frontiers
in Earth Science

ORIGINAL RESEARCH
published: 16 November 2020
doi: 10.3389/feart.2020.554845

A Stochastic Framework to Optimize Monitoring Strategies for Delineating Groundwater Divides

Jonas Allgeier[1], *Ana González-Nicolás*[2], *Daniel Erdal*[1,3], *Wolfgang Nowak*[2] *and Olaf A. Cirpka*[1]*

[1]Center for Applied Geoscience, University of Tübingen, Tübingen, Germany, [2]Institute for Modelling Hydraulic and Environmental Systems (LS3/SimTech), University of Stuttgart, Stuttgart, Germany, [3]Tyréns AB, Göteborg, Sweden

Surface-water divides can be delineated by analyzing digital elevation models. They might, however, significantly differ from groundwater divides because the groundwater surface does not necessarily follow the surface topography. Thus, in order to delineate a groundwater divide, hydraulic-head measurements are needed. Because installing piezometers is cost- and labor-intensive, it is vital to optimize their placement. In this work, we introduce an optimal design analysis that can identify the best spatial configuration of piezometers. The method is based on formal minimization of the expected posterior uncertainty in localizing the groundwater divide. It is based on the preposterior data impact assessor, a Bayesian framework that uses a random sample of models (here: steady-state groundwater flow models) in a fully non-linear analysis. For each realization, we compute virtual hydraulic-head measurements at all potential well installation points and delineate the groundwater divide by particle tracking. Then, for each set of virtual measurements and their possible measurement values, we assess the uncertainty of the groundwater-divide location after Bayesian updating, and finally marginalize over all possible measurement values. We test the method mimicking an aquifer in South-West Germany. Previous works in this aquifer indicated a groundwater divide that substantially differs from the surface-water divide. Our analysis shows that the uncertainty in the localization of the groundwater divide can be reduced with each additional monitoring well. In our case study, the optimal configuration of three monitoring points involves the first well being close to the topographic surface water divide, the second one on the hillslope toward the valley, and the third one in between.

Keywords: gaussian process emulation, preposterior data impact assessor, bayesian analysis, uncertainty quantification, optimal design of measurements, delineation, groundwater divide

OPEN ACCESS

Edited by:
Andrés Alcolea,
Independent Researcher, Winterthur, Switzerland

Reviewed by:
Jorge Jódar,
Instituto Geológico y Minero de España (IGME), Spain
Alberto Guadagnini,
Fondazione Politecnico di Milano, Italy

***Correspondence:**
Olaf A. Cirpka
olaf.cirpka@uni-tuebingen.de

Specialty section:
This article was submitted to Hydrosphere, a section of the journal Frontiers in Earth Science

Received: *23 April 2020*
Accepted: *08 October 2020*
Published: *16 November 2020*

Citation:
Allgeier J, González-Nicolás A, Erdal D, Nowak W and Cirpka OA (2020) A Stochastic Framework to Optimize Monitoring Strategies for Delineating Groundwater Divides. Front. Earth Sci. 8:554845. doi: 10.3389/feart.2020.554845

1. INTRODUCTION

Groundwater divides are curves separating different subsurface catchments. Water entering the subsurface on one side of the groundwater divide ends up in a different receptor than water infiltrating on the other side of the divide. Delineating groundwater divides is therefore important for the analysis of aquifer water budgets, for investigating contaminant fate, and other applications of groundwater management. Groundwater divides also represent attractive geometries for setting second-type boundaries of hydrogeological models, since the water flux across the divide is zero (e.g., Pöschke et al., 2018; Erdal and Cirpka, 2019; Qiu et al., 2019). Obviously, a natural stream network contains many nested surface water and groundwater divides of different order (i.e., a catchment can be subdivided into sub-catchments). That is why for the mentioned research areas, it is always important to define the scale of investigation to identify which groundwater divides are relevant and which sub-catchments can be attributed to a higher-order catchment.

A common assumption when delineating groundwater divides is that the groundwater table is a subdued representation of the surface topography (Tóth, 1963; Haitjema and Mitchell-Bruker, 2005). This simplifies the delineation to finding the surface water divides, which can be derived directly from digital elevation models using geographic information systems (Tarboton et al., 1991). However, the topography of a phreatic groundwater surface may substantially differ from the land surface so that the groundwater and surface water divides do not coincide (Haitjema and Mitchell-Bruker, 2005; Bloxom and Burbey, 2015; Han et al., 2019). In fact, Haitjema and Mitchell-Bruker (2005) reported on a whole class of aquifers naturally exhibiting such shifts between surface and subsurface water divides. They demonstrated under which conditions a groundwater table is mainly controlled by surface topography or by recharge. These authors concluded that a shifted groundwater divide may be caused by relatively high hydraulic conductivity in conjunction with a difference between the elevation of drainage points in neighboring valleys. Additional factors contributing to shifts in groundwater divides include tilted aquifer strata, spatial heterogeneity in the recharge rate, and anisotropy in hydraulic conductivity. Of course, anthropogenic influence (e.g., drinking water extraction wells) can also contribute to shifted groundwater divides.

The location of groundwater divides can be constrained by hydraulic-head measurements. Theoretically speaking, a very dense network of piezometers could be used to accurately interpolate the groundwater-table map, which could subsequently be analyzed by the same tools as used for delineating surface-water divides. In practice, this is not advisable as the number of observation wells is restricted by financial costs, labor intensity, and legal restrictions. That is, groundwater divides must be delineated with head measurements from a limited number of piezometers. A classical way of doing this is by calibrating groundwater flow-and-transport models to the head measurements, which explicitly uses all information fed into the model construction (e.g., the geometry and parameter ranges of geological units and boundary conditions) and leads to hydraulic-head fields that are consistent with conservation principles.

As only a limited number of observation wells is affordable, their placements should be specifically optimized for delineating a particular groundwater divide. Either, one wants to find the best possible piezometer configuration for a fixed number of wells, in which the optimum is defined by minimizing the uncertainty of the divide's position, or one wants to find the well configuration requiring the least number of wells for a fixed target uncertainty of the divide's location. In both cases, the objective is to maximize the information-to-costs ratio, which is a general problem well-known under the name of "optimal design of experiments" (Pukelsheim, 2006; Fedorov, 1972).

In this study, we solve the described optimization problem. We provide a framework to identify the best set of points to delineate a particular groundwater divide. The "goodness" of such a point set is defined by how much the uncertainty in the divide's location is reduced, if hydraulic-head measurements were available at these points. The best set of points might then be implemented as real-world monitoring wells, whose measurements can be used to calibrate a flow model for actually delineating the divide of interest.

Of course, during the stage of identifying promising measurement locations it is unknown which measurement values would be obtained at these locations. To circumvent this problem, we apply a specific technique of optimal experimental design, called Preposterior Data Impact Assessor (PreDIA, Leube et al., 2012). We feed it with a sample of steady-state groundwater models that is efficiently pre-selected to include only plausible subsurface flow fields (Erdal et al., 2020). By means of delineating the groundwater divide for each individual realization and virtually conducting all possible measurements, we can quantify both, the total uncertainty of the groundwater divide's location across the domain and by how much this scalar quantity can be reduced with a specific measurement configuration.

The main contributions of the present study are the formulation of the problem and the development of a suitable objective function for delineating a groundwater divide, as well as the combination of PreDIA with the pre-selection of plausible model results.

The motivation behind our work originates from a real field site. During the investigation of a floodplain, it was discovered that the observed lateral groundwater influxes from the hillslope are too small to drain the water quantities gained by the hillslope's expected recharge. This imbalance of in- and outfluxes has led to the conclusion that the groundwater divide underneath the hillslope is shifted in a way that the contributing area draining toward the floodplain is much smaller than expected, when considering the surface water divide as contributing boundary. The phenomenon of flow crossing surface water divides has been referred to as "interbasin groundwater flow". It needs to be quantitatively estimated, before detailed studies focusing on the hillslope or floodplain can be conducted. The information of whether or not such interbasin flow occurs in a domain and how pronounced it is can furthermore be of utter importance, for example if contamination occurs in one basin and a sensitive receptor (e.g., a drinking water supply well) is located in the other one.

We developed our framework for cases, where the (suspected) shift of a groundwater divide is the phenomenon of interest that needs to be quantified. In reality, such a shifted divide might additionally be subject to transient processes (i.e., it might move with time). This is not covered by our methodology, but we believe our analysis might still be useful in such cases (see **section 4.5**). We want to emphasize that a shifted divide does not imply its movement over time. A groundwater divide can very well be at a (quasi-)steady state while being shifted due to the geological setting, which does not change significantly over time scales relevant for groundwater management.

Section 2 introduces and explains the underlying framework. Real data from a site in Southwest Germany are used in **section 3** to test the method. We want to highlight that we separate our site-specific implementation details (*application*) from the general approach of our framework (*Methods*). The results of our example study are presented and discussed in **section 4**. Finally, we draw conclusions and give an outlook in **section 5**.

2. METHODS

2.1. Subsurface Flow Equations

The optimal experimental design method we use later on (**section 2.4**) is based on stochastic runs of a steady-state subsurface flow model. To model saturated and unsaturated parts of the subsurface, we solve the steady-state version of the Richards equation for variably saturated flow in porous media (Richards, 1931):

$$-\nabla \cdot \mathbf{q} = Q \tag{1}$$

$$\mathbf{q} = -\mathbf{K} k_{\mathrm{rel}}\left(h_p\right) \nabla h_{\mathrm{tot}} \tag{2}$$

$$h_p = h_{\mathrm{tot}} - z \tag{3}$$

in which $\mathbf{q}$ is the specific discharge vector (dim $\mathbf{q} = \mathrm{L\,T^{-1}}$), Q represents volumetric source ($Q < 0$) or sink ($Q > 0$) terms (dim $Q = \mathrm{T^{-1}}$), h_{tot} is the total head (dim $h_{\mathrm{tot}} = \mathrm{L}$), $\mathbf{K}$ is the hydraulic-conductivity tensor (dim $\mathbf{K} = \mathrm{L\,T^{-1}}$) under water-saturated conditions, k_{rel} is the dimensionless relative permeability, h_p is the pressure head (dim $h_p = \mathrm{L}$), and z is the geodetic height (dim $z = \mathrm{L}$).

The relative permeability k_{rel} and the dimensionless effective saturation S_e are parameterized by the Mualem/van-Genuchten relationships (Mualem, 1976; van Genuchten, 1980):

$$S_{\mathrm{e}} = \begin{cases} \left(1 + \left(\alpha \left|h_p\right|\right)^N\right)^{\frac{1-N}{N}} & \text{if } h_p < 0 \\ 1 & \text{otherwise} \end{cases} \tag{4}$$

$$k_{\mathrm{rel}}\left(S_{\mathrm{e}}\left(h_p\right)\right) = \sqrt{S_{\mathrm{e}}}\left(1 - \left(1 - S_{\mathrm{e}}^{\frac{N}{N-1}}\right)^{\frac{N-1}{N}}\right)^2 \tag{5}$$

$$\Theta_{\mathrm{w}} = \Theta_{\mathrm{r}} + (\Theta_{\mathrm{s}} - \Theta_{\mathrm{r}}) S_{\mathrm{e}} \tag{6}$$

in which Θ_{w}, Θ_{r}, and Θ_{s} are the actual, residual, and saturated dimensionless (volumetric) water contents, α is a van-Genuchten parameter similar to the inverse entry-pressure head (dim $\alpha = \mathrm{L^{-1}}$), and N is the associated dimensionless pore-distribution index.

By including the Mualem/van-Genuchten parametrization, the Richards equation holds for variably saturated flow (i.e., both the saturated and unsaturated zone). In the saturated zone ($h_p > 0$), both the effective saturation and the relative permeability become unity. Here, the Richards equation naturally simplifies to the groundwater-flow equation based on Darcy's law and the continuity equation. In the unsaturated zone ($h_p < 0$), the effective saturation and relative permeability are subject to nonlinear equations depending on the pressure head. The groundwater table is located at the transition from saturated to unsaturated zone ($h_p = 0$). Since the used parametrization does not define a clear entry pressure, there is no capillary fringe in a strict sense. However, the parameter α serves a similar purpose meaning that only if the capillary head (equals $-h_p$ in the unsaturated zone) is well above $\frac{1}{\alpha}$, the saturation drops significantly. That is, the model includes a zone above the groundwater table where the effective water saturation is close to unity, which resembles the capillary fringe. Using the Richards equation coupled to Mualem/van-Genuchten relationships to model saturated and unsaturated parts of the subsurface simultaneously has been common practice for decades (e.g., Tocci et al., 1998; Farthing et al., 2003; Suk and Park, 2019).

We apply the following boundary conditions:

$$h_{\mathrm{tot}} = h_{\mathrm{fix}} \qquad \text{on} \quad \Gamma_{\mathrm{D}} \tag{7}$$

$$\mathbf{n} \cdot \mathbf{q} = q_{\mathrm{fix}} \qquad \text{on} \quad \Gamma_{\mathrm{N}} \tag{8}$$

$$h_{\mathrm{tot}} = \min\left[h_{\mathrm{sim}}, z_{\mathrm{surf}}\right] \qquad \text{on} \quad \Gamma_{\mathrm{S}} \tag{9}$$

$$Q = \frac{C_{\mathrm{L}}}{V} \cdot \left(h_{\mathrm{tot}} - h_{\mathrm{riv}}\right) \qquad \text{on} \quad \Gamma_{\mathrm{L}} \tag{10}$$

$$Q = \begin{cases} \frac{C_{\mathrm{D}}}{V} \cdot \left(h_{\mathrm{tot}} - z_{\mathrm{surf}}\right) & \text{if } h_{\mathrm{tot}} - z_{\mathrm{surf}} > \Delta z \\ 0 & \text{otherwise} \end{cases} \qquad \text{on} \quad \Gamma_{\mathrm{T}} \tag{11}$$

in which h_{fix} is a known hydraulic head (dim $h_{\mathrm{fix}} = \mathrm{L}$), $\mathbf{n}$ is the dimensionless unit normal vector, q_{fix} is a known normal flux (dim $q_{\mathrm{fix}} = \mathrm{L\,T^{-1}}$), h_{sim} is the simulated head if the boundary was considered a no-flow boundary ($h_{\mathrm{sim}} = \mathrm{L}$), z_{surf} is the surface elevation (dim $z_{\mathrm{surf}} = \mathrm{L}$), C_{L} is a river conductance (dim $C_{\mathrm{L}} = \mathrm{L^2\,T^{-1}}$), V is the volume related to the source/sink term (dim $V = \mathrm{L^3}$), h_{riv} is a known river head (dim $h_{\mathrm{riv}} = \mathrm{L}$), C_{D} is a drainage conductance (dim $C_{\mathrm{D}} = \mathrm{L^2\,T^{-1}}$) and Δz is a pressure difference threshold (dim$\Delta z = \mathrm{L}$). Here, Γ_{D} denotes a Dirichlet boundary, Γ_{N} a Neumann boundary, Γ_{S} a seepage boundary, Γ_{L} a leaky (e.g., river) boundary and Γ_{T} a top drainage boundary.

The leaky boundary condition can account for interactions between groundwater and river water. The respective exchange flux is driven by the head difference $h_{\mathrm{tot}} - h_{\mathrm{riv}}$ and a conductance C_L:

$$C_{\mathrm{L}} = \frac{L_{\mathrm{riv}} \cdot w_{\mathrm{riv}}}{L_{\mathrm{sed}}} \cdot K_{\mathrm{sed}}, \tag{12}$$

where L_{riv} and w_{riv} are the associated river stretch length and width (dim $L_{\mathrm{riv}} =$ dim $w_{\mathrm{riv}} = \mathrm{L}$), L_{sed} is the thickness of the sediment bed (dim $L_{\mathrm{sed}} = \mathrm{L}$), and K_{sed} is the sediment's hydraulic conductivity (dim $K_{\mathrm{sed}} = \mathrm{L\,T^{-1}}$).

A similar conductance C_D regulates the drainage flux at surficial drainage boundary conditions:

$$C_{\mathrm{D}} = \frac{A}{L_{\mathrm{lay}}} \cdot K_{\mathrm{lay}}, \tag{13}$$

where A is the associated surface area (dim $A = \mathrm{L^2}$), L_{lay} is the thickness of the intermediate layer (dim $L_{\mathrm{lay}} = \mathrm{L}$) and K_{lay} is its hydraulic conductivity (dim $K_{\mathrm{lay}} = \mathrm{L\,T^{-1}}$).

After simulating subsurface flow, we use particle tracking to determine the groundwater divide as explained in **section 2.3**. Toward this end, we introduce particles at the land surface, track their advective movement according to the advective velocity **v**, and analyze on which side of the groundwater system they end. This approach is a common procedure for delineating subsurface water divides (e.g., Hunt et al., 2001; Han et al., 2019):

$$\frac{d\mathbf{x}_i}{dt} = \mathbf{v}\left(\mathbf{x}_i\left(t\right)\right) \tag{14}$$

$$\text{subject to } \mathbf{x}_i\left(t = 0\right) = \mathbf{x}_i^{\text{ini}} \tag{15}$$

$$\text{with } \mathbf{v} = \frac{\mathbf{q}}{\Theta_{\text{w}}} \tag{16}$$

in which $\mathbf{v}$ is the linear velocity (dim $\mathbf{v} = \mathsf{L}\,\mathsf{T}^{-1}$), $\mathbf{x}_i\left(t\right)$ is the position vector (dim $\mathbf{x}_i\left(t\right) = \mathsf{L}$) of particle i at time t (dim $t = \mathsf{T}$), and $\mathbf{x}_i^{\text{ini}}$ is the starting location (dim $\mathbf{x}_i^{\text{ini}} = \mathsf{L}$).

The approach of delineating the groundwater divide by particle tracking obviously implies that the divide is located within the modeling domain. This is in contrast to many practical groundwater-modeling studies, where the domain is bounded by the assumed groundwater divides. Under such conditions, these groundwater divides are fixed by the model choice. Since we want to study the uncertainty of the groundwater divide, we require a model domain where the divide is in the interior so that the model has the freedom to shift it.

2.2. Generation of a Plausible Model Sample

In order to capture the uncertainty of the divide's location (prior to any measurements and after hypothetical measurements), our framework makes use of ensemble-modeling. This implies the repeated simulation of the same conceptual model with different numerical representations. These can be formally identical, differing only, for example, in some material property values. They could also differ in more fundamental properties, like the internal structure. We call the final group of model entities a "sample", to avoid confusion with the term "ensemble" referring to such a group of infinite size. Each entity of the sample is termed a realization or sample member.

Formally, a sample member is defined both, by the formulation of the general model itself (common to all members) and by a member-specific set of parameters. In addition to that, the sample member also comprises its deterministic modeling results (after the model was evaluated), which can be reproduced from the general model by using the same parameter set. We denote these parameter sets **S**, a vector of all individual properties that differ between realizations. The vector **S** may include not only material properties, but also boundary conditions or geometric descriptors (for an example, we refer to our application in **section 3.2.3**).

In theory, we could create a sample of sufficient size just by drawing random parameter sets from appropriate prior distributions and subsequent numerical modeling of subsurface-flow. These prior distributions could be derived from measurements (e.g., pumping tests for hydraulic conductivities), other models (e.g., recharge rates) or expert knowledge (e.g., anisotropies). Afterward, particle-tracking would obtain one groundwater divide for each realization. In practice however, we need to exclude parameter sets that lead to implausible model results (e.g., wrong signs of fluxes across boundaries; more examples in context of our application, **section 3.3**), because that would ignore obvious insight into the correct system behavior and thus overstretch uncertainty. Conversely, we do not want to restrict the parameter ranges too much because we want to assess the full space of plausible model parameters. Therefore, we keep the prior parameter ranges untouched, but rely on the exclusion of models with obviously unrealistic results (denoted unbehavioral or implausible).

While excluding unbehavioral realizations is a conditioning step, we would not yet consider it a model calibration, but rather a plausibility check or pre-selection (see Erdal and Cirpka, 2019; Erdal and Cirpka, 2020; Erdal et al., 2020). In a rigorous conditioning step (i.e., "stochastic calibration") that could follow on this pre-selection, we would modify the parameters of sample members to better meet the exact measurement values. A potential method to do that would be an ensemble Kalman smoother. However, a full stochastic calibration on the existing data would be computationally expensive, but not informative about the quantity of interest, namely the position of the groundwater divide. The lack of hydraulic-head measurements that are informative about the delineation of the groundwater divide is the very reason why we perform the optimal design of experiments to begin with.

The decision about the plausibility and ultimately its acceptance or rejection of a candidate model is based on a set of criteria. Each plausibility criterion compares a scalar model outcome (e.g., the flux across a specific boundary) with a target value that must not be exceeded or fallen below. Only if a model realization fulfills all plausibility criteria, it will be included in the sample for further analysis.

A key problem of the pre-selection is that more than 94 % of randomly drawn parameter sets in our application miss at least one criterion. If we performed full runs of the numerical subsurface-flow model for each model candidate, we would thus waste more than 94 % of the computing time on model runs that must be discarded. To overcome this problem efficiently, we have adopted the pre-selection method of Erdal et al. (2020) (based on Erdal and Cirpka, 2019). It is based on relating the plausibility criteria with the model parameters **S** by means of interpolation, to estimate whether a new parameter set is likely to be plausible or not. Toward this end, it follows these steps:

(1) We create a small initial sample of **S** by Latin Hypercube sampling from appropriate priors and perform numerical subsurface-flow modeling for all sample members. We compute the respective values of the plausibility criteria for each realization.
(2) We train one Gaussian process emulator per plausibility criterion with the initial sample of full model runs. A Gaussian process emulator is a kriging interpolator in parameter space (a "proxy model" or "surrogate model") that estimates the expected value of the plausibility criterion and quantifies its estimation variance, provided that the

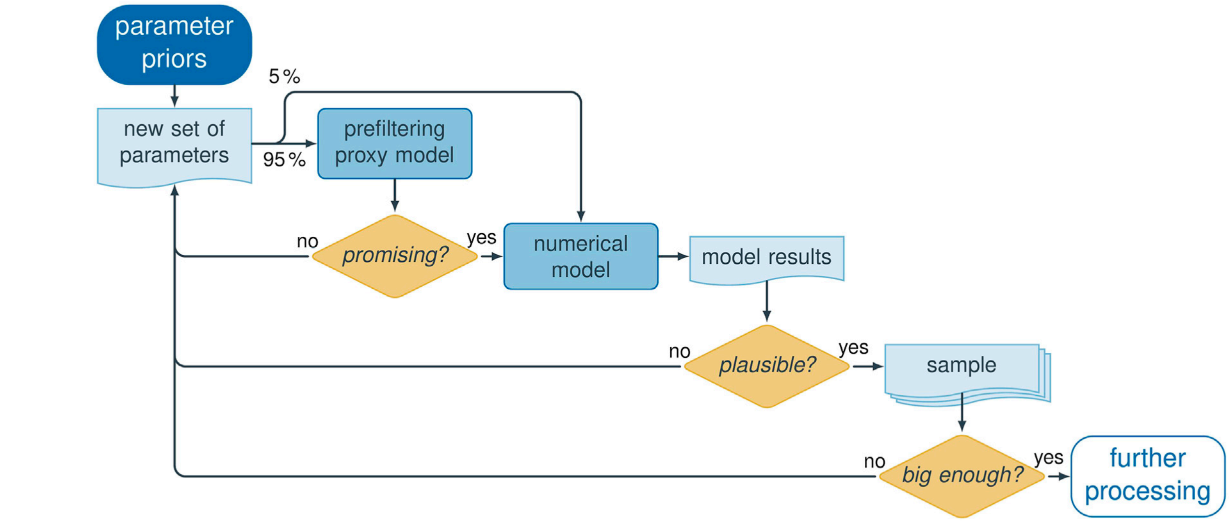

FIGURE 1 | Procedure used to generate a sample of physically plausible model realizations for the optimal experimental design analysis.

assumptions of kriging (e.g., statistical stationarity) hold. We want to emphasize here that this is not a spatial interpolation, but an interpolation of the model response to parameter values.

(3) We then draw further random samples of **S**. For each of them, we apply the Gaussian process emulators to compute the compliance probability with each plausibility criterion. If a realization's product of all individual compliance probabilities (i.e., its overall probability) does not exceed a certain threshold value (in our case 50 %), we discard it and draw a new sample. This evaluation is comparably quick (fraction of a second) and saves us modeling time that would be wasted by running a model that would probably need to be rejected due to implausible results.

(4) For a model candidate where this product exceeds the threshold probability (a "stage-1-accepted" realization), we perform the simulation of the full subsurface-flow model. A small percentage of sample members (we use 5 %) is run directly without checking against the Gaussian process emulator estimates first.

(5) If the model candidate also meets the plausibility criteria after running the full numerical model, it is "stage-2-accepted" (i.e., included in the sample of physically plausible models), and particle-tracking simulations are performed to obtain the groundwater divide. Otherwise, it is discarded.

(6) With an increasingly large set of full model runs, the Gaussian process emulator model is regularly retrained to improve its accuracy in predicting the behavioral status of subsequent model candidates.

With this procedure, we were able to increase the overall acceptance ratio, that is, the number of stage-2-accepted full-model runs over the total number of full-model runs. In the initial small sample (full Monte Carlo), only 6 % of the realizations passed the plausibility check (111 out of 2000). With the interpolation method, we were able to achieve an acceptance ratio of 69 % of realizations subject to a full model run (50,000 of 72,481 stage-1-accepted parameter sets; a large number of randomly drawn parameter sets was rejected in stage 1). **Figure 1** schematically illustrates the whole sample-generation procedure. It results in n_{sample} stage-2-accepted realizations that will actually be used in the following analysis.

2.3. Uncertainty in Delineating a Groundwater Divide

For each stage-1-accepted parameter realization (see step 4 in **section 2.2**), we determine the scalar model outcomes of the plausibility check. Additionally, we simulate virtual measurement values of hydraulic heads at all potential measurement locations, by determining the respective elevations of the groundwater table at these locations. The number and location of such potential measurements is known prior to the analysis and part of the problem statement.

Only for the n_{sample} stage-2-accepted realizations, we compute via particle tracking a vector **z** of particle fates for a regular map of starting locations: We introduce n_{par} particles at the model domain's surface. These particles are tracked through the domain until they exit the domain through a groundwater outlet. This tracking allows us to classify the particles into two categories summarized by the classification vector **z** with $z_i \in \{0, 1\}$ and $i = 1, \ldots, n_{\text{par}}$. A particle i that ends up in one outlet (A) is assigned the value $z_i = 1$, while a particle ending up in the other outlet (B) obtains a value of $z_i = 0$. Since each particle is related to a starting point in two-dimensional space, **z** represents what we call the binary particle-fate map. This binary classification is sufficient to delineate the boundary of a single subdomain, but it cannot be used to delineate all groundwater divides between more than two subdomains (e.g., due to groundwater extraction wells). In the appendix (**section**

5.1) we include a generalization to an arbitrary number of subdomains. In the following, we will focus on binary systems, because this is the most common scenario.

Other approaches than particle tracking for the delineation of groundwater divides exist. They are typically based on locating the "ridge of the groundwater table". However, they have been shown to be less reliable (Han et al., 2019).

The fate of a particle i depends on the parameter vector $\mathbf{S}$ (including all variable model decisions). The probability of z_i being one (that is, of the associated starting point to be within the catchment of outlet A) is computed by integrating over the space $\Omega_\mathbf{S}$ of the parameter vector $\mathbf{S}$, weighted with the probability density of $\mathbf{S}$:

$$\begin{aligned} P(z_i) &= \int_{\Omega_\mathbf{S}} z_i(\mathbf{S}) p(\mathbf{S})\, \mathrm{d}\mathbf{S} \\ &\approx \sum_{j=1}^{n_\text{sample}} z_i(\mathbf{S}_j) P(\mathbf{S}_j), \end{aligned} \tag{17}$$

in which $z_i(\mathbf{S})$ is the binary fate of particle i for the given parameter vector $\mathbf{S}$, $p(\mathbf{S})$ is the probability density of $\mathbf{S}$, and the second row of **Eq. 17** is the Monte-Carlo approximation of $P(z_i)$ by the sample of discrete $\mathbf{S}$-values with the probability $P(\mathbf{S}_j)$ given to the $\mathbf{S}$-value of the jth realization. In our initial sample, all accepted realizations are equally likely, implying $P(\mathbf{S}_j) = 1/n_\text{sample}\, \forall j$. Upon conditioning on (virtual) head measurements, $P(\mathbf{S}_j)$ will become a Bayesian weight (see below). Franzetti and Guadagnini (1996) and Hunt et al. (2001) used a similar approach to estimate the uncertainty of capture-zone delineations.

We can now compute the probability $P_\text{mc}(z_i)$ of misclassifying the fate of particle i:

$$P_\text{mc}(z_i) = 2P(z_i)(1 - P(z_i)). \tag{18}$$

This equation expresses the probability that particle i, which actually ends up in outlet A, is estimated to end up in outlet B or vice versa. P_mc ranges from zero (full certainty) to 0.5 (maximum uncertainty). The underlying assumption is that the decision threshold for classification is at 50 %. That is the reason for 0.5 being the largest value of P_mc. $P(\mathbf{z})$ and $P_\text{mc}(\mathbf{z})$ can be visualized as maps of probability all over the catchment. We integrate the probability of misclassification over all starting locations $\mathbf{x}_\text{ini}$ of particles to obtain an integral metric U of describing the uncertainty of the groundwater divide:

$$\begin{aligned} U(\mathbf{z}) &= \frac{1}{A_\text{2D}} \int_{A_\text{2D}} P_\text{mc}(z(\mathbf{x}^\text{ini}))\, \mathrm{d}\mathbf{x}^\text{ini} \\ &\approx \frac{1}{A_\text{2D}} \sum_{i=1}^{n_\text{par}} P_\text{mc}(z_i) A_i^\text{ini} \end{aligned} \tag{19}$$

in which A_2D is the two-dimensional top surface area of the model domain and A_i^ini is the contributing area of particle i, which may be computed by Voronoi tesselation of all starting locations (e.g., Brassel and Reif, 1979). Large values of $U(\mathbf{z})$ express that the outlet destination of particles is uncertain on a large fraction of the domain's surface, which is not desirable.

As discussed in the context of **Eq. 17**, the probability $P(z_i)$ of starting location $\mathbf{x}_i^\text{ini}$ being in the catchment of outlet A, and thus the associated probability of misclassification $P_\text{mc}(z_i)$ and ultimately the overall uncertainty $U(\mathbf{z})$, depends on the probabilities $P(\mathbf{S}_j)$ of individual parameter realizations j. This implies that conditioning the parameter vector $\mathbf{S}$ on head observations will change the overall uncertainty U of delineating the groundwater divide. The following optimal design analysis aims at minimizing U.

2.4. Prospective Optimal Experimental Design

To find the optimal placement of piezometers in order to delineate a groundwater divide, we apply the optimal experimental design method PreDIA (the Preposterior Data Impact Assessor, Leube et al., 2012), which we briefly review in the given context.

The scientific question of optimal design is to find the combination of measurements or experiments with the largest information content regarding a target quantity, before the experiment itself is carried out. Formally, the objective is to identify the single design $\mathbf{d}_\text{opt}$ of a set of n_des possible designs $\mathbf{d}$ in the design space $\mathbf{d} \in \mathbf{D}$ that maximizes a utility function $\phi(\mathbf{d})$ (Leube et al., 2012):

$$\mathbf{d}_\text{opt} = \underset{\mathbf{d} \in \mathbf{D}}{\arg\max} [\phi(\mathbf{d})] \tag{20}$$

A design in this notation is a vector containing information about how measurements are taken in time and space. The utility function $\phi(\mathbf{d})$ is a measure of the usefulness of data obtained with an experiment using design $\mathbf{d}$. The evaluation of ϕ obviously requires knowledge about the measurement results of a particular design, which is unknown at the stage of the optimal-experimental-design analysis. PreDIA can circumvent this problem by means of ensemble-based modeling.

As previously described, $\mathbf{S}$ denotes the input parameter vector, comprising all uncertain model decisions, such as material properties (e.g., hydraulic conductivity), boundary conditions (e.g., recharge), geometric parameters (e.g., thickness of geological units), or structural modeling parameters (e.g., presence of layers). As outlined above, we create a sample of members with physically plausible behavior. The variability in model input $\mathbf{S}$ leads to interdependent variability of model output, both with respect to simulated measurements and simulated target quantities (the particle-fate maps).

For a given realization $\mathbf{S}_i$, we can simulate virtual observations $\mathbf{f}_y(\mathbf{S}_i, \mathbf{d})$ for a specific design $\mathbf{d}$, in which $\mathbf{f}_y$ denotes the simulation outcome of the measured quantities. To account for measurement error, we add a random error term $\boldsymbol{\varepsilon}_y$ to $\mathbf{f}_y(\mathbf{S}_i, \mathbf{d})$ to obtain virtual measurements $\mathbf{y}_i(\mathbf{d})$ of a specific design $\mathbf{d}$ and parameter realization i:

$$\mathbf{y}_i(\mathbf{d}) = \mathbf{f}_y(\mathbf{S}_i, \mathbf{d}) + \boldsymbol{\varepsilon}_y \tag{21}$$

To answer the optimal-experimental-design question, we use the stage-2-accepted realizations to compute the $1 \times n_\text{par}$

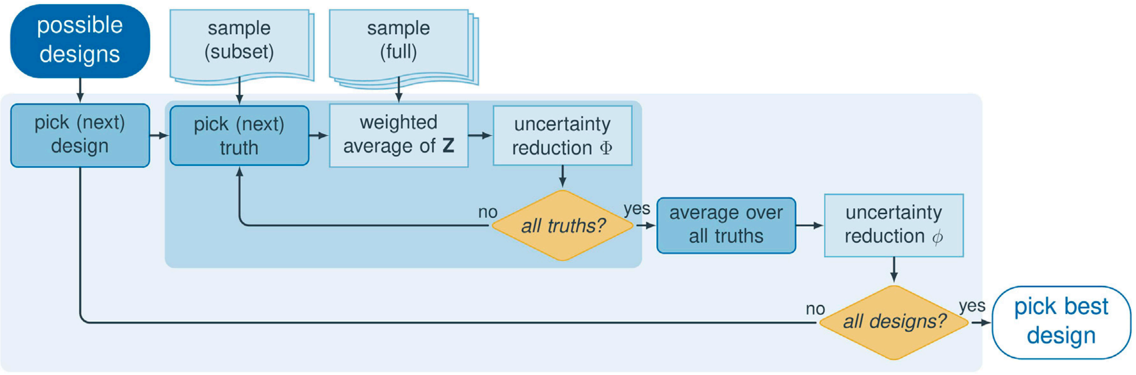

FIGURE 2 | Schematic illustration of the general preposterior data impact assessor procedure. Inner loop in dark blue, outer loop in light blue.

vector of prediction variables $\mathbf{z}$ (binary particle-fate map) as discussed above. The prediction solely depends on the input parameter vector $\mathbf{S}$ and is independent of the measurement design $\mathbf{d}$. In our particular application, the prediction variable is binary, namely whether a particle introduced into the subsurface at a given location belongs to one out of two catchments. The binary nature of $\mathbf{z}$ implies that the sample average of it equals the vector of probabilities that the individual elements of $\mathbf{z}$ are one.

After acquiring n_{sample} stage-2-accepted sample members of the parameter vector $\mathbf{S}$ and computing the associated virtual measurements and prediction variables, we have $n_{\text{sample}} \times n_{\text{des}}$ sets of $\mathbf{y}(\mathbf{d})$ and n_{sample} sets of $\mathbf{z}$ (which can be summarized in a $n_{\text{sample}} \times n_{\text{par}}$ matrix $\mathbf{Z}$). As illustrated in **Figure 2**, PreDIA proceeds in the following way to identify the best design:

(1) Compute the unconditional sample mean $P(z_i)$ of all target variables z_i by **Eq. 17** with equal probabilities of all realizations.

(2) Compute the vector of unconditional probabilities of misclassification $P_{\text{mc}}(z_i)$ by **Eq. 18** and the associated overall prior uncertainty of groundwater-divide delineation $U(\mathbf{z})$ by **Eq. 19**.

(3) Select a random subset of n_{sub} realizations used to define virtual truths. Its distribution of virtually measured values $\mathbf{y}$ should be similar to the corresponding distribution using the full sample (across all designs). When computationally feasible, select all n_{sample} sample members such that $n_{\text{sub}} = n_{\text{sample}}$.

(4) Loop over all designs $\mathbf{d}$:

   a. Loop over the n_{sub} realizations with index j:

      (1) Realization j with the virtual observations $\mathbf{y}_j(\mathbf{d})$ and the virtual prediction variable $\mathbf{z}_j$ is temporarily declared as truth.

      (2) Each realization $i \neq j$ of the full set of n_{sample} realizations is assigned a Bayesian weight depending on how close the respective observations $\mathbf{y}_i(\mathbf{d})$ are to $\mathbf{y}_j(\mathbf{d})$. The weights are computed by the likelihoods L_{ij} of observation $\mathbf{y}_i(\mathbf{d})$ using the observation $\mathbf{y}_j(\mathbf{d})$ as temporary truth:

$$w_{ij} = \frac{L_{ij}}{\sum_i L_{ij}} \tag{22}$$

$$L_{ij} = \begin{cases} \dfrac{1}{\sqrt{(2\pi)^{n_y}|\mathbf{R}_\varepsilon|}} \\ \exp\left(-\dfrac{1}{2}\left(\mathbf{y}_i(\mathbf{d}) - \mathbf{y}_j(\mathbf{d})\right)^T \mathbf{R}_\varepsilon^{-1}\left(\mathbf{y}_i(\mathbf{d}) - \mathbf{y}_j(\mathbf{d})\right)\right) & \text{if } i \neq j \\ 0 & \text{otherwise} \end{cases} \tag{23}$$

      in which n_y is the number of virtual measurements according to the current design $\mathbf{d}$, and $\mathbf{R}_\varepsilon$ is the $n_y \times n_y$ covariance matrix of measurement errors, here assumed to be a diagonal matrix, which implies that the measurement errors are uncorrelated.

      The weights are summarized in a $n_{\text{sample}} \times 1$ vector $\mathbf{w}_j$ of weights.

      (3) Compute the mean of all prediction variables in $\mathbf{Z}$, conditioned on the observations $\mathbf{y}_j(\mathbf{d})$ of the temporary true parameter set $\mathbf{S}_j$ according to the current design $\mathbf{d}$ by **Eq. 17** with the probability of realization i set to the weight w_{ij}:

$$P\left(\mathbf{z}\middle|\mathbf{y}_j(\mathbf{d})\right) = \mathbf{w}_j^T\mathbf{Z} = \sum_i^{n_{\text{sample}}} w_{ij}\mathbf{z}_i \tag{24}$$

      The $1 \times n_z$ vector $P(\mathbf{z}|\mathbf{y}_j(\mathbf{d}))$ is the vector of probabilities that the individual elements of $\mathbf{z}$ are one, conditioned on the vector of observations $\mathbf{y}_j(\mathbf{d})$ of realization j using the design $\mathbf{d}$.

      (4) Compute the conditional probability of misclassification $P_{\text{mc}}(\mathbf{z}|\mathbf{y}_j(\mathbf{d}))$ by substituting $P(\mathbf{z}|\mathbf{y}_j(\mathbf{d}))$ rather than the vector of unconditional probabilities $P(\mathbf{z})$, into **Eq. 18**.

      (5) From the vectors of conditional and unconditional probabilities of misclassification, $P_{\text{mc}}(\mathbf{z}|\mathbf{y}_j(\mathbf{d}))$ and $P_{\text{mc}}(\mathbf{z})$, respectively, compute a scalar metric $\Phi(\mathbf{y}_j(\mathbf{d}))$ summarizing the relative reduction of uncertainty U in classifying all elements of $\mathbf{z}$ by

considering the observations $\mathbf{y}_j(\mathbf{d})$ belonging to design $\mathbf{d}$:

$$\Phi\left(\mathbf{y}_j(\mathbf{d})\right) = 1 - \frac{U\left(\mathbf{z}\middle|\mathbf{y}_j(\mathbf{d})\right)}{U(\mathbf{z})} \tag{25}$$

by using **Eq. 19**. Steps 4. a (1) to 4. a (5) define the inner loop, illustrated by dark blue shading in **Figure 2**. In the inner loop, each of the n_{sub} virtual observations for the currently chosen design $\mathbf{d}$ are temporarily considered the truth. The inner loop results in n_{sub} objective-function values for a given design $\mathbf{d}$.

b. Marginalize the objective function over the n_{sub} realizations:

$$\phi(\mathbf{d}) = \frac{1}{n_{\mathrm{sub}}}\sum_{j=1}^{n_{\mathrm{sub}}}\Phi\left(\mathbf{y}_j(\mathbf{d})\right) \tag{26}$$

in which we have assumed that all "temporary truth" realizations j are equally likely. $\phi(\mathbf{d})$ is the utility function of design $\mathbf{d}$. Steps 4. a and 4. b define the outer loop over all designs $\mathbf{d} \in \mathbf{D}$, which is illustrated by light blue shading in **Figure 2**).

(5) Identify the design $\mathbf{d}_{\mathrm{opt}}$ maximizing $\phi(\mathbf{d})$ according to **Eq. 20**.

The two loops of PreDIA require large sample sizes to make reliable statements about design performances. To estimate whether the chosen sample is large enough for the results to be meaningful, one can use the averaged effective sample size AESS (Leube et al., 2012, adapted from; Liu, 2008). It is a measure of how many realizations actually contribute to the analysis, where low values indicate filter degeneracy, which needs to be mitigated by increasing the ensemble size.

PreDIA has fundamental advantages over other optimal-experimental-design techniques. It is applicable to inherently non-linear problems without the need of a linearization. It is also very versatile because it imposes few restrictions on the numerical model. Besides the definition and reading of some pre-run input and post-run output quantities, the actual numerical simulation code is independent of PreDIA. This independence makes it trivial to couple any numerical model with PreDIA. It can be seen as a post-processing routine for any modeling sample. PreDIA can capture all kinds of known or estimated uncertainties in boundary conditions, material properties, model structure, or any other model parameters due to its ensemble-based nature.

The disadvantage of PreDIA lies in its computational cost. The analysis requires large sample sizes (i.e., tens of thousands of model runs) and is computationally expensive itself. These difficulties, however, can be overcome with parallel computing techniques (i.e., running multiple realizations at the same time) and simplified models that are comparably quick.

2.5. Numerical Implementation

Our framework does not depend on the choice of any specific software, neither for the flow simulation nor for the optimal-design analysis. In the following application, we use HydroGeoSphere to solve for three-dimensional subsurface flow using standard finite elements on triangular prisms (Therrien et al., 2010; Brunner and Simmons, 2012). Because of the Richards equation's nonlinearity, we do not directly solve for steady-state flow. Instead, we use the transient solver of HydroGeoSphere with constant forcings over a simulation time of $3 \cdot 10^{12}\,\mathrm{s} \approx 100\,000$ years using adaptive discretization in time. It is reasonable to assume that steady state is achieved within this time.

The velocity field of HydroGeoSphere is transferred to Tecplot to perform advective particle tracking with Tecplot's streamtracing routine in its command line mode (Tecplot Inc., 2019).

The stochastic engine responsible for the sampling of the parameter space and performing the plausibility check of sample members by the Gaussian process emulator-based surrogate model is written in Matlab (The MathWorks Inc., 2019) and based on the code of Erdal and Cirpka (2019). We execute the stochastic sampler on a mid-size high-performance computing cluster with 24 Intel Xeon L5530 nodes (8 cores per node; 2.4 GHz and 8 MB per chip).

The optimal design analysis using PreDIA is implemented as a separate Matlab code that acts on the full sample of stage-2-accepted realizations after its acquisition.

3. APPLICATION TO A FIELD SITE

3.1. Description of the Study Site

We apply the presented framework to delineate the groundwater divide between the Ammer and Neckar catchments north and south of the Wurmlingen Saddle, respectively, close to Tübingen in South-West Germany. **Figure 3** shows a map of the area outlining the model domain (solid black line), the surface-water divide (dashed black line), and streams/drainage features (blue lines). The area of interest comprises a floodplain in the Ammer catchment, which is part of ongoing hydrogeologic and geophysical research (e.g., Martin et al., 2020). Previous modeling studies suggested a shift of the groundwater divide in this area toward the Ammer catchment in the north (Kortunov, 2018). This hypothesis was supported by the Neckar valley being about 10 m lower than the Ammer valley and dipping of the strata toward the south. However, no piezometers currently exist along the decisive hillslope so that the hypothesis of a shifted groundwater divide is fairly uncertain. Delineating the groundwater divide with higher certainty would help to determine the Ammer floodplain's water budget more accurately.

In order to test the hypothesis of a shifted groundwater divide, installing up to three piezometers is planned. Due to legal and logistical reasons, all new groundwater observation points need to be placed on a transect parallel to the street from Unterjesingen to Wurmlingen (see **Figure 3**). We use the presented method to determine the best configuration of piezometers along this transect.

The model domain contains parts of both the Ammer and Neckar catchments, so that the groundwater divide emerges from the model instead of being set as a boundary condition. The surface elevation ranges from approximately 330 m to 475 m

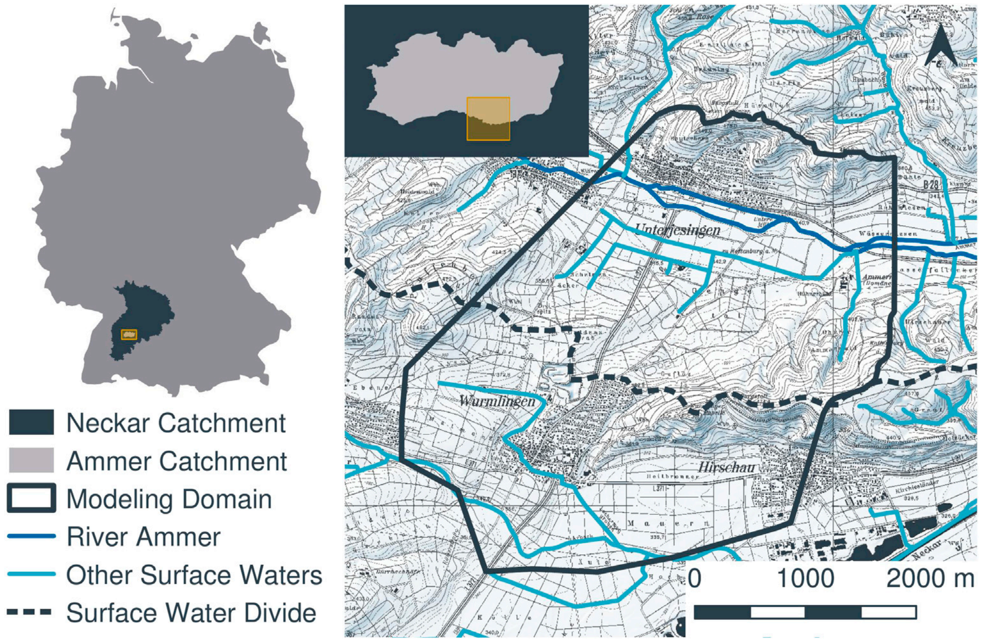

FIGURE 3 | Two-dimensional overview of the model domain and its location in Germany.

above sea level. In the East of the model domain, the surface-water divide is on a ridge (“Spitzberg”) formed by a sequence of mud- and sandstones that most likely does not allow groundwater recharge to the main aquifer. Likewise, in the West, the surface-water divide is on a plateau (“Pfaffenberg”). In the center of the model domain, by contrast, the topographic surface-water divide is a saddle with gentle slopes both toward the north and south.

The model domain includes the floodplain of the Ammer river with the river itself and a network of artificial drainage channels. The drainage features running south-north on the hillslope are typically dry, unless during storm events. On the southern side, the model domain includes parts of the Neckar floodplain, but does not reach River Neckar. The only surface water on this side of the hills is a small creek (“Arbach”). However, a dense network of observation wells in the Neckar valley allowed us to define a fixed-head boundary condition along the southern boundary of the model domain.

The bedrock geology in the area is governed by sequences of sandstones and mudstones belonging to the Upper Triassic Keuper formation (Aigner and Bachmann, 1992). The regional geology has been subject to many (hydro-)geological investigations (e.g., Kekeisen, 1913; Harreß, 1973; D’Affonseca et al., 2020). The Ammer and Neckar rivers have carved small basins into the bedrock (Martin et al., 2020), which are filled with Quaternary sediments forming the floodplains. In total, we distinguish twelve hydrostratigraphic units, which we briefly characterize in the following from bottom to top:

(1) lower *Erfurt formation* (kuE): The kuE unit is roughly 20 m thick. Being made of thin layers of mudstones and dolostones, it acts as an aquitard, separating the shallow groundwater system from the underlying middle Triassic Muschelkalk formation, a regional karstified aquifer (D’Affonseca et al., 2020).
(2) upper kuE: We divide the kuE into two subunits of similar thickness to account for its heterogeneity in hydraulic conductivity.
(3) unweathered *Grabfeld formation* (kmGr): The kmGr is a mudstone unit bearing gypsum, anhydrite, mudstones, and shales. It can reach thicknesses of up to 100 m (Schmidt et al., 2005). Its hydraulic properties vary strongly depending on its degree of weathering. The unweathered, anhydrite-bearing kmGr is considered tight but may be fractured to allow some water circulation.
(4) weathered kmGr: Water contact has transformed anhydrite to gypsum within the kmGr. Upon further weathering, the gypsum dissolves (Ufrecht, 2017), which can increase the hydraulic conductivity by orders of magnitude (Kirchholtes and Ufrecht, 2015). Due to the strong contrast in hydraulic conductivity, we divide kmGr into the unweathered and weathered rock.
(5) mud- and sandstone formations (km2345): We lump the remaining bedrock formations *Stuttgart formation* (kmSt), *Steigerwald/Hassberge/Mainhardt formation* (kmSw/kmHb/kmMh), *Löwenstein formation* (kmLw), and *Trossingen formation* (kmTr), which are made of interbedded mudstones, silty mudrocks, dolomite layers, sandstones, and clay conglomerates, to a single unit with uniform hydraulic properties. These strata occur only at the outskirts of our model domain where they cover the kmGr.
(6) hillslope-hollow fillings: hillslope hollows on the southern hillslopes of the Ammer valley are cut into the kmGr. They are partially filled with poorly sorted sediments deposited by mudflows.

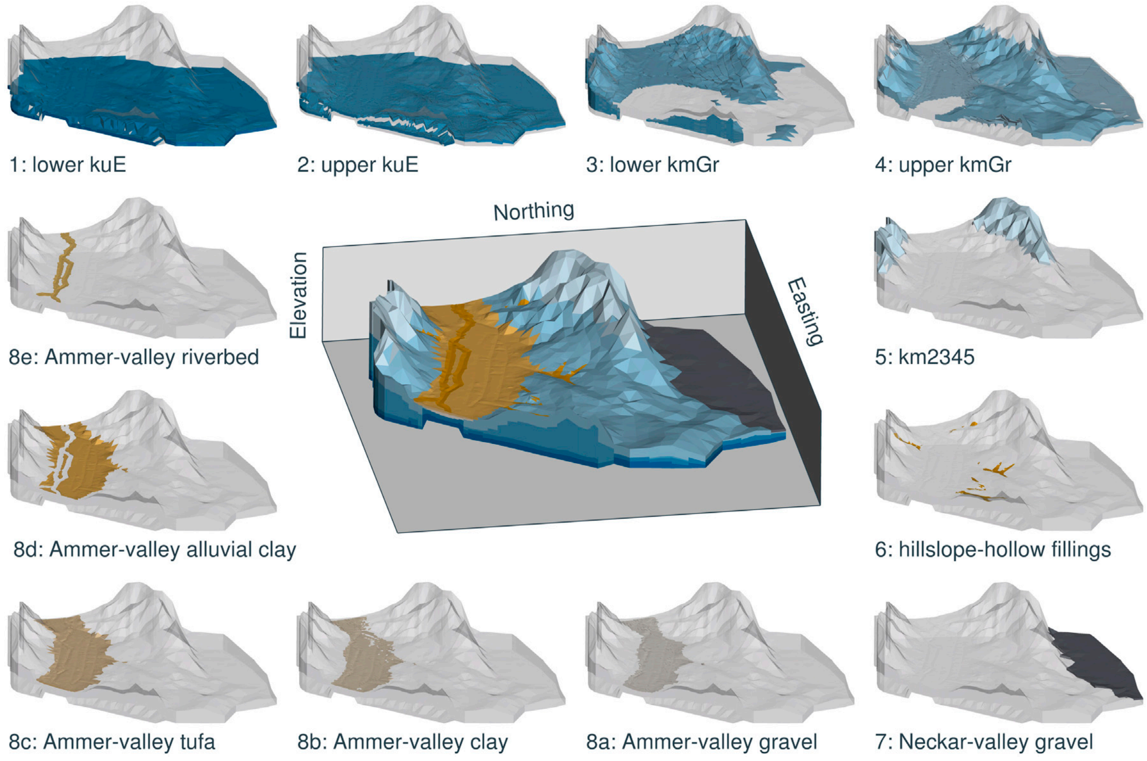

FIGURE 4 | Three-dimensional overview of the subsurface-flow model. The vertical dimension is exaggerated by a factor of five to enhance distinctness.

(7) Neckar-valley gravel: The floodplain material on the Neckar side mostly consists of Quaternary sandy gravel sediments of several meter thickness.

(8) Ammer-valley Quaternary: The Ammer floodplain comprises five distinct layers (Martin et al., 2020):

a. Ammer-valley gravel: The lowest floodplain unit in the Ammer valley consists of a Pleistocene clayey gravel body, acting as a local aquifer. Its thickness is in the range of 5 m to 10 m.

b. Ammer-valley clay: A clay unit of approximately 2 m to 3 m thickness forms an aquitard between the two floodplain aquifers.

c. Ammer-valley tufa: This Holocene unit consists mostly of autochthonous limestone aggregates. It has a thickness of several meters. Slug tests conducted by Martin et al. (2020) identified this layer as an aquifer.

d. Ammer-valley alluvial clay: The top of the Quaternary filling of the Ammer floodplain is a several meter thick colluvium of silty and clayey fines.

e. riverbed of the Ammer river: Underneath River Ammer, a layer of recent river sediments with different grain size than the surrounding sediments can be found. This layer could have an increased hydraulic conductivity, due to consisting of coarse sediments deposited by the river. However, it is also possible that this layer has a reduced conductivity due to colmation of clayey deposits.

Figure 4 illustrates the considered hydrostratigraphic units in three-dimensional renderings.

3.2. Details of the Subsurface-Flow Model

3.2.1. Discretization

Figure 5 shows a plan view of the model discretization and boundary conditions. The model domain covers an area of approximately 13 km^2. We discretize the two-dimensional area by 3,959 triangles arranged in a conforming unstructured grid. These triangles are extruded in the vertical dimension to generate triangular prisms. Using 35 prism layers from the bottom of the lower kuE formation to the surface elevation results in a grid of 138,565 three-dimensional elements with 74,412 nodes. The number of layers is constant throughout the domain, whereas the layer thicknesses vary. The topmost layers of the domain are discretized more finely, in order to better resolve the unsaturated zone. The chosen mesh is a compromise between numerical accuracy and computational effort. A comparison between models set up on this grid with models defined on an eightfold refined version revealed some deviations at the coarser parts (mostly on the Neckar side and in the deeper subsurface of the domain). However, we deem these acceptable because they occur where the exact hydraulic heads are of little interest to us anyway and because they are minor compared to the variance between different model realizations. For future applications we suggest to perform a grid convergence analysis with a range of different discretizations. The coarsest grid providing adequate accuracy should be selected.

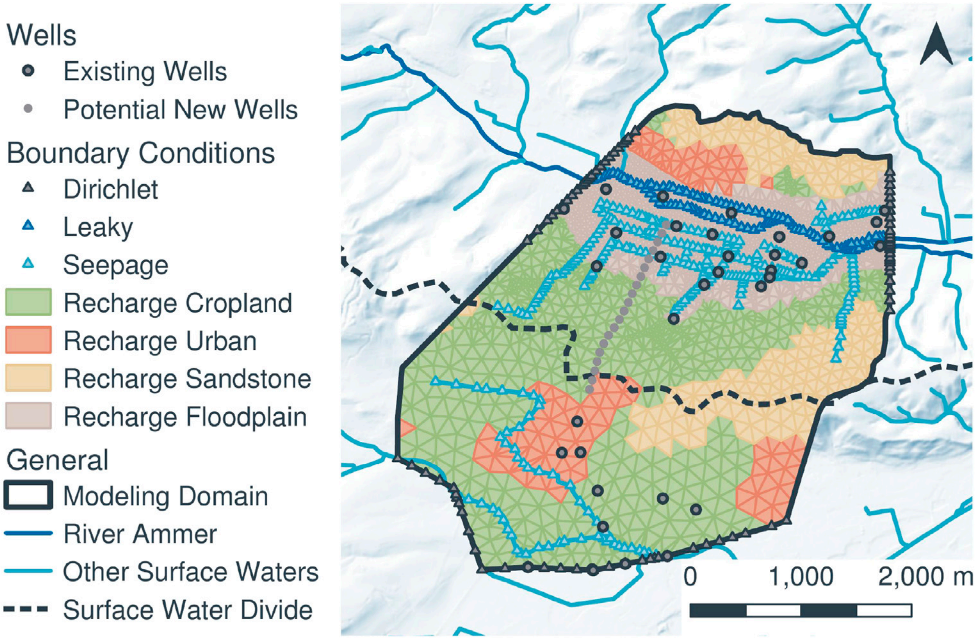

FIGURE 5 | Boundary conditions and two-dimensional discretization of the underlying subsurface flow model.

3.2.2. Boundary Conditions

Along different parts of the boundary, we apply different boundary conditions:

(1) If not specified otherwise, all outer mesh faces are assigned a no-flux (Neumann) boundary condition. These boundaries are either in formations of very low conductivity (particularly the bottom) or the boundaries are far away from the area of interest like the northern boundary, which is derived from a secondary surface water divide on the far side of the Ammer valley. The eastern and western boundaries are approximately parallel to the estimated flow field.

(2) Three fixed-head boundary sections are defined at the western, eastern, and southern sides of the domain to allow regional groundwater flow (see **Figure 5**). To obtain the fixed-head values, we interpolate between observation well data. In the Ammer valley, the Dirichlet boundaries extend over the Quaternary fillings, while on the Neckar side, they extend over the whole depth of the model, where the formation consists of a thin, highly conductive gravel that ends at the municipality of Wurmlingen. Because of the high hydraulic conductivity and the absence of significant vertical hydraulic gradients here, we average the interpolated head values over depth for the Dirichlet assignment.

(3) On the top surface of the domain, we apply recharge as a fixed-flux (Neumann) boundary condition across element faces. Recharge rates in different zones depend on land use (cropland, floodplain, urban areas, and km2345-covered parts). By providing recharge as a model boundary we lump the dynamic interaction of evaporation, transpiration, precipitation and soil water storage into a single stationary quantity, which is of course a simplification. However, since we are interested in the effective, long-term behavior and not the high-resolution fluctuations, we consider this simplification justified. We base our range of possible recharge rates on previous work conducted in our domain or in comparable aquifers in close proximity (Holzwarth, 1980; Wegehenkel and Selg, 2002; Selle et al., 2013).

(4) We use a leaky boundary condition to simulate the interaction between groundwater and the Ammer river.

(5) For the network of drainage ditches in the Ammer valley and the small surface water creek in the Neckar valley, we apply seepage boundaries.

(6) Drainage boundary conditions are applied to all other surface nodes, allowing water to drain whenever the groundwater table is above the ground surface. We distinguish between elements that belong to the Ammer floodplain (highlighted in light brown in **Figure 5**) and the remaining surface.

Note that there are no groundwater abstractions within the model domain so that we do not need to consider corresponding internal boundary conditions.

We tested different initial conditions for the flow solution (e.g., a hydraulic head field interpolated from measurements, hydraulic heads equaling the surface elevation, a constant depth to the water table). The choice of initial condition affected mostly the run time needed to reach convergence to steady-state, but influenced the steady-state flow field itself only marginally. We settled with initial hydraulic heads equal to the surface elevation. For other applications, we recommend a similar comparison procedure to identify a useful initial condition. Choices that are too far away from a realistic flow field (e.g., a completely dry domain) can lead to convergence problems due to the nonlinearity of Richards' equation.

3.2.3. Uncertain Parameters and Prior Information

Each discretized spatial element (i.e., triangular prism) has a set of parameters defining the hydraulic properties of its material. All elements belonging to the same hydrostratigraphic unit share the same set of parameters, including the horizontal and vertical hydraulic conductivities K_h and K_v, respectively, the van-Genuchten parameters α and N and the residual water saturation $S_\mathrm{wr} = \Theta_\mathrm{r}/\Theta_\mathrm{s}$. For the transient calculations, we also need storage-related parameters (i.e., porosity or specific storativity), but they do not affect the final steady-state solution.

Table 1 summarizes all material properties considered random. These parameters are the first part of the parameter set **S**, sampled by the stochastic engine. Prior to the pre-selection/conditioning, we assume a uniform distribution of each parameter between a minimum and a maximum value. These distributions reflect unbiased estimates within a range of plausibility based on hydrogeological knowledge about the formations and other uncertain expert knowledge.

The values in **Table 1** are grouped by horizontal saturated hydraulic-conductivity values K_h, anisotropy ratios $K_\mathrm{v}/K_\mathrm{h}$, and the van-Genuchten parameters α and N. The indices represent the hydrostratigraphic unit using the numbering scheme of **section 3.1**. In total, we consider 30 variable material properties (named #P1 to #P30), which is less than the number of units times the number of hydraulic properties ($12 \times 4 = 48$) because we chose some parameters to be identical in several geological units. The hydrostratigraphic units 1 to 6 share the same van-Genuchten properties, and the units 7 and 8a do not require these unsaturated properties because the gravel aquifers of the Neckar and Ammer valleys are always fully water saturated.

We do not treat the residual water saturations as random variables. Instead, we apply the following values in all model runs: $S_{\mathrm{wr},1-8} = 5\,\%$, $S_{\mathrm{wr},9} = 17\,\%$, $S_{\mathrm{wr},10} = 18\,\%$, $S_{\mathrm{wr},11-12} = 25\,\%$.

In total, we use nine random parameters (#B1 to #B9) related to boundary conditions, listed in **Table 2**. We again assume uniform priors within given bounds. Parameters #B1 to #B4 regulate the groundwater recharge R [$\mathrm{m\,s^{-1}}$] on the four types of land use. Here we take the random recharge rate R_cropland on undisturbed cropland as reference, which is reduced by random factors for the other land-use types (floodplain material, areas covered by mud-/sandstone, urban areas).

The parameters #B5 to #B8 modify the fixed-head values at Dirichlet and river boundaries. The base values for the fixed heads used on the southern boundary in the Neckar valley (h_Neckar) and the stage of River Ammer (h_Ammer) vary in space. In the stochastic setup, we consider random constant shifts of Δh_Neckar and Δh_Ammer to all nodes belonging to the respective boundaries. The fixed-head values on the groundwater in- and outflow faces in the Ammer floodplain are spatially constant but uncertain, so that the stochastic model directly treats these values, $h_\mathrm{Ammer,in}$ and $h_\mathrm{Ammer,out}$, as random variables. We have chosen the ranges of these values from time series of hydraulic head measured in existing piezometers close to the boundaries.

At last, #B9 represents the uncertain thickness of the drainage boundary in **Eq. 13** for all floodplain elements. The respective hydraulic conductivity is $K_\mathrm{8d,h}$. For the drainage boundaries outside of the floodplain, we assume a soil layer of 0.20 m thickness and a hydraulic conductivity of $1 \cdot 10^{-6}\,\mathrm{m\,s^{-1}}$. The river boundary condition (see **Eq. 12**) uses K_8e for its conductivity and the geometry parameters $L_\mathrm{riv} = 40\,\mathrm{m}$, $w_\mathrm{riv} = 3\,\mathrm{m}$, and $L_\mathrm{sed} = 0.5\,\mathrm{m}$.

Finally, we consider a total of five random parameters (#S1 to #S5) describing uncertain geometry of structural units. **Table 3** lists the ranges of the parameters. #S1 controls the maximum depth L_4 of the weathered kmGr formation (hydrostratigraphic unit 4): Wherever kmGr is the outcropping geological formation, the top layer with thickness L_4 is considered weathered, that is attributed to the hydrostratigraphic unit 4. The parameters #S2 and #S3 describe the three-dimensional extent of the hillslope-hollows. #S2 controls the lateral extent of the hollows by expanding or contracting their width by a constant factor. #S3 defines the bottom slope of the hollows, which thereby also controls their maximum depth. The total volume of the hydrostratigraphic unit 6 depends on both #S2 and #S3. The final two parameters #S4 and #S5 are converted to binary flags, deciding whether the hillslope hollows (#S4) and explicit river beds (#5) are considered at all. Negative values of #S4 and #5 indicate that the respective features are not considered, whereas positive values lead to realizations including these features. We have introduced these switches because the existence and hydraulic relevance of these hydrogeological elements is uncertain at the real field site. A full parameter set **S** is the concatenation of all #P, #B and #S values.

3.3. Plausibility Criteria for Model Pre-Selection

We define seven criteria to decide whether the flow solution of a model realization is plausible (i.e., stage-2-accepted). These criteria are listed in the following:

(1) To keep the realizations close to data observed in the field, the simulated hydraulic heads are compared to real head measurements obtained in the valleys (see **section 3.4**). As the model assumes steady-state flow, we time-average the available series of measured heads at 51 observation wells and compute the root mean square error (RMSE) of the corresponding simulated steady-state heads. For a model realization to be stage-2-accepted, its RMSE has to be smaller than 1.5 m. This reflects the order of magnitude of the measured annual fluctuations in hydraulic head, which are in the range of 0.5 m–2 m.

(2) The total groundwater flux Q_in crossing the fixed-head boundary at the western inflow end of the Ammer-floodplain aquifers must be positive.

(3) The total groundwater flux Q_out crossing the fixed-head boundary at the eastern outflow end of the Ammer-floodplain aquifers must be negative.

(4) The magnitude of the two fluxes, Q_in and Q_out, must be similar. It is unclear which of the boundaries exhibits the larger groundwater discharge at the field site. Both scenarios (increase of discharge from in-to outflow due to recharge and input from the hillslopes or decrease of discharge due to

TABLE 1 | Prior parameter ranges of random material properties of hydrostratigraphic units considered in the model.

ID	Name	Minimum	Maximum	Unit	Comment
#P1	$\log_{10}K_{1,\mathrm{h}}$	−8.0	−6.0	$\mathrm{m\,s^{-1}}$	–
#P2	$K_{2,\mathrm{h}}$	$1/250 \cdot K_{1,\mathrm{h}}$	$1/2 \cdot K_{1,\mathrm{h}}$	$\mathrm{m\,s^{-1}}$	–
#P3	$\log_{10}K_{3,\mathrm{h}}$	−9.0	−6.3	$\mathrm{m\,s^{-1}}$	–
#P4	$K_{4,\mathrm{h}}$	$K_{3,\mathrm{h}}$	$10^3 \cdot K_{3,\mathrm{h}}$	$\mathrm{m\,s^{-1}}$	–
#P5	$\log_{10}K_{5,\mathrm{h}}$	−8.3	−7.0	$\mathrm{m\,s^{-1}}$	–
#P6	$\log_{10}K_{6,\mathrm{h}}$	−9.0	−3.0	$\mathrm{m\,s^{-1}}$	–
#P7	$\log_{10}K_{7,\mathrm{h}}$	−5.3	−3.0	$\mathrm{m\,s^{-1}}$	–
#P8	$\log_{10}K_{8a,\mathrm{h}}$	−5.3	−3.0	$\mathrm{m\,s^{-1}}$	–
#P9	$\log_{10}K_{8b,\mathrm{h}}$	−10.0	−7.0	$\mathrm{m\,s^{-1}}$	–
#P10	$\log_{10}K_{8c,\mathrm{h}}$	−5.3	−3.0	$\mathrm{m\,s^{-1}}$	–
#P11	$\log_{10}K_{8d,\mathrm{h}}$	−9.0	−5.3	$\mathrm{m\,s^{-1}}$	–
#P12	$\log_{10}K_{8e}$	−8.0	−3.0	$\mathrm{m\,s^{-1}}$	–
#P13	$K_{1,\mathrm{v}}/K_{1,\mathrm{h}}$	1/15	1	–	–
–	$K_{2,\mathrm{v}}/K_{2,\mathrm{h}}$	1/15	1	–	Coupled to #P13
#P14	$K_{3,\mathrm{v}}/K_{3,\mathrm{h}}$	1/15	1	–	–
#P15	$K_{4,\mathrm{v}}/K_{4,\mathrm{h}}$	1/15	1	–	–
#P16	$K_{5,\mathrm{v}}/K_{5,\mathrm{h}}$	1/15	1	–	–
#P17	$K_{6,\mathrm{v}}/K_{6,\mathrm{h}}$	1/5	1	–	–
#P18	$K_{7,\mathrm{v}}/K_{7,\mathrm{h}}$	1/5	1	–	–
#P19	$K_{8a,\mathrm{v}}/K_{8a,\mathrm{h}}$	1/5	1	–	–
#P20	$K_{8b,\mathrm{v}}/K_{8b,\mathrm{h}}$	1/15	1	–	–
#P21	$K_{8c,\mathrm{v}}/K_{8c,\mathrm{h}}$	1/15	1	–	–
#P22	$K_{8d,\mathrm{v}}/K_{8d,\mathrm{h}}$	1/15	1	–	–
#P23	α_{1-6}	0.50	5.00	$\mathrm{m^{-1}}$	–
#P24	α_{8b}	0.01	0.10	$\mathrm{m^{-1}}$	–
#P25	α_{8c}	8.00	12.00	$\mathrm{m^{-1}}$	–
#P26	α_{8d}	0.50	0.70	$\mathrm{m^{-1}}$	–
#P27	$N_{1\text{-}6}$	1.50	6.00	–	–
#P28	N_{8b}	1.40	1.70	–	–
#P29	N_{8c}	1.80	2.20	–	–
#P30	N_{8d}	1.50	2.10	–	–

TABLE 2 | Prior ranges of parameters describing boundary conditions of the model.

ID	Name	Minimum	Maximum	Unit	Comment
#B1	R_{cropland}	$1.5 \cdot 10^{-9}$	$8.0 \cdot 10^{-9}$	$\mathrm{m\,s^{-1}}$	–
#B2	$R_{\mathrm{floodplain}}/R_{\mathrm{cropland}}$	0	1	–	Coupled to #B1
#B3	$R_{\mathrm{mud/sandstone}}/R_{\mathrm{cropland}}$	0	1	–	Coupled to #B1
#B4	$R_{\mathrm{urban}}/R_{\mathrm{cropland}}$	0.25	1	–	Coupled to #B1
#B5	$\Delta h_{\mathrm{Neckar}}$	−0.50	0.50	m	–
#B6	$\Delta h_{\mathrm{river}}$	−0.25	0.25	m	–
#B7	$h_{\mathrm{Ammer,in}}$	346.0	347.0	m	–
#B8	$h_{\mathrm{Ammer,out}} - h_{\mathrm{Ammer,in}}$	−8.6	−7.6	m	Coupled to #B7
#B9	L_{8d}	0.10	1.50	m	–

TABLE 3 | Prior ranges of structural parameters.

ID	Name	Minimum	Maximum	Unit	Comment
#S1	L_4	0	50	m	–
#S2	Size factor hollows	0.5	1.5	–	–
#S3	Bottom slope hollows	0.0	0.7	%	–
#S4	Switch hollows	−0.5	0.5	–	No hollows if <0
#S5	Switch riverbed	−0.5	0.5	–	No riverbed if <0

drainage into the rivers and channels) are possible. Therefore, we only evaluate the ratio γ of the absolute flux difference over the mean flux:

$$\gamma = 2\,\frac{\left||Q_{\mathrm{in}}| - |Q_{\mathrm{out}}|\right|}{\left(|Q_{\mathrm{in}}| + |Q_{\mathrm{out}}|\right)} \tag{27}$$

This ratio can take values between $\gamma = 0$ (both fluxes are identical) and $\gamma = 2$ (one flux is zero). For a stage-2-accepted model realization, we require $\gamma \leq 1$, which is equivalent to requiring $\frac{1}{3} \leq \frac{|Q_{\mathrm{in}}|}{|Q_{\mathrm{out}}|} \leq 3$.

(5) The sum of all exchange fluxes between the subsurface and rivers must be negative (i.e., net groundwater discharge into rivers). Field data on the exchange fluxes are difficult to obtain because the change of river discharge due to surface-water/groundwater exchange is very small along the investigated stretch. Nonetheless we expect that the rivers are net gaining as there are no groundwater abstractions within the domain. Losing conditions might occur only locally on short stretches of the rivers and channels.

(6) A typical behavior shown in many models with randomly drawn parameters is extensive flooding of the model domain. At the real floodplain, by contrast, we do not observe permanent flooding outside of ditches. To exclude flooding of the floodplain under steady-state flow conditions, we require that the total flux across all drainage nodes is small (see **section 3.2.2**). As plausibility we set that the total flux leaving at the surface must be smaller than 10 % of the total flux produced by the recharge boundaries.

(7) Finally, the water flux leaving at the drainage ditches should not be excessive. In the real floodplain, these ditches carry water only seasonally and in small quantities. Since the actual fluxes are unknown and hard to estimate, we require a stage-2-accepted realization to drain less than 50 % of the recharged water through the ditches.

3.4. Tested Experimental Designs

Currently, there are 35 piezometers already installed at the field site, for which a decent-quality dataset of hydraulic head in one or multiple depths is available. **Figure 5** shows the location of these observation wells by gray circular dots with black edges. Accounting for different depths in multi-level wells, hydraulic heads are measured at 51 points. However, there are no piezometers located on the hillslope between the two valleys. This lack of observation points results in high uncertainty regarding groundwater flow underneath the hillslope and in the location of the groundwater divide.

In order to fill this gap, the installation of up to three additional piezometers is planned on a transect. We identified twenty potential piezometer locations along this transect, coinciding with edges of the computational grid. These locations are marked in **Figure 5** as gray circular dots without an edge. The line of points extends longer on the North than the South, because we expect the divide to be shifted toward the North. This is so, because the northern valley is at a higher elevation than the southern valley, and also the geological units dip toward the

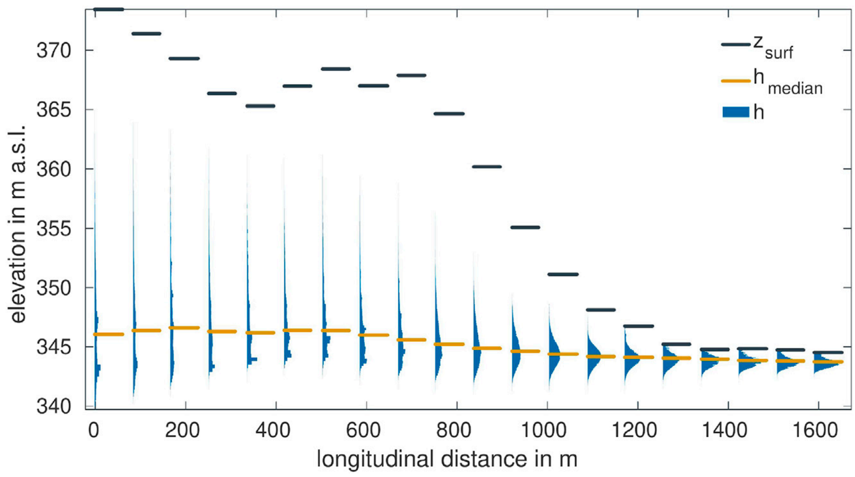

FIGURE 6 | Distributions of virtual hydraulic-head observations using the sample of stage-2-accepted realizations at all twenty potential locations along the transect.

south-west. Furthermore, a preliminary study conducted by Kortunov (2018) also suggested a shift in this direction.

The optimal experimental design analysis considers designs consisting of one, two, or three new wells, each placed on one of the twenty locations. Our design space **D** consists of all possible combinations. The total number of possible designs n_{des} for 1, 2, and 3 locations out of a set of n_{pts} can be evaluated by:

$$n_{\mathrm{des}} = n_{\mathrm{pts}} + \frac{1}{2}n_{\mathrm{pts}}\left(n_{\mathrm{pts}} - 1\right) + \frac{1}{6}n_{\mathrm{pts}}\left(n_{\mathrm{pts}} - 1\right)\left(n_{\mathrm{pts}} - 2\right), \qquad (28)$$

in which n_{pts} is the number of potential observation points. With $n_{\mathrm{pts}} = 20$, **Eq. 28** results in a total of $n_{\mathrm{des}} = 20 + 190 + 1140 = 1350$ individual designs, out of which we need to identify the best one.

While the optimal three-well design will obviously outperform the optimal two- and one-well designs, we want to investigate which information gain (e.g., reduction in uncertainty of delineating the groundwater divide) is achieved by installing more or fewer wells. However, we do not perform a full cost-benefit analysis, as the (financial) costs are difficult to compare to the benefit of reducing the uncertainty in the groundwater-divide delineation.

4. RESULTS AND DISCUSSION

Of 72,481 stage-1-accepted realizations, 20,600 needed to be rejected, because they yielded implausible results according to the given criteria. Another 1881 model runs were rejected, because they did not converge within 40 min of wall-clock time, set as limit to use the available computational resources efficiently. The remaining sample consists of $n_{\mathrm{sample}} = 50\,000$ accepted realizations. Among the successful realizations, the model run times roughly followed a log-normal distribution with a mean of 20.7 min, a median of 19.5 min, and a standard deviation of 6.9 min (not shown here). Due to parallelization of up to 57 simultaneous model runs, the total wall-clock time for all realizations was approximately three weeks. For computational speed-up, we only used $n_{\mathrm{sub}} = 10\,000$ realizations as virtual truths for the optimal design analysis. We checked the validity of this subset size by comparing the average binary fate maps of the whole sample and the subset. There were no significant deviations.

4.1. Uncertainty and Sensitivity of Head Observations to Parameters

Figure 6 shows the distributions of the simulated groundwater-table measurements at the twenty proposed locations. Each profile relates to one suggested observation-well location and includes, 1) a histogram of simulated head values of all 50,000 accepted sample members, 2) the median of the simulated head (h_{median}, yellow-brown dash markers), and 3) the position of the land surface (z_{surf}, black dash markers). The longitudinal distance is evaluated along the line connecting the proposed locations from south to north (i.e., the index zero corresponds to the first, southernmost investigated point).

At the southern end of the transect, which is close to the surface-water divide, the statistical distributions of the groundwater table are very wide, whereas at the northern end in the Ammer floodplain they become quite narrow. This behavior can be explained with the plausibility constraints put onto the model selection. As **Figure 5** shows, most existing observation wells are within the Ammer floodplain, restricting the variability of hydraulic heads by plausibility criterion 1. Also plausibility criterion 6, excluding realizations showing extended flooding, contributes to narrowing the variability of hydraulic heads within the floodplain. By contrast, there are no piezometers to constrain the models along the southern hillslope. Observation wells further away from the hydraulic-head-constraining floodplain show larger uncertainty than those close by, which reflects the uncertainty in groundwater recharge and transmissivity of the weathered part of the Grabfeld formation

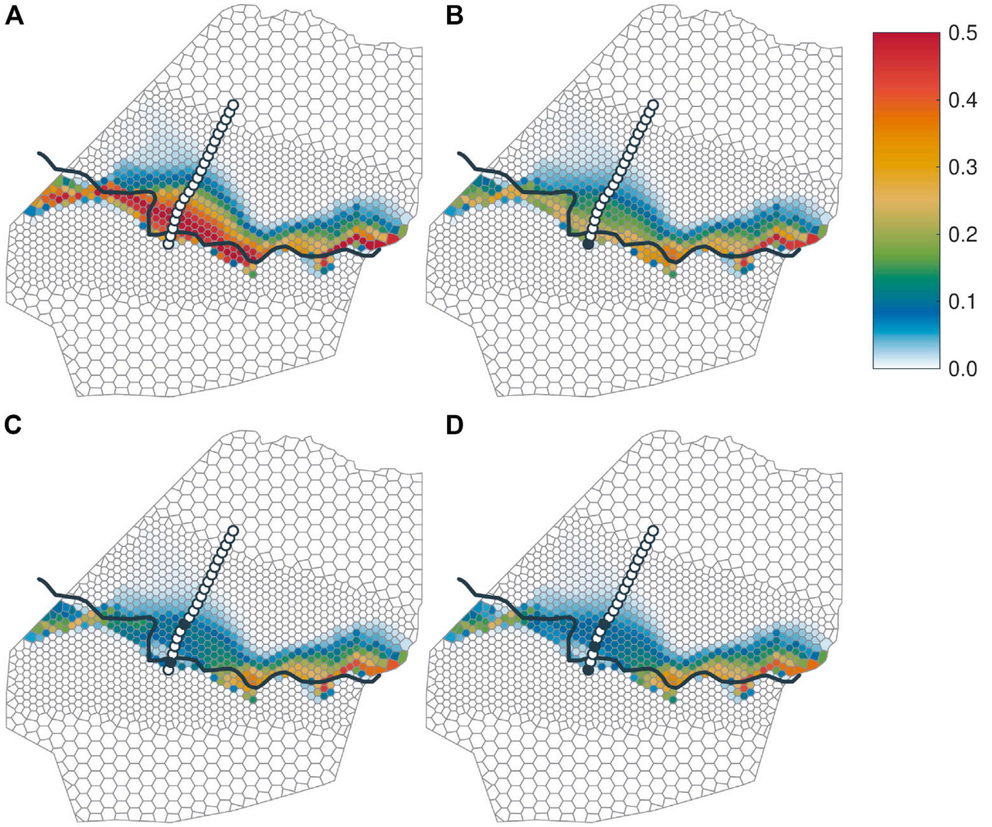

FIGURE 7 | Probability maps of misclassifying the attribution to the Ammer and Neckar subsurface-catchments $P_{mc} = 2P \cdot (1 - P)$. **(A)** Prior P_{mc}; **(B)**: P_{mc} for the best design with one additional piezometer; **(C)**: with two additional piezometers; **(D)**: with three additional piezometers.

kmGr. The conditioning by the pre-selection procedure might also explain why the shape of the head histograms in **Figure 6** transforms from near-Gaussian for the northern wells to multi-modal wide distributions toward the southern end.

As indicated by the black dashes in **Figure 6**, the topography along the transect is not strictly monotonic. At about one quarter along the length of the profile, a hillslope hollow oriented in the WSW-ENE direction crosses the transect. Along the transect, the median of the simulated hydraulic head follows the topography to some extent, but with a much smaller range. At the southern end, the median profile of hydraulic head drops toward the south along a distance of 200 m, whereas the surface elevation profile increases. The median groundwater table dipping toward the south of the transect might indicate that the groundwater divide is shifted toward the north, as hypothesized by Kortunov (2018). However, not all individual realizations show the same trend as the median, indicating that the general statement of Kortunov (2018) may be uncertain. This is why we performed the ensemble-based particle-tracking analysis to evaluate the location of the groundwater divide and its uncertainty in the following section.

To gain insights in how the head observations depend on the input parameters, we performed a global sensitivity analysis using the framework developed by Erdal et al. (2020) applying the method of active subspaces (Constantine et al., 2014; Constantine and Diaz, 2017) supported by a Gaussian process emulation of the target quantity. The active-subspace method results in activity scores, expressing the relative importance of all input parameters for a selected target variable. We performed this analysis for the simulated hydraulic-heads at the 20 potential locations for the new piezometers along the transect. At the 14 southern-most locations, which are all located along the hillslope in the weathered Grabfeld formation, the activity scores were the highest for the conductivities in the unweathered and weathered Grabfeld formation, the thickness of the weathering layer, and the recharge rate of cropland. At the six northern-most locations, located closer to/within the floodplain, we saw a shift toward conductivities of floodplain sediments and recharge in the floodplain. Similar observations on global sensitivity patterns have been made by Erdal and Cirpka (2019) in a study on a neighboring catchment with similar geology.

4.2. Maps of Misclassification Probability

Figure 7 shows maps of the misclassification probability P_{mc} according to **Eq. 18**. It quantifies how likely it is that any point on the map is considered part of one subsurface catchment while belonging in reality to the other one. The 1,526 polygons were constructed by Voronoi tesselation based on the set of starting points for particle tracking. The resolution is higher in a stripe within a few hundred meters north and south of the surface water divide (shown as a black line) because we suspect the groundwater divide to be within this area. The colors of the polygons reflect the misclassification probability P_{mc} of a particle released in the center of the polygon. As explained in **section 2.3**, P_{mc} ranges between zero and 0.5 (wrong attribution in half of the cases).

Figure 7A shows the map prior to installing any new piezometers. The highest values of the misclassification probability occur close to the surface-water divide. On the Neckar (southern) side of the surface-water divide, the misclassification probability drops rapidly. Here, all model realizations agree that these points belong to the Neckar subsurface catchment. On the Ammer (northern) side of the surface-water divide, by contrast, the misclassification probability decreases gradually, overall resulting in an uncertainty belt of the groundwater divide with a width ranging between 100 m and 800 m. This confirms the hypothesis of Kortunov (2018) that the groundwater divide might be shifted in this direction. At the foot of the hillslope within the Ammer valley, the misclassification probability is again practically zero, because these points belong to the Ammer subsurface catchment in almost all stage-2-accepted model realizations.

The width of the identified uncertainty belt is comparably small at the steeper hillslopes toward the east and at the very western end, where the topmost geological layer is the low conductive km2345 (see **Figure 4**, layer 5). In contrast to that, the width is large on the gentle saddle in the western and middle parts of the domain, where the top subsurface-layer consists of weathered kmGr, which has a higher hydraulic conductivity. This observation agrees with the findings of Haitjema and Mitchell-Bruker (2005), stating that groundwater and surface water divides are more likely to differ in aquifers with high transmissivities (for a given recharge rate and geometry). The transect of the twenty proposed piezometer locations crosses the broadest part of the uncertainty zone perpendicular to the course of the belt. This is fortunate for the optimal experimental design, since we can acquire information just within the most uncertain parts of the system.

Figures 7B–D show the maps of the misclassification probability after performing the optimal-experimental-design analysis for one, two, and three additional piezometers, respectively. In each of these figures, the identified optimal piezometer locations are marked by circles with black filling, while the unused potential piezometer locations are depicted as white-filled circles.

Figure 7B reveals how the misclassification probability is expected to be reduced by placing a single additional piezometer. The optimal location is the southernmost point along the transect close to the surface-water divide. Unsurprisingly, the location of this piezometer coincides with the location that shows the highest uncertainty of hydraulic heads in **Figure 6**. A comparison between **Figures 7A,B** shows that the misclassification probability is not only reduced in the direct vicinity of the chosen new piezometer, but essentially over the entire width of the Wurmlingen saddle, whereas the effect at the eastern end of the model domain is negligible. This pattern reflects the smoothness of hydraulic heads, but is strongly affected by the assumption that each lithostratigraphic unit has a uniform set of hydraulic parameters (only the groundwater-recharge values are subdivided by land-use). The latter implies that conditioning the model on a single observation point in a particular unit, here the weathered kmGr, affects the model outcome at all other points within this unit. However, if we had considered internal variability within the units, individual head measurements would not have reduced the uncertainty at distant points within that unit to the same extent. Consistent to these arguments, the eastern end of the uncertainty belt (where the topmost geological unit is km2345 rather than weathered kmGr) is not affected by placing a piezometer along the transect.

Further reduction of the misclassification probability can be achieved by placing a second additional piezometer at the northern fringe of the uncertainty belt (**Figure7C**), whereas the uncertainty pattern does not visually change when placing a third additional piezometer between the first and second piezometers (**Figure 7D**).

4.3. Performance of Designs

Figure 8 summarizes the performance of all 1,350 investigated piezometer configurations (grouped by one-, two- and three-additional-piezometer designs). All plots use the design number on the abscissa. In the following discussion, we use the notation "(first piez. | second piez. | third piez.)" to describe a given design, in which the numbers of the piezometer locations are sorted from south to north, and the missing piezometers in the one- and two-piezometer designs are marked by a dash. The designs are numbered in the following way: The first twenty designs contain only one additional piezometer, ranging from $(1|-|-)$ to $(20|-|-)$. The designs 21 to 210 are two-piezometer designs, starting with the combination $(1|2|-)$, incrementing the second location in steps of one to $(1|20|-)$, then moving from $(2|3|-)$ to $(2|20|-)$ and so forth, until $(19|20|-)$ is reached. In order to exclude replicates, the index of the second piezometer is always larger than that of the first. Finally, the designs 211 to 1,350 start with $(1|2|3)$ and increment the third location first, then the second, and then the first one, until reaching the final design $(18|19|20)$. Again we avoid replicates by requiring that the piezometer indices increase from the first to the third piezometer within all designs. **Figures 8D–F** visualize the piezometer designs by displaying the selected piezometers of each design as rectangles.

The top row of **Figure 8A-C** shows the values of the utility function $\phi(\mathbf{d})$ of the given designs $\mathbf{d}$ according to **Eq. 26**. It quantifies the expected relative reduction of the spatial mean of P_{mc} applying the measurement design $\mathbf{d}$. Theoretically, this metric can range between zero (no reduction of uncertainty at all) to one (perfect identification of the groundwater divide).

In the single-piezometer designs (**Figure 8A**), the performance declines with increasing design number (placing the new piezometer further north along the transect). While the first three designs result in a similar relative uncertainty reduction of $\approx 36\,\%$, $\phi(\mathbf{d})$ gradually decreases to a negligible low value of $\approx 3\,\%$ at location 20. The optimal design is $(1|-|-)$, resulting in a performance of $\phi = 36.6\,\%$. The best locations for placing a single piezometer coincide with the points at which the prior uncertainty of hydraulic head is the highest (see **Figure 6**), so that

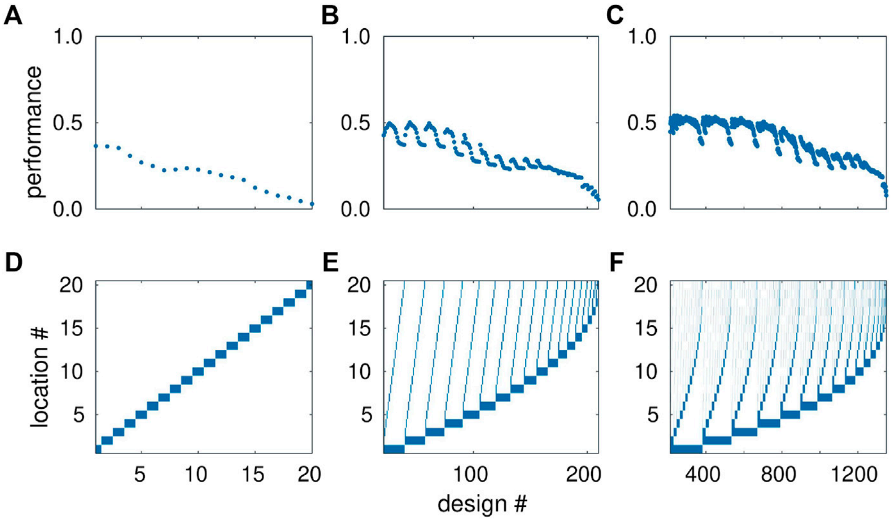

FIGURE 8 | Performance of all 1,350 investigated monitoring designs. Top row **(A–C)**: normalized utility function $\phi(\mathbf{d})$ of the given design according to **Eq. 26**; bottom row (**D–F**): piezometer combination of the given design. **(A**,**D)** Designs with one additional piezometer; **(B**,**D)** with two additional piezometers; **(C**,**F)**: with three additional piezometers.

constraining the model by taking a single head measurement at these points yields the highest information gain. As the hydraulic heads at the northern end of the transect are already constrained by the plausibility criteria of the model pre-selection, additional piezometers in this part of the transect hardly pay off.

Figure 8B shows the performances of all two-piezometer designs. Like in the one-piezometer designs, configurations including southern piezometer locations (design numbers 21 to ≈ 100) perform better than other designs. For a given first piezometer location, the performance depends on the distance between the two piezometers. At least for the well-performing designs 21 to 100, the optimal distance between the two piezometers is on the order of several hundred meters. Such a configuration performs better than designs in which the two new piezometers are further apart or closer to each other. The best two-piezometer configuration is (2|7|–), leading to an uncertainty reduction of $\phi = 50.2\,\%$.

The optimal two-piezometer designs may be explained by the combined effects of having the highest prior uncertainty of hydraulic head at the southern end of the transect (discussed in the context of the one-piezometer designs) and the inherent spatial correlation of hydraulic head caused by the groundwater-flow equation itself: One piezometer should be located at the most informative southern end; placing two piezometers to close to each other would yield redundant information (and observing a small head difference would drown in the measurement error), while placing the second piezometer at the northern end would be of little use because here the hydraulic heads are already constrained by the plausibility criteria.

In the three-piezometer designs (**Figure 8C**), this pattern is maintained, with the best location of the third piezometer being in the middle of the other two new observation wells. Thus, placing the third well further north, where the head-uncertainty is low, is less beneficial than refining the spatial resolution of head measurements in the southern third of the transect. The best three-piezometer configuration is (1|7|15) with $\phi = 54.2\,\%$, which is not drastically better than the best two-piezometer configuration. We conjecture that adding a fourth piezometer along the transect would yield an even lower increase of performance. Thus, in a practical application, it might be better to invest the money needed to install such a well in other investigations like elaborate well tests, or in entirely different locations (see **section 4.4**).

As a quality check, we determined the average effective sample size for the three optimal designs. The values are comparably large ($\mathrm{AESS}_1 = 859.7$, $\mathrm{AESS}_2 = 179.7$ and $\mathrm{AESS}_3 = 68.1$), which means the sample of $n_{\mathrm{sample}} = 50\,000$ was large enough to make reliable statements about the results.

Notably, all three optimal designs use very similar locations. Each larger optimal configuration basically includes the smaller ones as a subset (with the exception of switching between locations 2 and 1 in the two-location design). This means that, in the given application, one could decide whether and where to install the next observation well after installing the preceding ones, yielding essentially the same optimal designs. Such behavior is beneficial from a practical standpoint of view as, in real-world applications, the decision about extending a measurement network is often made only after realizing that the existing network is not (yet) sufficient. However, we cannot generalize that such a behavior occurs in all cases. In other applications, the optimal designs of many piezometers may not be a superset of the designs with fewer piezometers. Also, the information gained by the actual data value obtained by a first well could change the current state of knowledge, hence leading to (slightly) different later design decisions (Geiges et al., 2015). In such cases, deciding the number of observation wells would be necessary ahead of the first drilling in order to achieve optimal results.

We may compare the performance of the optimal designs with those of intuitive choices using the same number of new piezometers. When installing a single piezometer, one might place it on the middle of the transect using the design $(10|-|-)$. The uncertainty reduction of this particular design is $\phi = 22.9\,\%$, which is considerably smaller than the optimal performance of $\phi = 36.6\,\%$. When placing installing two piezometers, one could either maximize the distance along the full transect with design $(1|20|-)$ or subdivide the transect into three similarly long sections with the design $(7|14|-)$. The performances of these scenarios are $\phi = 37.2\,\%$ and $\phi = 25.1\,\%$, respectively, while the best two-piezometer design achieved $\phi = 50.2\,\%$. Actually, the best single-piezometer design performs almost as good as the intuitive two-piezometer design taken the two end points of the transect, and is considerably better than the intuitive design using identical section lengths. Finally, intuitive choices for the three-piezometer designs would be design $(1|10|20)$, which includes the two end points of the transect, and design $(5|10|15)$, subdividing the transect into sections of similar length. The respective uncertainty reductions are $\phi = 50.5\,\%$ and $\phi = 42.0\,\%$ compared to a reduction of $\phi = 54.2\,\%$ obtained by the optimal design. These calculations exemplify the benefit of an optimal-design-evaluation over intuitive choices.

4.4. Designs With the Third Piezometer being Placed Off the Transect

As shown in **Figure 7**, installing new piezometers along the suggested transect reduces the misclassification probability $P_{\mathrm{mc}}(\mathbf{x})$ on the hillslope parallel to the transect, but hardly affects $P_{\mathrm{mc}}(\mathbf{x})$ at the eastern end of the uncertainty belt. This part of the high-uncertainty belt is covered by the lithostratigraphic units km2345. Therefore, this uncertainty depends on the hydraulic properties and groundwater recharge of this model layer, and can only be reduced by observations that are sensitive to these properties. Because installing a third piezometer along the transect does not reduce $P_{\mathrm{mc}}(\mathbf{x})$ in this zone, the difference between the two- and three-piezometer designs is rather small. We thus hypothesize that placing a third piezometer somewhere else would yield a better performance. We tested this hypothesis by defining an alternative design space: we keep the best two piezometer locations along the transect fixed and then allow the third piezometer to be placed at any node of the two-dimensional computational grid. This resulted in 2067 additional designs.

Figure 9A shows which performance ϕ can be achieved as a function of the location of the third piezometer. The maximum performance of $\phi = 69.3\,\%$ is obtained by placing the third piezometer in the eastern part of the domain, roughly 400 m north of the highest-uncertainty region remaining after installing two piezometers (see 7C). This point is located in a hillslope hollow (see **Figure 5**) that collects groundwater recharged in the km2345 unit. The corresponding hydraulic head is sensitive to the hydraulic properties and groundwater recharge of the km2345 unit, which affects $P_{\mathrm{mc}}(\mathbf{x})$ in the eastern section of the uncertainty belt. The latter is confirmed by **Figure 9B**, displaying the resulting map of misclassification probability $P_{\mathrm{mc}}(\mathbf{x})$ for this newly defined optimal design, indicating that the new location of the third piezometer indeed reduces $P_{\mathrm{mc}}(\mathbf{x})$ in the eastern section of the uncertainty belt, which was hardly influenced by installing wells exclusively along the transect.

The average effective sample size of the optimal design in this substudy is comparably low ($\mathrm{AESS}_3^* = 4.4$). This drop is caused by the large information gain by the freely moving third well, so that only few realizations achieve significant likelihoods when compared to the hypothetical data values. Given this low number, a larger sample would be necessary to validate the statistical significance of the interpretations. However, given the high computational costs and because this is only a substudy offset from our actual objectives, we refrain from doing so.

Figure 9A includes an interesting and instructive artifact of the model: According to our model, hydraulic-head measurements on the northern hillslope appear to be beneficial for delineating the groundwater divide at the southern boundary of the Ammer valley. Most likely this is caused by the assumed uniformity of hydraulic parameters within each lithostratigraphic unit. In the very north of the model domain, the km2345 unit crops out, implying the same values of hydraulic conductivity and groundwater recharge as in the zone of interest at the souther boundary. Thus, a hydraulic-head measurement within this northern zone constrains model parameters of the km2345 unit, reducing the misclassification probability in the eastern part of the uncertainty belt. However, we are doubtful that this would be confirmed in a real-world application.

4.5. Strengths and Limitations of the Framework

Our framework is easily adaptable to other cases and applications, with the underlying groundwater-flow model being trivially exchangeable. This flexibility makes it convenient to apply the presented technique to other sites. Both interfaces, from the stochastic sampler to the numerical model, and from the numerical model to the optimal experimental design analysis, require only basic input/output operations of parameter values and virtual observations. While we have implemented the stochastic sampler and PreDIA as Matlab scripts, the approach could easily be transferred to other programming environments. However, a particle tracking tool is a necessary requirement for our framework to work.

Among the most labor-intensive parts of the framework is the initial model development, which is needed in quantitative hydrogeological consultancy anyway. Computationally, the creation of the plausible sample is the most costly step, but this can largely be parallelized. To obtain reasonable uncertainty estimates, several thousand model realizations are needed. This may not be affordable by everybody who might be interested in the uncertainty of groundwater-divide delineation. These computer-time limitations may be overcome by cloud computing.

In practical applications, the costs related to elaborate modeling in the planning phase of a new observation-well needs to be compared to the other expenses. This includes filing the application for legal approval, advertising for bids,

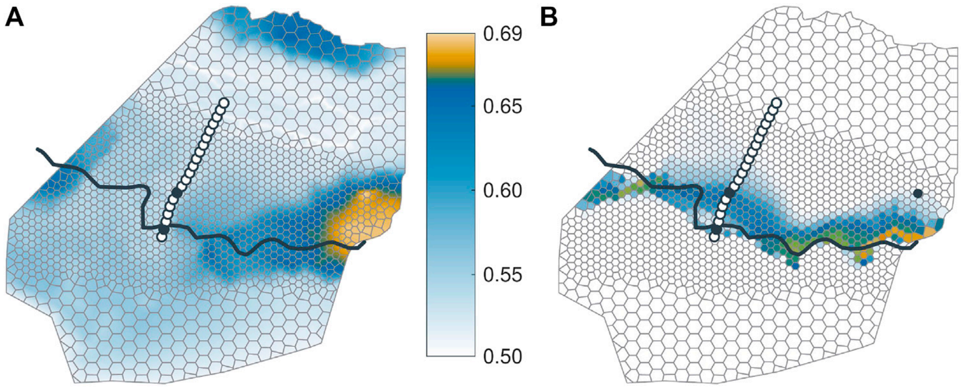

FIGURE 9 | Assessment of measurement designs with the third piezometer being placed outside the transect. **(A)** Performance ϕ of the design as a function of where the third observation well is placed; **(B)** map of the misclassification probability P_{mc} for the optimal three-piezometer design with one additional piezometer not being restricted to the transect (same colormap as in **Figure 7**).

planning of the fieldwork, and the drilling and completion expenses themselves. If the presented optimal-experimental-design method is initiated at the beginning of this process, it becomes an integral part of the decision-making process of how many new piezometers to install and where to place them.

The way we use the chosen optimal-design method PreDIA, we can only rank experimental designs within a given finite set. The number of elements in this set determines the computational costs of the optimal-design part of the analysis. In our application, we confined the design space by restricting the piezometer locations to a transect, reflecting the legal constraints at the given field site. With three piezometers at twenty potential locations, we had to consider 1,350 configurations. In the additional study presented in **section 4.4**, we removed the constraint to stay on the transect for one piezometer, considering 2067 potential locations. Allowing all three piezometers to be placed at any of these 2067 locations, would have resulted in more than $1.4 \cdot 10^9$ designs (see **Eq. 28**), which is computationally prohibitive. Tackling such a problem would need to involve an optimization algorithm around PreDIA to iteratively find a best-performing design without exhaustingly testing all of them. For the resulting search problem, the literature offers many suitable algorithms.

Our application was restricted to steady-state flow. Of course, real flow systems are never fully stationary, since they are always subject to transient forcings. Depending on the investigated site, this can include climatic influences, weather, tides or anthropogenic impacts (e.g., drinking water supply wells), all of which could affect the position of groundwater divides (e.g., Rodriguez-Pretelin and Nowak, 2018). Aquifers, where the expected movement of the groundwater flow divide over time is the main research question obviously need to account for this. Characteristics of such systems might be a significant abstraction of groundwater due to pumping wells, a known imbalance of the groundwate flow field or severe temporal fluctuations in groundwater recharge (e.g., Sanz et al., 2009). An interesting extension of our framework would be a transient analysis for such systems, by using transient simulations and time-dependent observations. Consequently, the underlying objective function would need to be redefined. We provide a possible extension toward dynamic systems in the appendix (**section 5.2**). However, the higher uncertainties related to inherently more complex transient models would require a larger sample and would most likely deteriorate the performance of the pre-selection method. In the context of transient data and models, a worthwhile avenue would be to combine optimal experimental design techniques with data-assimilation methods, but this is beyond the scope of the present study.

For most cases, where the divide is suspected to be shifted but not dramatically moving over time, our steady-state framework is applicable, with the interpretation of the steady-state as a “most representative state”. We also want to highlight that the goal of our framework is not to derive the position of groundwater divides themselves. Instead, we want to identify those locations that are best suited to conduct measurements providing insight for this delineation. The actual delineation, for example, can then be carried out by calibrating a groundwater flow model to the obtained measurement data. This second model can be more detailed, more finely discretized and even transient, as probably fewer model runs are necessary. If not already done, a rigorous grid convercence analysis should be performed ahead of the calibration to validate the numerical accuracy of the model.

As with every model, the performance of the method depends on the validity of underlying assumptions. In particular, we have assumed that the hydraulic parameters are uniform within each lithostratigraphic unit and that groundwater recharge is spatially uniform in zones defined by the topmost geological layer and land-use. Neglecting spatial variability within these zones expands the spatial ranges over which intended measurements are informative. We may also have missed discrete features altogether, which affect the position of the groundwater divide but do not influence the existing measurements. The latter would lead to a systematic bias.

The optimal-experimental-design method chosen in this study can accommodate any kind of uncertain parameters or uncertain model choices, provided that a prior uncertainty range is given. Both identifying the sources of uncertainty and defining the related prior distributions require expert knowledge, thus questioning the objectivity of the analysis. However, as with all Bayesian methods, such choices are at least made transparent. We have made good experience by initially setting fairly wide prior parameter ranges and then constraining the parameter space to behavioral models by the Gaussian-process-emulation supported pre-selection method (Erdal et al., 2020).

In the given application, we restricted the observations to hydraulic-head measurements, but this is not a limitation of the method. It is easy to augment the virtual observation vector by other data, such as hydraulic tests to be performed using the new observation wells, borehole dilution or tracer tests. Like with the extension to transient flow, the consideration of additional data types may also require more (uncertain) parameters. Systematically analyzing which type of data is most informative for which type of question is an ongoing issue of stochastic subsurface hydrology and optimal experimental design beyond the scope of the current study.

5. CONCLUSION

In this work we have presented a framework to identify the best piezometer configuration from a set of possible layouts to delineate local groundwater divides. Through the combination of filtered ensemble-based modeling of steady-state subsurface flow, particle tracking, and the application of the optimal-experimental-design technique PreDIA (Leube et al., 2012), we could identify the piezometer configuration for which we expect the largest reduction in the uncertainty of the groundwater divide. We have applied the method to an appropriate case study, which revealed the following insights:

(1) Configurations involving new measurement locations that are far away from existing ones perform better, because then the variability of hydraulic head, consistent with the existing data, is higher.
(2) In our application, a medium spacing of a few hundred meters between multiple new piezometers was optimal. Closer points would have led to redundant information due to the spatial auto-correlation of hydraulic head. Larger distances would have pushed observation points into non-informative regions close to existing measurements.
(3) The designs, defined as optimal by the presented framework, perform better than intuitive equidistant piezometer placements. In fact, the identified optimal design for a single piezometer provides similar information content as the tested intuitive equidistant placing of two piezometers, implying significant savings in real-world applications.
(4) Additional information obtained by adding more piezometers leads to further reduction of uncertainty, but the additional gain of information decreases with each new piezometer.
(5) Our procedure may be used to estimate whether the additional information gain is worth the effort of installing an additional observation well or not. The actual decision depends on the case at hand and involves a tradeoff between desired certainty and available resources. In our case, sequential optimization of one piezometer location after the other led to practically the same designs as jointly optimizing multiple piezometer designs, but this observation cannot be generalized.

A worthwhile follow-up study would be the extension of the presented framework to transient flow systems.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

JA set up the numerical flow and particle-tracking model, implemented the stochastic sampler, performed the computations, created the figures, and wrote the draft manuscript. AG performed the optimal experimental design analysis and contributed to manuscript revision. DE developed the stochastic sampler and the pre-selection method. WN and OC conceived the presented idea, supervised the work, provided funding, and revised the manuscript draft; all authors read and approved the submitted version.

FUNDING

This research was funded by the German Research Foundation (DFG) in the framework of the Research Training Group GRK 1829 "Integrated Hydrosystem Modelling" and the Collaborative Research Centre SFB 1253 CAMPOS - Catchments as Reactors (Grant Agreements GRK 1829/2 and SFB 1253/1 2017).

ACKNOWLEDGMENTS

We acknowledge support by the Open Access Publishing Fund of the University of Tübingen.

REFERENCES

Aigner, T., and Bachmann, G. H. (1992). Sequence-stratigraphic framework of the German Triassic. *Sediment. Geol.* 80, 115–135. doi:10.1016/0037-0738(92)90035-P.

Bloxom, L. F., and Burbey, T. J. (2015). Determination of the location of the groundwater divide and nature of groundwater flow paths within a region of active stream capture; the New River watershed, Virginia, USA. *Environ Earth Sci.* 74, 2687–2699. doi:10.1007/s12665-015-4290-1.

Brassel, K. E., and Reif, D. (1979). A procedure to generate Thiessen polygons. *Geogr. Anal.* 11, 289–303. doi:10.1111/j.1538-4632.1979.tb00695.x.

Brunner, P., and Simmons, C. T. (2012). HydroGeoSphere: a fully integrated, physically based hydrological model. *Ground Water.* 50, 170–176. doi:10.1111/j.1745-6584.2011.00882.x.

Constantine, P. G., and Diaz, P. (2017). Global sensitivity metrics from active subspaces. *Reliab. Eng. Syst. Saf.* 162, 1–13. doi:10.1016/j.ress.2017.01.013.

Constantine, P. G., Dow, E., and Wang, Q. (2014). Active subspace methods in theory and practice: applications to kriging surfaces. *SIAM J. Sci. Comput.* 36, A1500–A1524. doi:10.1137/130916138.

D'Affonseca, F. M., Finkel, M., and Cirpka, O. A. (2020). Combining implicit geological modeling, field surveys, and hydrogeological modeling to describe groundwater flow in a karst aquifer. *Hydrogeol. J.* [Epub ahead of print]. doi:10.1007/s10040-020-02220-z

Erdal, D., and Cirpka, O. A. (2019). Global sensitivity analysis and adaptive stochastic sampling of a subsurface-flow model using active subspaces. *Hydrol. Earth Syst. Sci.* 23, 3787–3805. doi:10.5194/hess-23-3787-2019.

Erdal, D., and Cirpka, O. A. (2020). Technical note: improved sampling of behavioral subsurface flow model parameters using active subspaces. *Hydrol. Earth Syst. Sci.* doi:10.5194/hess-2019-629.

Erdal, D., Xiao, S., Nowak, W., and Cirpka, O. A. (2020). Sampling behavioral model parameters for ensemble-based sensitivity analysis using Gaussian Process Emulation and Active Subspaces. *Stoch. Environ. Res. Risk. Assess.* doi:10.1007/s00477-020-01867-0

Farthing, M. W., Kees, C. E., and Miller, C. T. (2003). Mixed finite element methods and higher order temporal approximations for variably saturated groundwater flow. *Adv. Water Resour.* 26, 373–394. doi:10.1016/S0309-1708(02)00187-2.

Fedorov, V. V. (1972). *Theory of optimal experiments*. New York and London: Academic Press.

Franzetti, S., and Guadagnini, A. (1996). Probabilistic estimation of well catchments in heterogeneous aquifers. *J. Hydrol.* 174, 149–171. doi:10.1016/0022-1694(95)02750-5.

Geiges, A., Rubin, Y., and Nowak, W. (2015). Interactive design of experiments: a priori global versus sequential optimization, revised under changing states of knowledge. *Water Resour. Res.* 51, 7915–7936. doi:10.1002/2015WR017193.

Haitjema, H. M., and Mitchell-Bruker, S. (2005). Are water tables a subdued replica of the topography? *Ground Water.* 43 (6):781–786. doi:10.1111/j.1745-6584.2005.00090.x.

Han, P.-F., Wang, X.-S., Wan, L., Jiang, X.-W., and Hu, F.-S. (2019). The exact groundwater divide on water table between two rivers: a fundamental model investigation. *Water* 11, 685. doi:10.3390/w11040685.

Harreß, H. M. (1973). Hydrogeologische Untersuchungen Im Oberen Gäu (Tübingen). PhD thesis. Germany: University of Tübingen. Available at: https://rds-tue.ibs-bw.de/link?kid=1073957446.

Holzwarth, W. (1980). Wasserhaushalt Und Stoffumsatz Kleiner Einzugsgebiete Im Keuper Und Jura Bei Reutlingen-Tübingen (Tübingen). Dissertation. Germany: Universität Tübingen. Available at: https://rds-tue.ibs-bw.de/link?kid=1078052956.

Hunt, R. J., Steuer, J. J., Mansor, M. T. C., and Bullen, T. D. (2001). Delineating a recharge area for a spring using numerical modeling, Monte Carlo techniques, and geochemical investigation. *Ground Water.* 39, 702–712. doi:10.1111/j.1745-6584.2001.tb02360.x.

Kekeisen, F. (1913). Das Ammertal – Geologische Studie *(Rottenburg a. N.: Pfeffer & Hofmeister)*. Available at: https://rds-tue.ibs-bw.de/link?kid=1165633094.

H. J. Kirchholtes and W. Ufrecht (Editors) (2015). *Chlorierte Kohlenwasserstoffe im Grundwasser: untersuchungsmethoden, Modelle und ein Managementplan für Stuttgart*. Wiesbaden: Springer Vieweg.

Kortunov, E. (2018). Reactive transport and long-term redox evolution at the catchment scale. PhD thesis. Germany: University of Tübingen.

Leube, P. C., Geiges, A., and Nowak, W. (2012). Bayesian assessment of the expected data impact on prediction confidence in optimal sampling design. *Water Resour. Res.* 48, W02501. doi:10.1029/2010WR010137.

Liu, J. S. (2008). *Monte Carlo strategies in scientific computing*. New York, NY: Springer Science & Business Media.

Martin, S., Klingler, S., Dietrich, P., Leven, C., and Cirpka, O. A. (2020). Structural controls on the hydrogeological functioning of a floodplain. *Hydrogeol. J.* doi:10.1007/s10040-020-02225-8.

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

Pöschke, F., Nützmann, G., Engesgaard, P., and Lewandowski, J. (2018). How does the groundwater influence the water balance of a lowland lake? A field study from Lake Stechlin, north-eastern Germany. *Limnologica* 68, 17–25. doi:10.1016/j.limno.2017.11.005.

Pukelsheim, F. (2006). *Optimal design of experiments*. SIAM.

Qiu, H., Blaen, P., Comer-Warner, S., Hannah, D. M., Krause, S., and Phanikumar, M. S. (2019). Evaluating a coupled phenology-surface energy balance model to understand stream-subsurface Temperature dynamics in a mixed-use farmland catchment. *Water Resour. Res.* 55, 1675–1697. doi:10.1029/2018WR023644.

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

Rodriguez-Pretelin, A., and Nowak, W. (2018). Integrating transient behavior as a new dimension to WHPA delineation. *Adv. Water Resour.* 119, 178–187. doi:10.1016/j.advwatres.2018.07.005

Sanz, D., Gómez-Alday, J. J., Castaño, S., Moratalla, A., De las Heras, J., and Martínez-Alfaro, P. E. (2009). Hydrostratigraphic framework and hydrogeological behaviour of the mancha oriental system (SE Spain). *Hydrogeol. J.* 17, 1375–1391. doi:10.1007/s10040-009-0446-y.

Schmidt, M., Ohmert, W., Schreiner, A., and Villinger, E. (2005). Geologische Karte von Baden-Württemberg 1:25.000 – erläuterungen Zu Blatt 7420 Tübingen (Freiburg im Breisgau: regierungspräsidium Freiburg (Landesamt für Geologie, Rohstoffe und Bergbau)). 5th Edn.

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

Suk, H., and Park, E. (2019). Numerical solution of the Kirchhoff-transformed Richards equation for simulating variably saturated flow in heterogeneous layered porous media. *J. Hydrol.* 579, 124213. doi:10.1016/j.jhydrol.2019.124213.

Tóth, J. (1963). A theoretical analysis of groundwater flow in small drainage basins. *J. Geophys. Res.* 68, 4795–4812. doi:10.1029/JZ068i016p04795.

Tarboton, D. G., Bras, R. L., and Rodriguez-Iturbe, I. (1991). On the extraction of channel networks from digital elevation data. *Hydrol. Process.* 5, 81–100. doi:10.1002/hyp.3360050107.

Tecplot Inc (2019). User's manual Tecplot 360 EX 2019 release 1. (Bellevue, WA).

The MathWorks Inc (2019). Matlab (R2019b) (Natick, Massachusetts).

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

Tocci, M. D., Kelley, C. T., Miller, C. T., and Kees, C. E. (1998). Inexact Newton methods and the method of lines for solving Richards' equation in two space dimensions. *Comput. Geosci.* 2, 291–309. doi:10.1023/A:1011562522244.

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

van Genuchten, M. T. (1980). A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. *Soil Sci. Soc. Am. J.* 44, 892. doi:10.2136/sssaj1980.03615995004400050002x.

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

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer (AG) declared a past co-authorship with one of the authors (WN) to the handling editor.

APPENDIX

5.1. Generalization to Non-Binary Systems

In cases where one wants to delineate not only a particular (sub-) catchment's boundary, but the (potentially intersecting) groundwater divides between more than two of such catchments, the formulation of our objective functon (**Eq. 26**) based on binary particle fate maps (**Eq. 18**) is insufficient. Here, the particle fates cannot be described with the binary Bernoulli distributions, where the outcome for particle i is $z_i \in \{0, 1\}$. Instead, one could rely on categorical distributions, which can have more than two outcomes. For example, in a domain with three outlets the fate of particle i can be described with $z_i \in \{1, 2, \ldots, k\}$. Each of the outcomes would correspond to one outlet/subcatchment/receptor. We denote the total number of outcomes n_{fates}. To adapt our objective function to these cases, we need to formulate the overall probability of misclassifying the fate of a particle i. This can be done as described in the following.

We denote the probability that particle i belongs to the receptor k is $P(z_i = k)$. Then, the overall probability of misclassification becomes:

$$P_{\text{mc}}(z_i) = \sum_{k=1}^{n_{\text{fates}}} P(z_i = k) \cdot (1 - P(z_i = k)). \tag{29}$$

All other steps of the method remain as outlined above.

5.2. Possible Generalization to Transient Systems

A potential transient implementation of our framework would require a new formulation of the objective function. In such applications both, the modeled subsurface flow-field and the observations would change over time. This means that also the particle fate maps are transient, since the fate probabilities might change throughout the simulation period. This results in dynamic maps of misclassification probability, that is $P_{\text{mc}}(z)$ becomes $P_{\text{mc}}(z, t)$, which is a function of time t.

One potential way to define a metric quantifying the uncertainty of a transient groundwater divide would be to perform an additional integration/averaging over the simulation modeling duration τ.

$$\begin{aligned} U(\mathbf{z}) &= \frac{1}{\tau \cdot A_{2D}} \int_{\tau} \int_{A_{2D}} P_{\text{mc}}\left(z\left(\mathbf{x}^{\text{ini}}\right), t\right) \mathrm{d}\mathbf{x}^{\text{ini}} \, \mathrm{d}t \\ &= \frac{1}{\tau \cdot A_{2D}} \int_{A_{2D}} \int_{\tau} P_{\text{mc}}\left(z\left(\mathbf{x}^{\text{ini}}\right), t\right) \mathrm{d}t \, \mathrm{d}\mathbf{x}^{\text{ini}} \end{aligned} \tag{30}$$