Joint Optimization of Measurement and Modeling Strategies With Application to Radial Flow in Stratified Aquifers

R. Maier, A. Gonzalez-Nicolas, C. Leven, W. Nowak, and O. A. Cirpka

1Center for Applied Geoscience, University of Tübingen, Tübingen, Germany, 2Institute for Modelling Hydraulic and Environmental Systems (LS3)/SimTech, University of Stuttgart, Stuttgart, Germany

Abstract: When applying environmental models, the choice of model complexity and the design of field campaigns depend on each other and on the modeling/prediction goal. We propose jointly optimizing model complexity and data collection (design) by maximizing the expected performance for the modeling goal. We use ensembles of highly resolved virtual realities and of less complex modeling variants that differ in their degrees of upscaling and simplified parameterization. For each design under consideration, we simulate hypothetical measurement data (subject to noise) with all realizations of all models. To mimic model calibration with hypothetical data, we identify pairs of best fitting realizations between virtual reality and each model variant for each design. Then, we emulate model choice by selecting (across the model variants, for each design and for each virtual reality) the pair that shows the best predictive match in the modeling goal. Finally, we identify the model/design combination that offers, on average over all virtual realities, the best predictive match. As a test application, we consider a heterogeneous, stratified aquifer, in which the stratification enhances hydraulic anisotropy on the macroscale. We define two different modeling goals: (a) estimating the hydraulic conductivity tensor upscaled to the full aquifer thickness and (b) predicting the pumping rate needed to dewater a construction pit. Our results indicate that jointly optimizing observation networks and model selection can reduce the prediction uncertainty of parameters at lower experimental costs. We also show that the involved trade-offs between model complexity and required design depend on the target quantity.

1. Introduction

Various monographies and review articles have formalized the process of scientific modeling for environmental applications in general, and in the context of hydro(geo)logy specifically (Hill & Tiedeman, 2007; Oreskes, 1998; Walker et al., 2003; Woessner & Anderson, 1996, among others). Though differing in some details, all these formalizations agree that model building and data collection are intertwined and that the typical modeling cycle consists of clearly defining the modeling goal, constructing the conceptual model(s), transferring it/them to numerical model(s), collecting appropriate data, calibrating the model(s), validating the model(s) by application to data not included in the calibration, and selecting the best model, if several competing models have been formulated. This process typically involves iterations, as conceptual models are modified due to persistent misfit to data, and new data are collected due to persistent ambiguities of the models. For each of these steps, extended literature exists.

Let us highlight two specific challenges in this cycle. The first challenge is to decide on the most appropriate modeling strategy: how much detail is needed in the model, how much simplification is allowable, and how complex will the resulting model have to be? In statistics, this is a classical question of model selection (Beven, 2002; Raftery, 1995), with corresponding digestions for parameterizations of groundwater problems (Yeh & Yoon, 1981). The second challenge is to decide on appropriate field campaigns, installations, and/or lab experiments: How much and which data are sufficient and conclusive (for calibration, validation, and selection) while keeping the costs for experimental/field campaigns at a bearable level? This well-known problem is called optimal design of experiments. It is mostly formalized as a mathematical optimization problem under the uncertainty of the yet unmeasured data (e.g., Chaloner & Verdinelli, 1995; Leube et al., 2012; Müller, 2005).
Traditionally, these two challenges are treated as disjoint: either one selects a model with data being given or one optimizes a field campaign to feed the information needs of a given model. Therefore, an even more substantial question regarding the modeling cycle arises: should one design the measurement strategy based on the selected model or vice versa? Apparently, there is an interrelation of measurement and modeling strategies, which rather calls for a joint solution. So far a rational, objective approach that jointly assists both modelers and experimentalists to target the joint problem more systematically has been lacking.

In our study, we propose that rational model selection should go hand-in-hand with cost-efficient data collection to achieve the best possible model performance toward a given modeling goal for a given experimental budget. We hypothesize that this joint problem can be solved by a corresponding joint optimization. This optimization has to find the best combination of model choice and corresponding field campaign design, so that the best expected performance toward the given modeling goal can be achieved. As predictive model performance is only speculative during the design phase of models and field campaigns, the optimization needs a means of anticipating both predictive uncertainty and systematic predictive errors. To include both uncertainty and systematic errors in this step, we propose to use an ensemble of highly resolved virtual realities to generate realistic virtual measurements and prediction targets. While these virtual realities provide hypothetical truths, calibrating models with this level of detail would trigger infeasible data requirements. Along with the virtual realities, we thus generate ensembles of simplified models with different model complexity that undergo model selection for different measurement designs that are subject to choice. We will provide a more detailed rationale and develop and present the according framework for joint optimization in section 2.

To test and demonstrate our proposed framework, we apply it to a case, in which we model steady-state groundwater flow in a stratified, fluvial gravel aquifer to predict the dewatering flux required to keep construction pits dry. The sedimentary structure of such deposits typically reveals a bedded character comprising horizontal strata of materials with different grain size (Freeze & Cherry, 1979). The stratiﬁcation of the materials is associated with variations in hydraulic properties (Heinz et al., 2003). It introduces spatial variability of the hydraulic-conductivity tensor $K$ among individual strata, which is the heterogeneity we consider in this study. Typically, groundwater preferentially ﬂows along the bedding planes rather than across them (Bear, 1979), introducing hydraulic anisotropy. Thus, the effective behavior in stratified aquifers is that the principal components of $K_{\text{eff}}$ are aligned in the horizontal and vertical directions. The ratio of the horizontal to vertical hydraulic conductivity is denoted the anisotropy ratio which typically is larger than unity, unless the layers are tilted (Bear, 1979; Freeze & Cherry, 1979; Hvorslev, 1951).

Information on the directional dependence of hydraulic conductivity is relevant in various hydrogeological applications including the design of remediation systems, the determination of well capture zone geometries considering partially penetrating wells, and the prediction of contaminant plume spreading, among others (e.g., Bair & Lahm, 1996; Klammler et al., 2017; Paradis et al., 2016; Paradis & Lefebvre, 2013; Riva et al., 2006). On regional scales, the stratified character of fluvial deposits can induce anisotropy ratios of up to 100 (Freeze & Cherry, 1979). This is particularly relevant when applying dewatering measures near big rivers to keep a construction pit dry. Hydraulic anisotropy influences the spatial extent of the depression cone when lowering the water table upon pumping. While higher hydraulic conductivities in the horizontal direction increase the lateral extent of the depression cone, the comparatively lower vertical hydraulic conductivities decrease its vertical extent (Bair & Lahm, 1996).

Neglecting the hydraulic anisotropy on larger scales when calculating the dewatering flux required to inhibit water flow to a construction pit will result in substantial additional costs. Therefore, we address the challenge of quantifying the hydraulic anisotropy on larger scales induced by heterogeneity on comparatively smaller scales in a stratified aquifer. Additionally, we want to assess the dewatering flux affected thereof in an idealized setting. A frequently applied approach for resolving three-dimensional subsurface heterogeneity with emphasis on the horizontal variability of $K$ is the approach of hydraulic tomography (e.g., Bohling, 2009; Bohling et al., 2007; Sánchez-León et al., 2016). Various studies have addressed the challenge of resolving the vertical variability of $K$ in aquifers using hydrogeological investigation techniques such as direct-push injection logging or the direct-push in situ permeameter (e.g., Bohling et al., 2012, 2016; Chen et al., 2008, 2010; Klammler et al., 2017).
We simulate an approach similar to that of hydraulic tomography by inverting steady-shape pumping tests using partially screened extraction wells, with screens at various depths, and multilevel observation wells at various radial distances. For analyzing these data, we need to specify a model considered as an appropriate representation of the complex subsurface. Calibrating a model in which we attempt to fully resolve the $\mathbf{K}$-field of an effectively heterogeneous subsurface would either demand an infeasible data amount or leave us with enormous uncertainties. Simplifying the model requires important decisions on using a sequence of scalar, local conductivity values across some relevant number of strata, an effective macroscopic conductivity tensor over the entire aquifer formation, or something in between. Simultaneously, we have to address the challenge of designing the measurement strategy with which we want to obtain conclusive and sufficient data while reducing the number of necessary observation wells to minimize costs. In facing this challenge, designing the data collection and choosing the affordable model complexity are inherently interconnected, and the optimal combination of measurement design and model selection depends on the ultimate application of the calibrated model.

The remainder of this paper is organized as follows. In section 2, we suggest our framework for jointly optimizing the measurement and modeling strategies in light of a set modeling goal. We apply this framework to hydraulic tests in a stratified aquifer in order to identify the hydraulic conductivity and its anisotropy for the design of dewatering measures in section 3. We discuss our results in section 4 and close the paper with concluding remarks in section 5.

2. General Approach

Our approach combines elements of two existing, but previously disjoint research fields: model selection and optimal design of experiments.

Different methods for assessing, rating, or selecting models result from different mentalities for which we refer the interested reader to existing review papers (e.g., Foglia et al., 2013; Höge et al., 2018). The typical notion in advanced model selection schemes is that considering the reproduction of measured data only (e.g., by minimizing the sum of weighted squared differences between measurements and model results) is an insufficient criterion for model selection, as this would favor unnecessarily complex models containing poorly constrained parameters. The implicit assumption is that these poorly constrained parameters cause large uncertainty in model predictions. Thus, typical objective functions for model selection include a penalty term for the complexity of the model, which may be expressed by different terms, depending on the underlying theoretical considerations.

Many model selection schemes consider the ability of the model to reproduce the data used for calibration, eventually splitting the data into a subset used for calibration and a second subset used for validation. The validation to data of the same type as used for calibration is useful if the purpose of the model is indeed to predict states that have been observed in the past, like in flood forecasts. In most hydrogeological applications, however, models are developed to predict states under conditions that are truly new (e.g., the drawdown due to a pumped well that has not yet been built). We would argue that model choices (jointly with observation designs) should be optimized to reduce the uncertainty of the true model targets, rather than extending the calibration procedure by terms that penalize the model complexity.

None of the model-selection schemes provide immediate guidance which data one should collect. This is the area of optimal design of experiments. Most optimal-design schemes rank competing measurement designs by minimizing the uncertainty of parameters (Sciotino et al., 2002; Vrugt et al., 2002), the uncertainty of associated predictions (Herrera & Pinder, 2005; McKinney & Loucks, 1992; Tiedeman et al., 2003, 2004), or costs associated with uncertainty (Feyen & Gorelick, 2005; Howard, 1966), all for a given model, while only a few approaches seek to provide data for optimal model selection (e.g., Atkinson, 2008; Nowak & Guthke, 2016; Pham & Tsai, 2016) or for optimal prediction in the presence of uncertain model choice (e.g., Neuman et al., 2012; Nowak et al., 2010; Pham & Tsai, 2015).

Based on these considerations, our approach is to combine the quest for model choice and design of field campaigns. Specifically, we combine Akaikie-type ideas about “best predictive performance” from model selection with “minimum predictive uncertainty after calibration and model selection” from optimal design...
of experiments, but additionally equipped with “accuracy compared to a reference model” as in model reduction studies.

In the remainder of this section, we illustrate our resulting framework for the joint optimization of measurement and modeling strategies of a complex system. Our methodology includes five steps, illustrated in Figure 1, which are explained in the subsequent, corresponding subsections.

2.1. Definition of the Modeling Problem and Modeling Goals

Defining the basic conceptual model of a problem involves stating the governing equations with their initial and boundary conditions to simulate the system at hand. In this context, a key conceptual uncertainty lies in selecting the state variables and processes to be modeled, and in the spatial discretization of the material properties of the system. Additionally, conceptual uncertainties may arise from the initial and boundary

Figure 1. General approach for jointly optimizing the measurement and modeling strategy of a complex system.
conditions of the modeling problem (Tiedeman et al., 2003). In our application, the conceptual uncertainty is restricted to the scale and zonation of material properties only. We denote the true spatially variable parameter field as \( x \), which in reality exhibits variations on all spatial scales. In order to achieve identifiability in model calibration, one must typically discretize or simplify the parameter field, for example, by the introduction of internally homogeneous zones, or the definition of a limited set of spatial base functions used for spatial interpolation (e.g., RamaRao et al., 1995). In any case, this yields a vector \( x_0 \) of a few uncertain parameter values, in which the index \( k \) refers to a model choice (see below). Hence, we consider both the uncertainty of parameter values and the structural uncertainty regarding the spatial discretization of the parameter fields. As we will see in our application, these two uncertainties are not independent of each other—it is well known that a coarser representation of material properties requires partially upscaled (effective) parameter values (Rubin, 2003, Chapter 5), if not even effective governing equations that differ from the equations used at high resolution.

In order to calibrate the competing models, we need to define which quantity \( y \) could potentially be measured in a field test and can be computed by the model candidates. Choosing a measurement design, that is, a set of locations where to measure \( y \), will be discussed in Step 2.

The performance of the calibrated models, for given measurement designs and parameter representation, requires defining a quantitative and computable target quantity \( t \) associated with a given modeling goal. An optimized measurement and modeling strategy should lead to the best agreement in the target quantity between the prediction of the calibrated model and observations of the target quantity. At the moment when the model and experimental strategy are developed, however, true measurements of the target quantity do not exist yet, so that we have to simulate them with the given uncertain knowledge about the system.

The last aspect is how to handle the high prior uncertainty regarding the true system. The identified best measurement setup and conceptual model should be robust against this uncertainty. Therefore, we work with an ensemble of virtual realities at high spatial resolution and compute the performance of the measurement designs and model concepts over the entire ensemble. This ensemble of virtual realities allows us to mimic many possible outcomes of any proposed measurement design, and so to statistically anticipate the subsequent calibration, validation and model selection.

### 2.2. Definition of Possible Observation Points

To make the optimal-design problem a selection process among preset designs, we need to define a set of possible observation locations and times, here called points. These individual possible points are denoted \( p_i, i = 1...n_p \), where \( n_p \) is the total number of observation points considered. This set is larger than the number of observation points that could be assessed in a single, feasible field experiment. Thus, we can propose many preset designs as different subsets from this large set: We define permissible measurement combinations, denoted as the designs \( d_j, j = 1...n_d \) to be tested, in which \( j \) is an index to a particular design, and \( n_d \) is the total number of designs. These combinations are chosen to be realistic from a field-experimental perspective regarding the number of points and their spatial arrangement. Note that the number of observation points chosen in design \( d_j \), denoted \( n_{obs} \), can be design specific.

In classical optimal design, the problem would be to identify the best measurement design \( d_j \) within the set of proposed designs, given a conceptual model \( M_k \) (see below), minimizing an uncertainty metric of the target quantity \( t \). Instead, we are interested in minimizing an error metric between a yet-to-be selected model \( M_k \) and an assumed (yet uncertain) truth.

### 2.3. Definition of the Ensemble of Virtual Realities and of the Ensemble of Simplified Model Candidates

In the following, we work with several model ensembles. The ensemble of virtual realities \( M_v \) replaces the true field system. These models exhibit the highest possible resolution of the parameter field \( x \) and include all known processes. The discretized parameter field of the virtual reality is denoted as \( x_0 \). Because the prior uncertainty is fairly high, we generate many virtual realities, leading to an ensemble of \( x_0 \). This ensemble may need to undergo a plausibility check to rule out virtual realities that are in stark contrast to past field observations (e.g., Erdal & Cirpka, 2019), but they are not calibrated. In the following, \( x_0(\ell), \ell = 1...n_\ell \), is
For all virtual realizations $\ell$, the parameter set of realization $\ell$ associated with the virtual reality $M_\ell(l)$, and $n_\ell$ is the number of realizations.

In addition to the virtual realities $M_\ell$, we specify $n_m$ ensembles of feasible model candidates $M_k$, $k = 1 \ldots n_m$. These are conceptually simpler than the virtual realities and could actually be calibrated in real applications (Doherty & Christensen, 2011). In our application presented later, the individual ensembles $M_k$ differ from each other by the resolution (and hence scale) of the parameter fields $x$.

In principle, generating the $n_m$ ensembles of models $M_k$ could be independent of the virtual realities, but we construct them by upscaling the parameter vectors $x_0(\ell)$ of each virtual reality $\ell$ to a lower resolution, resulting in one upscaled parameter vector $x_k(\ell)$ for each simplified model $k$ and realization $\ell$. This procedure ensures that the different model ensembles are realistically similar, and it avoids the difficult procedure of specifying consistent parameter priors for models at different scales. Using all $x_0(\ell)$ and $x_k(\ell)$, we simulate hypothetical field data (measurable quantities $y$) at all possible observation locations $p_i$ considered. In the following, $y_0(p_i(d_j), \ell)$ is the (error-free) simulated measurement at point $i$ within design $j$ in virtual reality (or realization) number $\ell$, whereas $y_k(p_i(d_j), \ell)$ is the same simulated measurement in the simplified model number $k$.

The simulated measurements of the virtual realities need to be perturbed by systematic and random measurement errors that reflect possible contributions to errors in field measurements (Barlebo et al., 2004). In particular, we account for the uncertainty of placing the observations at the right place by implementing a random perturbation $\varepsilon_{p_i}(d_j, \ell)$ of the measurement locations $p_i(d_j)$. Additionally, we add white noise $\tilde{y}_0$ to each measurement at point $i$ and virtual reality $\ell$ drawn from a Gaussian distribution with zero mean and a standard deviation representing the standard measurement error. The latter corresponds to the resolution of the intended measurement device used in a field experiment.

In contrast to the high-resolution virtual realities, the simulated measurements based on the simplified model candidates for calibration are not perturbed: Only the virtual realities have the job to mimic the outcomes of proposed field campaigns, while the simplified models will be calibrated against these mimicked outcomes.

For all virtual-reality realizations $\ell$ and all conceptual models $M_k$, we also compute the corresponding target quantities $t(d_j, \ell)$ and $\tilde{y}_k(p_i(d_j), \ell)$, which will be important for model selection in Step 5.

### 2.4. Mimicking a Best Estimate Calibration

The next step lies in mimicking a best estimate calibration for each virtual reality $\ell$ and measurement design $d_j$, in which the simulated and perturbed data sets of the virtual realities are compared to measurement-error free predictions by the simplified conceptual models $M_k$. Rather than performing an iterative optimization, we make use of the existing ensembles of model candidates, each consisting of the same number $n_\ell$ of realizations. This is not a conceptual difference to calibration, but merely a concession to the computational cost that would result from a calibration problem repeated $n_d \times n_m \times n_\ell$ times. Hence, for each virtual reality $\ell$, design $d_j$, and simplifying model $M_k$, we determine the realization $\lambda$ of the simplified model with the smallest sum of squared differences $\Phi(d_j, k, \ell, \lambda)$ between the perturbed simulated measurements of the virtual reality and nonperturbed simulated measurements of model $M_k$:

$$\Phi(d_j, k, \ell, \lambda) = \sum_{i=1}^{n_{\text{obs}}(d_j)} \left( y_0(p_i(d_j)) + \varepsilon_{p_i}(d_j, \ell) \right)^2 + \tilde{y}_0 - y_k(p_i(d_j), \lambda)^2$$

Note that since we compare each virtual-reality realization $\ell$ with all available realizations of a model candidate, we need to introduce the additional index $\lambda$. The best realization of model $M_k$ is denoted $\lambda_{\text{opt}}(d_j,k,\ell)$:
This procedure is repeated with all virtual realities and all measurement designs, resulting in the number of realizations $n_r$ being sufficiently large. Due to the measurement errors included in the virtual-reality runs and the inaccuracies of model simplifications, the index of the optimal realization $\lambda_{opt}(d_j, k, \ell)$ may differ from the index of the virtual reality $\ell$ that has generated the data. Unlike in methods that aim at posterior distributions of parameters (e.g., preDIA, Leube et al., 2012), we do not exclude $\ell$ from the set of candidates, as the peak of the likelihood in a real best estimate calibration using the conceptual model $M_k$ may be very close to the upscaled parameter set $x_k(\ell)$. However, we additionally perform the best estimation calibration for the case that the simplified models are not able to encounter the true set of parameters, that is, by excluding the respective realization $\ell$ that has generated the data in the calibration process. The results are presented and discussed in the supporting information.

After having identified the best fit pairs, we quantify the quality of best fit with the root-mean-square error, $\text{RMSE}^t(d_j, k)$ for each design $d_j$ and conceptual model $M_k$ on average over all virtual-reality realizations $\ell$:

$$\text{RMSE}^t(d_j, k) = \sqrt{\frac{1}{n_{\text{obs}}(d_j)} \frac{1}{n_r} \sum_{\ell=1}^{n_r} \phi(d_j, k, \ell, \lambda = \lambda_{opt}(d_j, k, \ell))}$$

In our application, the individual conceptual models $M_k$ differ in the spatial resolution of the same parameter field $x$. Thus, one may expect that a model variant with a larger number of discretized parameters is better in meeting the perturbed measurements than a variant with fewer discretized parameters. That is, the corresponding error metric $\phi(d_j, k, \ell, \lambda_{opt}(d_j, k, \ell))$ should be smaller for $k$ representing a model with more model parameters. However, the simulated measurements are prone to error, so that the higher flexibility of the model with more parameters may be wasted on reproducing these errors without increase in predictive performance—a phenomenon known as overfitting (Vanlier et al., 2014). Hence, to truly test which combination of conceptual model and measurement design is best, we must evaluate their ability to predict the target quantity $t$ rather than their ability to reproduce the error-affected measurements on which they have been calibrated.

### 2.5. Optimization With Respect to the Modeling Goal

After determining all best fit models for each design and conceptual model, we analyze the performance of the selected models in predicting the target quantity $t$, that is, the modeling goal, over all virtual-reality realizations. Toward this end, we compute the relative error $E_{rel}^t(d_j, k, \ell)$ in the target quantity $t$ for each virtual reality $\ell$, measurement design $d_j$, and conceptual model $M_k$:

$$E_{rel}^t(d_j, k, \ell) = \frac{t_0(\ell) - t_k(\lambda_{opt}(d_j, k, \ell))}{t_0(\ell)}$$

Since we subsequently want to compare errors in target quantities among all realizations of virtual realities which may vary by several orders of magnitude we consider relative errors in Equation 4. However, this is not the case when calibrating each virtual reality $\ell$ using Equation 1 since errors relative to the magnitude of the true measurements would not change the result of determining the best realization $\lambda_{opt}(d_j, k, \ell)$ of each model $M_k$.

With Equation 4 we obtain empirical distributions depending on the measurement design $d_j$ and conceptual model $M_k$. We analyze these empirical distributions by three standard metrics: (1) the mean relative error $ME_{rel}^t(d_j, k)$, quantifying a bias in predicting the target quantity, (2) the standard deviation of the relative error $SDE_{rel}^t(d_j, k)$, quantifying the spread, and (3) the root-mean-square relative error $\text{RMSE}_{rel}^t(d_j, k)$, quantifying the overall misfit:
3. Application to Radial Flow in Stratified Aquifers

In this section, we describe a test case of radial flow in a stratified aquifer to which we apply the combined optimization of measurement design and conceptual model choice outlined in section 2.

3.1. Problem Statement

Our application is motivated by dimensioning the dewatering of construction sites in fluvial aquifers that can be approximated as stratified system. For simplicity, we consider flow toward a single well in both (1) the hydraulic tests to determine the hydraulic conductivity in the stratified aquifer and (2) in the target application to dewatering the construction site. Then, the steady-state groundwater-flow equation can be expressed in radial and vertical coordinates $r$ [L] and $z$ [L], respectively, as

$$K_r \left( \frac{\partial^2 h}{\partial r^2} + \frac{1}{r} \frac{\partial h}{\partial r} \right) + \frac{\partial}{\partial z} \left( K_z(z) \frac{\partial h}{\partial z} \right) = 0$$

in which $h$ [L] denotes the hydraulic head, whereas $K_r(z)$ [L T$^{-1}$] and $K_z(z)$ [L T$^{-1}$] are the horizontal and vertical hydraulic conductivities, respectively. The conductivities vary only in the vertical direction. In the framework described in section 2, the combined parameter fields $K_r(z)$ and $K_z(z)$ make up the parameter field $x$ with different spatial resolutions among the conceptual models to be calibrated and in the virtual realities.

In the following, we consider confined conditions and a flat aquifer base and top. The boundary conditions of the groundwater flow equation differ between the setup of the hydraulic tests used to identify the conductivity distribution and the application to the dewatering measure. The setup of both the hydraulic tests and the dewatering scenario are explained next.

3.2. Hydraulic Tests

In the hydraulic tests, we simulate quasi steady-state groundwater flow to a partially penetrating well of radius $r_w$ [L]. Water is extracted with a constant pumping rate successively from six different depths. As common in pumping-test analysis, the measurements are based on drawdown $s$ [L], that is, the difference between the steady-state hydraulic head without pumping to that established upon pumping. By this, the influence of ambient flow is removed. The drawdown meets the same partial differential equation as the hydraulic head:
\[
\frac{\partial^2 s}{\partial r^2} + \frac{1}{r} \frac{\partial s}{\partial r} + \frac{\partial}{\partial z} \left( K_z \frac{\partial s}{\partial z} \right) = 0
\]

(10)

with boundary conditions explained below. We mimic the steady-shape regime of a pumping test, in which the differences in hydraulic head between observation locations do not change anymore with time even though the actual head values do change. As a consequence, the measured quantity \( y \) according to the notation in section 2 consists of drawdown differences between piezometer screens rather than the absolute values. Quasi steady-state drawdowns are measured in several multilevel piezometers at different radial distances (Figure 2a).

The computational domain has a radius \( R \) [L] chosen large enough that it hardly affects drawdown differences. At this boundary, we consider zero drawdown:

\[
s(R, z) = 0 \quad \forall z
\]

(11)

At the aquifer base and top, we assume no-flow boundary conditions:

\[
\left. \frac{\partial s}{\partial z} \right|_{z = z_{\text{bot}}} = \left. \frac{\partial s}{\partial z} \right|_{z = z_{\text{top}}} = 0
\]

(12)

with the vertical coordinates \( z_{\text{bot}} \) [L] and \( z_{\text{top}} \) [L] at the aquifer base and top, respectively. At the radius of the well \( r_w \), we assign a no-flow boundary condition everywhere except for the active well screen \( i_{\text{act}} \) of length \( w \) [L], where we enforce the head to be constant along the screen and assume a constant total flux corresponding to the pumping rate \( Q_w \) [L^3 T^{-1}] (Table 1):

Figure 2. (a) Setup of the hydraulic test to estimate hydraulic conductivity: Radial-symmetric steady-state groundwater flow to a partially penetrating well. Water is successively extracted along six well screen sections \( i_{\text{scr}} \) at \( r = r_w = 0.1 \) m and drawdown is observed at all depths of water extraction in 3 out of 18 radial distances to the pumping well. Analysis of 816 designs, each design combining three multilevel piezometers. (b) Application to the dewatering of a construction pit: Radial symmetric steady-state groundwater flow (colored lines: head contours; black lines: streamlines) toward a construction pit within a large river. The base of the construction pit is at \( z_{\text{pit}} \), sheet piles pushed to depth \( z = z_{\text{sheet}} \) are placed at \( r = r_{\text{sheet}} \) to separate the construction pit from the surrounding river.
at \( r = r_w \):

\[
\begin{align*}
\frac{\partial s}{\partial r} &= 0 & \text{if } \frac{w}{2} \leq z \leq \frac{w}{2} \\
\int_{z_c(iscr)-w}^{z_c(iscr)+w} 2\pi r K \frac{\partial s}{\partial r} dz &= Q_w & \text{if } \frac{w}{2} \leq z \leq \frac{w}{2}
\end{align*}
\]

(13)

in which \( z_c(iscr) \) [L] is the vertical coordinate in the center of the screen section \( i_{scr} \in \{1;2;3;4;5;6\} \). Note that, as stated in Equation 13, water is extracted from each of the six well screens successively.

### 3.3 Hypothetical Dewatering Scenario

The hydraulic tests are performed to guide the design of dewatering a hypothetical cylindrical construction pit within a large river. As shown in Figure 2b, the construction pit is separated from the surrounding river at the radial distance \( r_{\text{sheet}} \) [L] by sheet piles extending to a depth \( z_{\text{sheet}} \) [L] that is somewhat deeper than the bottom pit \( z_{\text{pit}} \) [L]. Outside of the construction pit, the hydraulic head at the top equals the river stage \( h_{\text{river}} \) [L]. This implies the following fixed-head boundary conditions:

\[
h(r < r_{\text{sheet}}, z = z_{\text{pit}}) = z_{\text{pit}}
\]

(14)

\[
h(r > r_{\text{sheet}}, z = z_{\text{top}}) = h_{\text{river}}
\]

(15)

In addition, there are no-flow boundary conditions at the outer radial boundary of the domain, the bottom of the aquifer and along the sheet piles:

\[
\frac{\partial h}{\partial r} \bigg|_{r=R} = 0 \quad \forall z
\]

(16)

\[
\frac{\partial h}{\partial z} \bigg|_{z=Dae} = 0 \quad \forall r
\]

(17)

\[
\frac{\partial h}{\partial r} \bigg|_{r=r_{\text{sheet}}, z_{\text{sheet}} < z < z_{\text{top}}} = 0
\]

(18)

In real-world applications, the underlying assumption of radial symmetry may be violated and ambient flow may be needed to be considered as well, thus requiring a 3-D model domain. While such simulations pose no principal problems, the computational effort would be considerably higher. As we do not attempt to simulate the situation at a specific site, for which the corresponding geometries and boundary condition would be available, we restrict our analysis to the simplified case of radial flow toward a circular construction pit.

The key target quantity to predict (\( f \) according to the nomenclature of section 2.1) is the dewatering flux \( Q_d \) [L³ T⁻¹] needed to keep the water table in the pit at the bottom of the pit. This flux is computed by integrating the vertical specific discharge over the area of the construction pit:

### Table 1

**Parameters of the Tests Case**

| Hydraulic tests | Dewatering scenario |
|-----------------|---------------------|
| Pumping rate \( Q_w \) | 0.004 m³/s |
| Minimum/maximum mean of \( \ln K \) | \(-11.5/-4.6\) |
| Minimum/maximum variance of \( \ln K \) | 0.5/3 |
| Minimum/maximum correlation length of \( \ln K \) | 0.1 m/1.5 m |
| Vertical discretization | 0.025 m |
| Radial discretization | Logarithmically increasing from well radius \( r_w = 0.1 \) m to radius \( R = 50 \) m |
| \| | 0.1 m for \( r < r_{\text{sheet}} = 5 \) m; Logarithmically increasing from \( r_{\text{sheet}} \) to radius \( R = 50 \) m |

10.1029/2019WR026872
As computing the dewatering flux requires setting up and running an additional numerical model, we also test whether the fully upscaled, anisotropic hydraulic conductivity tensor (see section 3.5) is met by the estimated conductivity distribution resulting from the analysis of the hydraulic test. This results in two more target quantities, namely the effective horizontal and vertical conductivities \( K_{\text{eff}}^r \) \([L T^{-1}]\) and \( K_{\text{eff}}^z \) \([L T^{-1}]\), respectively.

3.4. Proposed Designs of Observation Wells in the Hydraulic Test

To suggest potential measurement designs, we need to define spatial locations \( p_i \) at which the drawdown may be measured. Trying to stay realistic regarding experimental feasibility, we limit the number of multilevel piezometers to three and the number of observation depths per multilevel piezometers to six, namely exactly at the depths of the screen sections of the extraction well. Figure 2a shows 18 potential distances at which the multilevel piezometer could be placed, including the multilevel extraction well itself. Therefore, the total number of observation points \( n_p = 108 \) points.

As mentioned in section 2.3, in the virtual realities we consider that piezometers may be misplaced. We simulate this by considering erroneous \( r \) and \( z \) coordinates of an intended measurement point in the virtual realities (but not in the simplified models to be calibrated against the data). Toward this end, we compute random values drawn from uncorrelated uniform distributions on the interval \([-0.025 \text{ m}; 0.025 \text{ m}]\) and \([-0.02 \text{ m}; 0.02 \text{ m}]\) for perturbing the \( r \) and \( z \) coordinates of a proposed observation point, respectively. In the field installation of piezometers the horizontal position may increasingly drift with depth, causing increasing uncertainty of the actual horizontal position with increasing depth. To account for this uncertainty, we additionally perturb the \( r \) coordinates of measurement points by the following double cumulative noise term:

\[
\varepsilon_{r, \text{cum}}(z_i) = 5 \times 10^{-3} |m| \sum_{j=1}^{i} \sum_{k=1}^{j} \omega_k
\]

in which the summation is performed over the layers of the numerical grid starting from the top, \( z_i \) is the vertical coordinate of \( i \)th grid layer from the top, and \( \omega_j \) is drawn from a standard normal distribution. Information on the spatial discretization is given in section 3.6.

An example of resulting deviations in intended and misplaced measurement locations is shown in Figure 3. Since we do not consider the pumping well to be affected from an imprecise placement of measurement location, we exclude such errors in the therein considered observation points.

As we use drawdown differences, each design includes \( 3 \times 6 - 1 = 17 \) observations. However, for designs involving the extraction well we exclude the pumped screen section itself as an observation point since in field applications the extractions screen is typically influenced by well-skin effects and pressure fluctuations caused by the pump. Thus, such designs only include 16 drawdown differences. There are in total \( n_d = 816 \) possibilities to combine three out of 18 multilevel piezometers, making up the individual measurement designs \( d_j \). We group the designs into sets numbered 0 to 15 according to the distance of the respective observation well closest to the pumping well (Figure 2a). Within one set, we sort the designs by shifting the observation wells further outward, starting with the outermost well until it reaches the last possible radial distance (Figure 4).
In each design \( d_j \), the lowermost observation point of the observation well closest to the pumping well constitutes the reference point for generating drawdown differences \( \Delta s \). For the particular case that water is extracted at the lowermost depth when analyzing designs of Set 0, we change the reference point to the second lowermost observation point for generating \( \Delta s \). Each simulated drawdown of the virtual realities is independently perturbed by a measurement error \( \varepsilon_s \) drawn from a normal distribution with zero mean and standard deviation \( \sigma_s = 3 \) mm corresponding to the maximum resolution of the intended measurement device in a field test.

### 3.5. Generating Realizations of Hydraulic Conductivity as Virtual Realities and for Model Calibration

The ensemble of virtual realities with the parameter sets \( x_0 \) contains \( n_r = 20,000 \) realizations of highly resolved vertical profiles of hydraulic conductivity. Toward this end, we generate multi-Gaussian one-dimensional fields of the horizontal log-hydraulic conductivity with an exponential covariance function of uncertain mean, variance, and correlation length. To address this uncertainty, we draw for each realization the mean, variance, and correlation length from a uniform distribution on the interval of the respective minimum and maximum (Table 1) before generating the autocorrelated vertical profile of \( \ln K_r(z) \).

Figure 5a shows the example of a highly resolved conductivity profile with the blue line marking the local horizontal conductivity and the red line the slightly smaller vertical conductivity. To obtain local values of the vertical hydraulic conductivity, we employ a random local anisotropy factor between 1 and 2 using the cumulative distribution function of a standard normal distribution shifted by 1.

Each highly resolved hydraulic conductivity field, representing a virtual reality, is partially upscaled to obtain the candidate parameter fields of the \( n_m \times n_r \) realizations of simplified models. The first conceptual model assumes a single homogeneous anisotropic layer, denoted 1-layer model. The second model assumes two layers of equal thickness, the third model three layers, and the fourth model six layers, denoted 2-, 3-, and 6-layer models, respectively. Upscaling the horizontal conductivity is done by taking the arithmetic...
average of the highly resolved $K_r$ field over the layer thickness, and upscaling the vertical conductivity by the harmonic average of the highly-resolved $K_r$ field:

$$K_{r}^{\text{eff}}(\text{layer}) = \frac{1}{\Delta z} \int_{z_{\text{top}}(\text{layer})}^{z_{\text{bot}}(\text{layer})} K_r(\xi) d\xi$$

(21)

$$K_{\xi}^{\text{eff}}(\text{layer}) = \Delta z \left( \int_{\bar{z}_{\text{top}}(\text{layer})}^{\bar{z}_{\text{bot}}(\text{layer})} \frac{1}{K_{\xi}(\xi)} d\xi \right)^{-1}$$

(22)

in which $z_{\text{top}}(\text{layer}) [\text{L}]$ and $z_{\text{bot}}(\text{layer}) [\text{L}]$ are the bottom and top vertical coordinates of the layer over which the averaging is performed, respectively, and $\Delta z = z_{\text{top}}(\text{layer}) - z_{\text{bot}}(\text{layer}) [\text{L}]$ is the corresponding layer thickness. Figure 5b illustrates the partially upscaled conductivity profiles according to the 1-, 2-, 3- and 6-layer models derived from the highly resolved profile shown in Figure 5a. $K_{r}^{\text{eff}}$ and $K_{\xi}^{\text{eff}}$ of the 1-layer model represent the fully upscaled radial and vertical conductivities and act both as a candidate in the calibration exercise and as alternative targets $t$ of the overall optimization.

3.6. Numerical Implementation and Computational Effort

The entire computational code is written in Matlab. The generation of highly resolved auto-correlated hydraulic-conductivity fields is done by the spectral approach of Dietrich and Newsam (1993), whereas the flow fields are computed by conforming finite elements using bilinear elements on a rectangular grid.

With the ensemble size of $n_r = 20,000$ realizations, we need to simulate $6 \times 20,000 = 120,000$ drawdown fields of the hydraulic tests using the virtual truths, one for each of the six extraction depths and for each realization. From this, we can construct all hypothetical measurement values for all 816 designs and 20,000 realizations. We also need to perform 20,000 simulations of the dewatering example using the virtual realities in order to obtain the correct values of the target quantity for Step 5. Information on the discretization of both numerical models is given in Table 1.

The computational effort for the simulations using the simplified conceptual models is 4 times larger because we test four conceptual models. However, the effort to compute 480,000 drawdown fields in order to mimic the best estimate calibrations (in Step 4) is still much smaller than performing $816 \times 20,000 \times 4 = 65,280,000$ real best estimate calibrations.

All other steps follow the procedure outlined in section 2:

1. We perturb the locations of all potential observation points in all realizations of the virtual truth, pick and perturb the simulated drawdown values at these locations with measurement noise, and assign the right drawdown differences to the right measurement designs to obtain the virtual measurements $y_0$ for all realizations $\ell$ of the virtual truth and all designs $d_j$ (part of Step 3 in Figure 1).

2. We compute the target quantities $K_r^{\text{eff}}(l), K_{\xi}^{\text{eff}}(l)$, and $Q_d(l)$ for all realizations $l$ of the virtual truth (part of Step 3 in Figure 1).

3. We pick the simulated drawdown values at all unperturbed potential observation points in all realizations of all four candidate models and assign the right drawdown differences to the right measurement designs to obtain the simulated measurements $y_0(p(d_j),l)$ of the different candidate models (part of Step 3 in Figure 1).

4. We mimic the best estimate calibration for each virtual reality $\ell$, measurement design $d_j$ and conceptual model $M_k$ according to Equation 2 (Step 4 in Figure 1).
5. We determine the combination of measurement design $d_j$ for a given conceptual model $M_k$ or the best combination of the latter two by finding the minimum of the root-mean-square relative errors $RMSE_{rel}(d_j, k)$ according to Equation 7 with $t$ being $K_{eff}^r$, $K_{eff}^z$, or $Q_d$ (Step 5 in Figure 1).

The overall computational effort on a Fujitsu P957/power High End PC was 336.1 hr wall clock time.

4. Results and Discussion

4.1. Reproduction of Measurements by the Calibrated Models

Figure 6 illustrates the root-mean-square error $RMSE_{rel}(d_j, k)$ of the best fit of the 1-, 2-, 3-, and 6-layer models in meeting the drawdown differences of the virtual reality as function of the measurement design. Black line: root-mean-square values of the measurements. Purple line: Measurement error expected from the measurement devices employed in field tests.

Figure 6. Root-mean-square error $RMSE_{rel}(d_j, k)$ of the best fit of the 1-, 2-, 3-, and 6-layer models in meeting the drawdown differences of the virtual reality as function of the measurement design. Black line: root-mean-square values of the measurements. Purple line: Measurement error expected from the measurement devices employed in field tests.

5. We determine the combination of measurement design $d_j$ for a given conceptual model $M_k$ or the best combination of the latter two by finding the minimum of the root-mean-square relative errors $RMSE_{rel}(d_j, k)$ according to Equation 7 with $t$ being $K_{eff}^r$, $K_{eff}^z$, or $Q_d$ (Step 5 in Figure 1).

The overall computational effort on a Fujitsu P957/power High End PC was 336.1 hr wall clock time.

4. Results and Discussion

4.1. Reproduction of Measurements by the Calibrated Models

Figure 6 illustrates the root-mean-square error $RMSE_{rel}(d_j, k)$ (see Equation 3) of the best fit realizations $\lambda_{opt}$ for each measurement design $d_j$ and conceptual-model candidate $k$ as colored lines (red: 1-layer model, green: 2-layer model, blue: 3-layer model, gold: 6-layer model). The $RMSE_{rel}(d_j, k)$ values quantify how well the measurements simulated by the calibrated simplified models agree with the simulated and perturbed measurements of the virtual realities after mimicking the calibration (Step 4). For comparison Figure 6 also contains a black line representing the root-mean-square value of the measurements in the virtual realities quantifying the total strength of the signal. Because Figure 6 shows the $RMSE_{rel}(d_j, k)$ values on a semilogarithmic scale, the distance between the colored and the black line can be interpreted as relative error of the calibration. For additional comparison, we computed the mean measurement error considered in the field tests (purple line in Figure 6).

First, we discuss general trends: As described in section 3.4, we sort the designs into sets comprising all designs with the same observation well closest to the pumping well. Set 0 includes the nonpumped well screens of the pumping well as observation points (see Figure 2a). As seen in Figure 6, the $RMSE_{rel}(d_j, k)$ values in general decrease with increasing distance between the pumping well and the first multilevel piezometer and approach the measurement error (see purple line of Figure 6). However, also the measured drawdown differences decrease (see black line of Figure 6) so that the relative misfits for a given conceptual model do not differ so strongly between the different design sets.
As can be seen by the illustrative drawdown contour lines in Figure 2a, the vertical variation in drawdown decreases with increasing distance to the pumping well so that a distinction between models with many or few layers becomes hardly possible when only distant multi-level piezometers are considered. For the same reason, the imprecise location of the observation points in the virtual realities has a higher effect on the absolute and relative measurement error for close-by than for far-away piezometers. However, we have not considered such misplacements for observations at the nonpumped well screens of the pumping well. The latter may explain why, contrary to the general trend, the design set 1 results in larger calibration errors than the design Set 0.

Now, we compare the models: Obviously, for all designs, the 6-layer model shows the smallest misfit of the measurements. This confirms the trivial expectation that a model with a higher spatial resolution of the hydraulic conductivity can reproduce measurements better than a model with a coarser resolution. In the design set including the extraction well as a piezometer, the 2-, and 3-layer models perform similarly well as the 6-layer model, whereas the RMSE$^2(d, k)$-curves of the model candidates separate when considering designs of larger set numbers. This also accounts for the 1-layer model which performs worst for all designs, followed by the 2- and 3-layer models. The increased spreading between model candidates with increased set number (pronounced by the logarithmic scaling in Figure 6) results from the decrease in signal (see black line in Figure 6) suggesting that increasing the spatial resolution of a model for reproducing measurements especially matters when measurements are afflicted with a small signal strength.

4.2. Predictive Performance of Calibrated Models for Given Measurement Designs Regarding Individual Target Quantities

While the 6-layer model is best in reproducing the observed drawdown differences, this does not guarantee that it is also best in predicting the desired target quantities, as there is the danger of overfitting. To assess the predictive performance of the models, we compute the relative error in predicting the target quantities of the dewatering flux $Q_d$, the effective radial hydraulic conductivity $K_{eff}^r$ upscaled over the full aquifer thickness, and the equally effective vertical hydraulic conductivity $K_{eff}^z$.

Figure 7. Fraction of all realizations $n_r$ in which a particular model candidate was determined as best among the available model candidates in minimizing $E_{rel}(d, k, e)$ of the target quantities: (a) $Q_d$, (b) $K_{eff}^r$, and (c) $K_{eff}^z$. 

As can be seen by the illustrative drawdown contour lines in Figure 2a, the vertical variation in drawdown decreases with increasing distance to the pumping well so that a distinction between models with many or few layers becomes hardly possible when only distant multi-level piezometers are considered. For the same reason, the imprecise location of the observation points in the virtual realities has a higher effect on the absolute and relative measurement error for close-by than for far-away piezometers. However, we have not considered such misplacements for observations at the nonpumped well screens of the pumping well. The latter may explain why, contrary to the general trend, the design set 1 results in larger calibration errors than the design Set 0.

Now, we compare the models: Obviously, for all designs, the 6-layer model shows the smallest misfit of the measurements. This confirms the trivial expectation that a model with a higher spatial resolution of the hydraulic conductivity can reproduce measurements better than a model with a coarser resolution. In the design set including the extraction well as a piezometer, the 2-, and 3-layer models perform similarly well as the 6-layer model, whereas the RMSE$^2(d, k)$-curves of the model candidates separate when considering designs of larger set numbers. This also accounts for the 1-layer model which performs worst for all designs, followed by the 2- and 3-layer models. The increased spreading between model candidates with increased set number (pronounced by the logarithmic scaling in Figure 6) results from the decrease in signal (see black line in Figure 6) suggesting that increasing the spatial resolution of a model for reproducing measurements especially matters when measurements are afflicted with a small signal strength.

4.2. Predictive Performance of Calibrated Models for Given Measurement Designs Regarding Individual Target Quantities

While the 6-layer model is best in reproducing the observed drawdown differences, this does not guarantee that it is also best in predicting the desired target quantities, as there is the danger of overfitting. To assess the predictive performance of the models, we compute the relative error in predicting the target quantities of the dewatering flux $Q_d$, the effective radial hydraulic conductivity $K_{eff}^r$ upscaled over the full aquifer thickness, and the equally effective vertical hydraulic conductivity $K_{eff}^z$. 

Figure 7. Fraction of all realizations $n_r$ in which a particular model candidate was determined as best among the available model candidates in minimizing $E_{rel}(d, k, e)$ of the target quantities: (a) $Q_d$, (b) $K_{eff}^r$, and (c) $K_{eff}^z$. 

As can be seen by the illustrative drawdown contour lines in Figure 2a, the vertical variation in drawdown decreases with increasing distance to the pumping well so that a distinction between models with many or few layers becomes hardly possible when only distant multi-level piezometers are considered. For the same reason, the imprecise location of the observation points in the virtual realities has a higher effect on the absolute and relative measurement error for close-by than for far-away piezometers. However, we have not considered such misplacements for observations at the nonpumped well screens of the pumping well. The latter may explain why, contrary to the general trend, the design set 1 results in larger calibration errors than the design Set 0.

Now, we compare the models: Obviously, for all designs, the 6-layer model shows the smallest misfit of the measurements. This confirms the trivial expectation that a model with a higher spatial resolution of the hydraulic conductivity can reproduce measurements better than a model with a coarser resolution. In the design set including the extraction well as a piezometer, the 2-, and 3-layer models perform similarly well as the 6-layer model, whereas the RMSE$^2(d, k)$-curves of the model candidates separate when considering designs of larger set numbers. This also accounts for the 1-layer model which performs worst for all designs, followed by the 2- and 3-layer models. The increased spreading between model candidates with increased set number (pronounced by the logarithmic scaling in Figure 6) results from the decrease in signal (see black line in Figure 6) suggesting that increasing the spatial resolution of a model for reproducing measurements especially matters when measurements are afflicted with a small signal strength.
We first report the empirical probability that a particular calibrated model $M_k$ performs best in meeting an individual target quantity $t$ for a given design $d_i$ over all realizations $\ell = 1...n_r$ of the virtual reality. Per individual realization $\ell$, we denote the selection of best performing model $k^*_{opt}(d_i, \ell)$:

$$k^*_{opt}(d_i, \ell) = \arg\min_k E_{rel}^i(d_i, k, \ell)^2$$

(23)

and then count the relative frequencies of model selection across the virtual-reality realizations. Figure 7 shows the resulting selection probability for the 1-, 2-, 3-, or 6-layer models, respectively. Each subplot is related to one of the three target quantities (dewatering flux $Q_d$ and fully upscaled horizontal and vertical hydraulic conductivities $K_{eff}^r$ and $K_{eff}^z$).

Looking at general trends across all target quantities, we see that all selection probabilities range between 5.4% and 49%. This indicates that none of the model candidates explicitly predominate model selection. In general, the selection probability of the 1-, 2-, 3-, and 6-layer models reach similar values with increasing design number, that is, with increasing distance between the pumping well and the observation wells. As already discussed in section 4.1, the vertical variation in drawdown decreases with increasing distance to the pumping well. This leads to a better fit in calibration. In the current context, this also means that a distinction between models becomes hardly possible when only distant multilevel piezometers are considered.

Looking at individual model errors, for all target quantities the homogeneous 1-layer model is selected the least. This even holds for the fully upscaled directional hydraulic conductivities $K_{eff}^r$ and $K_{eff}^z$, which are uniform effective parameters that should relate perfectly to the 1-layer conceptual model (per definition, see section 3.5). We may explain this discrepancy by the spatially nonuniform sensitivities of the drawdown measurements with respect to hydraulic conductivity even under uniform hydraulic conditions, which leads to a bias in the estimated conductivities of the 1-layer model. Since the 1-layer model considers the largest possible scale, its selection probability for the fully upscaled directional hydraulic conductivities $K_{eff}^r$ and $K_{eff}^z$ improves for piezometers at far radial distances.

Now, we discuss individual target quantities. Figure 7a shows that the 6-layer model has the highest fraction of best predicting the dewatering flux $Q_d$ among the model candidates, regardless of the measurement designs in Sets 0–10. We may explain this by the flow field in the vicinity of the construction pit shown in Figure 2b: The dewatering flux crucially depends on (a) the hydraulic conductivity tensor around the bottom end of the sheet pile and (b) the vertical hydraulic conductivity integrated between the bottom of the sheet pile $z_{sheet}$ and the bottom of the construction pit $z_{pit}$. The 6-layer model is suited best to capture these properties.

In contrast, for predicting $K_{eff}^r$ and $K_{eff}^z$, the 6-layer model is not chosen as frequently as the best model (Figures 7b and 7c). Here, first the 3-layer model and then the 2-layer model approximate the best performing model (predicting $K_{eff}^r$) or perform just as well as the 6-layer model (predicting $K_{eff}^z$) when increasing the distance of the multilevel piezometers to the pumping well. This can be explained with the effect of overfitting the erroneous measurements in the virtual reality suggesting that the model complexity of a 6-layer model is required for predicting $Q_d$, yet does not necessarily yield better results for predicting $K_{eff}^r$ and $K_{eff}^z$.

### 4.3. Predictive Errors

Finally, to provide quantitative statements, Figure 8 reports the root-mean-square relative errors $RMSE_{rel}^i$ in predicting the three target quantities in all 20,000 realizations considering the 816 different designs with the calibrated 1-, 2-, 3-, and 6-layer models. The individual subplots refer to the chosen conceptual models, whereas the line colors distinguish the target quantities. For all designs and models, the $RMSE_{rel}^i$ values strongly depend on the targeted parameter $K_{eff}^r$, $K_{eff}^z$, or $Q_d$, respectively. The errors vary between 0.13 and 3.66, indicating that the target quantities are underestimated or overestimated by 13% to 366%, depending on the selected measurement design, model candidate, and target quantity. For most measurement designs and model candidates, predicting the fully upscaled horizontal hydraulic conductivity $K_{eff}^r$ is easier than predicting the fully-upscaled vertical conductivity $K_{eff}^z$ and the dewatering flux $Q_d$. This can be
explained by the flow being predominantly horizontal, so that the designs are more informative about horizontal properties of the system.

Over all designs, the average magnitude of $R\text{MSE}^{rel}_{t}$ depends on the selected model, with the 1-layer model exhibiting the largest offsets between the correct target quantity values and the prediction by the calibrated model, whereas the most complex 6-layer model scores mostly the best, even though the prediction errors of the 3-layer model highly resemble those of the 6-layer model.

The $R\text{MSE}^{rel}_{t}$ values show characteristic trends among the analyzed measurement designs, which are similar among the model candidates but appear to be differently pronounced due to the given dependence on systematic model errors. In the following, we discuss these trends for the different target quantities.

In general, the highest error in predicting $K_{\text{eff}}^{r}$ (blue lines in Figures 8a–8d) is obtained by applying the first design within each set. These are the designs in which the three multilevel piezometers are the closest to each other. With increasing design number within a set, the errors in predicting $K_{\text{eff}}^{r}$ fluctuate showing a decreasing trend toward the last design of the associated set. In order to predict the fully upscaled horizontal hydraulic conductivity $K_{\text{eff}}^{r}$, it is best to include measurements with the largest horizontal offset possible, and in a given set this is achieved by taking the two outermost positions for piezometers two and three. However, it is neither preferable to choose a first piezometer location that is too close to the pumping well, where the effects of misplacing the piezometers are the biggest, nor too far away where the measured signal gets too small. This implies that the least uncertainty in predicting $K_{\text{eff}}^{r}$ is achieved with the first multilevel piezometer being placed at intermediate distances to the extraction well, and the two others as far away as possible.

In contrast to the $R\text{MSE}^{rel}_{t}$ values for $K_{\text{eff}}^{r}$, those for $K_{\text{eff}}^{z}$ and $Q_{d}$ show an increasing trend within each set (red and black lines in Figures 8a–8d), expressing that the prediction uncertainty is reduced when the second multilevel piezometer is placed close to the first rather than the third one. This can be explained by the necessity of strong vertical drawdown differences in the observations to obtain good estimates of the

Figure 8. Root-mean-square relative errors $R\text{MSE}^{rel}_{t}$ in predicting $K_{\text{eff}}^{r}$ (blue), $K_{\text{eff}}^{z}$ (red), and $Q_{d}$ (black) as function of the chosen measurement designs for individual model choices. (a) 1-layer model, (b) 2-layer model, (c) 3-layer model, and (d) 6-layer model.
vertical hydraulic conductivity, and strong vertical gradients are found closer to the extraction well. Similar to the error in predicting $K_{eff}^r$, the measurement design sets with the first observation well at a close, but not too close, distance to the pumping well seem to guarantee the smallest prediction error of $K_{eff}^z$ and $Q_d$.

### 4.4. Multiobjective Optimization

Equation 6 describes how to select the single best combination of measurement design $d_j$ and conceptual model $M_k$ to be calibrated to the data in order to meet an individual target quantity $t \in \{Q_d, K_{eff}^r, K_{eff}^z\}$. Depending on the target, we come to different optimal designs that are visualized in Figure 9a. In order to satisfy all three criteria, a multiobjective optimization is needed (resulting designs see Figure 9b).

We examine potential trade-offs in predicting the different target quantities with a certain model and measurement strategy following the Pareto principle. Toward this end, we consider the RMSE$_{rel}^t$ values for all target quantities from all available combinations $n_m \times n_d = 3,264$ of conceptual models and measurement designs and plot them against each other in Figure 10. In particular, Figures 10a–10c show pairwise Pareto plots of the relative prediction error in two of the three target quantities (Figure 10a: $K_{eff}^r$ and $K_{eff}^z$; Figure 10b: $K_{eff}^r$ and $Q_d$; Figure 10c: $K_{eff}^z$ and $Q_d$) for all combinations of calibrated model candidates and measurement designs, whereas Figure 10d shows a three-dimensional Pareto plot of all three RMSE$_{rel}^t$ values. The red, green, blue, and gold colors of points and spheres in Figure 10 relate to the 1-, 2-, 3-, and 6-layer models, respectively.

Figure 10 illustrates that there are no big trade-off effects when optimizing the model-and-design combinations for the fully upscaled horizontal conductivity $K_{eff}^r$ and either the corresponding fully upscaled vertical conductivity $K_{eff}^z$ (Figure 10a) or the dewatering flux $Q_d$ (Figure 10b). Meeting $K_{eff}^z$ and $Q_d$ clearly leads to a preference for the 6-layer model and 3-layer model, respectively, whereas meeting $K_{eff}^r$ alone does not show
such a preference. However, meeting the criterion of minimizing $RMSE_{rel}^t$ for $t = K_{eff}^r$ does not truly exclude meeting the other two criteria. We have computed the Pareto front (not shown), but it contained only a very small number of points, confirming the lacking trade-off between meeting $K_{eff}^r$ and $K_{eff}^z$ and $K_{eff}^r$ and $Q_d$, respectively.

As indicated by Figures 10a and 10b, Figure 10c shows that the 6-layer model is the preferred model for minimizing $RMSE_{rel}^t$ for $t = K_{eff}^z$ whereas for minimizing $RMSE_{rel}^t$ for $t = Q_d$ the 3-layer model predominates model selection. However, similar to their lacking trade-offs with $K_{eff}^r$, selecting a model-and-design combination for predicting the target $K_{eff}^z$ does not necessarily deteriorate the prediction of $Q_d$.

Instead of computing the Pareto front, we thus decided to set threshold values corresponding to the 10% best model-and-design combinations of the 3,264 choices ($RMSE_{rel}^t = 0.19$ for $t = K_{eff}^r$, $RMSE_{rel}^t = 0.43$ for $t = K_{eff}^z$, and $RMSE_{rel}^t = 0.54$ for $t = Q_d$). Figures 10a–10c show these threshold values as gray lines, and Figure 10d as translucent gray planes parallel to the axes. The spheres with clear colors in the foreground of Figure 10d belong at least in one of the three criteria to the 10% best model-and-design combinations, and those of them closest to the observer belong to the best regarding multiple criteria.

There are exactly two model-and-design combinations that belong to the 10% best with respect to $K_{eff}^r$, $K_{eff}^z$, and $Q_d$. Both suggest to analyze the data by the 6-layer model and to install the closest multilevel piezometer at 4-m distance to the pumping well, and the second closest at 6 m, whereas the location of the furthest multilevel piezometer is chosen either at 19 m or at 25 m distance (see also Figure 9b). We consider these two designs as the overall best ones.

5. Conclusions and Recommendations

We have presented a framework for jointly selecting measurement strategies and models of data interpretation to optimize the predictive skill of the calibrated model. The proposed framework is based on mimicking best fit model calibration using simulated and perturbed measurements in an ensemble of virtual realities by selecting the simulated measurements of ensembles of reasonable model candidates for a variety of feasible observation designs. For each ensemble member of the virtual reality and each measurement design, this
approach essentially imitates manual calibration, in which the best fit refers to a best match among manually chosen parameter sets rather than a global minimum.

Instead of judging the calibrated conceptual models and measurement designs by the ability of the models to reproduce the measurements in calibration, we define realistic prediction targets of the calibrated models that should be as close to the predictions of the virtual realities as possible. By this approach, overfitting of models is punished if and only if it deteriorates the predictive performance of the calibrated models. Unnecessary but harmless model complexity is not seen as a reason to reject a model candidate. This perspective is different to that of Bayesian model selection and to that of the theories behind information criteria, in which model complexity is penalized per se (e.g., Höge et al., 2018). The latter approaches, however, base the model selection exclusively on the measured data and not on prediction goals of the models beyond data reproduction.

In the given framework, jointly selecting measurement strategies and conceptual models is straightforward if a single prediction target is defined. In cases with several model purposes, methods of mult-objective optimization apply. In our application, however, two out of three model targets led to extremely similar choices of model-and-design combinations, and the third target was complementary rather than contradictory, so that a fairly clear best combination of model choice and measurement strategy could be identified without strong trade-off. This, of course, may be different in other applications, in which constructing the Pareto front may be advisable.

We have applied the suggested method to steady-state radial symmetric groundwater flow in stratified aquifers, in which a set of quasi steady-state pumping tests with partially penetrating wells and three multilevel piezometers are used to infer the macroscopic hydraulic anisotropy of the formation and the dewatering flux necessary to keep an exemplary construction pit dry. While the virtual realities exhibited a high resolution of the vertical hydraulic-conductivity profiles, the model candidates proposed and used for practice, considered 1, 2, 3, or 6 horizontal layers of uniform, anisotropic conductivity. The conceptual question of the model selection thus was whether one should estimate macroscopic hydraulic anisotropy, or hydraulic heterogeneity, or a combination of both, whereas the measurement-design question was whether multilevel piezometers should be placed particularly close to the pumping well (where drawdown signals are strong and exhibit strong vertical differences but are particularly prone to spatial misplacements of the piezometers) or further away (where both the signal and systematic measurement errors are smaller).

Not very surprisingly, the most complex conceptual model performed best with respect to meeting the “true” measurements during calibration. In our specific case, choosing the number of layers to be identical to the number of observation depths payed off, since this selection not only led to the best fit of observations but also to the best prediction of the fully upscaled vertical hydraulic conductivity and the dewatering flux necessary to keep the construction pit dry. For the fully upscaled horizontal conductivity chosen as target quantity, the superiority of the 6-layer model was less persistent over the different designs. In the discussed example we consider designs composed of three piezometers with six observation points each. We include the pumping well as a possible piezometer, yet exclude the observation point in the respective pumped screen section. This yields an overall of once 96 (designs including the pumping well) and once 102 (for designs excluding the pumping well) independent measured drawdown differences per design which showed little to no signs of overfitting when determining the $2 \times 6 = 12$ parameters of the 6-layer model. This may be different in applications with fewer independent measurements.

Quite clearly, ignoring the measurement design in model selection would deteriorate the model performance. In our application, placing observation wells at far radial distances to the pumping well hampers resolving hydraulic-conductivity values of several layers. Likewise, ignoring the model choice when optimizing the measurement strategy would make little sense. Our analysis also showed that the joint selection of the conceptual model and the measurement strategy depends on the ultimate objective of the model application. In our application, reducing the prediction uncertainty in the fully upscaled horizontal conductivity $K_{\text{H}}$ required observation points with large radial distance, whereas reducing the prediction uncertainty in the fully upscaled vertical conductivity $K_{\text{V}}$ and the dewatering flux $Q_d$ required multilevel observation wells at small radial distances, where the vertical hydraulic gradients are the biggest.
In our specific application we could achieve the best combined reduction in the relative prediction error of $K_i^0$, $K_i^1$, and $Q_d$ by employing the 6-layer model and collect data in one observation well close by and one observation well further away from the pumping well. The third observation well should be placed in between the two other observation wells, but rather close by to the first observation well than to the third one.

When transferring our joint optimization procedure to other applications, it is important to construct realistic virtual realities covering the unconditional uncertainty and variability of the system at hand. In our application, one may question whether the assumption of a perfectly stratified formation is realistic enough. Extending the virtual realities to exhibit 3-D, anisotropic heterogeneity would be possible in principle, but computationally more expensive. We further advise to consider all potential causes of erroneous measurements. In our application, this included to address the potential spatial misplacement of observation points, which may have a bigger and more systematic influence on the quality of the measurements than white-noise errors of the measurement device. Finally, the goal-oriented optimization of measurement design and model choice requires that the model targets are clearly defined and computable. Optimizing the model-and-design combination for one target and subsequently applying the model for other purposes could lead to suboptimal choices.

Data Availability Statement

The data generated by the model and the Matlab codes are available on the FDAT repository of the University of Tübingen (http://hdl.handle.net/10900.1/19bc1d8-b5ec-492f-919a-1ed23f81e5b1).

Acknowledgments

This work was supported by the Federal Waterways Engineering and Research Institute (BAW), Germany, and the Collaborative Research Center 1,253 CAMPOS (Project P8: Conceptual uncertainty, model legitimacy, and optimization of data acquisition strategies), funded by the German Research Foundation (DFG, Grant Agreement SFB 1253/12017). In the provided supporting information we present and discuss the results of mimicking the best estimate calibration for the case in which the true parameters sets are excluded in the calibration process.

References

Atkinson, A. C. (2008). DT-optimum designs for model discrimination and parameter estimation. Journal of Statistical Planning and Inference, 138(1), 56–64. https://doi.org/10.1016/j.jspi.2007.05.024
Bair, E. S., & Lahm, T. D. (1996). Variations in capture zone geometry of a partially penetrating pumping well in an unconfined aquifer. Ground Water, 34(5), 842–852. https://doi.org/10.1111/j.1745-6584.1996.tb02079.x
Bohling, H. C., Hill, M. C., & Rosbjerg, D. (2004). Investigating the Macrodispersion Experiment (MADE) site in Columbus, Mississippi, using a three-dimensional inverse flow and transport model. Water Resources Research, 40, W04211. https://doi.org/10.1029/2002WR001935
Bear, J. (1979). Hydraulics of groundwater. In McGraw-Hill series in water resources and environmental engineering. McGraw-Hill International Book Co. Retrieved from https://books.google.de/books?id=N-pOAAAAMAAJ
Beven, K. (2002). Towards a coherent philosophy for modelling the environment. Proceedings of the Royal Society of London. Series a: Mathematical, Physical and Engineering Sciences, 458, 2465–2484. https://doi.org/10.1098/rspa.2002.0986
Bohling, G. C. (2009). Sensitivity and resolution of tomographic pumping tests in an alluvial aquifer. Water Resources Research, 45, W02420. https://doi.org/10.1029/2008WR007249
Bohling, G. C., Butler, J. J. Jr., Zhan, X., & Knoll, M. D. (2007). A geostatistical analysis of scale hydraulic conductivity variations at the MADE site. Water Resources Research, 43, W04530. https://doi.org/10.1029/2006WR004932
Bohling, G. C., Liu, G., Dietrich, P., & Butler, J. J. Jr. (2016). Reassessing the MADE direct-push hydraulic conductivity data using a revised calibration procedure. Water Resources Research, 52, 8970–8985. https://doi.org/10.1002/2016WR019008
Bohling, G. C., Liu, G., Knobbe, S. J., Reboulet, E. C., Hyndman, D. W., Dietrich, P., & Butler, J. J. Jr. (2012). Geostatistical analysis of centimeter-scale hydraulic conductivity variations at the MADE site. Water Resources Research, 48, W02525. https://doi.org/10.1029/2011WR010791
Chaloner, K., & Verdinelli, I. (1995). Bayesian experimental design: A review. Statistical Science, 10(3), 273–304. https://doi.org/10.1214/ss/117709939
Chen, X., Burbach, M., & Cheng, C. (2008). Electrical and hydraulic vertical variability in channel sediments and its effects on streamflow depletion due to groundwater extraction. Journal of Hydrology, 352(3–4), 250–266. https://doi.org/10.1016/j.jhydrol.2008.01.004
Chen, X., Song, J., & Wang, W. (2010). Spatial variability of specific yield and vertical hydraulic conductivity in a highly permeable alluvial aquifer. Journal of Hydrology, 383(3–4), 379–388. https://doi.org/10.1016/j.jhydrol.2010.05.017
Dietrich, C. R., & Newsam, G. N. (1993). A fast and exact method for multidimensional Gaussian stochastic simulations. Water Resources Research, 29(8), 2861–2869. https://doi.org/10.1029/93WR00495
Doherty, J., & Christensen, S. (2011). Use of paired simple and complex models to reduce predictive bias and quantify uncertainty. Water Resources Research, 47, W12534. https://doi.org/10.1029/2011WR010763
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
Feyen, L., & Gorelick, S. M. (2005). Framework to evaluate the worth of hydraulic conductivity data for optimal groundwater resources management in ecologically sensitive areas. Water Resources Research, 41, W03019. https://doi.org/10.1029/2003WR002901
Foglia, L., Mehli, S. W., Hill, M. C., & Burlando, P. (2013). Evaluating model structure adequacy: The case of the Maggia Valley groundwater system, southern Switzerland. Water Resources Research, 49, 260–282. https://doi.org/10.1029/2011WR011779
Freeze, R. A., & Cherry, J. A. (1979). Groundwater. ISBN 0–13–365312-9. Englewood Cliffs, NJ: Prentice-Hall. Retrieved from https://books.google.de/books?id=8F7KwFnxKwUC, ISBN 0–13–365312-9.
Heinz, J., Kleinheidam, S., Teutsch, G., & Aigner, T. (2003). Heterogeneity patterns of Quaternary glacitectonic gravel bodies (SW-Germany): Application to hydrogeology. Sedimentary Geology, 158, 1–23. https://doi.org/10.1016/S0037-0738(02)00239-7

MAIER ET AL.
