{J\,kg^{-1}K^{-1}}$
Heat conductivity of water	K_w	0.592015	0.63122	0.657	$\mathrm{W\,m^{-1}K^{-1}}$
Heat conductivity of the solid	K_s	1.0	1.7	3.1	$\mathrm{W\,m^{-1}K^{-1}}$
Dynamic viscosity of water	μ	$4.237 \cdot 10^{-4}$	0.000633	0.0011	$\mathrm{Pa\,s}$
Intrinsic permeability	k_s	$2 \cdot 10^{-20}$	$3 \cdot 10^{-20}$	$2 \cdot 10^{-19}$	m^2
Initial temperature	T_0	292.15	293.15	294.15	K

this avenue. Therefore, an "ansatz" containing the parallel analysis of modeling and experimental data is the ideal choice, as it allows us to link model calibration with the analysis of parameter uncertainties. The applied approach is very closely linked to Bayesian history matching [39]. However, in contrast to the latter, the posterior function is taken from filtering the response of the direct sampled priors, whereas Bayesian approaches typically use a likelihood function (derived from the history-match error) to obtain Markov Chain Monte Carlo estimates.

Aside from the URL perspective, repository concepts including monitoring activities at least for the early post-closure phase constitute a potential basis for analyses incorporating history matching. Thus, if deemed suitable, this workflow can be of potential relevance for the design and monitoring of future repository systems.

Note in passing, that similar considerations apply to other geotechnologies.

Following the steps as presented by [38], we want to highlight some features that appear particularly interesting for the class of problems at hand as summed up in the previous section:

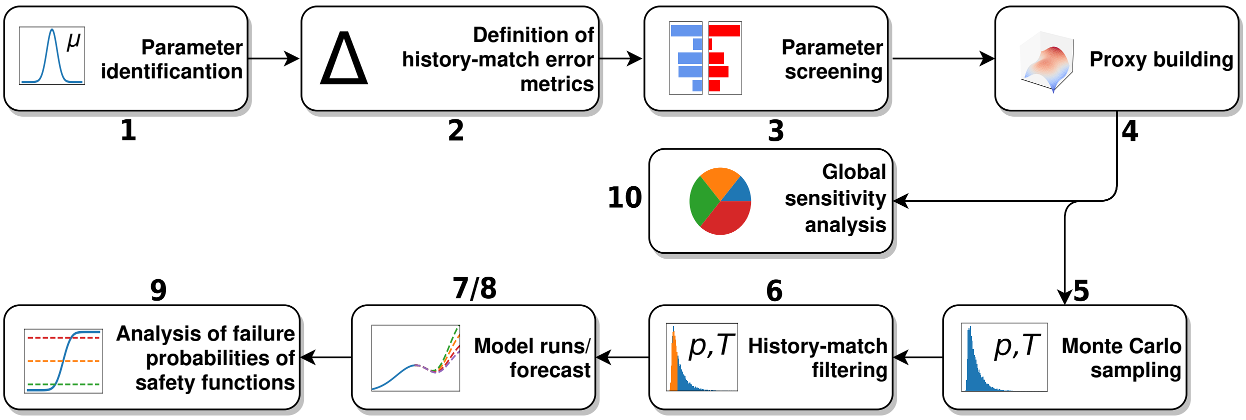


Figure 2: Schematic sketch of the workflow.

- history matching canonically incorporates experimental data enabling us to calibrate the numerical model
- the use of proxy models for the history match error allows for direct Monte-Carlo sampling with a statistically sufficient number of samples while keeping the computational burden manageable
- initial parameter screening prior to proxy building makes it possible to neglect insignificant parameters
- in contrast to many other more specialized uncertainty quantification methods in finite element modeling, it is generally applicable to non-linear and coupled problems

In Fig. 2, we present a sketch in which we transferred the workflow to our purposes of uncertainty quantification and sensitivity analysis.

1. The first step is devoted to problem framing. In order to assess the applicability of this approach to coupled THM models while keeping the computational burden minimal, we employ an analytical solution that retains most of the essential primary THM couplings for the model evaluations as well as for synthetic data generation. Another issue we want to address in this paper is the sensitivity of this methodology to particular input parameter distributions given only lower and upper bounds. Therefore, all parameters that are assumed to carry uncertainty are assigned to two different kinds of input distributions. We assume distribution types found for the Meuse/Haute Marne URL site (France; [40]), and compare

the results to data gathered based on uniform priors for all parameters. As we are dealing mostly with parameter uncertainties of the epistemic type, we are seeking to identify a method that is robust in terms of prior parameter distribution assumptions.

2. In the second step, based on response quantities of interest, history match error metrics are defined.
3. In the third step, we take the THM model and subject it to two different kinds of screening designs (one-variable-at-a-time and Placket-Burman) to obtain some initial understanding of sensitivity.
4. Consequently, a proxy is built based on space-filling Latin-hypercube sampling. However, before proceeding to real applications, we need to evaluate the size of the sampling space required and estimate the resulting computational effort.
5. After checking essential quality measures, we use the proxy to perform a Monte-Carlo sampling on the entire uncertainty space.
6. During the history matching step, only results are selected that follow observed (in our case, synthetic) experimental data.
7. In an attempt to account for proxy errors, the filtered parameter sets are used to conduct real model runs, that should support the agreement between the experimental and modeling results.
8. The history-matched parameter sets are used for forecasts. As the idea of our test case is to mimic a real-world experiment, we change the source term by increasing the power of the heat source for the forecast. Since the analytical solution is not capable of describing this change, we used a corresponding verified numerical models to conduct the forecast.
9. The uncertainty quantification in terms of percentiles of predefined performance or observation measures can then be done based on a forecast.
10. Finally, we present results from a global sensitivity analysis in terms of Sobol's indices that are based on the proxy estimate in order to identify the main factors contributing to the observed uncertainty. This is done both for prior and posterior distributions of input factors.

3.2. Workflow implementation: synthetic experimental data

The synthetic experimental data was created using a random seed selecting a set of input parameter values uniformly within the lower and upper bounds of each input parameter (Tab. 1) by evaluating the analytical model (AM). The data was gathered from a data sheet describing the characteristics of the Opalinus Clay at the Mt. Terri site in Switzerland [41]. The generated time-dependent response curves were varied with additional white noise (see Fig. 3).

3.3. Workflow implementation: parameter input for analysis

For system analysis, the same data with their corresponding input distributions were used (Tab. 1 and SI Tab. 1). The functional forms for the input parameter distributions were taken from a description of the DECOVALEX 2019 Task E specifications [42, 40]. Additionally, we also performed an analysis based on the same input limits but using only uniform input distributions to study the significance of the functional form of the input distributions. The water-related parameters a_{w}, ρ_{w}, K_{w}, c_{w}, and μ are relatively precisely known for given temperature and pressure. However, as they are assumed to be constant in the analytical model (AM), we treat them here as independent random variables with min/max values according to the temperature interval spanned by $T_{min} = 290$ K and $T_{max} = 380$ K. As a matter of fact, such a treatment expands the uncertainty space beyond the necessary scope as the underlying relation is known to be deterministic. Nevertheless, as we are are performing also a sensitivity analysis, it gives us also an insight into which parameters can be regarded as constant due to their low sensitivity and for which parameters we have to consider their deterministic relationship. In this study, we focused on five response quantities namely temperature, pressure, displacement, and radial as well as circumferential stress evaluated at an arbitrarily chosen observation point ($P = (0.5\,\mathrm{m}, 0.0)$) and compared to the corresponding (synthetic) experimental data obtained as time series at the same location. In this publication, we subject our investigation to a single observation point only, because of spherical symmetry (i.e. the problem has only one effective spatial dimension) and the fact that both, the synthetic experimental data as well as the history match-model (AM/NM) are of the same origin, i.e., the information obtained at a single location should be enough to perform a sufficient history match.

3.4. Workflow implementation: Design of Experiments (DoE)

In the applied workflow, methods of experimental design (DoE) are used to reduce the number of degrees of freedom in order to build a proxy model (PM) that can be used efficiently for Monte-Carlo sampling and history matching. The degree of a possible agreement between experimental and modeling (time-series) data permits the quantification of model uncertainties, whereas the probabilistic analysis of the forecasted history matched model (*NM*) allows for uncertainty quantification of model and parameter uncertainties combined. The relative significance of different input factors on the model output is investigated by means of a global sensitivity analysis based on the proxy model (*PM*) with particular reference to the history match error. The *PM* was built based on the space-filling Latin-Hypercube design. For the purpose of an initial parameter screening, the parameter space was sampled using a (folded) Placket-Burman design [43] and a One-Variable-at-A-Time (OVAT) design at the domain bounds and around the 'best' values. This procedure is often referred to as local sensitivity analysis.

3.5. Workflow implementation: used software and libraries

The entire workflow was implemented in Python using the pyDOE2[2] library for the experimental designs, GPy [44] for the proxy modeling of the history match error and SALib[3] for the global sensitivity analysis (GSA). The entire workflow is wrapped around the multiphysics simulator OpenGeoSys[4] [24, 45] and thereby ready for use with configurations of much greater complexity.

3.6. Problem Framing

The system under scrutiny is a greatly simplified model of a single canister of radioactive waste described by a point heat source in an infinite homogeneous isotropic porous fluid-saturated medium, as described in Section 2. In this case study, we analyzed the scalar quantities T and p as well as the u_r component of the displacement vector and the σ_{rr} and $\sigma_{\varphi\varphi}$ components of the stress tensor. As stated earlier, the response is only measured at one location, so we decided, therefore, to include both stress components as derived quantities in order to make the model sufficiently complex. The corresponding history match error metrics are then defined for each response quantity by

$$e^{\mathrm{HM}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(d_i^{\mathrm{obs}} - d_i^{\mathrm{sim}})^2}. \tag{12}$$

Here, n is the number of data/observation points in time, but also space, meaning that one error metric can, in principle, also be comprised of several observation locations. Under certain

[2] https://pypi.org/project/pyDOE2/
[3] https://salib.readthedocs.io/
[4] https://www.opengeosys.org/

(experimental) conditions, it might be useful to introduce an additional weighting factor for the data points. As we have only one observation point in this study, the sum is over a number of 2000 time steps ($\Delta t = 5000\,\mathrm{s}$) used here for the synthetic data generation as well as for (NM-)modeling. d_i^{obs} corresponds to an observed datum, whereas the d_i^{sim} corresponds to the simulated data. In this study, d_i^{obs} are synthetic experimental data, as described in the previous section.

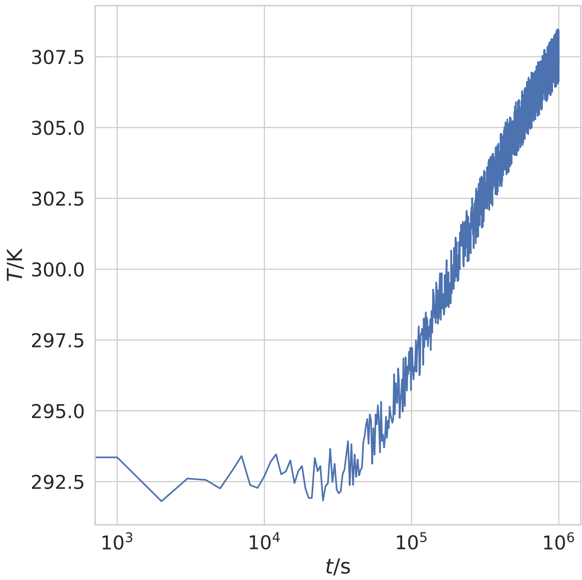
4. Results

4.1. Sensitivity Screening

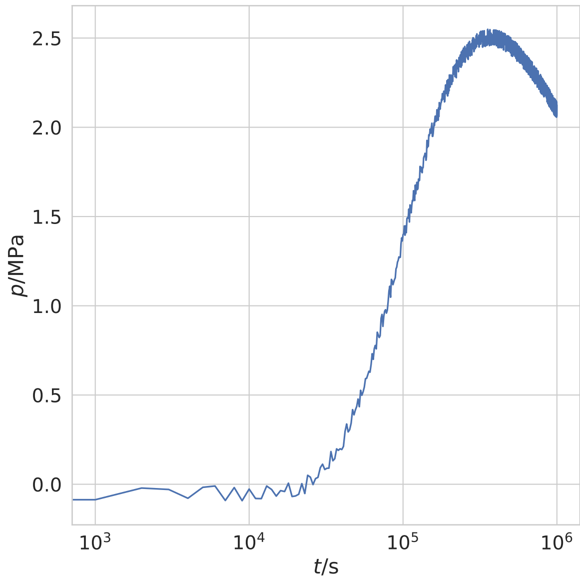
A local sensitivity analysis is conducted to get a first idea about the model (AM), i.e., its overall variability, its heavy hitters, and also any non-significant input parameters to be omitted during proxy building. We strive for a smaller number of inputs as they should result in better proxy quality if keeping the number of overall training samples fixed. Alternatively, seen from another perspective, we might be able to reduce the number of training samples resulting in a comparable proxy quality when having fewer inputs, i.e., a smaller uncertainty space. To serve this purpose, the size of the screening design should be taken much smaller compared to the runs needed for proxy training. More precisely, the required effort to neglect inputs should relate to a probable gain in time or proxy quality.

In this study, we start with a One-Variable-at-A-Time (OVAT) design to get a first insight and to create tornado plots (i.e., horizontal bar charts). The OVAT design was conducted in two ways. Both commence from a run with all parameters set to their baseline (referred to as 'best' in Tab. 1). The first type is intended to cover the entire parameter range, so changes in the response variables (e^{HM}) were obtained by changing the input parameters one at a time to their defined extreme values ('low'/'high' values in Tab. 1). In the second type of OVAT design, the changes around the baseline were reduced to a fraction of 100 and re-scaled afterward, i.e., the response was multiplied again by a factor of 100 to be comparable with the former design. However, the result of the latter type is more akin to a local sensitivity (tangent in the baseline), while the former evaluates the model (AM) at its actual extremes (secant). Additionally, we introduced a dummy parameter that is defined in the interval $[-1, 1]$ to obtain a clearer picture for distinguishing between significant and insignificant effects, which becomes important when applying statistical significance tests.

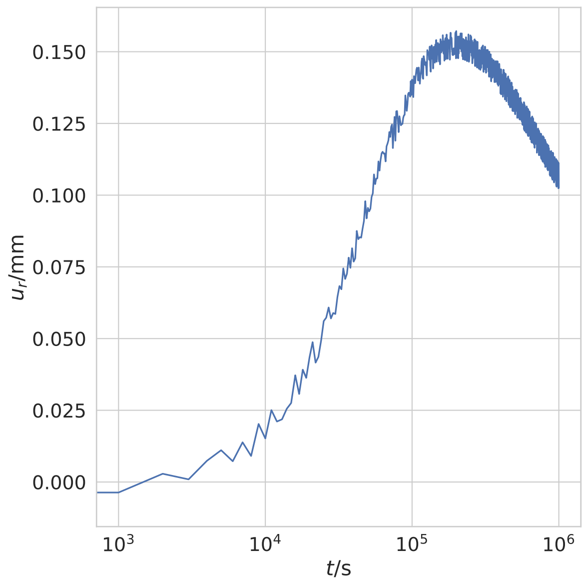
In Fig. 4 we show the tornado plots of the aforementioned types of OVAT designs for the pressure-related error metric e_p^{HM}. The tornado plots of the remaining response variables are



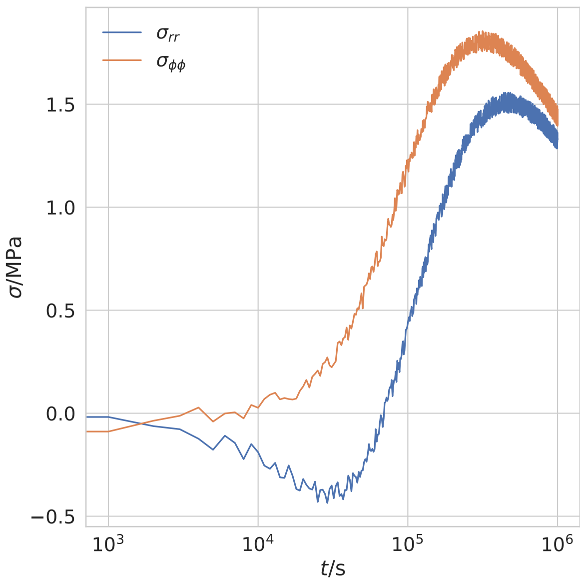
(a) Temperature



(b) Porepressure

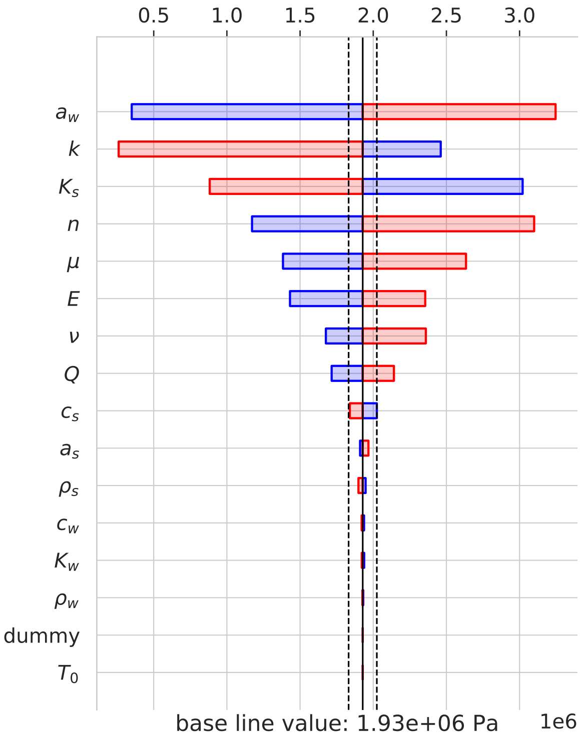


(c) Displacement

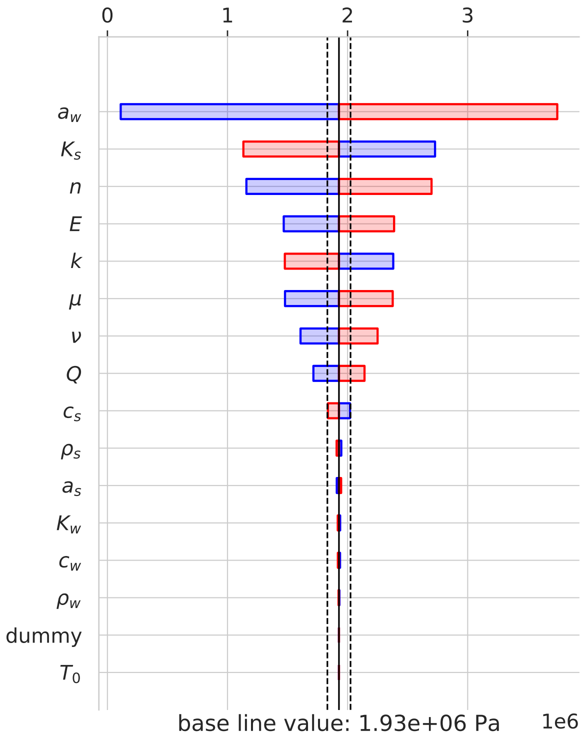


(d) Stress

Figure 3: Synthetic experimental response curves over the entire history match time span of 10^6 s (about 11 d) at Point $P = (0.5\,\mathrm{m}, 0.0)$.



(a) bounds screening



(b) narrow band screening

Figure 4: Tornado plots of the pore pressure history match error (Pa) at point $P = (0.5\,\mathrm{m}, 0.0, 0.0)$.

presented in Fig. 2-5 of the SI. Combining the information from all response variables, we can identify some parameters that have no or only a marginal impact (concerning the defined intervals) on the response variables. With a predefined triviality/significance margin of five percent, we find that ρ_{w}, ρ_{s}, c_{w}, K_{w}, and T_0 have only a small impact on all response variables, i.e., can be neglected later for proxy building.

The symmetry of the narrow band screening is due to the local perturbation of the base value and the subsequent re-scaling. Assymmetric effects due to a non-linear influence of model parameters or due to assymetric 'high'/'low' values are only visible in the bounds screening variant of the tornado plots (Fig. 4). We also see that shifts in ranking occur due to the strong impact of the actual parameter variability on its associated sensitivity, an effect that is not captured by a local tangent.

It is essential to mention that the presented OVAT design has several drawbacks. One is that the system is probed only locally at a fixed position around the intermediate values. The second is that we do not test interactions between variables. To obtain more accurate results, we applied a folded Plackett-Burman design to screen the main effects as suggested in [38]. The sensitivity screening is then done employing a t-test on the regression coefficients using linear regression of a linear model (LRM)[5]:

$$e^{\mathrm{HM}} = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \epsilon. \tag{13}$$

If studying also two-way (or higher) interactions, different designs need to be employed. Suitable experimental designs might be D-optimal and fractional-factorial designs keeping in mind that the minimum number of runs is given by the number of unknown coefficients of the regression equation.

While the Pareto chart in Fig. 5 shows a very similar behavior qualitatively as the tornado plot (Fig. 4), it appears as if only one parameter reached the critical t-value, i.e., can be considered as significant. However, a closer look at the coefficient of determination, and the F-statistics revealed that the sampling is not sufficient to exclude insignificant parameters because interaction terms were sampled but not included in the linear regression model (LRM).

As mentioned above, it is possible to switch to larger designs and also to account for interactions.

[5]For more on t-test and hypothesis testing in the context of regression analysis cf. *Regression Analysis by Example* [46].

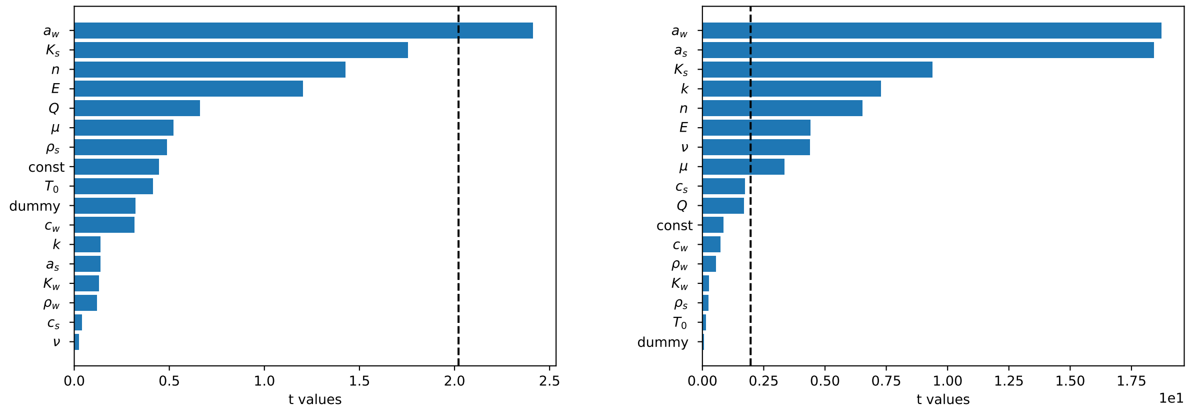


Figure 5: Pareto charts based on regression of a linear model for e_p^{HM}. Analysis of t values was conducted using a folded Placket-Burman design (left) and latin-hypercube sampling (right) of the model domain (400 samples). The vertical line corresponds to the critical t-value for a significance level of $p = 0.05$.

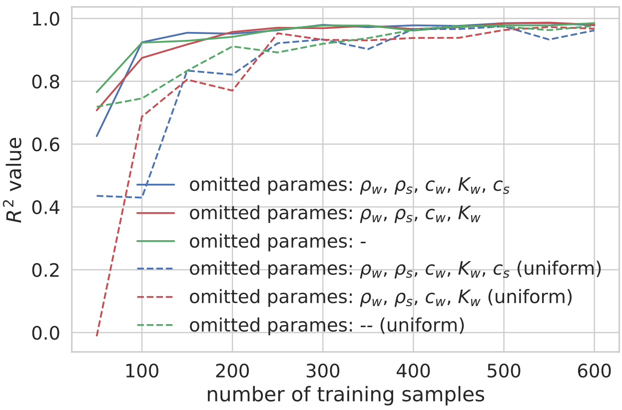
As we are conducting this *initial* sensitivity screening for achieving gains in proxy performance, we require to stay well below one order of magnitude of the space-filling design we later use for proxy-building – otherwise the reduction of the uncertainty space would not pay off. However, we need to perform a space-filling design like Latin hypercube sampling (LHS) of the order of several hundred runs either way. Therefore, we can use the conducted design for proxy-building to perform an additional sensitivity screening in order to confirm our results so far. Using an ordinary LHS design with a size of 400 sampling points, we applied the t-test as presented above using a linear model (LRM) accounting only for main effects. Contrary to Placket-Burman screening, the F-statistics support the claim of a significant influence of the linear fitting in general. Very much in agreement with the results from the OVAT screening, we find that ρ_{w}, ρ_{s}, K_{w} and c_{w} to have t-values below t_{crit} ($p \leq 0.05$) for all response error metrics, i.e. can be regarded as non-influential for all response parameters based on their given bounds of variation. When comparing both diagrams (Fig. 4 and Fig. 5, respectively), we find that the order of sensitivity of most parameters changes, which comes from different contributions of certain parameters at different locations when using a space-filling design. Another point, we want to stress here, is that in all cases, the dummy parameter ranks above other parameters while staying well below the critical t-value, which gives us confidence that those parameters do not have a significant impact on the model AM. The fact that the t-value is not precisely zero is not a surprise as we are trying to fit a linear model (LRM) with a manageable amount of samples. Even though using a space-filling design, we might overlook some effects that come from, e.g., a smiley face pattern or effects that stem from interactions of input parameters. However, at this stage only a preliminary screening is intended and a more thorough sensitivity analysis will be conducted later using the proxy (PM) and sampling-based methods that also provide much better coverage of the uncertainty space [29, 12] and that allow a better quantification of relative effects.

4.2. Proxy Building

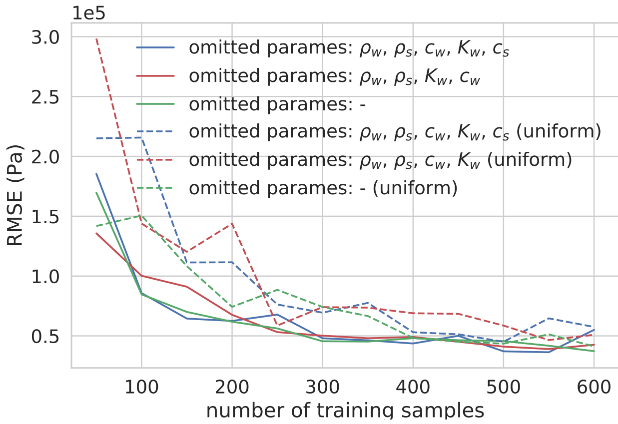
For proxy building, as stated in the previous paragraph, a space-filling design like Latin hypercube sampling (LHS) needs to be employed. As proxies, typical choices are polynomial, splines, kriging, or neural networks. A very suitable choice for our purpose is Gaussian process regression (often referred to as kriging in the geosciences) as every training point can be recovered precisely by the proxy, so it is thereby expected to provide better accuracy compared to parametric proxies. Comparisons for different types of proxies can be found elsewhere [38, 47]. At this stage, we take all parameters with assigned uncertainty with us, as we want to study what kind of influence the omission of some of them will have on the proxy quality. The proxy was built based on 50, 100, 150, ..., 400 training samples plus 200 testing samples of a space-filling Latin-hypercube design to observe the convergence behavior of the coefficient of determination in order to determine a required number of training samples.

In Fig. 6(c), we plotted the time needed to build and apply the proxy to the testing samples, together with the R^2 (Fig. 6(a)) and RMSE quality measures (Fig. 6(b)) for e_p^{HM} over a varying number of training samples. Different curves correspond to different classes of a priori input distributions (non-uniform vs. uniform assumptions) and different sets of omitted input parameters. What we see at first glance for the proxy quality measures, is that the R^2/RMSE values improve quite strongly between 50 and 200 training samples, while changes tend to converge for higher numbers. Keeping in mind that for each training and testing sample, a full run of the typically very costly NM-model is required, the search for a sample number compromising between accuracy and computational effort becomes obvious. Therefore, one would prefer the application of a sequential sampling strategy [48] and stop increasing the number of training samples when a justifiable R^2/RMSE value is reached.

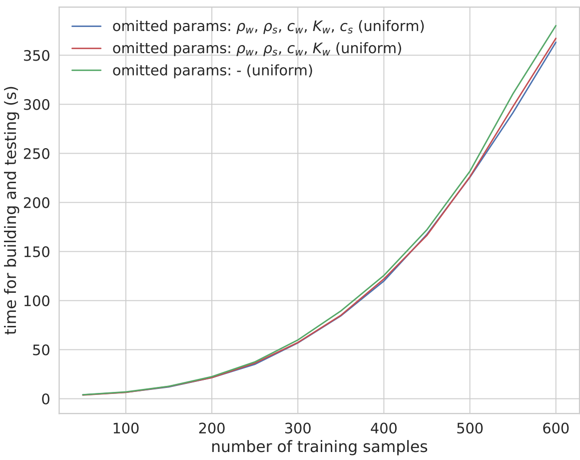
As can also be seen from the plots, we can clearly distinguish between uniform and non-uniform parameter distributions, although the differences vanish for a higher number of sampling points.



(a) Coefficient of determination (p)

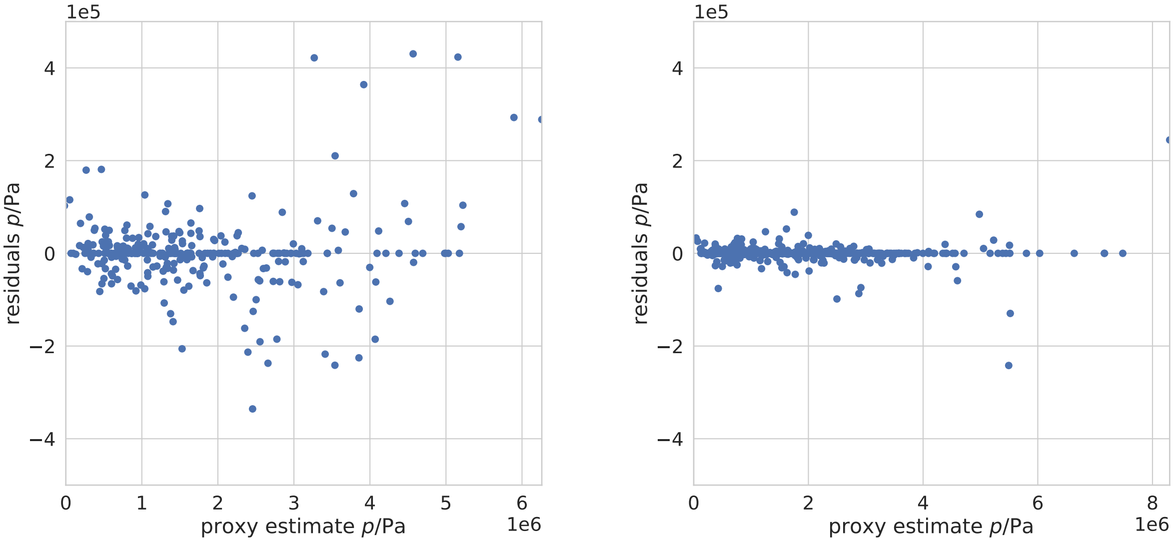


(b) Root-mean-square-error (p)



(c) Time for proxy building.

Figure 6: Performance of the e_p^{HM} kriging proxy for a varying number of input parameters and training samples.



(a) 100 training samples

(b) 400 training samples

Figure 7: Residuals of the e_T^{HM} error metric based on 100 and 400 training samples.

This means, with respect to proxy quality, that the drawback of not knowing the exact distribution can be compensated by choosing the size of the Latin-hypercube design big enough. On the other hand, no clear trend is visible for the omission of non-influential parameters. Indeed, the difference between uniform and non-uniform parameter distributions is somewhat expected, as in the uniform case, a greater parameter space is covered, leading to a lower sampling density on the average. The not-shown response quantities also confirm these trends (see Fig. 8-9 SI for further details). Similar to the quality measures, the time required for proxy building and testing is also not very sensitive to parameter omission. As R^2 and RMSE give us a rough sense of the proxy quality, we also had a look at how the residuals are, in fact, distributed (Fig. 7). The residuals are plotted versus their associated proxy estimates for 100 and 400 training samples. Both plots show only slight heteroscedastic behavior, mainly an increasing variance for a higher proxy estimate. While greater heteroscedasticity could pose a real problem, slight heteroscedasticity is often unavoidable. Another important measure of proxy quality is whether the proxy preserves essential mathematical and physical properties. One such property of the history match error is that it is strictly positive. In Fig. 8 one sees that a small fraction, especially for smaller sample sizes, does not fulfill this criterion. However, for the σ_{rr}- proxy built using 400 training samples only a fraction of $< 10^{-3}$

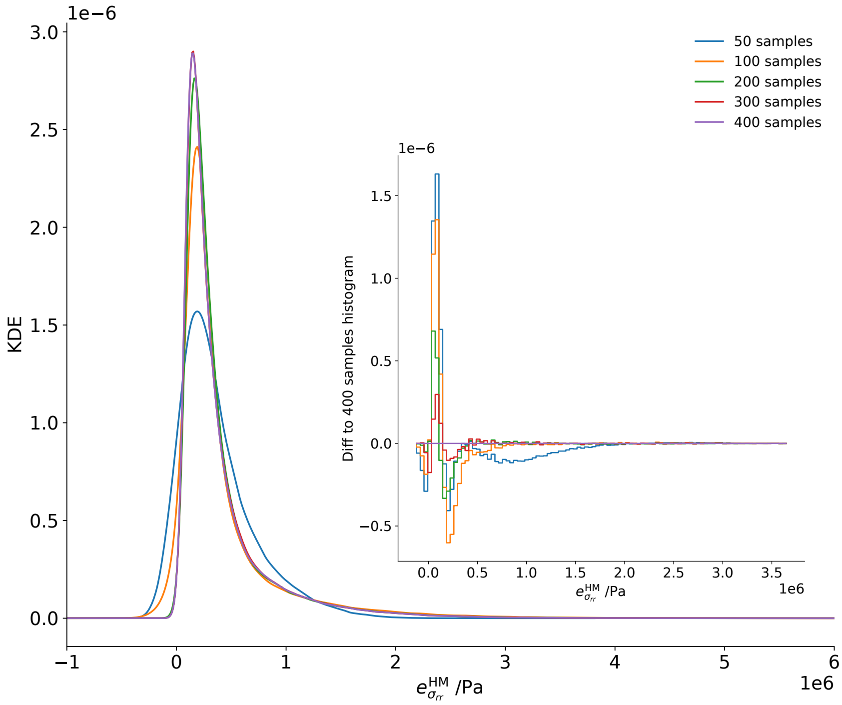


Figure 8: Kernel density estimate of the $e^{\mathrm{HM}}_{\sigma_{rr}}$ proxy error metric and their histogram of differences to 400 samples reference based on $200,000$ MC samples.

is negative, keeping in mind that the expectation value of $e^{\mathrm{HM}}_{\sigma_{rr}}$ is $\approx 10^5$ Pa. To conclude our analysis, a sample number of 400 without omitting any input seems to be a reasonable choice for our further investigation as no significant improvement could be achieved by either increasing the sample numbers or omitting sets of marginally influential parameters.

4.3. Monte-Carlo Sampling and history match Filtering

Based on the choice of the previous paragraph, the proxy model (PM) is ready for Monte-Carlo sampling. 200,000 Monte-Carlo samples are drawn from the prior 'non-uniform' and 'uniform' distributions and evaluated for each proxy error metric.

For testing purposes, we applied three increasingly strict filter criteria (Tab. 2) to the sampling results and analyzed the output (Fig. 9) as well as the corresponding history-matched input parameter distributions (Figs. 10 and 11) for non-uniform and all uniform priors. Thereby, we selected only samples that satisfied the history match criteria for all variables simultaneously. Whereas blue corresponds to the prior distributions, orange, green, and red denote the different conditions as defined in Tab 2. All histograms are re-scaled after filtering in order to be visible in the figures.

Taking first a look at Fig. 9, we see that except for the temperature, all history match curves are lying between values around zero and the cut-off condition smaller than the prior histogram reaching its maximum. The temperature-'anomaly' can be explained by the fact that the range of possible solutions is comparable to the noise of the synthetic experimental curve (cf. Fig. 3).

Therefore, the history match error is very much affected by the noise, which also leads to the shift of around 0.5 K in the histogram. Comparing the results to the data obtained from uniform prior distributions, we see that the posterior distributions seem to behave very similarly. In other words, if the prior uncertainty space is much bigger than the space obtained by the filter, then the exact forms of the prior parameter distributions become irrelevant when sampled appropriately. Another more subtle point is that the condition to satisfy all criteria simultaneously leads to a significant reduction of the number of matched curves compared to the histogram's area that would remain after application of the individual quantity's criterion alone.

variable		**test cond. 1**	**test cond. 2**	**test cond. 3**	**final cond.**	**RMSE**	**unit**
e_T^{HM}	$<$	3	2.5	2.0	2.0	$1.5 \cdot 10^{-3}$	K
e_p^{HM}	$<$	$1.5 \cdot 10^5$	$1.0 \cdot 10^5$	$0.5 \cdot 10^5$	$0.3 \cdot 10^5$	$0.5 \cdot 10^5$	Pa
$e_{u_r}^{\text{HM}}$	$<$	$1.5 \cdot 10^{-5}$	$1.0 \cdot 10^{-5}$	$0.5 \cdot 10^{-5}$	$0.3 \cdot 10^{-5}$	$0.5 \cdot 10^{-5}$	m
$e_{\sigma_{rr}}^{\text{HM}}$	$<$	$1.5 \cdot 10^5$	$1.0 \cdot 10^5$	$0.5 \cdot 10^5$	$0.3 \cdot 10^5$	$0.35 \cdot 10^5$	Pa
$e_{\sigma_{\varphi\varphi}}^{\text{HM}}$	$<$	$1.5 \cdot 10^5$	$1.0 \cdot 10^5$	$0.5 \cdot 10^5$	$0.3 \cdot 10^5$	$0.35 \cdot 10^5$	Pa

Table 2: History match filtering conditions for all response parameters. The testing conditions were used to investigate the effect of varying filter sizes using the *AM*-model. The final condition together with the *PM*-RMSE is used to demonstrate the workflow functionality including model runs for history match and forecast(AM/NM).

In a second step, we analyzed the input parameter combinations that were used to generate the response proxies that survived the history match filtering ('filtered priors'). For each parameter and filter condition, the corresponding distributions are depicted in Figs. 10 and 11. Here, we see that for some parameters like E, the filter does not seem to have any impact on the distributions, whereas for others (like K_s, a_s, k or μ) the filter seems to restrict the domain of definition. Indeed, the latter parameters were shown to be heavy hitters in sensitivity screening. For these parameters, we also see a better agreement between the input used to generate the experimental data and the posterior distribution with the tightest criterion. Parameters that retain their distributions after filtering were found to be less influential, which is very much in agreement with what one would assume. It is also worth noting that the posterior distributions found for all uniform priors agree very well with the ones for non-uniform priors for the influential parameters. This gives rise to the conclusion that the exact form of the prior distributions is not very relevant for the history match results. Another thing that attracts our attention is that, for k and μ, we find a kind of multimodal behavior after history matching that we see, especially for non-uniform priors. This behavior, as well as the fact that the sampling maxima of the tightest match criteria do not coincide with the input of the synthetic model (AM) run, we attribute to parameter interactions that cannot be disregarded. In general, the history matching workflow is also quite suitable to serve as an alternative to optimization algorithms for the purpose of parameter estimation. However, the latter discussed behavior also makes clear that the reverse problem of parameter estimation of such a highly non-linear model is ill-posed. For further analysis of parameter interactions, we allude to the global sensitivity analysis.

The impact of different prior distributions as well as the effect of different filter sizes on the direct values of the response functions are also investigated using the AM-model. For this analysis, we used the last time step of the history match and calculated the corresponding CDFs based on the testing filters defined in the previous section. The cumulative distribution functions (CDF) were evaluated at point $P = (0.5\,\mathrm{m}, 0.0, 0.0)$ and time $t = 5 \cdot 10^6\,\mathrm{s}$. In Fig. 12, the CDFs were presented based on prior distributions before and after applying the three different filter conditions. For all three conditions and all response functions, we find that the co-domain is significantly reduced, and we are able to give percentiles for each quantity from which we can select representative models. One major property that such a workflow needs to satisfy in order to demonstrate its robustness is that its results should not depend too much on arbitrarily chosen values. While it is quite obvious, that the filter size might have a greater impact on p10 and p90 percentiles, we assume, that this influence tends to be rather small for the p50 percentile. To assess the validity of this assumption, we calculated the relative change of the p50 value for each filter with respect to no filter as follows:

$$p^X = \frac{|X_{p50}^{\text{filter}} - X_{p50}^{\text{no filter}}|}{X_{p50}^{\text{no filter}}} \tag{14}$$

The corresponding values are given in Tab. 3. Here, we see that for non-uniform priors, the p50 value varies up to six percent, while the relative changes for all uniform priors are all systematically greater and amount up to 9 percent for the displacement. However, it is not obvious that the expectation for all uniform priors is subject to greater fluctuations in all cases, nor does our example prove that we get more reliable results for non-uniform priors. While these changes can be regarded here as quite small, they might matter in some cases and remind us that the choice of the filter size is somehow subjective, and it necessitates a more thorough analysis when discussing concrete safety functions.

After having obtained an overview, we are able to define the actual history match criteria with which we will perform real model runs (AM/NM) and consequently, the forecast. To define

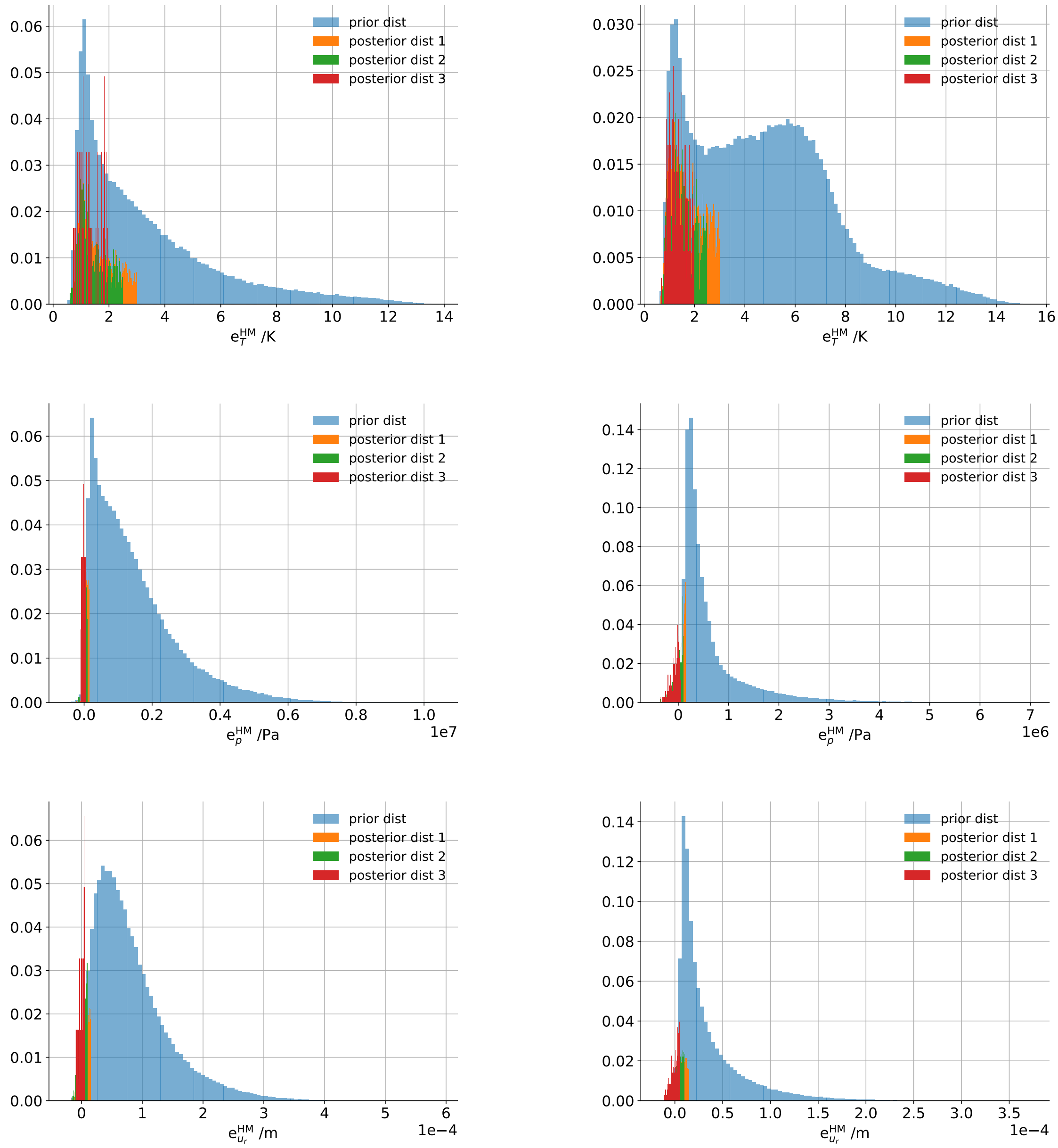


Figure 9: Distributions of the history match error responses for non-uniform (left) and all uniform priors (right) and the filter conditions listed in Tab. 2.

these conditions, one has to take also into consideration that several error sources should also have an impact on the defined confidence interval. Most important (i) is the (AM/NM) model error (e.g., heterogeneities or other processes that have to be taken into account), (ii) followed by the proxy error and (iii) last but not least, the sampling error. As the model error cannot be precisely determined, the choice of the history match thresholds coming from this contribution is

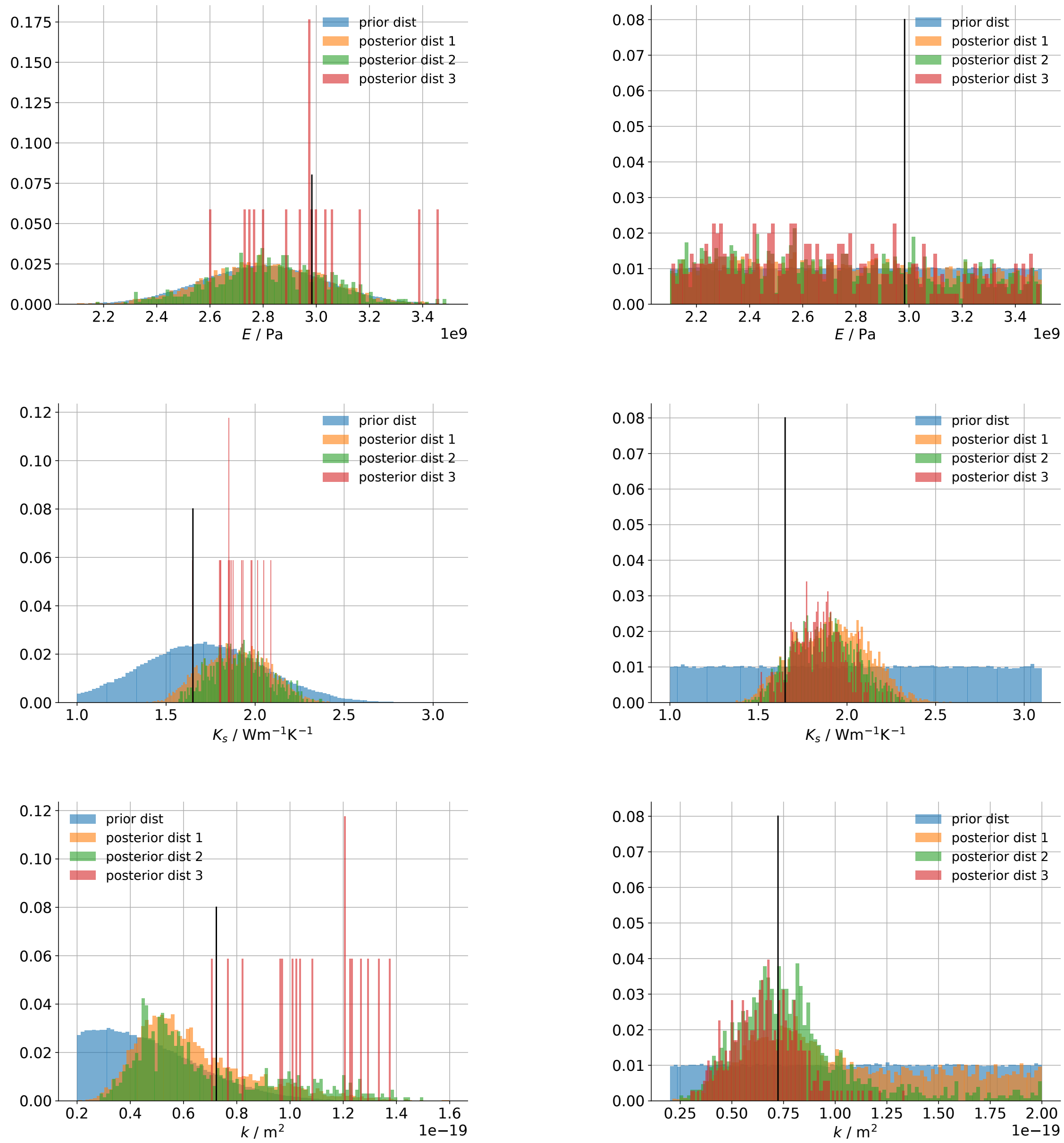


Figure 10: Parameter estimation for a selection of input quantities from Monte-Carlo sampling and after history match filtering for non-uniform (left) and all uniform (right) priors with the filter conditions listed in Tab. 2. The black line marks the input parameter for the creation of the synthetic experimental data.

somehow subjective. For our problem case, the criteria are given by the fifth column in Tab 2 (final conditions). Additionally, the proxy error was taken into account by incorporating the RMSE (sixth column, Tab 2) twice into the definition of the threshold [38]. One more technical criterion is that we need a sufficient number of surviving models after the intersection of all error metric thresholds

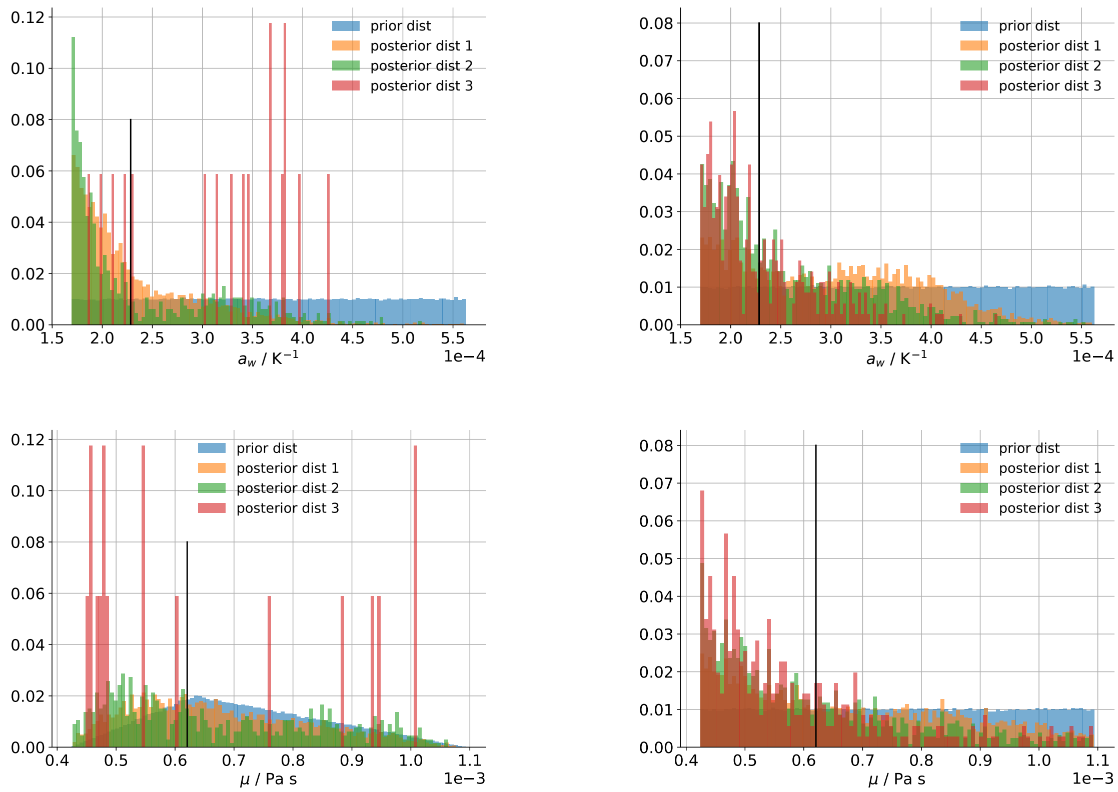


Figure 11: Parameter estimation for a selection of input quantities from Monte-Carlo sampling and after history match filtering for non-uniform (left) and all uniform (right) priors with the filter conditions listed in Tab. 2. The black line marks the input parameter for the creation of the synthetic experimental data.

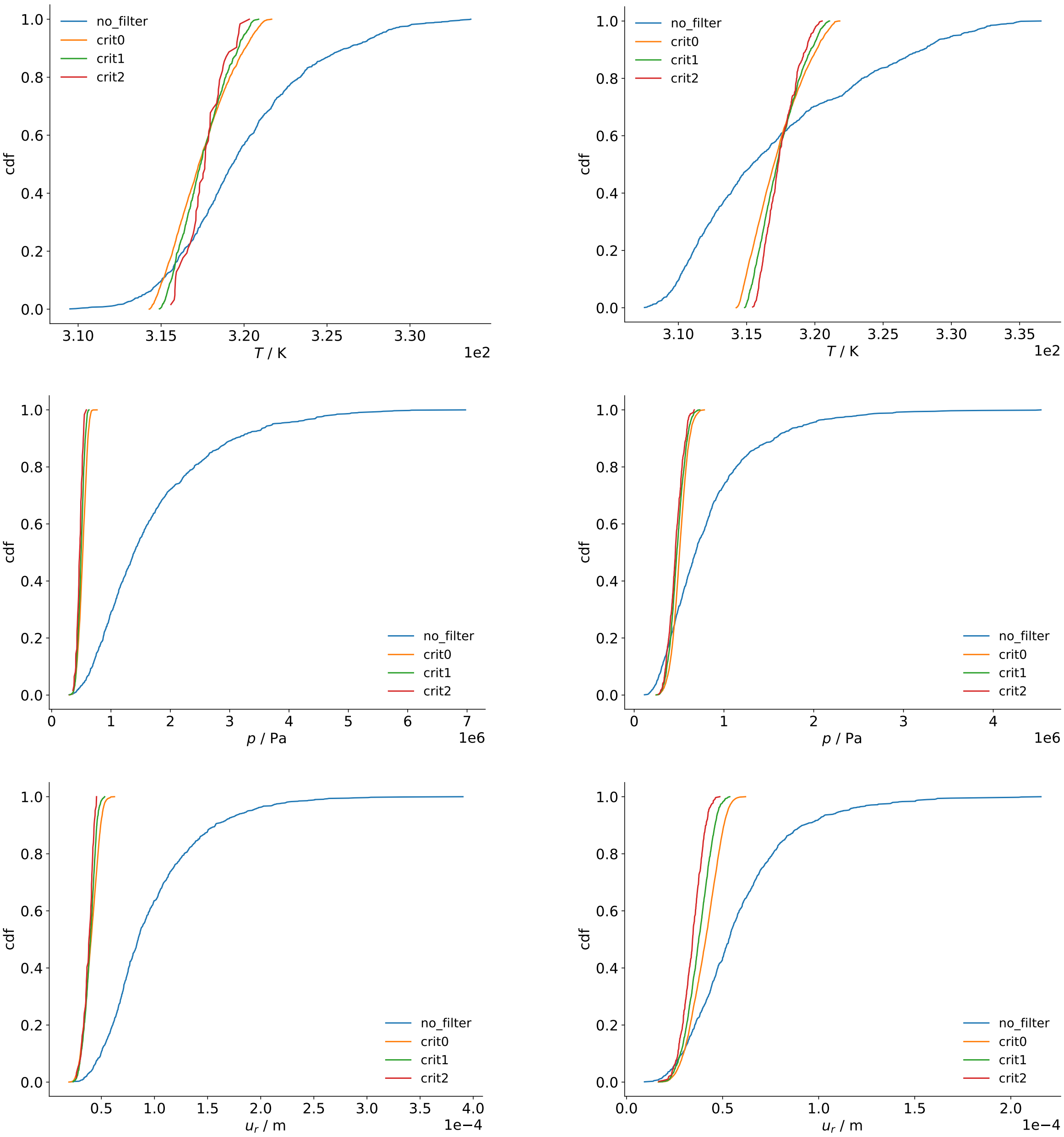


Figure 12: Cumulative distribution functions of the last history match time step obtained from Monte-Carlo sampling and after applying the testing history match conditions, see Tab. 2.

Table 3: Relative p50 values for all three testing filters with respect to the p50 value obtained from Monte-Carlo sampling.

	p^T		p^p		p^{u_x}	
	non-uniform	uniform	non-uniform	uniform	non-uniform	uniform
crit0	0.006866	0.003734	0.632036	0.266558	0.514512	0.251206
crit1	0.006673	0.004797	0.641845	0.286974	0.518569	0.315963
crit2	0.007089	0.006498	0.655329	0.319417	0.577460	0.340364

to perform a statistical analysis afterwards if probabilities are to be attached to the outcome. With the matched criteria, 'full' model (AM) runs were conducted again, to check whether the proxy estimates and history match criteria are in sufficient agreement with our assumptions and our conception of the history match (Fig. 13).

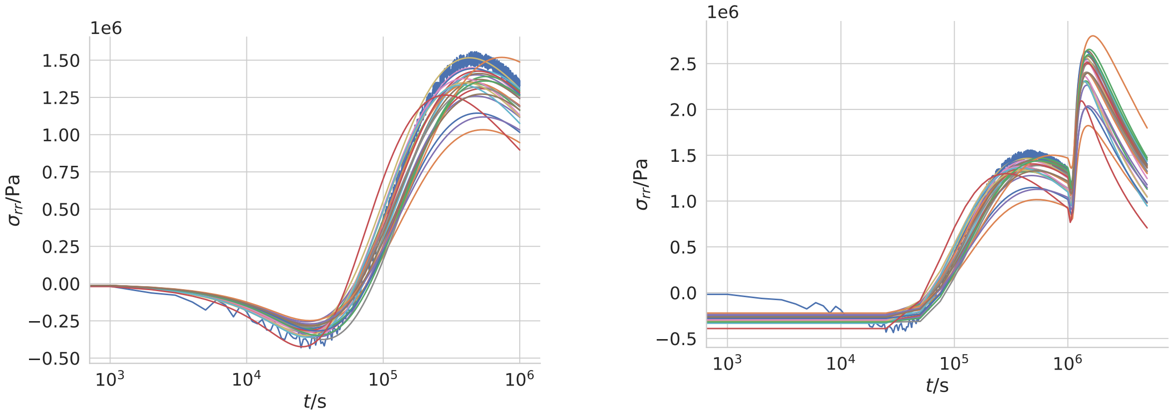
4.4. Forecast

Analogous to the (AM)-model runs of the history match, we repeated the same calculation this time with an equivalent numerical model (NM) with a subsequent forecast time ($1 \cdot 10^6\,\mathrm{s} - -5 \cdot 10^6\,\mathrm{s}$). From $1 \cdot 10^6\,\mathrm{s}$ to $1.1 \cdot 10^6\,\mathrm{s}$ the power of the heat source is linearly increased from 300 W to 600 W and kept constant until the end of the run. The corresponding stress response over the entire simulation time is presented in Fig. 13. The changes due to the altered source term conditions are clearly visible and can be confirmed or rejected *a posteriori* by experiments or monitoring data. In case of a rejection, the model would require further adaptations. In this study, we assume for simplicity that the chosen observation quantities are somewhat representative of the safety functions to be monitored. The corresponding CDFs of the last forecast step, yield analogous results as shown in Fig. 12 and are not subject to further analysis in this study.

4.5. Proxy-Based Global Sensitivity Analysis

Now that model uncertainty has been investigated, it is of practical interest to attribute this uncertainty to individual parameters or their combinations. In other words, one would like to understand how variations of the model input affect the history match error of the response functions. For that purpose, we analyzed the PM in terms of Sobol's indices by using their Monte-Carlo estimates.

For this purpose, we sampled the input space between their *min* and *max* values using the sampling scheme of Saltelli [29]. First and second-order indices were calculated using their Monte-



(a) history match (analytical model)

(b) history match and forecast obtained from numerical model

Figure 13: σ_{rr} as function of time at point $P = (0.5\,\mathrm{m}, 0.0, 0.0)$. a) Curves from AM-model satisfying all history match criteria. b) σ_{rr} results obtained from numerical model (NM) for the same parameter set including forecast with modified source term. In both cases the experimental curve is given in blue in background.

Carlo estimates. For all response parameters, we used a sampling size of $10,000$, which was shown to be sufficient for the width of the 0.95 confidence interval to be well below 0.025 for all indices (cf. Fig. 10 SI).

All in all, the global sensitivity analysis for our five error metric proxies provides a clear picture of the influence of single parameters on the model output. The results for both first- and second-order indices are given in Fig. 14. For the second-order indices, we present only values that exceed the error margin of 0.025. The bounds for the posterior analysis were estimated by the extreme values found during parameter estimation after applying the final filter condition (cf. Fig. 10 and Fig. 11). One general trend that attracts our attention is that more parameters become influential after applying the filter condition. Whereas the thermal conductivity K_s overwhelmingly dominates the influence on the temperature, after filtering, also other factors become important on the temperature. The analysis of the main effects is very much in agreement with the findings during parameter screening (Sec. 4.1). There, we found that ρ_w, ρ_s, K_w, c_w are non-influential and can be neglected for uncertainty analysis. The analysis also confirms the findings of Sec. 4.3: heavy hitters like K_s, k, or a_w changed their behavior after filtering. Looking at the second-order indices also reveals that the shift for K_s and the 'multimodal' behavior, we found for k, μ or a_w can be very much attributed to interaction effects between different parameters. At this point, it is important to note again that due to the challenges the analytical solution poses in terms of

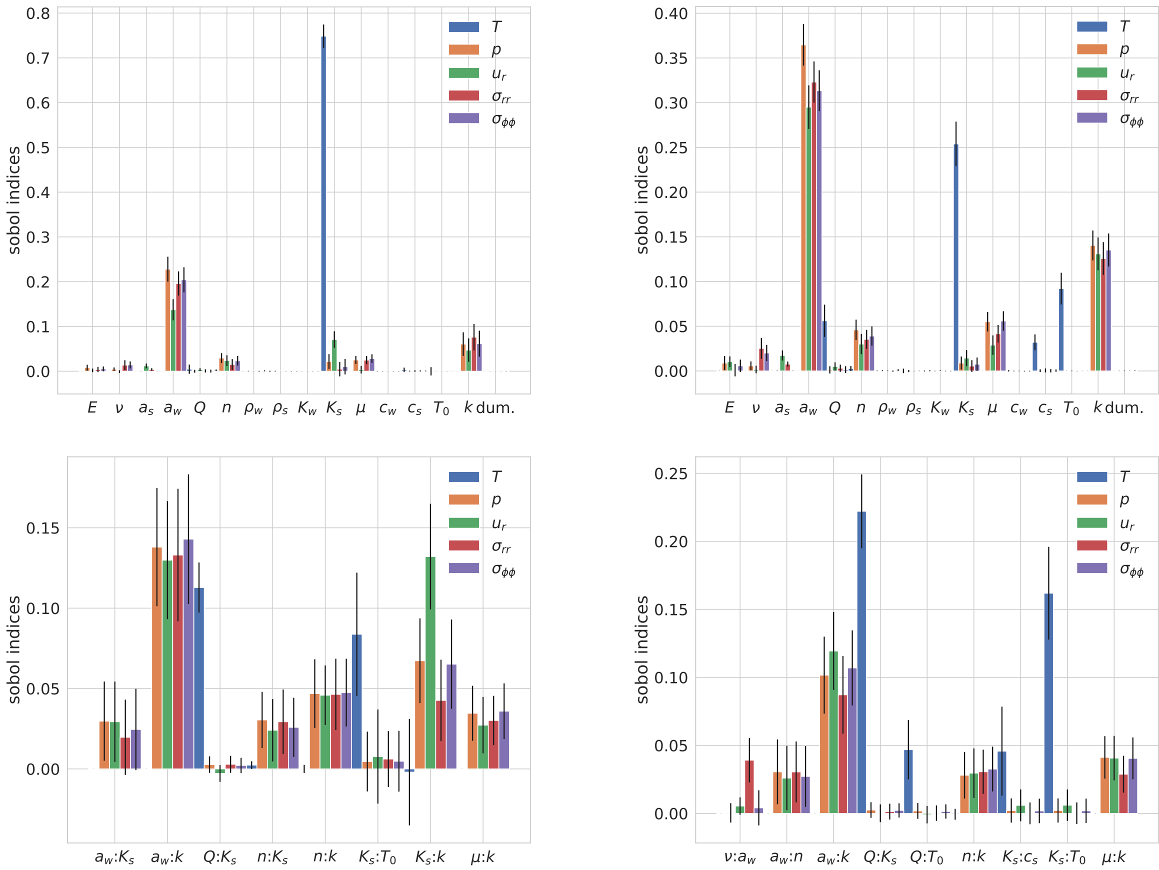


Figure 14: Sobol indices based on prior assumptions (left) and posterior (right) bounds. The upper graphs contain only indices of first order effects, while the lower contain second order interactions. For the second order terms, only combinations with effects greater than 0.025 are shown.

constant parameter input, we treated all of the water-related constants as uncertain themselves. From the GSA we can also conclude that we can treat ρ_w, K_w, and c_w as constants, while the functional dependence of μ and a_w seems to be relevant for further analysis utilizing finite element software.

5. Conclusions

In our work, we scrutinized the applicability of the Design of Experiments (DoE)-based assisted history matching for uncertainty quantification in linear coupled thermo-hydro-mechanical models.

In our manuscript, we used an analytical model of a simplified geotechnical problem in the form of a disposal cell containing heat-emitting radioactive waste emplaced in an isotropic fluid-saturated medium under realistic parameter conditions.

As these parameters often cannot be described in terms of a known probability distribution function (PDF), we compared PDFs based on expert opinion with a case of all uniform input parameter distributions. The most important findings of the presented work are:

- While in order to find a good history match, the filtered response must be covered by a valid input parameter range, it was shown in the present study that the exact form of the input parameter distribution becomes less critical. The result that the determining factors are the filtering conditions instead of the input parameter distribution makes the approach particularly interesting for our purpose in dealing with epistemic uncertainties as precise distribution forms are not known for the entire system.
- It was found that a detailed exploration of input parameter distributions before modeling is beneficial in reducing the uncertainty space and improving the proxy quality. At the same time, precise knowledge of all parameters is not required in order to obtain good history matched results. However, this comes at a cost in computing time as a broader sampling might be required.
- One disadvantage of the workflow, at first sight, is the somehow subjective choice of the history match filtering conditions. As we could show, the filter size can have an effect on the stochastic outcome (like percentiles), which can be crucial. The filter size describes the effect of uncertainty reduction due to the agreement of model output with experimental data. Therefore, care should be taken in quantifying these discrepancies for defining the history match thresholds. In case of doubt, one should assume them to be less tight. However, this is also part of a more general problem when dealing with the quantification of uncertainties that are of epistemic origin and involve model predictions: The meaning of exact numerical values is often overrated, and the usage of rigorous mathematical concepts often obscures the fact that the underlying problem eludes an exact quantification.
- As the history matching procedure reduces the uncertainty space significantly, it also affects the sensitivity, i.e. the relative ranking of input parameters. Therefore, conclusions drawn regarding sensitivity analyses prior to an experiment may have to be re-evaluated after (more) experimental data has become available.

In our manuscript, we showed that the workflow is particularly suitable for uncertainty quantification, sensitivity analysis, and model validation in geotechnical applications like radioactive waste repositories. However, before turning directly to real-world applications, the conceptual validity and computational feasibility of even more complex models incorporating non-linear phenomena (e.g. equations of state, material behavior) and spatial heterogeneities need to be demonstrated.

Acknowledgments

This work was funded in parts by the German Federal Ministry of Education and Research (BMBF) as part of the project Integrity of nuclear waste repository systems – Cross-scale system understanding and analysis – iCross (02NUK053E) and Impulse and Networking Funds of the Helmholtz Association (SO-093). We are also very grateful to the OpenGeoSys developer team for their enthusiastic, continuous work on further developing and improving the OGS platform for the scientific community.

References

[1] J. C. Helton, J. D. Johnson, C. J. Sallaberry, C. B. Storlie, Survey of sampling-based methods for uncertainty and sensitivity analysis, Reliab Eng Syst Saf 91 (10-11) (2006) 1175–1209.

[2] C.-F. Tsang, O. Stephansson, J. Hudson, A discussion of thermo–hydro–mechanical (THM) processes associated with nuclear waste repositories, Int J Rock Mech Min Sci 37 (1-2) (2000) 397–402.

[3] F. Bernier, F. Lemy, P. De Cannière, V. Detilleux, Implications of safety requirements for the treatment of THMC processes in geological disposal systems for radioactive waste, J Rock Mech Geotech Eng 9 (3) (2017) 428–434.

[4] X.-L. Li, J. Lanru, P. Blaser, Impact of Thermo-Hydro-Mechanical- Chemical (THMC) processes on the safety of underground radioactive waste repositories. Proceedings of the European Commission TIMODAZ-THERESA International COnference, Luxembourg, 29 September - 1 October 2009, in: Proceedings of the European Commission TIMODAZ-THERESA International Conference. Luxembourg: Publications Office, no. October, 2009, pp. 489–494.

[5] C. F. Tsang, J. D. Barnichon, J. Birkholzer, X. L. Li, H. H. Liu, X. Sillen, Coupled thermo-hydro-mechanical processes in the near field of a high-level radioactive waste repository in clay formations, Int J Rock Mech Min Sci 49 (2012) 31–44.

[6] J. C. Helton, Uncertainty and sensitivity analysis techniques for use in performance assessment for radioactive waste disposal, Reliab Eng Syst Saf 42 (2-3) (1993) 327–367.

[7] R. G. Ghanem, P. D. Spanos, Stochastic Finite Elements: A Spectral Approach, 1991.

[8] G. Stefanou, The stochastic finite element method: Past, present and future, Comput Methods Appl Mech Eng 198 (9-12) (2009) 1031–1051.

[9] M. Eiermann, O. G. Ernst, E. Ullmann, Computational aspects of the stochastic finite element method, in: Computing and Visualization in Science, 2007.

[10] F. Tonon, A. Bernardini, A. Mammino, Reliability analysis of rock mass response by means of Random Set Theory, Reliab Eng Syst Saf 70 (3) (2000) 263–282.

[11] I. M. Sobol, Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates, Math Comput Simul 55 (1-3) (2001) 271–280.

[12] A. Saltelli, P. Annoni, I. Azzini, F. Campolongo, M. Ratto, S. Tarantola, Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index, Comput Phys Commun 181 (2) (2010) 259–270.

[13] A. L. Eide, L. Holden, E. Reiso, S. I. Aanonsen, Automatic history matching by use of response surfaces and experimental design, in: ECMOR IV-4th European conference on the mathematics of oil recovery, 1994.

[14] R. D. Hurrion, An example of simulation optimisation using a neural network metamodel: Finding the optimum number of kanbans in a manufacturing system, J Oper Res Soc 48 (11) (1997) 1105–1112.
[15] G. I. Schuëller, H. J. Pradlwarter, P. S. Koutsourelakis, A critical appraisal of reliability estimation procedures for high dimensions, Probabilistic Eng Mech 19 (4) (2004) 463–474.
[16] E. Zio, G. E. Apostolakis, Two methods for the structured assessment of model uncertainty by experts in performance assessments of radioactive waste repositories, Reliab Eng Syst Saf 54 (2-3) (1996) 225–241.
[17] K. Sentz, S. Ferson, Combination of Evidence in Dempster- Shafer Theory, Vol. 4015, Citeseer, 2002.
[18] Bounding uncertainty in civil engineering: Theoretical background and applications, Vol. 539, Springer Science & Business Media, 2012.
[19] H. F. Schweiger, G. M. Peschl, Reliability analysis in geotechnics with the random set finite element method, Comput Geotech 32 (6) (2005) 422–435.
[20] C. F. Tsang, O. Stephansson, L. Jing, F. Kautsky, DECOVALEX Project: From 1992 to 2007, Environ Geol 57 (6) (2009) 1221–1237.
[21] 25 years of DECOVALEX - Scientific advances and lessons learned from an international research collaboration in coupled subsurface processes, Int J Rock Mech Min Sci 122 (2019) 103995.
[22] S. Olivella, A. Gens, J. Carrera, E. E. Alonso, Numerical formulation for a simulator (CODE_BRIGHT) for the coupled analysis of saline media, Eng. Comput. (Swansea, Wales) 13 (7) (1996) 87–112.
[23] J. Rutqvist, C. Tsang, TOUGH-FLAC: a numerical simulator for analysis of coupled thermal-hydrologic-mechanical processes in fractured and porous geological media under multi-phase flow conditions, in: Proceedings of the TOUGH Symposium, no. July, Lawrence Berkeley Natl. Lab. Berkeley, CA, 2003, pp. 1–9.
[24] O. Kolditz, S. Bauer, L. Bilke, N. Böttcher, J. O. Delfs, T. Fischer, U. J. Görke, T. Kalbacher, G. Kosakowski, C. I. McDermott, C. H. Park, F. Radu, K. Rink, H. Shao, H. B. Shao, F. Sun, Y. Y. Sun, A. K. Singh, J. Taron, M. Walther, W. Wang, N. Watanabe, Y. Wu, M. Xie, W. Xu, B. Zehner, OpenGeoSys: An open-source initiative for numerical simulation of thermo-hydro-mechanical/chemical (THM/C) processes in porous media, Environ Earth Sci 67 (2) (2012) 589–599.
[25] B. M. Adams, K. R. Dalbey, M. S. Eldred, D. M. Gay, L. P. Swiler, W. J. Bohnhoff, J. P. Eddy, P. D. Hough, DAKOTA, A Multilevel Parallel Object Oriented Framework for Design Optimization, Parameter Estimation, Uncertainty Quantification, and Sensitivity Analysis, Manual (December 2009) (2010) User Manual.
[26] E. Patelli, H. Pradlwarter, COSSAN-X : A General Purpose Code for Computational Stochastic Structural Analysis COSSAN-X : A General Purpose Code for Computational Stochastic Structural Analysis Institute of Engineering Mechanics University of Innsbruck, in: IV European Conference on Computational Mechanics, Paris, France, EU, no. August 2018, 2010.
[27] X. Li, Entwicklung der Softwareplattform RESUS : Repository Simulation , Uncertainty propagation and Sensitivity Analysis Dissertation, Clausthal University of Technology, Clausthal-Zellerfeld, Lower Saxony, Germany, 2015.
[28] D. Draper, A. Pereira, P. Prado, A. Saltelli, R. Cheal, S. Eguilior, B. Mendes, S. Tarantola, Scenario and parametric uncertainty in GESAMAC: A methodological study in nuclear waste disposal risk assessment, Comput Phys Commun 117 (1) (1999) 142–155.
[29] A. Saltelli, S. Tarantola, On the relative importance of input factors in mathematical models: Safety assessment for nuclear waste disposal, J Am Stat Assoc 97 (459) (2002) 702–709.
[30] S. J. Wang, K. C. Hsu, Stochastic Analysis of a Thermal Uncoupled Thermal-Hydraulic-Mechanical Model, in: Poromechanics 2017 - Proceedings of the 6th Biot Conference on Poromechanics, American Society of Civil Engineers (ASCE), 2017, pp. 787–794.
[31] L. Nguyen-Tuan, T. Lahmer, M. Datcheva, E. Stoimenova, T. Schanz, A novel parameter identification approach for buffer elements involving complex coupled thermo-hydro-mechanical analyses, Comput Geotech.
[32] N. Watanabe, Finite element method for coupled thermo-hydro-mechanical processes in discretely fractured and non-fractured porous media, Ph.D. thesis, Dresden University of Technology (2011).
[33] R. Lewis, B. Schrefler, The Finite Element Method in the Static and Dynamic Deformation and Consolidation of Porous Media, 1st Edition, J. Wiley and Sons, 1998, an optional note.
[34] W. Ehlers, J. Bluhm, Porous Media: Theory, Experiments and Numerical Applications, 1st Edition, Springer, 2002.
[35] J. R. Booker, C. Savvidou, Consolidation around a point heat source., Int J Numer Anal Methods Geomech 9 (2) (1984) 173–184.
[36] A. A. Chaudhry, J. Buchwald, O. Kolditz, T. Nagel, Consolidation around a point heat source (correction and verification), Int J Numer Anal Methods Geomech 43 (18) (2019) 2743–2751.
[37] H. R. Müller, B. Garitte, T. Vogt, S. Köhler, T. Sakaki, H. Weber, T. Spillmann, M. Hertrich, J. K. Becker,

N. Giroud, V. Cloet, N. Diomidis, T. Vietor, Implementation of the full-scale emplacement (FE) experiment at the Mont Terri rock laboratory, in: Swiss Journal of Geosciences, Vol. 110, Springer, 2017, pp. 287–306.
[38] B. Li, E. W. Bhark, S. J. Gross, T. C. Billiter, K. Dehghani, Best practices of assisted history matching using design of experiments, SPE J 24 (4) (2019) 1435–1451.
[39] C. Maschio, D. J. Schiozer, Bayesian history matching using artificial neural network and Markov Chain Monte Carlo, J Pet Sci Eng 123 (2014) 62–71.
[40] C. Plúa, V. Manon, D. Seyedi, G. Armand, J. Rutqvist, J. Birkholzer, H. Xu, R. Guo, K. Tatcher, A. Bond, W. Wang, T. Nagel, H. Shao, O. Kolditz, lBNL-2001265 (2020).
[41] M. Bossart, P. Thurry, Characteristics of the Opalinus Clay at Mont Terri, Reports of the Swiss Geological Survey no. 3 (3) (2008) 26.
[42] G. Armand, F. Bumbieler, N. Conil, R. de la Vaissière, J. M. Bosgiraud, M. N. Vu, Main outcomes from in situ thermo-hydro-mechanical experiments programme to demonstrate feasibility of radioactive high-level waste disposal in the Callovo-Oxfordian claystone, J Rock Mech Geotech Eng 9 (3) (2017) 415–427.
[43] R. L. Plackett, J. P. Burman, The Design of Optimum Multifactorial Experiments, Biometrika 33 (4) (1946) 305.
[44] GPy, GPy: A gaussian process framework in python (2014).
URL `http://github.com/SheffieldML/GPy`
[45] L. Bilke, B. Flemisch, T. Kalbacher, O. Kolditz, R. Helmig, T. Nagel, Development of open-source porous-media simulators: principles and experiences., Transp Porous Media.
[46] S. Chatterjee, A. Hadi, Regression Analysis by Example, Wiley Series in Probability and Statistics, Wiley, 2006.
[47] O. Dubrule, Comparing splines and kriging, Comput Geosci 10 (2-3) (1984) 327–338.
[48] F. Xiong, Y. Xiong, W. Chen, S. Yang, Optimizing latin hypercube design for sequential sampling of computer experiments, Eng Optim 41 (8) (2009) 793–810.