Water Resources Research

RESEARCH ARTICLE
10.1029/2023WR034453

Surrogate-Model Assisted Plausibility-Check, Calibration, and Posterior-Distribution Evaluation of Subsurface-Flow Models

Jonas Allgeier[1,2] and **Olaf A. Cirpka**[1]

[1]Department of Geosciences, University of Tübingen, Tübingen, Germany, [2]Now at BoSS Consult GmbH, Stuttgart, Germany

Key Points:

- We use Gaussian Process Regression as proxy models to expedite the calibration of an expensive process-based hydrogeological model
- We estimate the full posterior parameter distribution by Markov-Chain Monte Carlo sampling using the proxy models
- We compare this distribution to results obtained by Neural Posterior Estimation

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

O. A. Cirpka,
olaf.cirpka@uni-tuebingen.de

Citation:

Allgeier, J., & Cirpka, O. A. (2023). Surrogate-model assisted plausibility-check, calibration, and posterior-distribution evaluation of subsurface-flow models. *Water Resources Research*, *59*, e2023WR034453. https://doi.org/10.1029/2023WR034453

Received 7 JAN 2023
Accepted 18 JUN 2023

Author Contributions:

Conceptualization: Jonas Allgeier, Olaf A. Cirpka
Funding acquisition: Olaf A. Cirpka
Investigation: Jonas Allgeier
Methodology: Jonas Allgeier
Software: Jonas Allgeier
Supervision: Olaf A. Cirpka
Visualization: Jonas Allgeier
Writing – original draft: Jonas Allgeier
Writing – review & editing: Olaf A. Cirpka

Abstract

Modern physics-based subsurface-flow models often require many parameters and computationally costly simulations. This prohibits traditional ensemble-based conditioning. We expedite the calibration of such models by using surrogate/proxy models based on Gaussian Process Regression (GPR). In an iterative procedure, we use the proxy models to (a) estimate the direction of steepest descent, (b) propose only parameter combinations for full-model runs that are likely to lead to plausible results, and (c) preselect proposed parameter combinations by their predicted performance. This method yields an ensemble of full-model runs covering the full plausible parameter space, but at higher resolution close to the optimum. This is the basis for Markov-Chain Monte Carlo (MCMC) simulations using GPR to estimate the posterior parameter distribution. We tested several variants of the scheme on a 3-D variably-saturated steady-state subsurface-flow model and compared it to a Neural Posterior Estimation (NPE) scheme, which requires samples of the prior distribution only. While the estimated posterior distributions of the two approaches were similar, the GPR-based MCMC approach reproduced the data better than samples from the NPE-based posterior distributions.

1. Introduction

Process-based numerical modeling of subsurface flow is an important tool in hydrogeological research and groundwater-resources management. The computational power (in terms of hardware and software) has improved drastically over the last decades. At the same time, models have increased in complexity, as new numerical methods have become feasible, more processes can been considered and evermore detail can be represented in the models (e.g., Venkataraman & Haftka, 2004; Y. Zhou & Li, 2011; Jakob, 2014). A side effect of this development is that modern models tend to have many adjustable parameters. The process of estimating values of model parameters so that the model output reasonably agrees with measured data is known as *model calibration*, *inverse modeling*, or *parameter inference* (e.g., Carrera et al., 2005; Hill & Tiedeman, 2006; H. Zhou et al., 2014). Among the large number of calibration philosophies and methods, the most classical approach is *manual calibration*, in which the modeler tests different parameter sets until a satisfactory agreement with the observations is achieved. Albeit still being used regularly in practice, it is generally regarded as tedious, inefficient, irreproducible, intransparent, and unable to truly find optimal parameter sets (Beckers et al., 2020; Carrera et al., 2005). Consequently, automated-calibration procedures have become more popular with the rise of computational abilities (e.g., Solomatine et al., 1999; Yeh, 1986). An arbitrarily chosen example is the Parameter ESTimation suite (PEST; Doherty, 2015; Doherty et al., 1994; Doherty & Hunt, 2010), a general-purpose calibration toolbox based on the Levenberg-Marquardt method (Levenberg, 1944; Marquardt, 1963) with extensions and variants, that is especially popular within the community of hydrogeology (e.g., Selle et al., 2013).

Automated calibration schemes like PEST aim to minimize a scalar metric quantifying the differences between simulations and observations (i.e., an objective function). Gradient-based methods, like the Gauß-Newton method and its descendants (e.g., the Levenberg-Marquardt method (Levenberg, 1944; Marquardt, 1963) and the trust-region reflective methods (Conn et al., 2000; Powell, 1970a, 1970b)) are comparably efficient in finding a minimum of the objective function. Unfortunately, they can neither guarantee that this is the global minimum, nor do they provide reliable uncertainty estimates if the functional dependence between model parameters and observables is nonlinear. Ensemble-based methods, like genetic algorithms (Das & Suganthan, 2011; Gen & Cheng, 1999; Goldberg, 1989) and Markov-Chain Monte Carlo (MCMC) methods (S. Brooks et al., 2011; Gilks et al., 1995) are better in finding the global minimum and may provide a good approximation of the parameter distribution conditioned on the measurements, but they require orders of magnitude more model runs than parameters, which can be prohibitive for computationally expensive models with long run times.

For these reasons, a special branch of calibration research tries to develop global calibration schemes that are as efficient as possible (see Haftka et al., 2016, for a detailed review). These methods are typically based on proxy models (also referred to as surrogate models), which might be a coarsened version of the original model or a black-box-type approximation relating input parameters sets to model-simulated observations in a simplified way (e.g., through machine-learning methods or by interpolation in parameter space). The underlying assumption is that the proxy model is considerably quicker to evaluate than the original model (at the cost of accuracy). The proxy models are used to select the most promising point(s) in parameter space for an evaluation with the full model. After evaluation, the proxy model is retrained and the next iteration starts. The variants of such proxy-model-assisted calibration schemes differ mostly in the utilized proxy model, the mechanism of proposing points, and the criteria to select points for model evaluation.

A key issue in proxy-model supported global-estimation methods is to strike a balance between parameter-space *exploration* and *exploitation*. In this context, exploration refers to drawing and testing points, where the expected model outcome is most uncertain in order to learn more about the model behavior. Most of the time, these points tend to be in regions of the parameter space which are far away from all previously evaluated points. In contrast, exploitation tries to make use of the available points in such a way, that new points are drawn and evaluated where the model is expected to show the best performance (i.e., have the smallest value of the objective function). An algorithm solely focusing on exploration might take forever to find a reasonably good point to satisfy the convergence criteria, while an algorithm only based on exploitation might quickly get stuck in a local minimum.

One of the earliest examples making use of a proxy-model assisted global optimization method is the Efficient Global Optimization algorithm developed by Jones et al. (1998). In each iteration it identifies a single promising point by optimization of the "expected improvement" metric, which balances the predicted objective function value with the estimated uncertainty. Regis and Shoemaker (2007) introduced the *surrogate-distance metric* to select points in parameter space to be tested with the full model from a randomly generated set of points, in which the proxy-model predicted value of the objective function and the distance to all previous points are merged. Regis and Shoemaker (2009) parallelized the approach. In each iteration, multiple points are iteratively selected from a set of randomly proposed points. All points are then evaluated in parallel using a computing cluster. Wang and Shoemaker (2014) and Xia et al. (2021) extended this general concept by more sophisticated point proposals, where the number of adjustable parameters is restricted over the course of the calibration.

In this study, we propose and apply another variant of proxy-model assisted calibration on the basis of the scheme of Regis and Shoemaker (2009). We extend the method to account for model plausibility, in which plausibility criteria are typically smoothed inequality constraints on secondary outputs of the model, such as the direction of flow at a fixed-head boundary (Allgeier et al., 2020; Erdal & Cirpka, 2019; Erdal et al., 2020). We also test different approaches of constructing the proxy model, and restricting the search direction from previous points considering the gradient of the objective function as estimated by the proxy model.

Global model calibration is associated with finding a single optimal parameter set for a given model and a set of observations. In many cases, especially for models with numerous parameters, there are different points in parameter space that can produce similar or even identical results, which makes the problem of global calibration ill-posed (e.g., H. Zhou et al., 2014). This is known as equifinality problem (Beven, 2006). Bayesian calibration methods address this by estimating the distribution of the calibrated parameters conditioned on the observations (e.g., Beckers et al., 2020; Mohammadi et al., 2018). Markov-Chain Monte Carlo methods represent a particularly popular example in the Bayesian toolbox. However, also these techniques suffer from the large number of model evaluations, which may make them unaffordable for complex process-based models.

In this study, we circumvent this problem in two different ways to still construct full posterior parameter distributions for our calibration problem. First, we define a proxy-model from those runs of the original model that were conducted during the global calibration. We then apply MCMC sampling to the much faster proxy variant rather than the original model. As an alternative, we generate a full posterior parameter distribution through a likelihood-free Simulation-Based Inference (SBI) method (Cranmer et al., 2020; Lueckmann et al., 2021; Tejero-Cantero et al., 2020), namely NPE. This tool is based on machine learning and requires only a collection of model inputs and outputs sampled from the prior distribution.

We compare the two resulting posterior distributions with each other and to the results of the optimization-based global model calibration. This allows us to (a) gain insights beyond a best-estimate parameter set, like the assessment of parametric uncertainty and correlation, and (b) assess which posterior construction method is more appropriate/reliable for our problem.

The remaining of this paper is structured as follows. In Section 2 we explain our methodology. In Section 3 we give a brief introduction into the test application. We then present and discuss the results in Section 4. Finally we draw conclusions and give suggestions for further research in Section 5.

2. Methods

2.1. Model

We denote the full model $\mathcal{M}$ and the associated n-tuple of parameters $\mathbf{p}$, where the latter has the dimension d. A model realization with a specific parameter set $\mathbf{p}$ results in the model outcome:

$$\{\boldsymbol{\vartheta}^{\bullet}, \boldsymbol{\varphi}^{\bullet}\} = \mathcal{M}(\mathbf{p}), \tag{1}$$

which is split into two parts: $\boldsymbol{\vartheta}^{\bullet}$ contains n_{obj} quantities needed for the evaluation of an objective function (explained in the following), whereas $\boldsymbol{\varphi}^{\bullet}$ consists of all n_{plaus} quantities necessary for a plausibility assessment (also explained in the following). We use the symbol "•" to indicate the output of a full model run, whereas outputs of the proxy-model are marked with the symbol "∘."

In our application, the model is a variably-saturated subsurface-flow model, solving the nonlinear Richards equations at steady state. The primary unknowns are pressure heads at nodes of the computational grid, but the sets of model outcomes $\boldsymbol{\vartheta}$ and $\boldsymbol{\varphi}$ may also include derived quantities, like fluxes across a boundary. The parameters describe hydraulic properties of subsurface units, boundary conditions, or the geometry of geological features. In the following, a particular parameter set $\mathbf{p}$ will be referred to as a point in parameter space.

2.2. Parameter Transformation

Directly working with the "physical" parameter values of the model is disadvantageous because different parameters have different units and also range over very different scales. To make parameters comparable and allow for meaningful interpretation of parameter combinations, we introduce a double parameter transformation on the individual parameters. The first one converts any physical parameter p_i to its prior cumulative probability $p_{\text{cdf},i}$, requiring a prior marginal probability function of that parameter, which can be chosen according to expert knowledge (normal, log-normal, scaled and shifted uniform distribution, …). $p_{\text{cdf},i}$ ranges between 0 and 1. Applied to all parameters, we denote this transformation $\mathbf{f}_{\text{cdf}}$ and its inverse $\mathbf{f}_{\text{cdf}}^{-1}$:

$$\mathbf{p}_{\text{cdf}} = \mathbf{f}_{\text{cdf}}(\mathbf{p}) \tag{2}$$

$$\mathbf{p} = \mathbf{f}_{\text{cdf}}^{-1}(\mathbf{p}_{\text{cdf}}). \tag{3}$$

We introduce a second transformation via the inverse logit transform, to re-scale the cumulative distribution values onto the unbounded interval $(-\infty, \infty)$:

$$\tilde{p}_i = f_{\text{scl}}(p_{\text{cdf},i}) = \log\left(\frac{p_{\text{cdf},i}}{1 - p_{\text{cdf},i}}\right). \tag{4}$$

This leads to a homogenized vector $\tilde{\mathbf{p}}$ with entries ranging from $-\infty$ to ∞ following a standard logistic distribution. By this our calibration methods do not need to account for boundaries and can explore the parameter space freely. By construction, the origin of the re-scaled parameter space is located at the prior median of all physical parameters. This makes it easy to interpret positive or negative numbers, as they mean "larger than" or "smaller than" the median.

For the re-scaling, any distribution that maps the space (0, 1) to $(-\infty, \infty)$ (e.g., the inverse Gaussian distribution) is equally applicable. We prefer the unscaled logit transformation, because it shows slightly more tailing than a similarly scaled probit transformation. It also has the advantage of a trivial formulation for both forward and backward transformation. A disadvantage is that conditioning techniques that require Gaussian distributions, like Kalman Filtering, are not applicable. As the mapping onto the standard normal distribution is denoted "Gaussian anamorphosis" (Everitt & Skrondal, 2010; Wackernagel, 2003), the transformation suggested here might be denoted "logistic anamorphosis."

In summary, we have three equivalent and effortlessly convertible ways to express parameter sets. We use $\tilde{\mathbf{p}}$ for calibration-internal calculations and transform these via $\mathbf{p}_{\mathrm{cdf}}$ to $\mathbf{p}$, whenever a full model run is desired.

2.3. Objective Function

The objective function f_{obj} compares the $\boldsymbol{\vartheta}$-part of the model output (which could be a proxy-model outcome $\boldsymbol{\vartheta}^{\circ}$ or a full-model output $\boldsymbol{\vartheta}^{\bullet}$) to a target data set $\boldsymbol{\vartheta}^{*}$ and provides a scalar quantity of agreement y. In this study, we use the common sum of squared normalized residuals as objective function:

$$y(\mathbf{p}) = f_{\mathrm{obj}}\left(\boldsymbol{\vartheta}^{*}, \boldsymbol{\vartheta}(\mathbf{p})\right) = \sum_{i=1}^{n_{\mathrm{obj}}} \left(\frac{\vartheta_i^{*} - \vartheta_i(\mathbf{p})}{\sigma_{\mathrm{obj},i}}\right)^2, \tag{5}$$

in which $\sigma_{\mathrm{obj},i}$ is the uncertainty of the ith element of $\boldsymbol{\vartheta}^{*}$. For the purpose of finding the best estimate, we will set $\sigma_{\mathrm{obj},i}$ to unity for all observations, so that the objective function becomes the sum of squared errors, whereas in the estimation of the full posterior distribution of parameters we will estimate a common value σ_{obj} for all measurements.

The best-estimate calibration aims to minimize $y(\mathbf{p})$:

$$\mathbf{p}_{\mathrm{best}} = \underset{\mathbf{p}}{\operatorname{argmin}}\, y(\mathbf{p}), \tag{6}$$

where $\mathbf{p}_{\mathrm{best}}$ is the best parameter set found (ideally the global optimum).

2.4. Plausibility Function

Ideally, the model calibration should result in a parameter set $\mathbf{p}_{\mathrm{best}}$ that not only produces a model realization with optimal agreement of measured and modeled observations, but is also *plausible*. We consider (im)plausibility a property of a model realization that would be more or less obvious to a human modeler when looking at the output. An implausible model might, for example, have fluxes across model boundaries that are orders of magnitudes off or even have the wrong sign. We want to exclude parameter sets that lead to implausible results altogether. We believe that restricting the calibration to plausible points is also helpful in minimizing the objective function.

In order to constrain the parameter space to its plausible subset, we define a plausibility function $f_{\mathrm{plaus}}(\mathbf{p})$ that takes the $\boldsymbol{\varphi}$-part of the model output (which could be $\boldsymbol{\varphi}^{\circ}$ or $\boldsymbol{\varphi}^{\bullet}$) and assigns a scalar quantity of plausibility based on some internal rules:

$$z = f_{\mathrm{plaus}}(\boldsymbol{\varphi}) = \prod_{i=1}^{n_{\mathrm{plaus}}} z_i(\varphi_i), \tag{7}$$

where z is the total plausibility score and $z_i \in [0, 1]$ is the plausibility score for the ith element of $\boldsymbol{\varphi}$, where $z_i = 0$ indicates a fully implausible and $z_i = 1$ a completely plausible behavior. Using the product of individual probability scores ensures that even if only a single criterion is not met, the overall assessment is "implausible." In some cases, the plausibility criteria are binary (e.g., when evaluating the sign of a flux), whereas in other cases it is difficult to define a hard criterion, so that we use smooth transitions between zero and one to indicate a plausibility fringe.

2.5. Proxy Models

For spatially distributed subsurface-flow models, like the one used in this study, a single model run may take several hours or even days. Calibration in the sense of global optimization typically requires many model runs, which means available time and/or computational resources can become a limitation, even when using computer clusters. We alleviate this problem by the use of proxy models. The actual calibration algorithm could then operate on the proxy-model, with occasional feedback between the full model and the proxy variant.

We use $\mathcal{P}$ to denote the proxy model. In our case, the proxy model is constructed (i.e., trained) from a set of known pairs of model input $\tilde{\mathbf{p}}$ and model output $\{\boldsymbol{\vartheta}^\bullet, \boldsymbol{\varphi}^\bullet\}$. The trained proxy model can provide a predicted model outcome for a specific re-scaled parameter vector $\tilde{\mathbf{p}}$ by approximating the outcome of the full model:

$$\{\boldsymbol{\vartheta}^\circ, \boldsymbol{\varphi}^\circ\} = \mathcal{P}(\tilde{\mathbf{p}}) \approx \mathcal{M}(\mathbf{p}) = \{\boldsymbol{\vartheta}^\bullet, \boldsymbol{\varphi}^\bullet\}, \tag{8}$$

where we use $\mathbf{p} = \mathbf{f}_{\text{cdf}}^{-1}\left(\mathbf{f}_{\text{scl}}^{-1}(\tilde{\mathbf{p}})\right)$. In analogy to the full model, the predicted model outcome consists of objective-function related quantities $\left(\boldsymbol{\vartheta}^\circ = \left[\vartheta_1{}^\circ, \ldots, \vartheta_{n_{\text{obj}}}{}^\circ\right]\right)$ and plausibility-function related quantities $\left(\boldsymbol{\varphi}^\circ = \left[\varphi_1{}^\circ, \ldots, \varphi_{n_{\text{plaus}}}{}^\circ\right]\right)$. To distinguish the proxy-model predictions from a full model output, we use the symbol "∘."

In our approach, we make use of Gaussian Process Regression (GPR) for proxy-modeling (Erdal et al., 2020; Jones et al., 1998; Kitanidis, 1997; Rasmussen & Williams, 2006). This technique is based on interpolating outcomes of full model runs in parameter space. The underlying concept and mathematical description is formally identical to kriging (Cressie, 1990; Krige, 1951; Matheron, 1963), but instead of interpolating in the two- or three-dimensional physical space, it is carried out in the higher-dimensional re-scaled parameter space (i.e., with $\tilde{\mathbf{p}}$).

The mean value μ of a prediction at a query point $\tilde{\mathbf{p}}^{\text{test}}$ based on GPR is given by:

$$\mu\left(\tilde{\mathbf{p}}^{\text{test}}\right) = \beta + \mathbf{r}^{\mathrm{T}}\mathbf{Q}^{-1}\left(\mathbf{y}^{\text{train}} - \mathbf{1}\beta\right), \tag{9}$$

where β is a trend coefficient, that can be directly inferred from the training data (input matrix $\tilde{\mathbf{P}}^{\text{train}}$, output vector $\mathbf{y}^{\text{train}}$), $\mathbf{1}$ denotes a vector of ones, whereas $\mathbf{Q}$ and $\mathbf{r}$ contain covariance values obtained from a covariance function C:

$$Q_{ij} = C\left(\tilde{\mathbf{P}}_i^{\text{train}}, \tilde{\mathbf{P}}_j^{\text{train}}, \boldsymbol{\ell}, \sigma^2\right), \tag{10}$$

$$r_i = C\left(\tilde{\mathbf{p}}^{\text{test}}, \tilde{\mathbf{P}}_i^{\text{train}}, \boldsymbol{\ell}, \sigma^2\right), \tag{11}$$

where σ^2 and $\boldsymbol{\ell}$ are hyperparameters optimized for any given training data set. In this study, we use the Matérn covariance function of order 3/2 (Matérn, 1960; Stein, 1999):

$$C\left(\tilde{\mathbf{p}}_a, \tilde{\mathbf{p}}_b, \boldsymbol{\ell}, \sigma^2\right) = \sigma^2\left(1 + \sqrt{6\sum_{i=1}^{d}\left(\frac{\tilde{p}_{i,a} - \tilde{p}_{i,b}}{\ell_i}\right)^2}\right)\exp\left(-\sqrt{6\sum_{i=1}^{d}\left(\frac{\tilde{p}_{i,a} - \tilde{p}_{i,b}}{\ell_i}\right)^2}\right), \tag{12}$$

implying that we can derive the partial derivatives of the mean predicted outcomes with respect to all rescaled parameters, that is the Jacobian of the proxy model, analytically (see Erdal et al., 2020).

We test two variants of the proxy model. In the first one, a single proxy model is used to directly relate input parameters to the value of the objective and plausibility functions, respectively, used in calibration:

$$y^\circ = \mathcal{P}_{\text{obj}}(\tilde{\mathbf{p}}) \approx f_{\text{obj}}\left(\boldsymbol{\vartheta}^*, \mathcal{M}(\mathbf{p})\right) = y^\bullet, \tag{13}$$

$$z^\circ = \mathcal{P}_{\text{plaus}}(\tilde{\mathbf{p}}) \approx f_{\text{plaus}}(\mathcal{M}(\mathbf{p})) = z^\bullet, \tag{14}$$

in which $\mathcal{P}_{\text{obj}}$ and $\mathcal{P}_{\text{plaus}}$ indicate that the GPR-interpolated variable is the value of the objective and the plausibility function, respectively. This approach is trivial to implement and computationally cheap, but it neglects large parts of the full-model behavior altogether. Two distinct input parameter sets might result in completely different model outputs $\boldsymbol{\vartheta}$, but rather similar objective function values y. The underlying structure is hidden from the proxy model, and the available information of the full-model runs (i.e., the raw outputs $\boldsymbol{\vartheta}$) is not utilized to the full extent. This holds for the plausibility function accordingly. Ultimately, this could result in suboptimal predictions of the proxy model.

In the second approach, we train one Gaussian Process Emulator (GPE) for each observation ϑ_i or for each contribution to the plausibility score φ_i and compute the estimated objective function from that:

$$y^\circ = f_{\text{obj}}\left(\boldsymbol{\vartheta}^*, \mathcal{P}_{\text{obs}}(\tilde{\mathbf{p}})\right) \approx f_{\text{obj}}\left(\boldsymbol{\vartheta}^*, \mathcal{M}(\mathbf{p})\right) = y^\bullet, \tag{15}$$

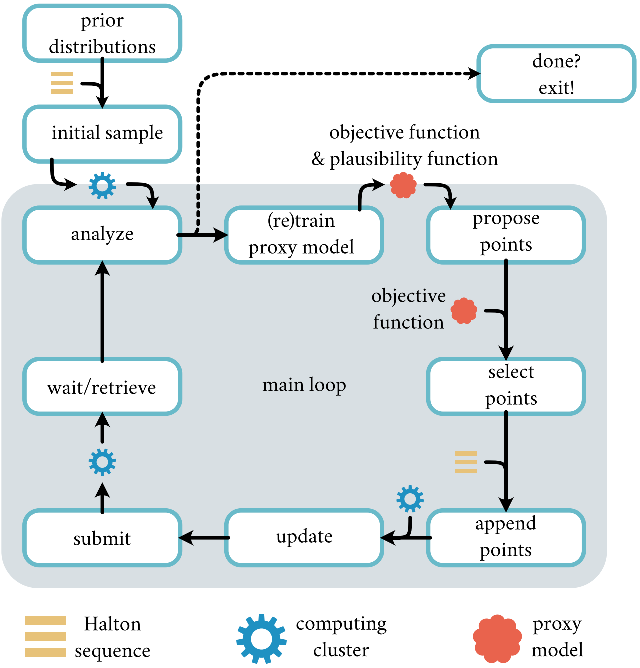

Figure 1. Flow chart of the procedure.

$$z^{\circ} = f_{\text{plaus}}(\mathcal{P}_{\text{crit}}(\tilde{\mathbf{p}})) \approx f_{\text{plaus}}(\mathcal{M}(\mathbf{p})) = z^{\bullet}, \tag{16}$$

in which $\mathcal{P}_{\text{obs}}$ and $\mathcal{P}_{\text{crit}}$ indicate that the proxy model predicts the observations and the plausibility criteria, respectively. As this (set of) proxy model(s) is more detailed, we expect it to yield a better approximation of the objective function for the same number of full-model runs. Obviously, this benefit comes at additional computational costs by training multiple rather than a single GPE which, however, can be trivially parallelized. For a limited number n_{obj} of independent observations (e.g., about one hundred or fewer), like in our steady-state subsurface-flow model, we deem these additional costs acceptable. This would be different in the calibration of transient models, where the objective function depends on time series. Here, additional data-reduction techniques would become necessary.

There is also an increase in the computational effort in the prediction step of the proxy model as multiple rather than a single GPR are performed at each probing point. Given the computational effort for running the full model and training the GPE, however, these extra costs are negligible.

2.6. Calibration Scheme

In order to find the best estimate of model parameters, we adapt the general parallelized proxy-model assisted calibration procedure of Regis and Shoemaker (2009), making use of parallel-computing capabilities. While we use different variants of the scheme, they are all based on the following general procedure, which is visually summarized in Figure 1:

1. *Prior Definitions*: For each model input parameter, a prior probability density function is defined. We do not consider prior correlations.
2. *Initial Sample* (see Section 2.6.1): We start with an initial set of points (in our case of size 120) drawn from the prior distributions in a space-filling way. The respective model realizations are generated and submitted to a computer cluster for evaluation.
3. *Main Loop*: This loop is executed until some termination criterion (e.g., based on convergence or computational budget) is met. In our case, we run all variations until 3070 model realizations were simulated with the full model. We call each iteration of this loop a *cycle*.
   (a) *Analyze*: The objective and plausibility functions are evaluated for all new model realizations. If the termination criterion is met, the main loop is terminated.
   (b) *(Re)train Proxy-Model(s)*: We keep track of all model inputs and outputs for all full model realizations. We train a proxy-model for both the objective function and the plausibility score.
   (c) *Propose Points* (see Section 2.6.2): We generate a set of $n_{\text{pro}} = 10^4$ plausible points scattered around those locations that have proven to be "good" so far (judged by the objective-function value).
   (d) *Select Points* (see Section 2.6.3): We iteratively apply the surrogate-distance metric for all (remaining) proposal points to select $n_{\text{set}} = 39$ points.
   (e) *Append Points* (see Section 2.6.2): We extend the selected set with 11 additional points not constrained by the estimated objective and plausibility functions, which results in 50 new model realizations that are to be evaluated with the full model.
   (f) *Update*: We import all model realizations that have finished in the meantime. This ensures that the next iteration step does not solely import points from a previous cycle.
   (g) *Follow-up Sample*: We submit the 50 points in the selected set for evaluation with the full model on the cluster.
   (h) *Wait*: The status of all currently running model realizations is continuously checked. When enough models are ready (in our case 39), we continue with the next cycle. The remaining realizations are kept running in the background.

This procedure differs from the one of Regis and Shoemaker (2009) in that we (a) consider plausibility as an additional model constraint, (b) do not perform intermediate resets to keep all calibration runs comparable, (c) use slightly different methods to propose and select points, and (d) add random points in each iteration that are not restricted by the surrogate-distance metric to expand global exploration of the parameter space.

Because individual runs can take quite different times to finish, we submit more points than we require for the next cycle. We then wait until a desired number of newly evaluated points and the respective information is available and continue with the calibration procedure, while the remaining submitted models are still executed. This ensures that the cluster resources are used meaningfully, while the algorithm is busy with finding the next points to be evaluated.

In the following, we explain the individual steps.

2.6.1. Initial Sample

We start the calibration procedure with constructing an initial sample of a given size (e.g., five times the number of input parameters d) using a Halton sequence (Berblinger & Schlier, 1991; Halton, 1960), which is a sequence of low discrepancy that can also be used to define space-filling designs. The main advantage of the Halton sequence over a Latin hypercube design is that it is not necessary to specify the total number of points in advance. Instead each next point of the sequence is close to optimally spaced compared to all previous points. By basing our initial sample on the Halton sequence, we provide an excellent basis not only for the calibration algorithm and the proxy-model, but also for continued global exploration of the parameter space.

The Halton sequence produces points of arbitrary dimension (in our case d) with coordinates ranging from zero to one. We treat these as points in the parameter space of the form $\mathbf{p}_{\mathrm{cdf}}$, which we can convert to physical parameter sets $\mathbf{p}$ to run the full model and to the rescaled parameter sets $\tilde{\mathbf{p}}$. This result in a denser sampling where the prior probability density of the parameters is higher, implying that computational resources are used to explore those regions of the parameter space in more detail that are a priori considered to be more likely.

2.6.2. Proposing New Plausible Points

A key step in the outlined algorithm is to propose points to be evaluated with the full model. Ideally, these points should result in plausible model realizations that either produce small values of the objective function or help restraining the proxy model. The general idea behind our approach of proposing n_{pro} new points (that are eventually subsampled to n_{set} points for actual model runs) is the following:

1. Rank all points that have been evaluated with the full model already by the respective values of the objective function.
2. Apply some random scattering around the good points.
3. Eliminate those points that are predicted to lead to implausible models.
4. Repeat the scattering and elimination until the desired number of points ($n_{\mathrm{pro}} = 10^4$) is reached.

To objectively select the set of good points among the already performed full model runs, we draft a weighted random subset from the set of all points. The respective weight w_i of each realization is based on its objective function value y_i in comparison to the best y_{min} and median objective function value $\bar{y}$ of all points:

$$w_i = \frac{\omega_i}{\sum \omega} \tag{17}$$

$$\omega_i = \max\left(0, \frac{\bar{y} - y_i}{\bar{y} - y_{\mathrm{min}}}\right)^2. \tag{18}$$

By comparing to the minimum and median objective function value and setting negative weight factors to zero, we make sure that “less than average” performing realizations are not used for generating new points. The squaring leads to a smooth transition at objective function values close to the median. Using the median compared to the arithmetic mean provides robustness toward realizations with very large objective-function values.

For the random scattering around the good points, we generalize the approach of Regis and Shoemaker (2009) who added random offsets $\Delta\tilde{\mathbf{p}}$ in parameter space. In our case, we split these offsets into a unit direction vector $\mathbf{e}$ and a magnitude c that are generated independently:

$$\Delta\tilde{\mathbf{p}} = c\,\mathbf{e}. \tag{19}$$

We base the magnitude c on a scaling factor σ_{scale}, which takes the role of the standard deviation in Regis and Shoemaker (2009). We start with $\sigma_{\text{scale}} = 1.5$ and increase it by 50% after three consecutive cycles that succeeded in finding a better point. Similarly, we reduce σ_{scale} by 50% after three consecutive cycles without an improvement. To avoid extreme scales, we restrict σ_{scale} to be in the range from 0.005 to 2.5. We sample c from the following probability density function $f(c|\sigma_{\text{scale}}, d)$

$$f(c|\sigma_{\text{scale}}, d) = \frac{2^{1-d/2}}{\Gamma\left(\frac{d}{2}\right)\sigma_{\text{scale}}}\left(\frac{c}{\sigma_{\text{scale}}}\right)^{d-1}\exp\left(-\frac{1}{2}\left(\frac{c}{\sigma_{\text{scale}}}\right)^{2}\right), \tag{20}$$

which is a scaled and transformed χ^2-distribution that describes the length of a vector where each entry is drawn from an independent normal distribution with standard deviation σ_{scale}. This mimics the random magnitudes of Regis and Shoemaker (2009). Our generalization lies in how we produce random offset directions $\mathbf{e}$.

Starting point in the choice of randomized directions is to estimate the direction of steepest descent $\mathbf{a}$ of the objective function value y with respect to $\tilde{\mathbf{p}}$ using the GPR-based proxy model. Equations 10 and 12 imply a smooth and analytically differentiable function $\vartheta_i^\circ(\tilde{\mathbf{p}})$ for each proxy-model outcome ϑ_i° (for the latter see Erdal et al., 2020). Then the partial derivative of y with respect to the rescaled parameter $\tilde{p}_j$ is:

$$\left.\frac{\partial y}{\partial \tilde{p}_j}\right|_{\tilde{\mathbf{p}}=\tilde{\mathbf{p}}_{\text{test}}} = \sum_{i=1}^{n_{\text{obj}}}\frac{\partial y}{\partial \vartheta_i^\circ}\left.\frac{\partial \vartheta_i^\circ}{\partial \tilde{p}_j}\right|_{\tilde{\mathbf{p}}=\tilde{\mathbf{p}}_{\text{test}}} = 2\sum_{i=1}^{n_{\text{obj}}}\frac{\vartheta_i^* - \vartheta_i^\circ(\tilde{\mathbf{p}}_{\text{test}})}{\sigma_{\text{obj},i}}\left.\frac{\partial \vartheta_i^\circ}{\partial \tilde{p}_j}\right|_{\tilde{\mathbf{p}}=\tilde{\mathbf{p}}_{\text{test}}}, \tag{21}$$

in which we have used the definition of the objective function of Equation 5. This is performed for all parameters, and the direction-of-steepest-descent vector $\mathbf{a}$ is the negative resulting gradient, $\mathbf{a} = -\nabla_{\tilde{p}} y$.

For new points, we might require that the scalar product between $\mathbf{e}$ and the estimated direction of the steepest descent $\mathbf{a}$ is positive:

$$\mathbf{a} \cdot \Delta\tilde{\mathbf{p}} > 0. \tag{22}$$

This restricts the direction vector to be pointing at least toward that side of the parameter space, where the proxy-model expects an improvement. We propose a generalization of this approach in Supporting Information S1, where $\mathbf{e}$ is allowed to differ by any pre-determined angle from the estimated direction of steepest descent.

Finally, we want to ensure that mostly such points are generated, that ultimately lead to plausible results with the full model. To do so, we utilize the proxy model(s) for the plausibility criteria/score and perform plausibility-based rejection-sampling at this stage of constructing the proposal points. A point with a predicted plausibility score of one is immediately added to the set of n_{pro} proposal points that will be used for selecting new full model runs (see Section 2.6.3). A point with a predicted plausibility of zero is immediately rejected and will not be part of the proposal set. For points with an intermediate plausibility, the score z is compared with a random number drawn from the standard uniform distribution. If the plausibility score is larger than the random number, it is added to the set of proposal points. As a result, the set of proposal points only contains suggestions that are predicted to be at least on the fringe of plausibility. A higher priority is given to points closer to full plausibility.

2.6.3. Selecting New Points

Similar to the approaches of Regis and Shoemaker (2007, 2009), and Xia et al. (2021) we use the surrogate-distance metric to balance exploitation and exploration. This criterion uses a weighted average between two normalized metrics:

- The exploration metric: a normalized measure of the minimal dimensionless distance (in parameter space) between a proposed point and all points that have already been evaluated/selected. The minimal distances of all points in the currently proposed set are linearly normalized by the largest and smallest value that occur in the set:

$$m_i^{\text{explore}} = \frac{\max(\mathbf{d}) - d_i}{\max(\mathbf{d}) - \min(\mathbf{d})}, \tag{23}$$

where m_i^{explore} is the exploration metric of the ith point in the proposal set and $\mathbf{d}$ is the vector of minimum Euclidean distances between the proposed points and all previous points. This metric ranges from zero to one, where zero indicates largest minimal distance and one smallest minimal distance. A pure exploration scheme would always strive for the point with the smallest value, as it is the one furthest away from all previous points.

- The exploitation metric: a normalized measure of performance as it is predicted by the proxy-model. The predicted objective function values are linearly normalized by the largest and smallest values of the current set of proposed points:

$$m_i^{\text{exploit}} = \frac{y_i^\circ - \min(\mathbf{y}^\circ)}{\max(\mathbf{y}^\circ) - \min(\mathbf{y}^\circ)}, \tag{24}$$

  where m_i^{exploit} is the exploitation metric of the ith point in the proposal set and $\mathbf{y}^\circ$ are the predicted objective function values of the proposed points. This also results in numbers ranging from zero to one, where zero indicates minimal predicted objective function value and one indicates the opposite. A pure exploitation scheme would always go for the point with the smallest value, as it is predicted to perform best.

To select n_{set} points from a large set based on this metric, we use n_{set} weights ξ that linearly scale between zero and one. For each weight ξ we select the evaluation point with the lowest weighted average of the two measures:

$$i_{\text{select}} = \underset{i}{\text{argmin}} \left[\xi \cdot m_i^{\text{explore}} + (1-\xi) \cdot m_i^{\text{exploit}}\right], \tag{25}$$

where i_{select} is the index of the point that is selected for a full model evaluation. After selecting a point, the exploitation metric is re-evaluated for all remaining points in the set and the next point is selected based on the next weight. We start with a focus on exploitation ($\xi = 0$) and go toward exploration ($\xi = 1$).

2.6.4. Appending Points

We append the set of proposed and subsequently selected points by a number of additional unconstrained points. These are taken directly from the respectively next points in the Halton sequence. This ensures a continued unbiased global exploration of the parameter space. We include this step to maintain a certain minimum density of points even in those parts of the parameter space that are predicted to perform poorly or lead to implausible results.

2.6.5. Variants

In total we apply the presented calibration scheme in four different variants:

- "single": We use only a single GPE as proxy model for the objective function, and another single GPE for the plausibility function. We do not restrict the space of point proposals. This variant is conceptually closest to the original implementation of Regis and Shoemaker (2009).
- "multiple": We use multiple GPEs as a meta-proxy model for both, the objective function and the plausibility function. We expect the proxy-model predictions to be more accurate, which should enhance the calibration. We still do not restrict the space of point proposals.
- "multiple + direction": In addition to using multiple GPEs for objective and plausibility functions, we restrict the space of proposed points by requiring a positive scalar product between step direction and estimated direction of steepest descent (as outlined in Section 2.6.2). This should avoid stepping into the completely wrong direction, while still allowing enough randomness to find good new points. As a result, we expect a further improvement of the calibration.
- "uninformed": For comparison, we also apply a naïve global exploration variant of the presented scheme. It does not use the outlined methods for point proposal and selection, but instead just continues with the next samples of the Halton sequence. As this approach focuses solely on exploration and does not make an effort for exploitation, we expect it to perform poorly compared to the other variants. We also iteratively train multiple GPEs for this case, but the respective prediction information is not used during the calibration itself.

We compare both, the progression of these algorithms over the course of the individual cycles, and the final outcome after more than 3,000 full model evaluations.

2.7. Construction of Posterior Distributions

2.7.1. Proxy-Model Based Markov-Chain Monte Carlo Method

The calibration schemes outlined above aim at finding a single best performing point in the parameter space. Often, it is desirable to also estimate the uncertainty of the parameters, ideally in terms of a full joint posterior distribution of all parameters. The likelihood-based MCMC method with Metropolis-Hastings sampling

(Hastings, 1970; Metropolis et al., 1953) is known to converge toward the true posterior parameter distribution in the limit of an infinite sample. However, the computational effort of the full model is too high to use it in the MCMC framework, which may require at least $\mathcal{O}\left(10^4\right)$ model runs. Instead, we run an MCMC scheme on a quick-to-evaluate GPE-based proxy model. For that, we use the final GPE(s) of the calibration-scheme variant with the best performance. We then apply a classical Metropolis-Hastings MCMC algorithm, which is briefly summarized in the following:

1. We randomly sample $n_{\text{chains}} = 12$ points from the prior distribution to initiate different chains. These points are treated as "trial points."
2. We evaluate the proxy-model predictions for the trial points to emulate model runs and obtain approximated simulated observations $\boldsymbol{\vartheta}^\circ$.
3. We determine the log-likelihood $\log \mathcal{L}_{\text{post}}$ of each trial point by comparing the virtual observations with the calibration target $\boldsymbol{\vartheta}$*:

$$\log \mathcal{L}_{\text{lik}} = -\frac{1}{2}\left(\boldsymbol{\vartheta}^\circ - \boldsymbol{\vartheta}^*\right)\mathbf{C}^{-1}\left(\boldsymbol{\vartheta}^\circ - \boldsymbol{\vartheta}^*\right)^{\mathsf{T}}, \tag{26}$$

   where $n_{\text{obj}} \times n_{\text{obj}}$ is a covariance matrix of measurement errors $\mathbf{C}$ and all constant terms were omitted, as they are not necessary for the following calculations. We assume independent Gaussian measurement errors with the same variance σ^2_{obs} for all observations $\left(\mathbf{C} = \sigma^2_{\text{obs}}\mathbf{I}\right)$. If the model was able to meet all observations with the right set of parameters, σ_{obs} would only reflect the uncertainty of the measured observations (in the order of a few centimeters for hydraulic-head measurements). In realistic cases, however, σ_{obs} should also account for model structural errors (i.e., all reasons why the model can only approximate the observed values). As a result, σ_{obs} needs to be artificially inflated, because otherwise the likelihoods would drop rapidly if not evaluated at the best parameter sets. On the other hand, if the uncertainty is inflated too much, all realizations will be considered approximately equally likely, because even very large deviations between modeled and measured values would be considered acceptable. We obtain our value of σ_{obs} by considering that the expected value of the sum of squared residuals for a Gaussian likelihood equals the variance σ^2_{obs} times the degrees of freedom. We solve this expression for σ_{obs} by considering the sum of squared residuals at the best point found among the calibration variants:

$$\sigma_{\text{obs}} = \sqrt{\frac{\sum_i^{n_{\text{obj}}}\left(\vartheta_i^* - \vartheta_i{}^\circ(\mathbf{p}_{\text{best}})\right)^2}{n_{\text{obj}} - d}}. \tag{27}$$

4. We determine the prior log-probability $\log \mathcal{L}_{\text{prior}}$ of each trial point via the probability-density function of the logistic distribution:

$$\log \mathcal{L}_{\text{prior}} = \sum_{i=1}^{d} -\tilde{p}_i - 2 \cdot \log(1 + \exp(-\tilde{p}_i)). \tag{28}$$

5. The logarithm of the posterior probability density P_{post} is then given by the sum of the two terms:

$$\log P_{\text{post}} = \log \mathcal{L}_{\text{lik}} + \log \mathcal{L}_{\text{prior}}. \tag{29}$$

6. A comparison between the posterior densities of the current trial points and the previous points in the chains decides whether the trial points should be accepted or rejected:

$$\delta = \exp\left(\log P_{\text{post}}^{\text{trial}} - \log P_{\text{post}}^{\text{previous}}\right) - \nu, \tag{30}$$

   where ν is a random number drawn from the uniform distribution. Whenever this difference is positive ($\delta > 0$), the trial point is accepted. Whenever it is negative, the trial point is rejected and the previous point is re-accepted (i.e., repeated). As the initial points do not have precursors, they are all accepted.
7. The current list of n_{chains} (re-)accepted points is perturbed to generate new trial points. We use trivial perturbations with offsets generated from scaled standard normal distributions. The associated scaling factor is dynamically tuned over the course of the MCMC procedure to maintain an acceptance rate of about 30%.

8. With the new set of trial points, the procedure is repeated from the second step until convergence between the chains is achieved. For that we require that the criterion developed by Gelman and Rubin (1992) needs to be smaller than 1.1 for all parameters.

The final points represent an MCMC-based sample from the posterior distribution.

2.7.2. Simulation-Based Inference

We compare the proxy-model-based MCMC approach to a likelihood-free alternative for estimating the posterior parameter distribution using SBI. In particular, we apply Sequential Neural Posterior Estimation (SNPE) (Greenberg et al., 2019; Lueckmann et al., 2017; Papamakarios & Murray, 2016), which is based on training a deep neural network. In contrast to the GPE proxy model, the SNPE method treats model outputs (i.e., observations) as an input. For any queried set of observations, SNPE provides a description of a multi-dimensional distribution of parameter values that are assessed as "likely to produce the queried observations." It is trivial to sample this distribution to infer information about parameter correlations and uncertainties.

We use the single-round version of SNPE (i.e., NPE), because tests with our data have shown that iterative re-training of the neural network does not improve the quality of the posterior distribution. To be precise, we use the SNPE-C implementation (also known as automatic posterior transformation) with atomic loss (Greenberg et al., 2019) and a Masked Autoregressive Flow density estimation neural network (Papamakarios et al., 2017). We make use of the flexible interface of the SBI toolbox of Tejero-Cantero et al. (2020) to perform NPE on pre-simulated data. In summary, the procedure for our case works as follows (see Greenberg et al., 2019, for a comprehensive description):

1. Sample n_{SBI} fully plausible model realizations from the prior distribution. This results in n_{SBI} input parameter sets $\tilde{\mathbf{p}}$ and n_{SBI} corresponding model outputs $\boldsymbol{\vartheta}^{\bullet}$. To retrieve only plausible model realizations, we use the GPE(s) of the best calibration scheme in a post-processing open loop step to scan the Halton sequence for parameter sets that are likely to produce plausible results. Promising points (with a probability score >0.5) are evaluated with the full model. As soon as n_{SBI} model realizations are available, we continue with the next step. This first step can be interpreted as an adjustment of the prior distribution to omit implausible parts of the parameter space. We include it to avoid bias by sampling implausible parts of the parameter space in the training of the NPE.
2. Formulate or choose an approximate d-dimensional posterior density distribution description $P_{\mathrm{post}}(\boldsymbol{\Psi}, \tilde{\mathbf{p}})$ (in our case Masked Autoregressive Flow), which has a set of parameters $\boldsymbol{\Psi}$ and assigns a scalar posterior probability density P_{post} to any parameter set $\tilde{\mathbf{p}}$.
3. Define a neural network f_{NN} that relates the density estimator's parameters $\boldsymbol{\Psi}$ to given model outputs $\boldsymbol{\vartheta}^{\bullet}$:

$$\boldsymbol{\Psi} = f_{\mathrm{NN}}(\boldsymbol{\Phi}, \boldsymbol{\vartheta}^{\bullet}), \tag{31}$$

   where $\boldsymbol{\Phi}$ are the weights (i.e., coefficients) of the neural network.
4. Train the neural network on the sampled data, by adjusting the weights $\boldsymbol{\Phi}$ to minimize the following term:

$$\boldsymbol{\Phi}^{*} = \underset{\boldsymbol{\Phi}}{\operatorname{argmin}} \left[-\sum_{i=1}^{n_{\mathrm{SBI}}} \log\left(P_{\mathrm{post}}(f_{\mathrm{NN}}(\boldsymbol{\Phi}, \boldsymbol{\vartheta}_i^{\bullet}), \tilde{\mathbf{p}}_i) \right) \right]. \tag{32}$$

5. For a given set of observations $\boldsymbol{\vartheta}^{*}$, the posterior density is then given by:

$$P_{\mathrm{post}}\left(f_{\mathrm{NN}}\left(\boldsymbol{\Phi}^{*}, \boldsymbol{\vartheta}^{*} \right), \tilde{\mathbf{p}} \right). \tag{33}$$

There are two reasons, why one should be careful with the interpretation of the NPE results in our application: NPE, like Bayesian approaches, works with (a) stochastic models under the assumption (b) that the target data can be generated with the model and the right parameters. In our case, however, the subsurface flow model is deterministic and there is little hope that any parameter set would provide a perfect agreement between modeled and measured data. The latter has various reasons, including model structural errors and the comparison of a steady state model with a snapshot of real-life transient data. As a result, we apply NPE slightly outside of its original purpose. Nonetheless, the results might still be interpretable and useful.

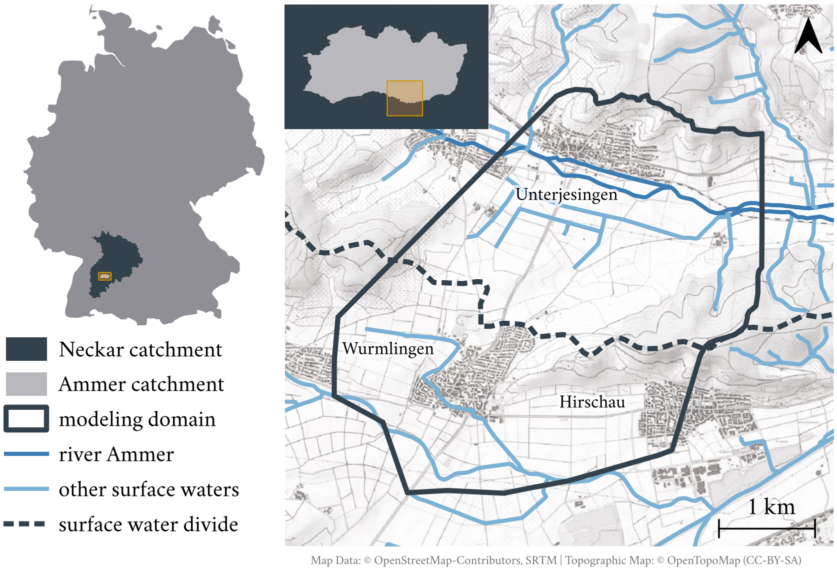

Figure 2. Two-dimensional overview of the model domain.

3. Application to a 3-D Subsurface-Flow Model

3.1. General Problem Statement

The model to be calibrated in this study describes the Ammer floodplain site close to Tübingen in South-West Germany (Martin et al., 2020). An earlier version of this model has been investigated in a previous publication (Allgeier et al., 2020). The calibration aims to align the model output with hydraulic-head observations recorded in a period of stable flow conditions. A summary of the model is given in the following.

The model describes the long-term steady-state, subsurface flow field at the field site. The model domain covers large parts of the Ammer floodplain, toward the south it also extends across the Ammer catchment's surface-water divide to include parts of the adjacent Neckar catchment. A map view of the domain is shown in Figure 2.

The two-dimensional domain is extended to a conforming three-dimensional mesh consisting of 22 layers of triangular prisms. The layer thickness is not uniform, as the vertical discretization is finer toward the top. The final mesh contains 102,784 prism elements (4,672 in each model layer) and 55,752 three-dimensional vertices. Each mesh element is assigned to 1 of 12 lithostratigraphic units. Each unit uses a set of uniform material properties. The lithostratigraphic units considered in this model are (from bottom to top):

- Erfurt formation (`kuE`): Bedrock unit of about 20 m thickness composed of thin sandstone and claystone layers containing dolomite beds and carbonate banks (Geyer & Gwinner, 2011; Hagdorn & Nitsch, 2009; Kirchholtes & Ufrecht, 2015).
- Grabfeld formation (`kmGr`): Bedrock mudstone unit of up to 100 m thickness bearing gypsum, anhydrite and claystones (Geyer & Gwinner, 2011; Kleinert, 1976).
- Weathering zone of Grabfeld formation: As the gypsum-rich rocks of the Grabfeld formation are prone to weathering (Kirchholtes & Ufrecht, 2015; Ufrecht, 2017), we include a weathering zone in the upper parts of the formation.
- Lumped sand- and mudstone formations (`km2345`): Summarized bedrock sandstone units that only exist on hills in the domain. These are the Stuttgart, Steigerwald, Hassberge, Mainhardt, Löwenstein, and Trossingen formations according to the official nomenclature of the German Geological Survey.
- Hillslope hollows: Colluvial hillslope fillings on the northern and southern flanks of the Ammer valley that reach thicknesses around 10 m (Martin et al., 2020).
- Gravel in the Neckar valley: Quaternary floodplain material on the Neckar side consisting of sandy gravel that can reach thicknesses of several meters.
- Clayey gravel in the Ammer floodplain: Lowermost Quaternary floodplain material on the Ammer side consisting of clayey gravel than can reach thicknesses of 5–10 m (Klingler et al., 2020, 2021).

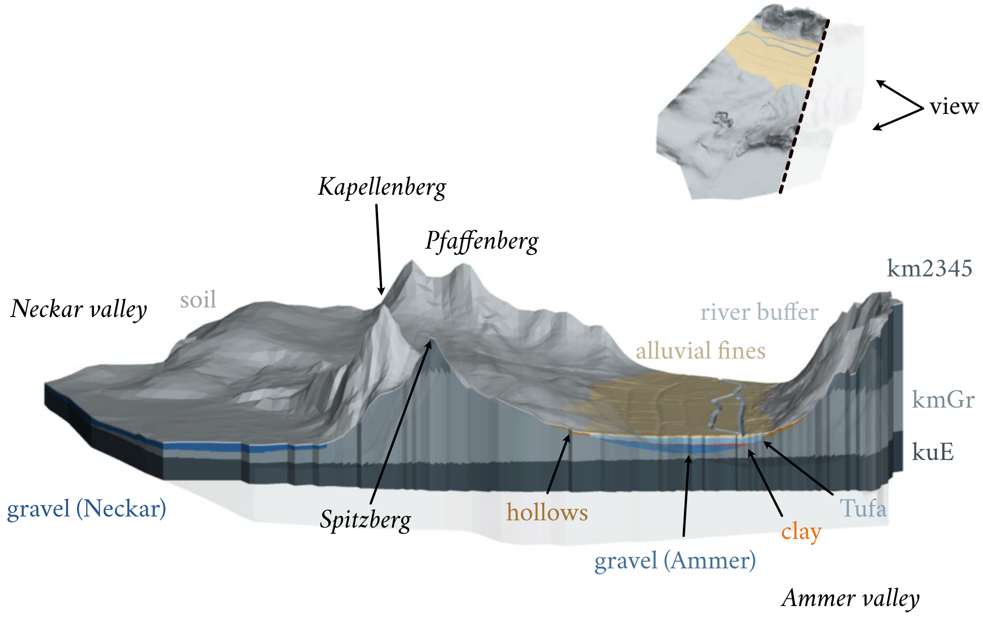

Figure 3. Three-dimensional rendering of the geology in the model domain.

- Clay layer in the Ammer floodplain: Silty clay unit (between 0.5 and 3 m thickness) that separates the Ammer gravel aquifer from the overlaying groundwater storey.
- Tufa layer in the Ammer floodplain: Upper groundwater storey in the Ammer floodplain consisting of Holocene autochthonous calcareous material (Klingler et al., 2020; Martin et al., 2020).
- Alluvial fines in the Ammer floodplain: Confining unit in the Ammer floodplain consisting of alluvial silt and clay (Klingler et al., 2020).
- River buffer zone around River Ammer: River deposits surrounding River Ammer (Martin et al., 2020).
- Top soil layer: Uppermost unit outside of the Ammer floodplain with an assumed thickness of 1.5 m.

Figure 3 provides a three-dimensional rendering of the geology considered within the model. The weathering zone in the Grabfeld formation is not highlighted, as it varies in thickness between model realizations.

To simulate variably saturated flow, we use the nonlinear, modified Richards equation in its stationary form (Richards, 1931; Richardson, 1922):

$$-\nabla(\mathbf{K}k_{\mathrm{rel}}(h)\nabla h) = Q, \tag{34}$$

with

$$\mathbf{K} = I_3\left[K_{xy}, K_{xy}, K_z\right]^{\mathsf{T}}, \tag{35}$$

where h in L is the total hydraulic head, K_{xy} and K_z in $\mathsf{L/T}$ are hydraulic conductivities in the horizontal and vertical direction, respectively, k_{rel} is the dimensionless relative permeability and Q in $1/\mathsf{T}$ represents volumetric source and sink terms. For the description of the unsaturated zone by means of $k_{\mathrm{rel}}(h)$, we employ standard constitutive relationships (R. H. Brooks & Corey, 1964; Mualem, 1976; van Genuchten, 1980).

The equations are solved with HydroGeoSphere (Brunner & Simmons, 2012; Therrien et al., 2010), a fully-integrated hydrogeological modeling environment based on the Finite Element Method. We solve the problem numerically with the transient solver of HydroGeoSphere with constant boundary conditions until steady state is reached. The flow field is subject to the following boundary conditions:

- A specified recharge rate (Neumann boundary) is applied at all top faces of the model domain outside of the Ammer Quaternary.
- Lateral leaky boundary conditions (Robin boundaries) are used to describe the lateral connections of the Ammer Quaternary to upstream and downstream aquifers (groundwater inlet and outlet).

Table 1
Prior Distribution Definitions for All Model Parameters

#	Parameter	Unit	Type	c_1	c_2	c_3	Support interval
1	Weathering depth	m	$\log_{10}\mathcal{N}$	1.46	0.11		$(0.00, \infty)$
2	Anisotropy (bedrock)	–	$\mathcal{B}$	1.52	3.04		(0.00, 1.00)
3	Anisotropy (Quaternary)	–	$\mathcal{B}$	3.04	1.52		(0.00, 1.00)
4	Anisotropy (soil)	–	$\mathcal{B}$	3.04	1.52		(0.00, 1.00)
5	N (bedrock)	–	$\log_{10}\mathcal{N}_\mathrm{T}$	0.2	0.06		$(1.00, \infty)$
6	N (Quaternary)	–	$\log_{10}\mathcal{N}_\mathrm{T}$	0.2	0.06		$(1.00, \infty)$
7	N (soil)	–	$\log_{10}\mathcal{N}_\mathrm{T}$	0.2	0.06		$(1.00, \infty)$
8	K_{xy} (kuE)	m/s	$\log_{10}\mathcal{N}$	−5.27	0.68		$(0.00, \infty)$
9	K_{xy} (lower kmGr)	m/s	$\log_{10}\mathcal{N}$	−8.77	1.16		$(0.00, \infty)$
10	K_{xy} (upper kmGr)	m/s	$\log_{10}\mathcal{N}$	−5.17	0.73		$(0.00, \infty)$
11	K_{xy} (km2345)	m/s	$\log_{10}\mathcal{N}$	−7.51	1.19		$(0.00, \infty)$
12	K_{xy} (hollows)	m/s	$\log_{10}\mathcal{N}$	−6.16	1.00		$(0.00, \infty)$
13	K_{xy} (Neckar gravel)	m/s	$\log_{10}\mathcal{N}$	−3.43	0.65		$(0.00, \infty)$
14	K_{xy} (soil)	m/s	$\log_{10}\mathcal{N}$	−5.19	0.49		$(0.00, \infty)$
15	K_{xy} (gravel)	m/s	$\log_{10}\mathcal{N}$	−4.35	0.84		$(0.00, \infty)$
16	K_{xy} (clay)	m/s	$\log_{10}\mathcal{N}$	−8.11	0.62		$(0.00, \infty)$
17	K_{xy} (Tufa)	m/s	$\log_{10}\mathcal{N}$	−5.09	0.66		$(0.00, \infty)$
18	K_{xy} (alluvial fines)	m/s	$\log_{10}\mathcal{N}$	−7.00	1.02		$(0.00, \infty)$
19	K_{xy} (river buffer)	m/s	$\log_{10}\mathcal{N}$	−5.19	0.49		$(0.00, \infty)$
20	h_leaky (north, inlet)	m	$\mathcal{N}_\mathrm{T}$	350.50	0.75		$(346.69, \infty)$
21	h_leaky (north, outlet)	m	$\mathcal{N}_\mathrm{T}$	335.50	0.75		$(-\infty, 337.26)$
22	h_fixed (south)	m	$\mathcal{N}$	326.25	0.76		$(-\infty, \infty)$
23	Q (river)	m^3/s	$\mathcal{E}_\mathrm{T}$	0.5	0.05	0.13	(0.34, 1.59)
24	Recharge rate	m/s	$\mathcal{B}_\mathrm{S}$	2.06	6.25		$(0.00, 1.91 \cdot 10^{-8})$

Note. $\mathcal{N}$, normal distribution with the mean value c_1 and the standard deviation c_2; $\log_{10}\mathcal{N}$, A log-normal distribution where the base-10 logarithm of the parameter has the mean c_1 and the standard deviation c_2; $\mathcal{B}$, beta distribution with the two shape parameters c_1 and c_2; $\mathcal{B}_\mathrm{S}$, linearly scaled beta distribution according to the given support interval; $\mathcal{E}$, generalized extreme-value distribution with location parameter c_1, scale parameter c_2 and shape parameter c_3; "T", truncation according to the given support interval.

- A fixed-head boundary (Dirichlet boundary) is used to describe the lateral connection of the Neckar Quaternary material to the downstream aquifers.
- River Ammer is implemented as a Dirichlet boundary condition. The corresponding heads are inferred from a surface-water model and a prescribed river discharge.
- Drain boundaries are used to implement drainage ditches and are also applied to the top surface to avoid too high groundwater pressures.

In the following, we describe the model parameters that vary between the realizations.

3.2. Parameters

We assume independence between all model parameters because little credible information about parameter correlations is available prior to considering the measured data. This implies that we can describe the joint prior distribution of all parameters by d individual distributions, each with its own support interval and cumulative density function $f_\mathrm{cdf}(p_i)$. Table 1 provides an overview of the prior distributions of our parameters, of which we consider $d = 24$ in total.

We have one structural parameter (#1), which describes the thickness of the weathering zone of the Grabfeld formation. Material properties make up 18 parameters (#2–#19). For each lithostratigraphic unit we have one parameter describing the saturated horizontal hydraulic conductivity. For the anisotropies (expressed as dimensionless ratios K_z/K_{xy}) and the dimensionless parameter N defining the constitutive unsaturated zone relationship we have grouped the units into bedrock, Quaternary and soil layers. Finally, five parameters (#20–#24) describe boundary conditions. Here, we have three hydraulic-head offsets, the Ammer-river discharge and the recharge rate. A detailed reasoning for our choice of all prior distribution definitions is given in Supporting Information S1.

3.3. Plausibility Function

We base the plausibility function of a model realization on the individual fluxes across boundaries. These are available as raw output from HydroGeoSphere in form of the water balance summary. In our case, the plausibility function is the product of five contributions ($f_{\text{plaus}}(\boldsymbol{\varphi}) = \varphi_1 \cdot \varphi_2 \cdot \varphi_3 \cdot \varphi_4 \cdot \varphi_5$), which are explained in the following:

- Two binary criteria state that the fluxes Q_{inlet} and Q_{outlet} across the groundwater inlet/outlet boundaries on the Ammer side should be positive/negative:

$$\varphi_1 = \begin{cases} 1 & \text{if } Q_{\text{inlet}} > 0 \\ 0 & \text{otherwise} \end{cases}; \quad \varphi_2 = \begin{cases} 1 & \text{if } Q_{\text{outlet}} < 0 \\ 0 & \text{otherwise} \end{cases} \tag{36}$$

- Two criteria state that Q_{inlet} and Q_{outlet} should be of similar magnitude. We accept all realizations where the ratio of one flux magnitude to the other (e.g., $r_Q = \frac{|Q_{\text{outlet}}|}{|Q_{\text{inlet}}|}$) is less than two, and we reject all those where the inlet/outlet carries more than four times the flux of the outlet/inlet. Between these two limits we define a gradual transition by the smoothing function $f_s(x) = 3x^2 - 2x^3$:

$$\varphi_3 = \begin{cases} 1 & \text{if } r_Q < 2 \\ 0 & \text{if } r_Q > 4 \\ f_s\left(\dfrac{4}{r_Q} - 1\right) & \text{otherwise} \end{cases}; \quad \varphi_4 = \begin{cases} 1 & \text{if } \dfrac{1}{r_Q} < 2 \\ 0 & \text{if } \dfrac{1}{r_Q} > 4 \\ f_s(4r_Q - 1) & \text{otherwise} \end{cases} \tag{37}$$

- A smooth criterion states that the Ammer river should not be a major source of water. We require that its net contribution Q_{river} to the total flux of the water balance Q_{tot} (the flux across all boundaries) is less than 10%:

$$\varphi_5 = \begin{cases} 1 & \text{if } Q_{\text{river}} < 0 \\ 0 & \text{if } Q_{\text{river}} > 0.1 \cdot Q_{\text{tot}} \\ f_s\left(1 - 10\dfrac{Q_{\text{river}}}{Q_{\text{tot}}}\right) & \text{otherwise} \end{cases} \tag{38}$$

The smoothing between 0 and 1 for the last three criteria is implemented to alleviate the effects of the arbitrarily chosen thresholds. The third-order smoothing function f_s ensures that the transition at the thresholds has a continuous first derivative.

4. Results and Discussion

We first describe and discuss the results of running the best-point calibration variants for 3070 full-model realizations. Then, we analyze the full posterior distributions (derived with MCMC and SBI). Finally, we take a look at the residuals between modeled and measured observations.

4.1. Calibration-Scheme Variants

Figure 4 compares how the smallest value of the objective function obtained so far decreases with the iteration cycles among the different calibration variants. We can observe that the uninformed sampling strategy converges very slowly, as expected, even with the space-filling Halton sequence. After nearly 60 cycles it has only reached an objective function value of about 6.0 m^2, which was achieved by the other three schemes already within the first two cycles. After more than 3,000 realizations (59 cycles), the informed schemes that rely on multiple internal GPEs have found points with objective function values close to each other (“multiple”: 2.90 m^2; “multiple + direction”:

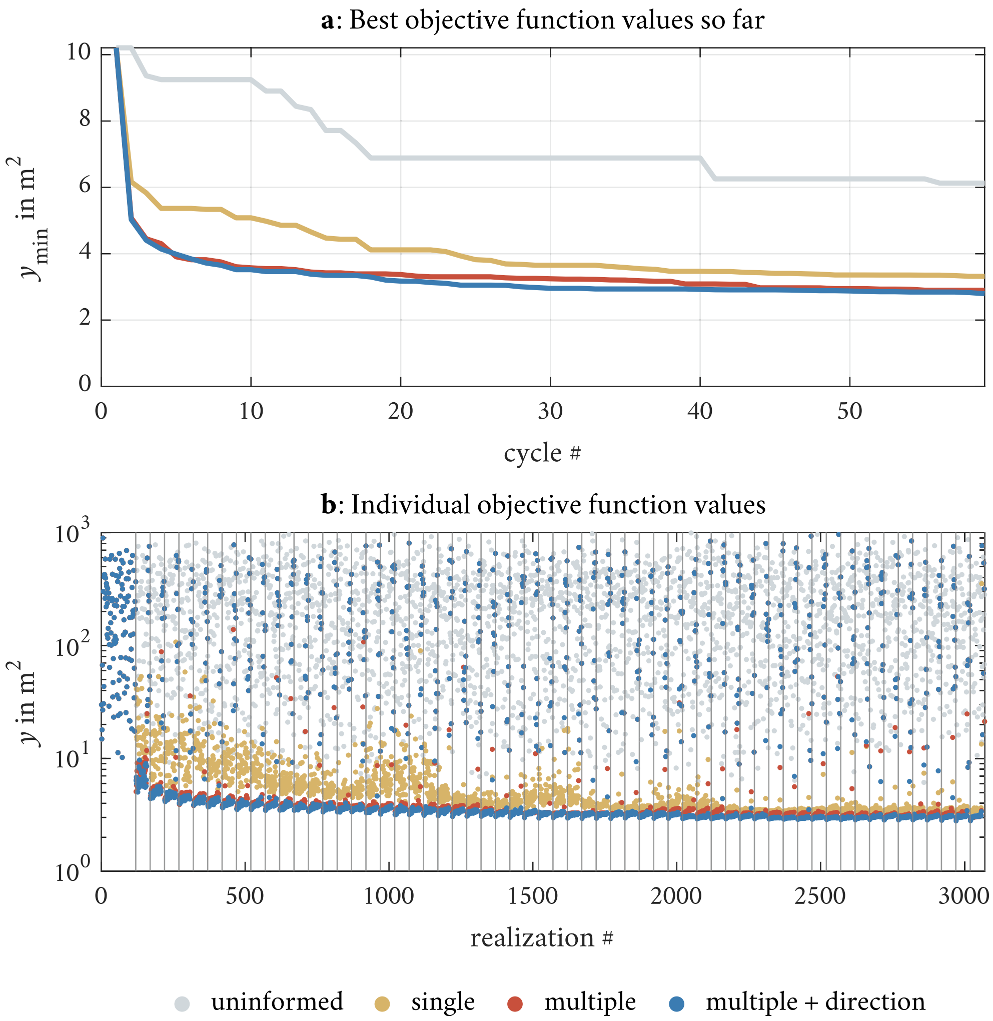

Figure 4. Objective function values y for the different calibration variants. (a) Best objective function value of all previous iteration cycles. (b) All individual objective function values on a logarithmic scale. The vertical lines separate the cycles.

2.80 m^2), while the variant based on a single, lumped GPE produced results between these two and the uninformed scheme (3.32 m^2). In general, this ranking seems to be stable over the cycles: The informed schemes are able to find good regions in parameter space much faster than the uninformed one, and multiple GPEs accelerate the calibration even more. For the same number of cycles, the multi-GPE variant that uses the direction information leads to smaller values of the objective function than the one using the unrestricted point proposal. Therefore, using the direction information seems to help finding better points in parameter space, but the effect is only marginal.

Figure 4b exemplifies the sampling strategies:

- The uninformed variant consistently produces values of the objective function across the range from 6.0 to 1,000 m^2. As expected, there is no systematic improvement over time and all cycles produce similar results.
- After a sufficiently large number of cycles, all informed variants show a systematic improvement of the model fit. This can be attributed to finding good regions in parameter space. Within in each cycle, especially in later ones, it can be observed that early realizations yield smaller values of the objective function than later realizations. This is a direct outcome of the linearly decreasing exploitation/exploration weights that initially promote points that are predicted to be favorable (i.e., in the sense of having low objective function values). Increased values of the objective function result from shifting the weights toward regional parameter-space exploration. The last 11 realizations of each cycle correspond to the unbiased global sampling according to the Halton sequence. These realizations are recognizable by typically much larger values of the objective function, as they are not restricted to be close to good previous sampling points.

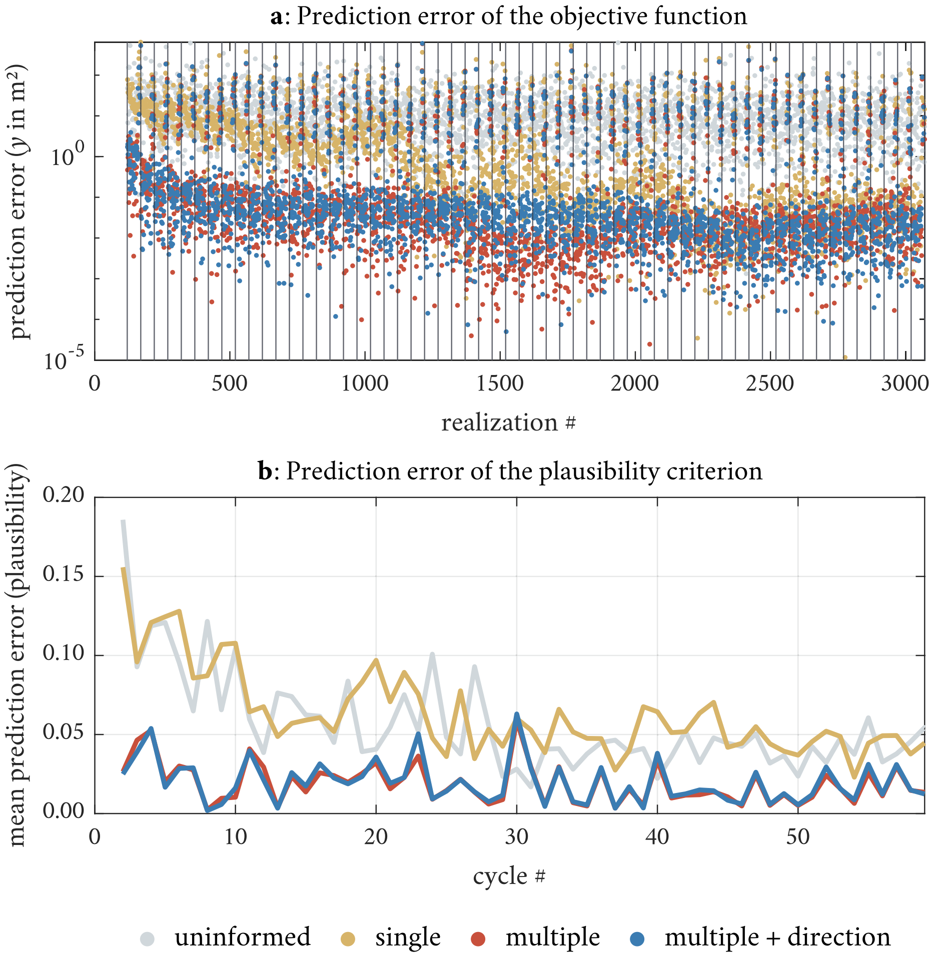

Figure 5. Absolute difference between proxy-model predictions and results of the full model runs, (a) with respect to the value of the objective function, (b) with respect to the plausibility score.

- All variants make use of the same Halton sequence. This is visible for the first cycle, where all four variants start with the same set of 120 initial points, and for the last 11 realizations of all subsequent cycles, where the points of the three informed schemes fall on top of each other.

In summary, the performance of all informed schemes is comparable with respect to the final value of the objective function, but using multiple GPEs helps to find and exploit good regions in parameter space faster.

Figure 5 shows the prediction errors of the proxy models with respect to the objective function (Figure 5a) and the plausibility criterion (Figure 5b) over the course of the calibration. The prediction error is computed as the absolute difference between the objective-function (or plausibility score) value obtained by running the full model and the predicted by the proxy model, respectively.

The prediction errors of the proxy models with respect to the objective-function value presented Figure 5a reveals several insights:

- Even though the uninformed approach uses multiple GPEs for predicting the objective function value, the predictions are of comparably low quality. The prediction error does not drastically decrease over time, which means that the prediction quality increases only marginally even though new information is appended to the GPEs. This illustrates the vastness of the parameter space, as even thousands of space-filling points result in such a low point density that GPE-based interpolation is obviously difficult.
- The informed variant based on single GPEs roughly starts with similar prediction errors as the uninformed case. With increasing number of new points, however, the prediction quality increases, which is the result of concentrating the sampling in regions of low values of the objective function, resulting in an increase of the

sampling density over time such that the interpolation improves. In contrast to that, the space-filling Halton design of the uninformed scheme tries to maintain a uniform density throughout the parameter space resulting in a very slow improvement of the interpolation quality.

- The calibration scheme variants based on multiple GPEs immediately achieve smaller prediction errors for the objective function value, which highlights the drastically improved prediction quality. Over the course of the calibration, the prediction error also decreases until the three informed schemes achieve a similar prediction quality after about 2,300 realizations.

The prediction quality of the plausibility score, shown in Figure 5b, generally confirms these observations, but also shows some differences:

- With mean prediction errors in plausibility of $\lessapprox 0.05$ (from 30 cycles onward), all four variants achieve a decent average prediction quality (a value of 1 would indicate that all plausibility predictions of a cycle were completely wrong; a value of 0 would mean perfect plausibility predictions).
- The uninformed variant and the single-GPE variant perform worse than the multi-GPE cases, but the initially large difference decreases over time. Again, this is probably related to the low point-density, where each new point is very far from all previous points.
- It is interesting to see that the uninformed scheme can also achieve a significant improvement for the prediction of the plausibility score. This indicates that predicting plausibility is easier (i.e., requires a smaller point density) than predicting the objective function value. This might be partly related to the fact that our plausibility function depends on fewer variables than our objective function.
- The informed variants with multiple internal GPEs produce conspicuously similar mean prediction errors. The reason for that can be found in the contribution of the individual realizations to these mean prediction errors. While the two multi-GPE schemes achieve a nearly perfect plausibility prediction of points selected by the surrogate-distance metric (i.e., a prediction error of 0), nearly the entire mean prediction error stems from the Halton sequence points, which are identical for both.

Figure 6 visualizes the estimated logit-score transformed parameters of the final best points of the four calibration variants. To give an impression of the calibration course, the full sequence of all intermediate best points after the first cycle (i.e., all the parameter sets that have contributed to Figure 4a) is shown, too. It is important to note that these point sets do not form an interpretable distribution (at least not a meaningful posterior distribution), as the optimization schemes are not designed to sample the full posterior distribution, but rather aim at finding the single global optimum. The corresponding ranges therefore can also not be related to parameter uncertainties.

Figure 6 shows that the best points found with the uninformed scheme obviously are a subset of the prior distribution, but already here some patterns emerge. For example, some parameters exhibit a systematic shift toward positive (e.g., parameters #17 and #18) or negative (e.g., parameter #13) values, implying best estimates of the original parameters that are larger and smaller than their prior median, respectively.

The best points found with the informed schemes share some common properties. Four commonalities in Figures 6b–6d stand particularly out:

- The parameters related to hydraulic conductivities show similar trends among the informed schemes. For example, the best estimate for parameter #13 is negative. Similarly, parameters #15–#18 form a consistent visual pattern from left to right: The scaled parameter values seem to shift from being more or less strongly negative (#15) to around zero (#16) to slightly positive (#17, with exception of "multiple + direction") to large and positive (#18).
- The scaled parameter values of parameter #8 (hydraulic conductivity of the lowermost Erfurt formation) change only slightly over the course of the calibration.
- Of the three parameters related to hydraulic-head offsets, the two related to the Ammer valley (parameters #20 and #21) are positive, while the head offset in the adjacent Neckar valley boundary (parameter #22) scatters around zero.
- Positive parameter values seem to be preferred for parameter #1 across all calibration scheme variants.

Conversely, there are also significant differences between the best points found by the single-GPE and multi-GPE schemes:

- While in all cases the values for parameter #8 are barely changed during the calibration, the absolute values of this parameter are different between the variants. The single-GPE scheme clearly shifts toward positive values, while the other variants produce negative values.

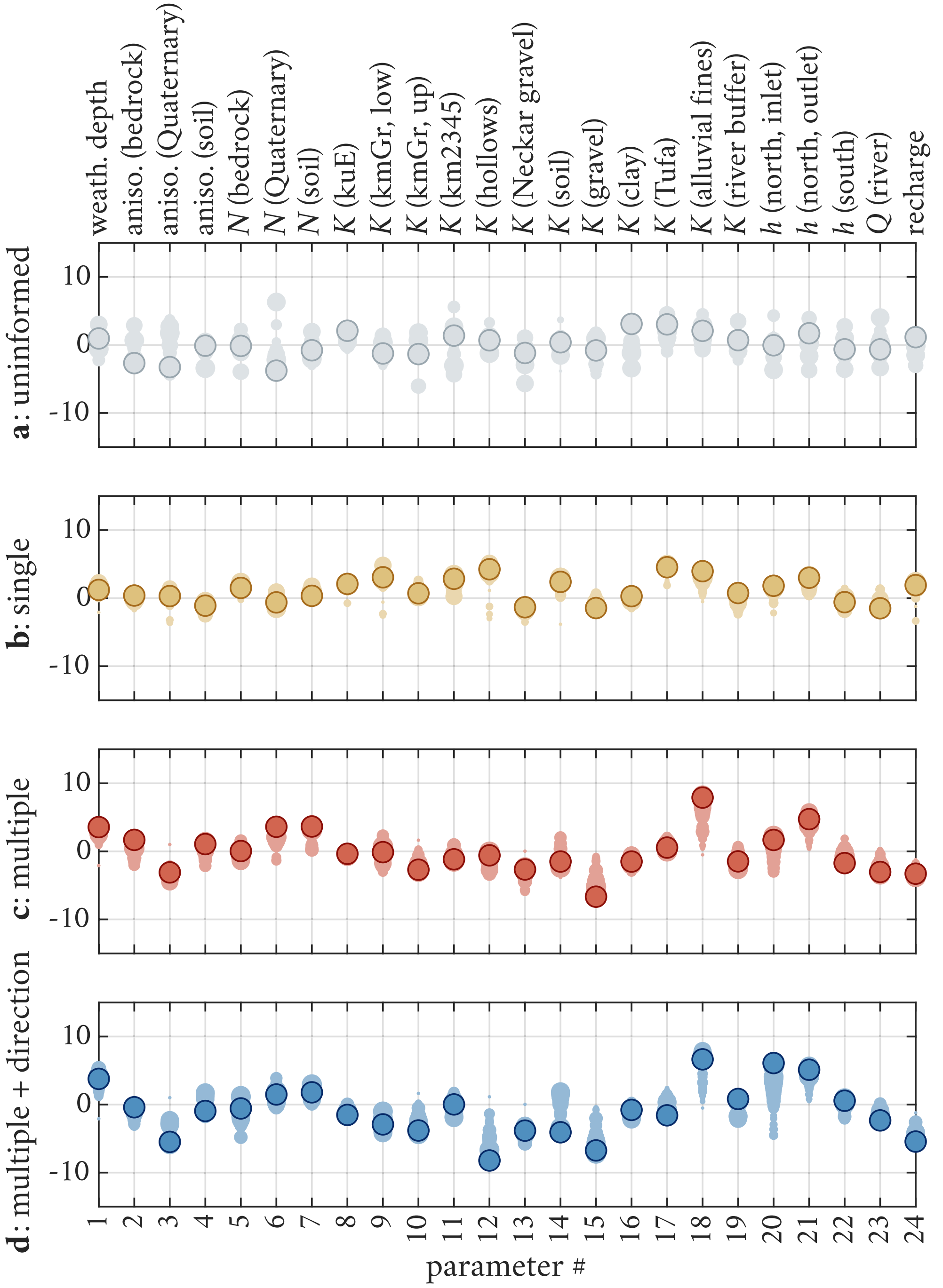

Figure 6. Visualizations of the best found parameter sets in the dimensionless $\tilde{p}$-space as large, dark, outlined circles. Intermediate best points are shown as faded smaller circles, where the circle size corresponds to calibration progress (smaller circles occur in earlier cycles).

- Similarly, the values for parameter #24 (recharge rate) disagree. Again, the point found by the single-GPE scheme is a positive normalized value, while the other variants have found good points only with negative values of this parameter.
- Finally, there are large discrepancies for parameter #12 (hydraulic conductivity of hillslope hollows). Positive, neutral and strongly negative values are all present in the three variants.

When comparing the best estimates of the transformed parameter values among the calibration variants, one should not forget that the best points found by them correspond to different values of the objective function. In particular, the uninformed sampling scheme shows the least agreement between simulated and measured data. By contrast, the objective-function values of the best points found with the three informed variants are reasonably similar. Nonetheless, the obtained best estimates of the parameters differ, advising an assessment of parameter uncertainty by estimating the full posterior distributions of the parameters.

4.2. Analysis of Posterior Distributions

4.2.1. Sample Construction

For the proxy-model based MCMC approach, we obtained an inflated standard deviation of measurements of $\sigma_{\text{obs}} = 0.41$ m at the point with the smallest sum of squared residuals of 2.80 m^2. This result was obtained with the "multiple + direction" variant. We therefore use the respective GPEs in the MCMC process. With σ_{obs}, the MCMC scheme outlined in Section 2.7.1 converged after about 285,000 proxy-model realizations. We omitted the first 5,000 realizations as burn-in period. From the rest we randomly chose 10,000 points serving as a posterior sample that can be compared to a sample drawn from the posterior distribution estimated by SBI. Finally, we run the full model for 250 randomly selected points of this set.

As alternative to the MCMC approach, we applied NPE as outlined in Section 2.7.2. For the first step, we generated a full-model sample of size $n_{\text{SBI}} = 5{,}000$ from the (plausibility-adjusted) prior distribution. The neural network training step of NPE involves a random split of the input data into a training and a test data set. To obtain robust results, we applied NPE 10 times and averaged the resulting posterior distributions. We then sampled 10,000 points from that distribution, of which a randomly chosen subset of 250 were used for full-model runs.

It is important to note that the comparison between MCMC and SBI/NPE is not entirely fair. The MCMC approach has access to the trained GPEs of the best calibration scheme variant. It is therefore (at least theoretically) able to use the underlying information from well-performing full-model runs. The SBI approach on the other hand can only process a sample from the (plausibility-adjusted) prior distribution, whose model realizations (although screened for plausibility) perform quite poorly in general.

4.2.2. Marginal Distributions

Both methods estimate multi-dimensional parameter distributions. Figure 7 shows the marginal distributions as violin plots for each parameter side by side. Figure 7a uses the full set of points generated by the uninformed calibration variant to give an impression of the prior distribution. We see that the scaled parameter values mostly are within the range between −5 and +5, with a higher density around 0. This is a direct outcome of the prior definition through the logit transformation. None of the parameters stands visually out, confirming the unbiasedness of the Halton sequence applied. Figure 7b shows the marginal distributions obtained with SBI, and Figure 7c with the proxy-model based MCMC approach. In Supporting Information S1 we also provide the marginal cumulative distribution functions of all unscaled (i.e., physical) parameter values (i.e., p) as prior and posterior distributions.

In Figure 7, some remarkable features of the marginal posterior distributions are immediately recognizable (and in agreement with the best estimates shown in Figure 6). For instance, parameter #8 (hydraulic conductivity of the Erfurt formation) has the narrowest dimensionless parameter distribution in both cases (Figures 7b and 7c), with a significant shift of similar magnitude toward larger values. Just as observed in Figure 6, we also see a distinct difference between the prior and posterior values for parameters #15–#18. In all four cases a parameter shift in the same direction can be observed for both estimates of the posterior distribution and the global calibration outcomes. The minor second peak for parameter #18 in the MCMC case is most likely an artifact of the underlying GPEs.

In general, the most striking differences between prior and posterior distributions occur for hydraulic-conductivity values. This is not surprising because hydraulic conductivities have a major control on hydraulic heads and the hydraulic-conductivities are defined on a logarithmic scale. However, not all hydraulic-conductivity values are equally important for the observed hydraulic heads. For instance, the conductivities of the soil layer, hillslope hollows and lumped top lithostratigraphic units are hardly affected by the inference. They obviously exert a low influence on the simulated heads at the measurement locations, which are mostly affected by conductivities in the direct vicinity of the measurement location and by those conductivities that determine the overall flow field. The former is the case for parameters #15 and #17 (most observation wells are installed in the corresponding Ammer gravel and Tufa layers); the latter applies for the hydraulic conductivities of the Erfurt formation (parameter #8;

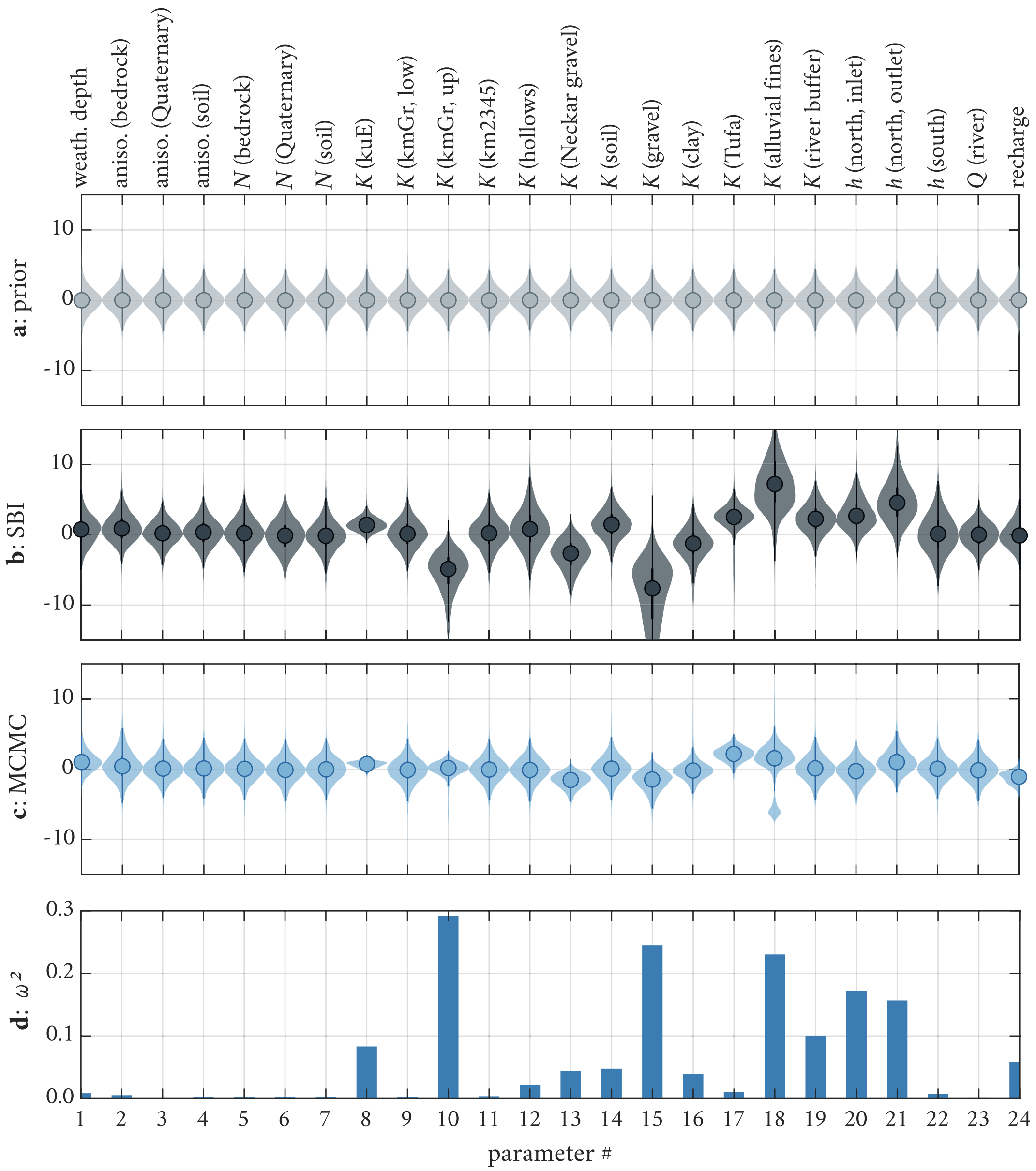

Figure 7. (a)–(c) Comparison of prior and posterior parameter distributions, where circles highlight the median values and the shaded areas show point density (violin plots). All parameter values are displayed in the dimensionless $\tilde{p}$-space. (a) Prior definitions of parameters. (b) Posterior parameter distribution as obtained by the Simulation-Based Inference (SBI) procedure. (c) Posterior parameter distribution as obtained by the Markov-Chain Monte Carlo (MCMC) procedure. (d) Dissimilarity between the SBI and MCMC marginal posterior distributions expressed by the Cramér-von-Mises criterion ω^2 according to Equation 39.

bottom-most lithostratigraphic unit, that extends across the entire domain) and the alluvial fines (parameter #18; connection between aquifers and surficial drainage network).

It is a bit surprising that the parameters related to the Ammer river itself (parameters #19 and #23) are not much affected by the inference. This could indicate that the connection between the groundwater system and the drainage ditch network is more important for the head observations than the connection to the river.

The prior and posterior distributions of the three parameters related to the van-Genuchten coefficient N are basically identical. This indicates a low sensitivity of the virtual hydraulic head observations to these parameters, which is not surprising, as these parameters have very little effect on the steady-state flow field and the observations.

Although producing qualitatively similar results, it is already obvious by visual comparison that the deviations between the SBI posterior distributions and the prior distributions are more extreme than in the MCMC case (with respect to shape distortion and offset). One way of objectively comparing the MCMC and SBI marginal posterior distributions of a parameter is the Cramér-von-Mises criterion ω^2 (Anderson, 1962; Cramér, 1928; von Mises, 1928):

$$\omega^2 = \int_{-\infty}^{\infty} (F(x) - G(x))^2 \, \mathrm{d}x, \tag{39}$$

where $F(x)$ and $G(x)$ are the individual empirical cumulative distributions functions of the samples. Larger values of ω^2 indicate a greater dissimilarity, and a value of $\omega^2 = 0$ would indicate a perfect agreement.

The resulting ω^2-values are displayed in Figure 7d. Especially for parameters #10, #15, and #18 to #21 (the hydraulic conductivities of the upper Grabfeld formation, gravel, alluvial fines, the river buffer and the hydraulic head offsets at the northern inlet), the posterior distributions disagree considerably. We attribute these differences mostly to the different “information access” of the posterior construction methods. The MCMC-related posterior distribution is based on estimating the likelihood of the model outcome using GPEs that have a good coverage close to the global minimum and are to the largest extent constrained by the plausibility criteria. For each new point, the MCMC considers both, the respective prior probability and the GPE-predicted likelihood. This allows stepping into the right direction, while avoiding drifts too far away from the reasonable parameter ranges. The SBI-related posterior distribution on the other hand, is directly constructed from model outcomes based on a sample of the (plausibility-adjusted) prior alone. It cannot utilize any feedback information from a proxy-model. It also does not have access to the true continuous prior distribution, which means that it can drift further apart from it more easily.

4.2.3. Posterior Parameter Correlation

The two approaches of estimating the full d-dimensional posterior parameter distribution allow us to compute correlation coefficients of the uncertainty of all 24 parameters. We visualize the resulting matrices by color coding in Figure 8. Here, the upper diagonal triangle relates to the correlation coefficients determined by SBI and the lower triangle those of the proxy-model based MCMC approach.

Visually most striking is a nearly perfect correlation between parameter #8 (hydraulic conductivity of the Erfurt formation) and parameter #24 (recharge rate) in both posteriors. The former is also correlated to most of the Quaternary hydraulic conductivities (gravel, clay, Tufa, alluvial fines). This also implies correlations between the recharge rate and these conductivities and a notable correlation between these hydraulic conductivities themselves. Considering that we analyze hydraulic-head measurements, the correlation between hydraulic conductivities and recharge is not surprising: If you increase both the recharge rate and all hydraulic conductivities, the resulting heads remain identical. In this context the Erfurt formation acts as a main drainage path of the system: Groundwater recharged in the topographic Ammer valley flows toward the valley filling and leaves to a substantial extent the valley via the Erfurt formation underneath the overlaying Grabfeld formation in the direction of the Neckar valley. Because the Erfurt formation acts as the major “groundwater valve” of the Ammer valley, its hydraulic conductivity is correlated with groundwater recharge the most.

Apart from these stronger correlations, there is some disagreement between the SBI- and MCMC-based parameter distributions and associated correlations. The largest discrepancy might be a moderately strong negative correlation between the hydraulic conductivities of Tufa and alluvial fines in the SBI case, which does not occur for the MCMC case. One could argue that these formations are the two top hydrostratigraphic units in the Quaternary and a negative correlation could indicate a compensation mechanism: if one of the two conductivities is larger, while the other one is lower, the effective average hydraulic conductivity (which could reflect the connectivity of the Quaternary to the drainage network) stays the same.

Overall, the MCMC correlation matrix has most correlations scattering closely around zero, indicating the absence of a parameter relationship. In the SBI case, slightly larger correlation coefficient magnitudes are visible throughout the correlation coefficient matrix. However, it is difficult to draw further conclusions from that, as many correlation coefficients are still comparably small in magnitude and therefore imply weak relationships that might just be governed by random noise. The Supporting Information S1 contains visualizations of bivariate posterior parameter distributions indicating no obvious nonlinear parameter dependencies.

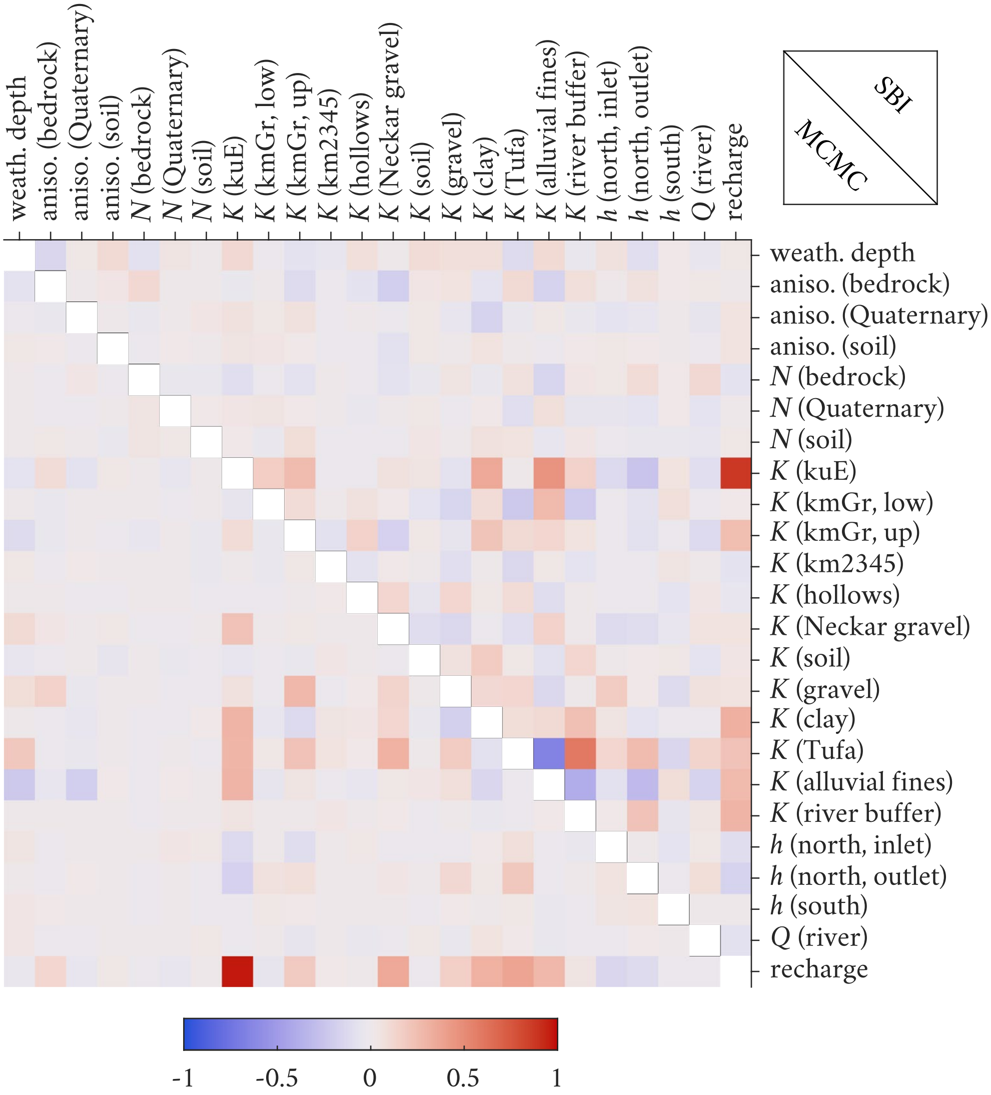

Figure 8. Correlation coefficients between scaled posterior parameters $\tilde{p}$ as determined by the Simulation-Based Inference (upper triangular matrix) and Markov-Chain Monte Carlo (lower triangular matrix) approaches.

4.3. Reproduction of Measurements by Samples of the Posterior Parameter Distributions

In this section, we test how well samples drawn from the posterior parameter distributions reproduce the field measurements $\boldsymbol{\vartheta}^*$, when applied in the full model. Figures 9a–9c shows 250 samples each of $\boldsymbol{\vartheta}^\bullet$ plotted against $\boldsymbol{\vartheta}^*$ for (a) the prior parameter distribution, (b) the posterior distribution estimated by SBI, and (c) that estimated by the proxy-model based MCMC. Figure 9d displays the calibration outcome of the best parameter set found across all calibration variants. We also include (d) the single best estimate obtained in the calibration procedure.

In all cases, the simulated heads scatter about the field measurements, which can be quantified by the root mean-square error RMSE = $\sqrt{\|\boldsymbol{\vartheta}^* - \boldsymbol{\vartheta}^\bullet\|_2^2/n_{\mathrm{obj}}}$. However, the amount of scattering is not identical for all observations. In particular, the simulated head observations far away from boundaries obtained with the parameters drawn from the prior distribution show a large scatter. By constraining the parameter distributions upon calibration or by SBI, this scatter is drastically reduced.

If the posterior distribution of the MCMC-based sample was estimated correctly, the RMSE should be similar to the inflated observation standard deviation σ_{obs} of 0.41 m, whereas the unconstrained prior is expected to lead to much larger values of the RMSE. Indeed the mean RMSE among the prior calculations is 2.10 m. As is already visually

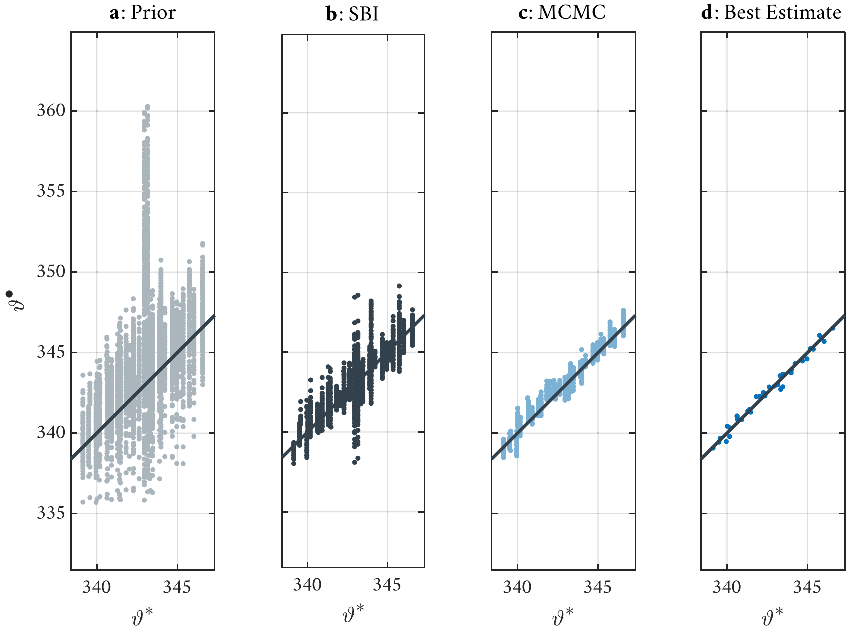

Figure 9. Comparison of full model-run results ($\vartheta^\bullet$) with field measurements (ϑ^*) for samples from the different parameter distributions. (a) samples drawn from the prior distribution, (b) samples drawn from the posterior distribution obtained by Simulation-Based Inference, (c) samples drawn from the posterior obtained through MCMC-sampling, (d) single best calibration estimate.

observable, the scatter of the simulated observations generated from SBI-derived posterior samples is considerably larger than that of the MCMC-based sample, with corresponding RMSE-values of 0.60 versus 0.37 m, respectively. In fact, from the perspective of reproducing the measurements, the posterior distribution estimated by likelihood-free SBI must be considered insufficient. The simulated values do get closer to the identity line, but the likelihood-based construction of the posterior using MCMC does a much better job, while providing a reasonable parameter uncertainty.

The performance of the best parameter set identified by the calibration procedure is remarkably good, compared to the scattering of the prior and posterior distributions, with a single RMSE of 0.26 m. The differences between modeled hydraulic heads and the corresponding field measurements (i.e., the residuals) are small and do not show any obvious visual patterns. For instance, the number of points with positive and negative residuals does not drastically differ. There is also no single observation that has a much larger residual than the other ones, and the residuals do not seem to depend on the observed value itself. This indicates a favorable lack of obvious bias. Obviously, not all measurements are met with the accuracy of typical measurement uncertainties (in the order of few centimeters). However, this is not surprising, considering (a) that we use a steady-state model with homogeneous layers to simulate a snapshot from a transient, heterogeneous, real system, and (b) that the measured elevation of the piezometers themselves might be inaccurate.

5. Conclusions and Outlook

In this study we outlined a strategy of global parameter optimization and construction of the posterior parameter distribution using GPE as proxy models to improve and accelerate the procedure. In the best-estimate calibration part of the overall approach, we suggest an extension of the parallelized approach of Regis and Shoemaker (2009), which is based on the surrogate-distance metric of Regis and Shoemaker (2007). We extend the method to account for model plausibility. Using multiple internal GPEs during the calibration helped to improve and accelerate the calibration. Accounting for the proxy-model based direction of steepest descent of the objective function in the proposal of new points has led to an additional minor enhancement. As strategy to estimate the full posterior parameter distribution we suggest an MCMC approach with Metropolis-Hastings sampling using the internal GPEs rather than full model runs. We successfully applied this approach to a steady-state subsurface flow model of the Ammer floodplain and the adjacent section of the Neckar catchment.

We also applied NPE/SBI to infer a full posterior distribution, which helped us to understand the connection between the calibrated parameter sets and also provided estimates on parameter uncertainty. The main properties of this distribution were in accordance with another posterior distribution derived from classical Bayesian MCMC sampling conducted with a GPE-based proxy model. Considering that the NPE only used 5,000 comparably low-performing realizations sampled from the prior distribution, these results are impressive. However, some deviations were noticeable, and realizations drawn from the MCMC-posterior outperformed their SBI equivalents in terms of agreement with measured data. Depending on the calibration problem at hand, the application of NPE should therefore be evaluated carefully.

For instance, if the purpose of constructing a posterior distribution is to obtain a rough idea about parameter values, uncertainties, and relationships (e.g., as part of a preliminary study) performing NPE on a model sample generated from prior parameter distributions is straightforward and might be sufficient. If details of the posterior distribution are important, or if well-performing samples of the posterior distribution are required, it might be worthwhile to (a) produce additional realizations in promising regions of the parameters space (e.g., by means of a global calibration scheme), (b) train a high-quality proxy model, and (c) perform MCMC sampling with that proxy model. Using GPEs as proxy models would allow applying more advanced MCMC schemes than applied in our study, like Hamiltonian MCMC methods (e.g., Hoffman & Gelman, 2014; Neal, 2011), as computing the gradient of the likelihood comes at very low additional costs. Another possible route could be the application of other SBI algorithms (e.g., Hermans et al., 2020; Papamakarios et al., 2019), which were beyond the scope of this study.

The presented multi-GPE calibration variants are not transferable to transient model calibration, because each time point for each well would essentially require a single GPE, as GPEs can only predict scalar quantities. This is unfeasible, as meta proxy-model training time and prediction time increase linearly with the number of GPEs and even with parallel computing there is a limit that is quickly reached already with a very coarse temporal resolution. By contrast, NPE and SBI have already been successfully applied to transient data and high-dimensional model outputs (Gonçalves et al., 2020; Lueckmann et al., 2017). This makes the field of SBI a promising candidate for inferring posterior parameter values of a transient version of the presented model. One key to that issue might be the development of applicable and valid summary metrics from time series data, as they are used for example, in Lueckmann et al. (2017). This could potentially re-enable the use of the presented GPE-assisted schemes.

Data Availability Statement

The model source files, the calibration code, the raw data supporting the conclusions of this study and Matlab codes used to generate the figures are publicly accessible in form of a repository at https://osf.io/4nr8e/ (Allgeier & Cirpka, 2023).

Acknowledgments

This research was funded by the German Research Foundation (DFG) in the framework of the Research Training Group GRK 1829 “Integrated Hydrosystem Modelling” (Grant Agreement GRK 1829/2). We acknowledge support by the Open Access Publishing Fund of the University of Tübingen.
We are grateful to Alexandra Gessner who suggested applying Simulation-Based Inference to our problem and appreciate the technical help of Michael Deistler in this regard. Open Access funding enabled and organized by Projekt DEAL.

References

Allgeier, J., & Cirpka, O. A. (2023). Raw data, code and plotting scripts for “Surrogate-model assisted plausibility-check, calibration, and posterior-distribution evaluation of subsurface-flow models”. *OSF*. https://doi.org/10.17605/OSF.IO/4NR8E

Allgeier, J., González-Nicolás, A., Erdal, D., Nowak, W., & Cirpka, O. A. (2020). A stochastic framework to optimize monitoring strategies for delineating groundwater divides. *Frontiers in Earth Science*, *8*, 554845. https://doi.org/10.3389/feart.2020.554845

Anderson, T. W. (1962). On the distribution of the two-sample Cramer-von Mises criterion. *The Annals of Mathematical Statistics*, *33*(3), 1148–1159. https://doi.org/10.1214/aoms/1177704477

Beckers, F., Heredia, A., Noack, M., Nowak, W., Wieprecht, S., & Oladyshkin, S. (2020). Bayesian calibration and validation of a large-scale and time-demanding sediment transport model. *Water Resources Research*, *56*(7), e2019WR026966. https://doi.org/10.1029/2019WR026966

Berblinger, M., & Schlier, C. (1991). Monte Carlo integration with quasi-random numbers: Some experience. *Computer Physics Communications*, *66*(2–3), 157–166. https://doi.org/10.1016/0010-4655(91)90064-R

Beven, K. (2006). A manifesto for the equifinality thesis. *Journal of Hydrology*, *320*(1), 18–36. https://doi.org/10.1016/j.jhydrol.2005.07.007

Brooks, R. H., & Corey, A. T. (1964). *Hydraulic properties of porous media (No. 3)*. Colorado State University.

Brooks, S., Gelman, A., Jones, G., & Meng, X.-L. (2011). *Handbook of Markov chain Monte Carlo*. CRC Press.

Brunner, P., & Simmons, C. T. (2012). HydroGeoSphere: A fully integrated, physically based hydrological model. *Groundwater*, *50*(2), 170–176. https://doi.org/10.1111/j.1745-6584.2011.00882.x

Carrera, J., Alcolea, A., Medina, A., Hidalgo, J., & Slooten, L. J. (2005). Inverse problem in hydrogeology. *Hydrogeology Journal*, *13*(1), 206–222. https://doi.org/10.1007/s10040-004-0404-7

Conn, A. R., Gould, N. I., & Toint, P. L. (2000). *Trust region methods*. Society of Industrial and Applied Mathematics (SIAM).

Cramér, H. (1928). On the composition of elementary errors. *Scandinavian Actuarial Journal*, *1928*(1), 13–74. https://doi.org/10.1080/03461238.1928.10416862

Cranmer, K., Brehmer, J., & Louppe, G. (2020). The Frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, *117*(48), 30055–30062. https://doi.org/10.1073/pnas.1912789117

Cressie, N. (1990). The origins of kriging. *Mathematical Geology*, *22*(3), 239–252. https://doi.org/10.1007/BF00889887

Das, S., & Suganthan, P. N. (2011). Differential evolution: A survey of the state-of-the-art. *IEEE Transactions on Evolutionary Computation*, *15*(1), 4–31. https://doi.org/10.1109/TEVC.2010.2059031

Doherty, J. (2015). *Calibration and uncertainty analysis for complex environmental models*. Watermark Numerical Computing Brisbane.

Doherty, J., Brebber, L., & Whyte, P. (1994). *PEST: Model-independent parameter estimation* (Vol. 122, p. 336). Watermark Computing.

Doherty, J., & Hunt, R. (2010). *Approaches to highly parameterized inversion: A guide to using PEST for groundwater-model calibration* (Scientific Investigations Report). US Department of the Interior, US Geological Survey.

Erdal, D., & Cirpka, O. A. (2019). Global sensitivity analysis and adaptive stochastic sampling of a subsurface-flow model using active subspaces. *Hydrology and Earth System Sciences*, *23*(9), 3787–3805. https://doi.org/10.5194/hess-23-3787-2019

Erdal, D., Xiao, S., Nowak, W., & 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), 1813–1830. https://doi.org/10.1007/s00477-020-01867-0

Everitt, B. S., & Skrondal, A. (2010). *The Cambridge dictionary of statistics*. Cambridge University Press.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472. https://doi.org/10.1214/ss/1177011136

Gen, M., & Cheng, R. (1999). *Genetic algorithms and engineering optimization*. John Wiley & Sons.

Geyer, O. F., & Gwinner, M. P. (2011). Geologie von Baden-Württemberg. In M. Geyer, E. Nitsch, & T. Simon (Eds.), *Schweizerbart'sche Verlagsbuchhandlung*. Retrieved from https://www.schweizerbart.de/publications/detail/isbn/9783510652679/Geologie_von_Baden_Wurttemberg

Gilks, W. R., Richardson, S., & Spiegelhalter, D. (1995). *Markov chain Monte Carlo in practice*. CRC Press.

Goldberg, D. E. (1989). *Genetic algorithms in search, optimization and machine learning* (1st ed.). Addison-Wesley Longman Publishing Co., Inc.

Gonçalves, P. J., Lueckmann, J.-M., Deistler, M., Nonnenmacher, M., Öcal, K., Bassetto, G., et al. (2020). Training deep neural density estimators to identify mechanistic models of neural dynamics. *eLife*, *9*, e56261. https://doi.org/10.7554/eLife.56261

Greenberg, D. S., Nonnenmacher, M., & Macke, J. H. (2019). Automatic posterior transformation for likelihood-free inference. In *Proceedings of machine learning research* (pp. 2404–2414).

Haftka, R. T., Villanueva, D., & Chaudhuri, A. (2016). Parallel surrogate-assisted global optimization with expensive functions – A survey. *Structural and Multidisciplinary Optimization*, *54*(1), 3–13. https://doi.org/10.1007/s00158-016-1432-3

Hagdorn, H., & Nitsch, E. (2009). Triassic of Southwest Germany – 175th anniversary of the Foundation of the Triassic System by Friedrich von Alberti.

Halton, J. H. (1960). On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals. *Numerische Mathematik*, *2*(1), 84–90. https://doi.org/10.1007/BF01386213

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 13–109. https://doi.org/10.1093/biomet/57.1.97

Hermans, J., Begy, V., & Louppe, G. (2020). Likelihood-free MCMC with amortized approximate ratio estimators. In *Proceedings of machine learning research* (p. 12).

Hill, M. C., & Tiedeman, C. R. (2006). *Effective groundwater model calibration: With analysis of data, sensitivities, predictions, and uncertainty*. John Wiley & Sons.

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*, 1593–1623.

Jakob, C. (2014). Going back to basics. *Nature Climate Change*, *4*(12), 1042–1045. https://doi.org/10.1038/nclimate2445

Jones, D. R., Schonlau, M., & Welch, W. J. (1998). Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, *13*(4), 455–492. https://doi.org/10.1023/A:1008306431147

Kirchholtes, H. J. & Ufrecht, W. (Eds.). (2015). *Chlorierte Kohlenwasserstoffe im Grundwasser: Untersuchungsmethoden, Modelle und ein Managementplan für Stuttgart*. Springer Vieweg.

Kitanidis, P. K. (1997). The minimum structure solution to the inverse problem. *Water Resources Research*, *33*(10), 2263–2272. https://doi.org/10.1029/97WR01619

Kleinert, K. (1976). Das Grundwasser im Kiesaquifer des oberen Neckartales zwischen Tübingen und Rottenburg (Unpublished doctoral dissertation). Eberhard Karls Universität Tübingen.

Klingler, S., Cirpka, O. A., Werban, U., Leven, C., & Dietrich, P. (2020). Direct-push color logging images spatial heterogeneity of organic carbon in floodplain sediments. *Journal of Geophysical Research: Biogeosciences*, *125*(12), e2020JG005887. https://doi.org/10.1029/2020JG005887

Klingler, S., Martin, S., Cirpka, O. A., Dietrich, P., & Leven, C. (2021). Kombination geophysikalischer und hydrogeologischer Methoden zur gezielten Erkundung feinkörniger Talfüllungen. *Grundwasser*, *26*(4), 379–394. https://doi.org/10.1007/s00767-021-00494-y

Krige, D. G. (1951). A statistical approach to some mine valuation and allied problems on the Witwatersrand (Unpublished doctoral dissertation). University of the Witwatersrand.

Levenberg, K. (1944). A method for the solution of certain non-linear problems in least squares. *Quarterly of Applied Mathematics*, *2*(2), 164–168. https://doi.org/10.1090/qam/10666

Lueckmann, J.-M., Boelts, J., Greenberg, D. S., Gonçalves, P. J., & Macke, J. H. (2021). Benchmarking simulation-based inference. *Proceedings of Machine Learning Research*, *130*, 14.

Lueckmann, J.-M., Goncalves, P. J., Bassetto, G., Öcal, K., Nonnenmacher, M., & Macke, J. H. (2017). Flexible statistical inference for mechanistic models of neural dynamics. In *Nips'17: Proceedings of the 31st international conference on neural information processing systems* (pp. 1289–1299).

Marquardt, D. W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, *11*(2), 431–441. https://doi.org/10.1137/0111030

Martin, S., Klingler, S., Dietrich, P., Leven, C., & Cirpka, O. A. (2020). Structural controls on the hydrogeological functioning of a floodplain. *Hydrogeology Journal*, *28*(8), 2675–2696. https://doi.org/10.1007/s10040-020-02225-8

Matérn, B. (1960). *Spatial variation: Stochastic models and their application to some problems in forest surveys and other sampling investigations* (Vol. 49). Meddelanden från statens Skogsforskningsinstitut.

Matheron, G. (1963). Principles of geostatistics. *Economic Geology*, *58*(8), 1246–1266. https://doi.org/10.2113/gsecongeo.58.8.1246

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, *21*(6), 1087–1092. https://doi.org/10.1063/1.1699114

Mohammadi, F., Kopmann, R., Guthke, A., Oladyshkin, S., & Nowak, W. (2018). Bayesian selection of hydro-morphodynamic models under computational time constraints. *Advances in Water Resources*, *117*, 53–64. https://doi.org/10.1016/j.advwatres.2018.05.007

Mualem, Y. (1976). A new model for predicting the hydraulic conductivity of unsaturated porous media. *Water Resources Research*, *12*(3), 513–522. https://doi.org/10.1029/WR012i003p00513

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In S. Brooks, A. Gelman, G. Jones, & X. Meng (Eds.), *Handbook of Markov chain Monte Carlo* (pp. 113–162).

Papamakarios, G., & Murray, I. (2016). Fast ϵ-free inference of simulation models with Bayesian conditional density estimation. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 29). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper/2016/file/6aca97005c68f1206823815f66102863-Paper.pdf

Papamakarios, G., Pavlakou, T., & Murray, I. (2017). Masked autoregressive flow for density estimation. In I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. V. N. Vishwanathan, et al. (Eds.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper/2017/file/6c1da886822c67822bcf3679d04369fa-Paper.pdf

Papamakarios, G., Sterratt, D., & Murray, I. (2019). Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows. In *Proceedings of the twenty-second international conference on artificial intelligence and statistics* (pp. 837–848). PMLR. Retrieved from https://proceedings.mlr.press/v89/papamakarios19a.html

Powell, M. J. D. (1970a). A hybrid method for nonlinear equations. In P. Rabinowitz (Ed.), *Numerical methods for nonlinear algebraic equations* (pp. 87–114). Retrieved from https://ci.nii.ac.jp/naid/10006528967/

Powell, M. J. D. (1970b). A new algorithm for unconstrained optimization. In J. B. Rosen, O. L. Mangasarian, & K. Ritter (Eds.), *Nonlinear programming* (pp. 31–65). Academic Press. https://doi.org/10.1016/B978-0-12-597050-1.50006-3

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. MIT Press.

Regis, R. G., & Shoemaker, C. A. (2007). A stochastic radial basis function method for the global optimization of expensive functions. *INFORMS Journal on Computing*, *19*(4), 497–509. https://doi.org/10.1287/ijoc.1060.0182

Regis, R. G., & Shoemaker, C. A. (2009). Parallel stochastic global optimization using radial basis functions. *INFORMS Journal on Computing*, *21*(3), 411–426. https://doi.org/10.1287/ijoc.1090.0325

Richards, L. A. (1931). Capillary conduction of liquids through porous mediums. *Physics*, *1*(5), 318–333. https://doi.org/10.1063/1.1745010

Richardson, L. F. (1922). *Weather prediction by numerical process*. The University Press. Retrieved from http://archive.org/details/weatherpredictio00richrich

Selle, B., Rink, K., & Kolditz, O. (2013). Recharge and discharge controls on groundwater travel times and flow paths to production wells for the Ammer catchment in southwestern Germany. *Environmental Earth Sciences*, *69*(2), 443–452. https://doi.org/10.1007/s12665-013-2333-z

Solomatine, D. P., Dibike, Y. B., & Kukuric, N. (1999). Automatic calibration of groundwater models using global optimization techniques. *Hydrological Sciences Journal*, *44*(6), 879–894. https://doi.org/10.1080/02626669909492287

Stein, M. L. (1999). *Interpolation of spatial data: Some theory for kriging*. Springer Science & Business Media.

Tejero-Cantero, A., Boelts, J., Deistler, M., Lueckmann, J.-M., Durkan, C., Gonçalves, P. J., et al. (2020). Sbi: A toolkit for simulation-based inference. *Journal of Open Source Software*, *5*(52), 2505. https://doi.org/10.21105/joss.02505

Therrien, R., McLaren, R., Sudicky, E., & Panday, S. (2010). *HydroGeoSphere: A three-dimensional numerical model describing fully-integrated subsurface and surface flow and solute transport*. Groundwater Simulations Group, University of Waterloo.

Ufrecht, W. (2017). Zur Hydrogeologie veränderlich fester Gesteine mit Sulfatgestein, Beispiel Gipskeuper (Trias, Grabfeld-Formation). *Grundwasser*, *22*(3), 197–208. https://doi.org/10.1007/s00767-017-0362-3

van Genuchten, M. T. (1980). A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. *Soil Science Society of America Journal*, *44*(5), 892–898. https://doi.org/10.2136/sssaj1980.03615995004400050002x

Venkataraman, S., & Haftka, R. (2004). Structural optimization complexity: What has Moore's law done for us? *Structural and Multidisciplinary Optimization*, *28*(6), 375–387. https://doi.org/10.1007/s00158-004-0415-y

von Mises, R. (1928). *Wahrscheinlichkeit, Statistik und Wahrheit*. Springer. Retrieved from https://www.webofscience.com/wos/alldb/full-record/BCI:BCI19310500016490

Wackernagel, H. (2003). *Multivariate geostatistics: An introduction with applications*. Springer Science & Business Media.

Wang, Y., & Shoemaker, C. A. (2014). A general stochastic algorithmic framework for minimizing expensive black box objective functions based on surrogate models and sensitivity analysis. arXiv:1410.6271. Retrieved from http://arxiv.org/abs/1410.6271

Xia, W., Shoemaker, C., Akhtar, T., & Nguyen, M.-T. (2021). Efficient parallel surrogate optimization algorithm and framework with application to parameter calibration of computationally expensive three-dimensional hydrodynamic lake PDE models. *Environmental Modelling & Software*, *135*, 104910. https://doi.org/10.1016/j.envsoft.2020.104910

Yeh, W. W.-G. (1986). Review of parameter identification procedures in groundwater hydrology: The inverse problem. *Water Resources Research*, *22*(2), 95–108. https://doi.org/10.1029/WR022i002p00095

Zhou, H., Gómez-Hernández, J. J., & Li, L. (2014). Inverse methods in hydrogeology: Evolution and recent trends. *Advances in Water Resources*, *63*, 22–37. https://doi.org/10.1016/j.advwatres.2013.10.014

Zhou, Y., & Li, W. (2011). A review of regional groundwater flow modeling. *Geoscience Frontiers*, *2*(2), 205–214. https://doi.org/10.1016/j.gsf.2011.03.003

References From the Supporting Information

Ammer, U., Einsele, G., Arnold, W., Klee, O., Agerer, R., Agster, G., et al. (1983). Wasserhaushalt, Stoffeintrag, Stoffaustrag und biologische Studien im Naturpark Schönbuch bei Tübingen. *Forstwissenschaftliches Centralblatt*, *102*(1), 282–324. https://doi.org/10.1007/BF02741862

Archer, N., Bonell, M., Coles, N., MacDonald, A., Auton, C., & Stevenson, R. (2013). Soil characteristics and landcover relationships on soil hydraulic conductivity at a hillslope scale: A view towards local flood management. *Journal of Hydrology*, *497*, 208–222. https://doi.org/10.1016/j.jhydrol.2013.05.043

BfG. (2003). Hydrologischer Atlas Deutschland. Retrieved from https://geoportal.bafg.de/mapapps/resources/apps/HAD/index.html?lang=de

Butscher, C., Huggenberger, P., Zechner, E., & Einstein, H. H. (2011). Relation between hydrogeological setting and swelling potential of clay-sulfate rocks in tunneling. *Engineering Geology*, *122*(3), 204–214. https://doi.org/10.1016/j.enggeo.2011.05.009

Capuano, R. M., & Jan, R. Z. (1996). In situ hydraulic conductivity of clay and silty- clay fluvial-deltaic sediments, Texas Gulf Coast. *Groundwater*, *34*(3), 545–551. https://doi.org/10.1111/j.1745-6584.1996.tb02036.x

Carsel, R. F., & Parrish, R. S. (1988). Developing joint probability distributions of soil water retention characteristics. *Water Resources Research*, *24*(5), 755–769. https://doi.org/10.1029/WR024i005p00755

D'Affonseca, F. M., Finkel, M., & Cirpka, O. A. (2020). Combining implicit geological modeling, field surveys, and hydrogeological modeling to describe groundwater flow in a karst aquifer. *Hydrogeology Journal*, *28*, 2779–2802. https://doi.org/10.1007/s10040-020-02220-z

D'Affonseca, F. M., Rügner, H., Finkel, M., Osenbrück, K., Duffy, C. E., & Cirpka, O. A. (2018). *Umweltgerechte Gesteinsgewinnung in Wasserschutzgebieten* (Tech. Rep.). Universität Tübingen.

Gudera, T., & Morhard, A. (2015). Hoch aufgelöste Modellierung des Bodenwasserhaushalts und der Grundwasserneubildung mit GWN-BW. *Hydrologie und Wasserbewirtschaftung*, *59*(5), 205–216. https://doi.org/10.5675/HYWA2015,51

Harreß, H. M. (1973). Hydrogeologische Untersuchungen im Oberen Gäu. Retrieved from https://rds-tue.ibs-bw.de/link?kid=1073957446

Hiller, T., Romanov, D., Kaufmann, G., Epting, J., & Huggenberger, P. (2012). Karstification beneath the Birs weir in Basel/Switzerland: A 3D modeling approach. *Journal of Hydrology*, *448–449*, 181–194. https://doi.org/10.1016/j.jhydrol.2012.04.040

Huch, M., & Geldmacher, H. (2013). *Ressourcen-Umwelt-Management: Wasser · Boden · Sedimente*. Springer-Verlag.

Kehrer, W. (1935). Ein Beitrag zur Hydrologie der Umgebung von Tübingen. Retrieved from https://rds-tue.ibs-bw.de/link?kid=1092730990

Keim, B., & Pfäfflin, H. (2006). *Grundwasserbilanzmodell für die Brunnen der ASG im Neckartal bei Kiebingen* (Tech. Rep. No. A240-1). Ingenieurgesellschaft Prof. Kobus und Partner GmbH.

Kortunov, E. (2018). Reactive transport and long-term redox evolution at the catchment scale (Unpublished doctoral dissertation). University of Tübingen. https://doi.org/10.15496/publikation-25162

Lessoff, S. C., Schneidewind, U., Leven, C., Blum, P., Dietrich, P., & Dagan, G. (2010). Spatial characterization of the hydraulic conductivity using direct-push injection logging. *Water Resources Research*, *46*(12), W12502. https://doi.org/10.1029/2009WR008949

LGRB. (2010). *Geologische Untersuchungen von Baugrundhebungen im Bereich des Erdwärmesondenfeldes beim Rathaus in der historischen Altstadt von Staufen i. Br.* (Tech. Rep.). Landesamt für Geologie, Rohstoffe und Bergbau. Retrieved from https://produkte.lgrb-bw.de/schriftensuche/sonstige_produkte/1203/

LGRB. (2012). *Zweiter Sachstandsbericht zu den seit dem 01.03.2010 erfolgten Untersuchungen im Bereich des Erdwärmesondenfeldes beim Rathaus in der historischen Altstadt von Staufen i. Br.* (Tech. Rep.). Landesamt für Geologie, Rohstoffe und Bergbau. Retrieved from https://www.lgrb-bw.de/geothermie/staufen/schadens_fall_staufen_bericht_2012

Li, S. (2011). Concise formulas for the area and volume of a hyperspherical cap. *Asian Journal of Mathematics and Statistics*, *4*(1), 66–70. https://doi.org/10.3923/ajms.2011.66.70

LTZ. (2021). *Agrarmeteorologie Baden-Württemberg*. Landwirtschaftliches Technologiezentrum Augustenberg. Retrieved from https://www.wetter-bw.de/Agrarmeteorologie-BW/Wetterdaten/Stationen-nach-Region/Tuebingen/BWAM146

Lu, C., Qin, W., Zhao, G., Zhang, Y., & Wang, W. (2017). Better-fitted probability of hydraulic conductivity for a silty clay site and its effects on solute transport. *Water*, *9*(7), 466. https://doi.org/10.3390/w9070466

LUBW. (2021). Daten aus dem Umweltinformationssystem (UIS) der LUBW Landesanstalt für Umwelt Baden-Württemberg. Retrieved from https://udo.lubw.baden-wuerttemberg.de/public/q/6dAtdEkpPEGhOGYi1c1PTe

Maier, U., Flegr, M., Rügner, H., & Grathwohl, P. (2013). Long-term solute transport and geochemical equilibria in seepage water and groundwater in a catchment cross section. *Environmental Earth Sciences*, *69*(2), 429–441. https://doi.org/10.1007/s12665-013-2393-0

Marsaglia, G. (1972). Choosing a point from the surface of a sphere. *The Annals of Mathematical Statistics*, *43*(2), 645–646. https://doi.org/10.1214/aoms/1177692644

Minasny, B., Hopmans, J. W., Harter, T., Eching, S. O., Tuli, A., & Denton, M. A. (2004). Neural networks prediction of soil hydraulic functions for alluvial soils using multistep outflow data. *Soil Science Society of America Journal*, *68*(2), 417–429. https://doi.org/10.2136/sssaj2004.4170

Muller, M. E. (1959). A note on a method for generating points uniformly on n-dimensional spheres. *Communications of the ACM*, *2*(4), 19–20. https://doi.org/10.1145/377939.377946

Nagarajarao, Y., & Mallick, S. (1980). Comparison of experimentally determined and calculated hydraulic conductivities for two alluvial sandy loam profiles. *Zeitschrift für Pflanzenernährung und Bodenkunde*, *143*(6), 679–683. https://doi.org/10.1002/jpln.19801430609

Rawls, W. J., & Brakensiek, D. L. (1989). Estimation of soil water retention and hydraulic properties. In H. J. Morel-Seytoux (Ed.), *Unsaturated flow in hydrologic modeling* (pp. 275–300). Springer Netherlands. https://doi.org/10.1007/978-94-009-2352-210

Schlosser, T., Schmidt, M., Schneider, M., & Vermeer, P. (2007). *Potenzial der Tunnelbaustrecke des Bahnprojektes Stuttgart 21 zur Wärme- und Kältenutzung*. Studie des Zentrums für Energieforschung Stuttgart.

Schollenberger, U. (1998). Beschaffenheit und Dynamik des Kiesgrundwasssers im Neckartal bei Tübingen (Unpublished doctoral dissertation). Eberhard Karls Universität Tübingen.

Schweizer, D., Prommer, H., Blum, P., & Butscher, C. (2019). Analyzing the heave of an entire city: Modeling of swelling processes in clay-sulfate rocks. *Engineering Geology*, *261*, 105259. https://doi.org/10.1016/j.enggeo.2019.105259

Schweizer, D., Prommer, H., Blum, P., Siade, A. J., & Butscher, C. (2018). Reactive transport modeling of swelling processes in clay-sulfate rocks. *Water Resources Research*, *54*(9), 6543–6565. https://doi.org/10.1029/2018WR023579

Wegehenkel, M., & Selg, M. (2002). Räumlich hochauflösende Modellierung der Grundwasserneubildung im Neckartal bei Tübingen. *Grundwasser*, *7*(4), 217–223. https://doi.org/10.1007/s007670200033

Willscher, B., Rausch, R., & Selg, M. (2002). *Quantifizierung des Wasserhaushalts im Einzugsgebiet der Brunnen Kiebingen im Neckartal bei Rottenburg* (Vol. 15). Landesamt für Geologie, Rohstoffe und Bergbau Baden-Württemberg.

Wittke, W. (2014). Swelling rock. In *Rock mechanics based on an anisotropic jointed rock model (AJRM)* (pp. 181–208). John Wiley & Sons, Ltd. https://doi.org/10.1002/9783433604281.ch8