STOCHASTIC MODELING IN HYDROGEOLOGY

EDITED BY: J. Jaime Gómez-Hernández, Liangping Li, Teng Xu and Andrés Alcolea

PUBLISHED IN: Frontiers in Earth Science and Frontiers in Environmental Science

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88971-037-9
DOI 10.3389/978-2-88971-037-9

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

STOCHASTIC MODELING IN HYDROGEOLOGY

Topic Editors:
J. Jaime Gómez-Hernández, Universitat Politècnica de València, Spain
Liangping Li, South Dakota School of Mines and Technology, United States
Teng Xu, Hohai University, China
Andrés Alcolea, Independent researcher, Switzerland

Dr. Andres Alcolea is employed by Geo-Energie Suisse AG and is the funder and CEO of HydroGeoModels. All other Topic Editors declare no competing interests with regards to the Research Topic subject

Citation: Gómez-Hernández, J. J., Li, L., Xu, T., Alcolea, A., eds. (2021). Stochastic Modeling in Hydrogeology. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88971-037-9

Table of Contents

frontiers in Earth Science

EDITORIAL
published: 17 May 2021
doi: 10.3389/feart.2021.698925

Editorial: Stochastic Modeling in Hydrogeology

J. Jaime Gómez-Hernández[1]*, *Liangping Li*[2], *Teng Xu*[3] and *Andrés Alcolea*[4]

[1]Institute for Water and Environmental Engineering, Universitat Politècnica de València, Valencia, Spain, [2]Department of Geology and Geological Engineering, South Dakota School of Mines, Rapid City, SD, United States, [3]State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering, Hohai University, Nanjing, China, [4]HydroGeoModels AG, Winterthur, Switzerland

Keywords: heterogeneity, uncertainty, Groundwater, simulation, hydrology

Editorial on the Research Topic

Stochastic Modeling in Hydrogeology

The call for this research topic asked for papers that would support the use of stochastic modeling in Hydrogeology, mainly focusing on proved applications of these techniques. We succeeded in attracting many expressions of interest that finally materialized in the collection of ten papers that makes up this section. Not all ten articles show real applications of stochastic Hydrogeology, but they all highlight the importance of uncertainty quantification in groundwater flow and mass transport modeling, and the need to use stochastic techniques to do it in an appropriate, systematic, and traceable manner.

If you were to read just one paper on this research topic, we suggest you the one by White et al. since it meets all the target objectives. It stresses the importance of parameter estimation (PE) and uncertainty quantification (UQ) and demonstrates it with a fully worked-out example in the Edwards aquifer, Texas, United States. The authors not only guide the reader on the different steps to perform PE and UQ in an aquifer that is complex to model but also provide all the scripts used so that anyone can inspire from their work and apply them to other cases.

Of the other papers demonstrating the importance of adequately handling heterogeneity in aquifer modeling, two focus on facies heterogeneity since this heterogeneity is more critical than the intrinsic heterogeneity of hydraulic conductivity within facies. The large contrasts in hydraulic conductivity are not due to its inherent spatial variability but to hydrofacies heterogeneity. Consequently, there is a need for robust methods for the generation of categorical realizations resembling the geological aquifer architecture. Both Carle and Fogg and Jorreto-Zaguirre et al. propose ways of handling soft or uncertain data in categorical simulation, each one using a different categorical simulation technique. Carle and Fogg demonstrate their findings in the Savanah River, South Carolina, United States and the Llagas basin, California, United States, and Jorreto-Zaguirre et al. in the Andrax Delta, Spain. Of particular interest is the discussion by Carle and Fogg of current methods and the outlook about the future.

Another interesting application of stochastic Hydrogeology to a real case is the paper by Colombo et al. who demonstrate the applicability of backward tracking to identify pollutant sources in the metropolitan area of Milano.

There are two more papers with real case applications. The first one, by Hemmings et al. discusses the importance of early uncertainty quantification aimed to maximize the efficiency of modeling in the context of decision support. The authors discuss the importance of identifying whether the cost of expensive history matching is worth it given the available data. They propose a decision support modeling workflow and demonstrate it in the Wairarapa Valley, New Zealand. The second one, by Cromwell et al. presents a clever use of deep neural networks to estimate hydraulic conductivities in a catchment using integrated surface-subsurface modeling and demonstrates its application in Rock Creek, Colorado, United States.

OPEN ACCESS

Edited by and reviewed by:
Nick Van De Giesen,
Delft University of Technology,
Netherlands

***Correspondence:**
J. Jaime Gómez-Hernández
jgomez@upv.es

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

Received: *22 April 2021*
Accepted: *03 May 2021*
Published: *17 May 2021*

Citation:
Gómez-Hernández JJ, Li L, Xu T and Alcolea A (2021) Editorial: Stochastic Modeling in Hydrogeology. Front. Earth Sci. 9:698925. doi: 10.3389/feart.2021.698925

The last application by Allgeier et al. addresses an optimization problem under uncertainty: the optimal selection of a monitoring network for the delineation of groundwater divides. The authors use the Preposterior Data Impact Assessor as their optimal experimental design method and demonstrate it to delineate the groundwater divide between the Ammer and Neckar river catchments in Germany.

There are three more papers that explore different facets of stochastic Hydrogeology in very different environments. Jeong et al. address the general problem of optimization under uncertainty and compare three different ensemble-based stochastic gradient methods for the optimal well placement for brine extraction in a synthetic carbon storage reservoir. In their conclusions, they provide recommendations on when and how to use each one of the three methods analyzed. Lam et al. discuss the application of an ensemble smoother for stochastic inverse modeling of groundwater flow parameters using transient hydraulic heads and flow rates as data. They limit their analysis to multi-Gaussian distributions for the flow parameters and use a synthetic case, which resembles the French Underground Research Laboratory site, to demonstrate the performance of the approach. Their final results are clearly influenced by the uneven coverage of the model domain by the observation locations. Finally, Williams et al. focus on a very specific problem related to the loss of resolution in the velocity field computed on an equivalent porous media derived from a fractured domain. The authors demonstrate the problem and propose a downscaling approach to recover the effects of subgrid heterogeneity in the context of radionuclide transport through fractured media. They demonstrate their approach on a synthetic brittle fault zone model.

The more than 13,000 views, in the first four months after the first paper was accepted, proves the interest of the scientific community in stochastic modeling in Hydrogeology. We hope that you enjoy reading this collection.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Conflict of Interest: AA was employed by HydroGeoModels AG.

The remaining 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.

frontiers in Earth Science

ORIGINAL RESEARCH
published: 28 February 2020
doi: 10.3389/feart.2020.00050

Toward Reproducible Environmental Modeling for Decision Support: A Worked Example

Jeremy T. White[1]*, *Linzy K. Foster*[1], *Michael N. Fienen*[2], *Matthew J. Knowling*[3], *Brioch Hemmings*[3] and *James R. Winterle*[4]

[1] U.S. Geological Survey, Austin, TX, United States, [2] U.S. Geological Survey, Middleton, WI, United States, [3] GNS Science, Wellington, New Zealand, [4] Edwards Aquifer Authority, San Antonio, TX, United States

A fully worked example of decision-support-scale uncertainty quantification (UQ) and parameter estimation (PE) is presented. The analyses are implemented for an existing groundwater flow model of the Edwards aquifer, Texas, USA, and are completed in a script-based workflow that strives to be transparent and reproducible. High-dimensional PE is used to history-match simulated outputs to corresponding state observations of spring flow and groundwater level. Then a hindcast of a historical drought is made. Using available state observations recorded during drought conditions, the combined UQ and PE analyses are shown to yield an ensemble of model results that bracket the observed hydrologic responses. All files and scripts used for the analyses are placed in the public domain to serve as a template for other practitioners who are interested in undertaking these types of analyses.

Keywords: decision-support, groundwater modeling, uncertainty quantification, parameter estimation, reproducible, scripting

OPEN ACCESS

Edited by:
J. Jaime Gómez-Hernández, Universitat Politècnica de València, Spain

Reviewed by:
Brian Alan Smith, Barton Springs/Edwards Aquifer Conservation District, United States
Sean McKenna, IBM Research, Ireland

***Correspondence:**
Jeremy T. White
jwhite@usgs.gov

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

Received: *12 November 2019*
Accepted: *13 February 2020*
Published: *28 February 2020*

Citation:
White JT, Foster LK, Fienen MN, Knowling MJ, Hemmings B and Winterle JR (2020) Toward Reproducible Environmental Modeling for Decision Support: A Worked Example. Front. Earth Sci. 8:50. doi: 10.3389/feart.2020.00050

1. INTRODUCTION

The importance of uncertainty quantification (UQ) in the context of environmental modeling for decision support is widely recognized (e.g., Anderson et al., 2015; Doherty, 2015a). So too is the importance of parameter estimation (PE), which, herein, we regard as the process of reducing uncertainty through history matching the simulation outputs to their state observation counterparts (a process often referred to as "calibration"). Together, UQ and PE represent critical analyses for model-based resource management decision support as they provide estimates of uncertainty in important simulated outcomes and reduce this uncertainty, respectively.

However, implementing high-dimensional UQ and PE in real-world modeling analyses can be difficult, from both a theoretical understanding standpoint (related to the depth and breadth of topical knowledge required), as well as from a mechanics/logistics standpoint arising from the preparation, implementation, and post-processing of these analyses. In the authors' experience, the difficulties commonly encountered when implementing UQ and PE for decision-support-scale modeling can preclude their application in many cases, especially when project time lines are short and funding is limited.

There are also strong calls for modeling based analysis (including UQ and PE analyses) to move toward more transparent, reproducible, and accountable processes. The reasons for this push are self-evident; several groups have called for increased transparency and reproducibility in computational science (Goecks et al., 2010; Stodden, 2010; Peng, 2011; Sandve et al., 2013; Liu et al., 2019) and in environmental simulation specifically Fienen and Bakker (2016). Some

authors have put forward examples of increasing the reproducibility of the forward environmental model construction process (e.g., Fisher et al., 2016). To that end, some script-based tools have been developed for practitioners to increase the reproducibility of the forward model construction process (Olsthoorn, 2010; Fisher, 2014; Bakker et al., 2016). However, these tools are focused on the forward model rather than the UQ and PE process; in a decision-support setting, UQ and PE analyses are critical to the robust deployment of a model, and are therefore likely just as important as the forward model.

Ironically, the need for PE and UQ can be, in many contexts, in competition with the need for reproducibility. This is because the PE and UQ analyses require many additional subjective conceptual choices and bring many more operations and steps into the implementation of the modeling analysis, and these additional complications can substantially decrease the reproducibility of a modeling analyses. This decrease of reproducibility, especially in the outcomes of the PE and UQ analyses, can reduce the credibility of the model as a decision support tool and may hamper resource management efforts.

Herein, we present a step toward reproducible UQ and PE analysis through a script-based workflow. We use the term "reproducible" to mean giving readers access to the datasets and scripting tools needed to reproduce our results (e.g., figures, **Supplementary Material** and associated data release; White et al., 2020) and the findings based on them. Readers are referred to Plesser (2018) and Kitzes et al. (2017) (and the references cited therein) for a more nuanced and detailed discussion of what "reproducibility" means in the context of computational science.

Several open-source software tools were used to implement the UQ and PE workflow, including:

- The python package FloPy (Bakker et al., 2016) was used to programmatically load, process, and manipulate an existing groundwater flow model;
- The python package pyEMU (White et al., 2016) was used to programmatically construct a high-dimensional PEST interface (Doherty, 2015b) around the forward model and the generate the prior parameter ensembles;
- The iterative ensemble smoother PESTPP-IES (White, 2018) was used to evaluate the prior parameter ensembles (for UQ) and to also perform formal, high-dimensional PE.

Within this scripted workflow, we programmatically construct a high-dimensional truncated multi-variate (log-)Gaussian prior parameter distribution (hereinafter referred to as the "Prior") and associated ensembles. The scripting is also used to define a subjective, management-focused likelihood function for the PE analysis. Additionally, we use scripting to post-process the results into the figures and **Supplementary Material** presented herein. In this way, we demonstrate that high-dimensional UQ and PE in real-world environmental modeling settings are achievable and can be both efficient and reproducible. Furthermore, given the increased interest in UQ and PE analyses in environmental simulation, the workflow presented herein provides the capability to efficiently and repeatably apply UQ and PE analyses to models that were previously constructed.

The rest of this manuscript is organized as follows. First, we briefly present the existing model, then we discuss the formulation of the Prior and the definition of the likelihood function used for PE, followed by the reproducible implementation and workflow presentation. Then the UQ and PE analysis results are presented, and finally, we discuss some nuances and implications of a script-driven UQ and PE workflow.

2. THE EDWARDS AQUIFER MODEL AND PURPOSE OF THE ANALYSIS

Herein, we use an existing model of the Edwards aquifer, Texas, USA from the work of Liu et al. (2017), based on the model of Lindgren et al. (2005). Briefly, the model is a MODFLOW-2005 (Harbaugh, 2005) model with 1 layer, 370 rows, and 700 columns arranged on a regular grid with a spacing of 1,340 feet; the geographic location of the model domain and features of interest are shown in **Figure 1**. Water enters the model domain as diffuse, areal recharge and as concentrated, stream-bed recharge—both of these recharge processes are simulated with the Recharge (RCH) package. Water leaves the model domain as spring flow (simulated with the Drain (DRN) package) and as extraction wells (simulated with the Well (WEL) package). Faults that are thought to function as barriers to flow are represented with the Horizontal Flow Barrier (HFB) package. Readers are referred to Liu et al. (2017) for more details regarding the model and specific simulations.

The model has been temporally discretized into two simulation time periods:

- *history-matching simulation*: simulates the period 2001–2015 with monthly stress periods. This simulation is used for PE (i.e., history matching) of observed spring flows and groundwater levels;
- *scenario simulation*: simulates the period 1947–1958 (known as the "drought of record") with monthly stress periods. This simulation is used to make a hindcast of simulated states of primary interest to groundwater resource managers, namely spring flow at Comal and San Marcos springs and groundwater level at index wells J-17 and J-27.

Both simulations use the same static (i.e., time-invariant) properties of hydraulic conductivity, storage, HFB conductances, and DRN boundary elements (stage and conductance). This is the mechanism for PE to reduce uncertainty in the scenario-simulation outputs of primary interest to groundwater resource managers. If these outputs are sensitive to the static properties and, through PE, the uncertainty in the static properties is reduced, then the uncertainty in the scenario-simulation outputs may also be reduced.

For the PE analysis, observations of spring flow and groundwater level are used for history matching from 6 and 336 locations, respectively, with a total of 1,060 spring flow state observations and 6,809 groundwater-level state observations.

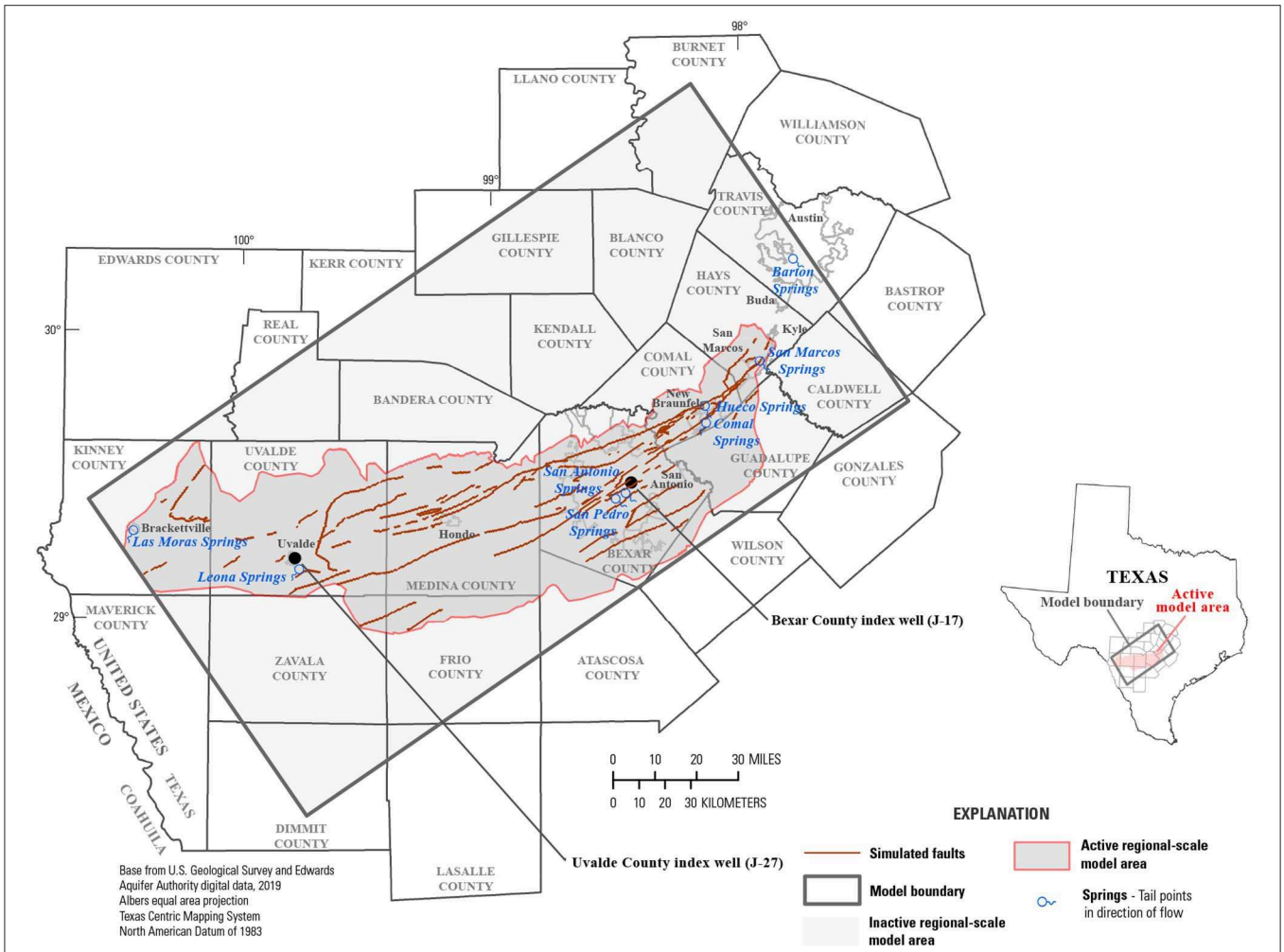


FIGURE 1 | The geographic extent of the active model domain (gray) and locations of springs and groundwater-level observation locations of primary interest to groundwater resource managers. Modified from Brakefield et al. (2015).

2.1. Model Purpose

The flow from Comal and San Marcos springs and the groundwater levels at index wells J-17 and J-27 during the scenario (e.g., drought) simulation are of particular interest to groundwater resource management and are the primary focus of the UQ and PE analyses presented herein. Therefore, we focus the PE analysis on reproducing the observed spring flow and water levels listed above as robustly as possible during the history-matching simulation. Logically, reproducing these observed states during history-matching should improve the ability to simulate these observed states during the scenario simulation. We note that state observations of spring flow and groundwater level are also available for the scenario hindcast simulation at Comal and San Marcos springs and at index wells J-17 and J-27, respectively. However, we use these state observations only to verify the robustness (or otherwise) of the various workflow components, and, more generally, of the workflow itself; these scenario-period observations are not used for history-matching purposes.

2.2. Parameterization and the Prior

Herein, we use a Bayesian uncertainty framework (Tarantola, 2005) to represent uncertainty in parameters and outputs of primary interest to groundwater resource managers. A critical part of any Bayesian uncertainty quantification (UQ) analysis is definition of the Prior. We use a high-dimensional parameter space (Doherty et al., 2011) with the aim of achieving robust estimates for the hindcast of simulated states of primary interest to groundwater resource managers, while also attempting to avoid any ill-effects arising from under-parameterization (White et al., 2014; Knowling et al., 2019). Specifically, we used 337,482 and 339,449 parameters to represent model input uncertainty in the history-matching and scenario simulations, respectively (including the shared static property parameters, outlined above).

In the high-dimensional parameter space, we defined a truncated, multi-variate (log-)Gaussian distribution as the Prior; we used the existing history-matching and scenario simulation model inputs of Liu et al. (2017) as the first moment (e.g., mean vector) of the Prior and a block-diagonal covariance matrix for

the second moment. Parameter variances were defined using expert knowledge and previous modeling analyses of the Edwards aquifer. The blocks in the prior parameter covariance matrix represent spatially- and temporally-correlated parameters, such as grid-scale and pilot-point (Doherty, 2003) parameters and time-varying parameters associated with well extraction rates. These correlations between spatially and temporally distributed parameters were specified using exponential variograms with the following ranges:

- 13,200 feet: grid-scale parameters, including hydraulic conductivity, specific storage, specific yield, initial conditions, HFB conductances; spatially-distributed well extraction rate parameters;
- 180 days: time-varying well extraction rate parameters; and
- 39,600 feet: pilot-point parameters, including hydraulic conductivity, specific storage, specific yield, initial conditions.

These ranges were selected so that the resulting spatially distributed model inputs had sufficient stochastic character in accordance with expert knowledge and previous modeling analyses of the Edwards aquifer.

Using the previously history-matched model inputs as the mean of the Prior is not standard practice in a purely Bayesian context because the same state observations will be used for conditioning herein. However, using the existing model inputs in this way allows us to take advantage of expert and institutional knowledge that has previously been assimilated into the model. Furthermore, using a very high-dimensional parameter space in combination with an ensemble framework allows us to account for the null-space contribution to uncertainty (Moore and Doherty, 2005) surrounding this history-matched location in parameter space.

We use a multi-scale parameterization strategy (McKenna et al., 2019) to explicitly represent different spatial scales of uncertainty and also to help understand how information is transferred from observed states to parameters (at different scales) in the PE analysis. For hydraulic conductivity, specific storage, specific yield, and initial conditions, three spatial scales of parameterization were used:

- a single, domain-wide ("global") multiplier parameter;
- pilot point multiplier parameters (Doherty, 2003) at a spacing of 39,600 feet; and
- grid-scale multiplier parameters (one parameter per active computational cell).

Recharge was parameterized using time-varying domain-wide multiplier parameters in conjunction with time-varying multiplier parameters for each of the 25 unique recharge "zones"—for each stress period, a domain-wide multiplier parameter and a multiplier parameter for each zone was specified. In this way, we attempt to account for spatial uncertainty as well as temporal uncertainty in the recharge estimates. See Brakefield et al. (2015; Figure 15) for an example of the recharge zonation and Puente (1978) for a description of the Edwards aquifer recharge estimation process. Readers are referred to the **Supplementary Material** for a graphically summary of the multi-scale parameterization.

Well extraction rates were also parameterized to account for spatial and temporal uncertainty in the well extraction rate estimates. A single set of extraction rate multiplier parameters (one per well) was applied across all stress periods. This set of spatially distributed multiplier parameters were used with a set of temporally-distributed multiplier parameters (one for each stress period). We note that, while groundwater extraction rates were metered during the history-matching period, the simulated groundwater extraction in the model is nevertheless still uncertain as these metered rates may not capture all of the groundwater extraction that occurred and because of uncertainty (e.g., error) induced through spatial and temporal discretization.

Because the exact hydrologic disposition and function of the simulated HFBs is unknown, these were also parameterized at the grid scale. The conductance of each HFB cell was treated as uncertain but was spatially correlated with nearby HFB cells using a geostatistical variogram with a range of 13,200 feet.

A summary of the parameterization and prior parameter variances is presented in the **Supplementary Material**.

Note that separate temporal parameters (recharge, well extraction and initial conditions) are used for the history-matching and scenario simulations. All other parameters are shared between the two simulations.

2.3. The Likelihood

Given the intended management purposes(s) of this modeling analysis, we focused the PE analysis on reproducing the observed states from the history-matching period that most resemble the outputs from the scenario period of primary interest to groundwater resource managers (Beven and Binley, 1992; Doherty and Welter, 2010; White et al., 2014). Specifically, we defined a subjective $\mathcal{L}_2$ norm likelihood function—expressed through observation weights—to focus the PE analysis on reproducing observed states from the following four locations:

- Comal springs flow;
- San Marcos springs flow;
- index well J-17 groundwater levels;
- index well J-27 groundwater levels.

The model outputs of primary interest to groundwater resource managers during the scenario simulation are of the same character (i.e., observed hydrologic state types, spatial locations) as the state observations used in the focused likelihood function. We therefore expect that reproducing the observed states at these four locations during the history-matching simulation should improve the model's ability to simulate these observed states during the scenario simulation (e.g., Doherty and Christensen, 2011; White et al., 2014).

The focused likelihood function was implemented by subjectively specifying weights (e.g., the inverse of observation variance) on these four state observation series that are two orders of magnitude higher than the weights on the other state observation series. In this way, we focus the PE analysis toward preferentially reproducing these four state observation series, with the expectation that better reproduction of these observed states in the history-matching simulation will lead to reduced uncertainty in the hindcast of simulated states at these

locations. Including the remaining state observations into the likelihood function with a lower weight (e.g., focus) helps to ensure physically plausible simulation results in the posterior history-matching ensemble. This subjective weighting scheme was applied using the mean residuals from the initial, prior Monte Carlo analysis (discussed below).

3. IMPLEMENTATION AND WORKFLOW

The UQ and PE analyses outlined above were implemented within a python-based scripting workflow; the workflow is contained entirely within the python script `eaa.py` and is implemented as functions within this script. We note the initial history-matching and scenario simulation model input files are preserved "as-is"—the scripting process does the only file handling.

At the highest-level, the workflow follows these steps (function names shown in parentheses):

1. (`setup_models_parallel`): Process the model input files for both history-matching and scenario simulations and generate high-dimensional PEST interface (Doherty, 2015b). Tasks include programmatically switching the MODFLOW model input formats to support free-format and external files, as well as rectifying the WEL files so that the same number of well entries are present in each stress period, which is important for parameterizing well extraction rates. This means including additional extraction well entries with extraction rate equal to zero for consistency. Define the geostatistical prior parameter covariance matrix and generate the prior parameter ensembles of 100 realizations for each simulation using the Prior distribution.
2. (`prep_for_parallel, run_condor`): Evaluate prior ensembles (with parallel computation) for both simulations.
3. (`reweight_ensemble`): Use the history-matching simulation prior ensemble mean residuals to define the focused likelihood function. Eliminate realizations that yield implausible outputs.
4. (`build_localizer, prep_for_parallel, run_condor`): Construct a localizing matrix for temporal parameters (discussed later). Perform the PE analysis using PESTPP-IES (White, 2018).
5. (`transfer_hist_pars_to_scenario`): Transfer the static (time-invariant) final (i.e., posterior) history-matching ensemble values to the scenario prior parameter ensemble, effectively forming the scenario posterior ensemble.
6. (`prep_for_parallel, run_condor`): Evaluate the scenario posterior ensemble.
7. (`plot_parallel`): Post-process the results of the UQ and PE analyses into figures and **Supplementary Material**.

We chose 100 realizations for the UQ and PE analyses as a trade-off between the need to express uncertainty and the need to minimize the computational burden both during these analyses and during follow-on scenario analyses to support resource management decision making.

The PEST interface construction is the most complex portion of the workflow as it involves setting up a multi-scale multiplier parameter process, the PEST control file, as well as template files and instruction files to interface with the model. Furthermore, because we are working in an ensemble framework, it is important to record all possible model outputs of interest within the PEST interface given that acquiring new model outputs requires a re-evaluation of the entire ensemble (as opposed to a deterministic setting where only a single model run is needed). To achieve this goal, we used the python modules FloPy (Bakker et al., 2016) and pyEMU (White et al., 2016) to automate the PEST interface construction process. These two python modules, when used together, can reduce (or eliminate as is the case here) instances where a practitioner must create or modify files in a manual fashion (e.g., "by hand") (Barchard and Pace, 2011). Furthermore, using pyEMU to automate the PEST interface construction and geostatistical-prior ensemble generation can greatly reduce the cognitive burden on practitioners and also facilitate UQ and PE analyses at earlier stages within the larger modeling analysis.

The prior parameter ensembles were evaluated in parallel using the iterative ensemble smoother PESTPP-IES (White, 2018); this code was also used to perform the PE for the history-matching simulation. The results of each PESTPP-IES analysis were post processed using the above-referenced plotting functions to produce the figures presented in the Results section. Two iterations of PESTPP-IES were used to history-match the prior parameter ensemble to the observed states from history-matching simulation.

The high-throughput run manager `HTCondor` (Thain et al., 2005; Fienen and Hunt, 2015) was used to coordinate starting the PESTPP-IES parallel "agents" on a distributed computing cluster, as well as the "master" instance (through the function `run_condor`). However, the analyses presented herein can also be completed using the function `run_local`, which starts parallel agents and the master instance using only locally available (on a single machine) computational resources.

Localization was used in PESTPP-IES to mitigate for the effects of spurious correlation issues that can accompany the use of ensemble (smoother) methods (Chen and Oliver, 2016). Here, we localize temporal parameters—using an 18-month window between temporal parameters and state observations such that only apparent cross-correlations between observations that occur within the 18 months following the application of a temporal parameter are allowed in the PESTPP-IES solution scheme. In short, temporal localization effectively prevents non-physical (i.e., backward in time) cross-correlations and also eliminates long-term cross-correlations that are not expected in the Edwards aquifer, which is a karst system that responds rapidly to changes in forcing conditions. The localization matrix was constructed by the function `build_localizer`.

We note that the algorithm encoded in PESTPP-IES implements a "regularized" parameter adjustment equation (e.g., Hanke, 1997; Chen and Oliver, 2013, 2016) that enforces regularization penalties individually for each realization to prefer each realization remain close to the prior-generated initial values. This regularization, used in conjunction with localization, attempts to retain maximum parameter variance in the posterior parameter ensemble.

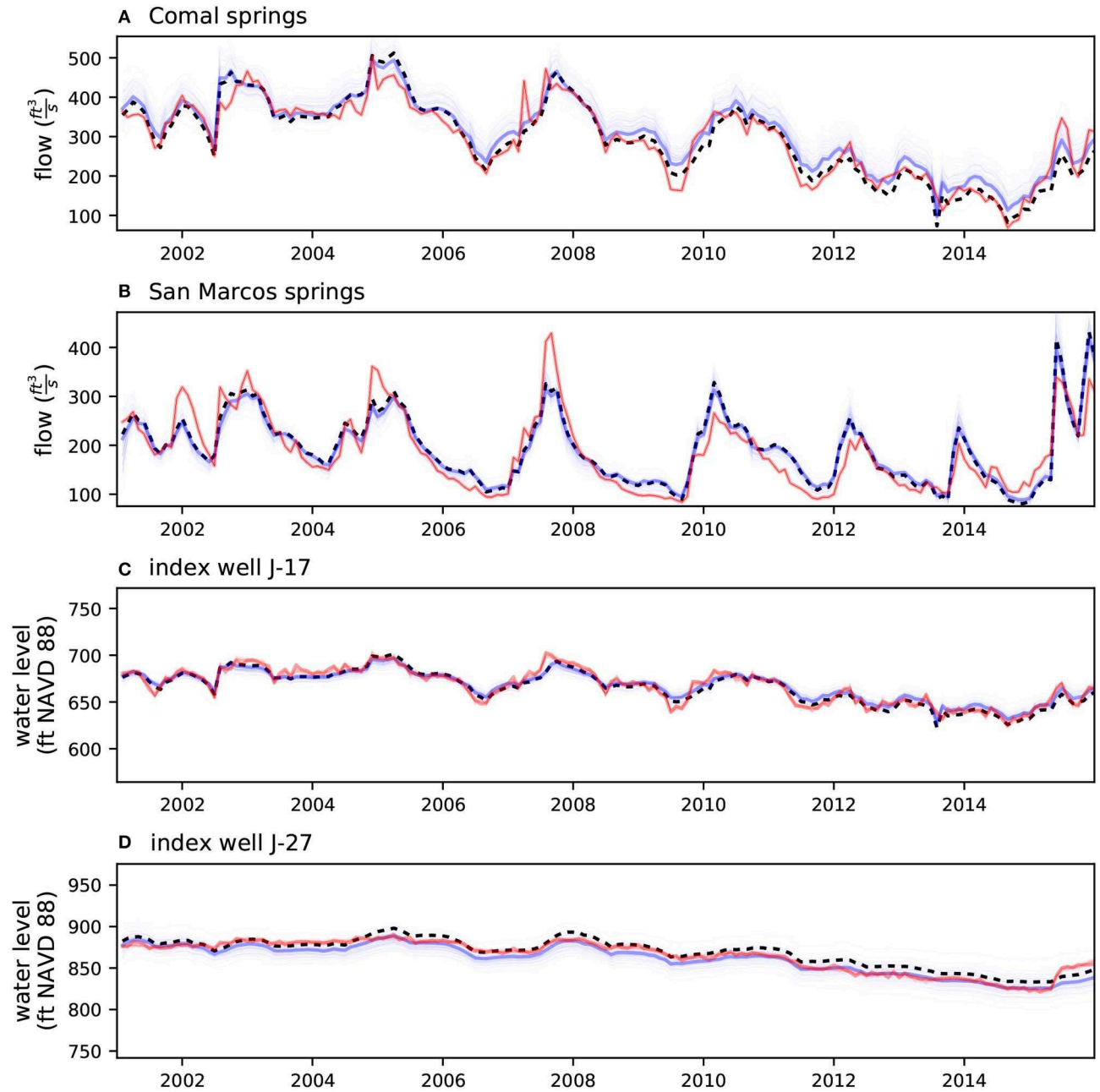


FIGURE 2 | Observed vs. simulated plots for the four locations of primary interest to groundwater resource managers for the history-matching simulation. Red trace is the observed series, dashed black trace is the simulation results from the existing model from Liu et al. (2017), thick blue trace is the *maximum a posteriori* simulation result (the posterior realization corresponding to the existing model inputs), light blue traces are the posterior ensemble simulation results. In general terms, the posterior ensemble closely reproduces the observed series.

4. RESULTS

During evaluation of the history-matching simulation prior parameter ensemble, 13 realizations were removed due to excessive run times and 5 realizations were removed for yielding "dry" model cells for locations where groundwater level have been measured, leaving 82 realizations for use in the PE analysis—these 82 realizations were used to evaluate prior and posterior scenario simulation uncertainty. In total, the history-matching simulation was evaluated 310 times; the scenario simulation was evaluated 182 times.

In general terms, the prior and posterior ensembles bracket the observed states behavior both for the history-matching and scenario simulations at the four locations of primary interest to groundwater resource managers (**Figures 2**, **3**). The posterior ensemble is tightly clustered around the observed states at

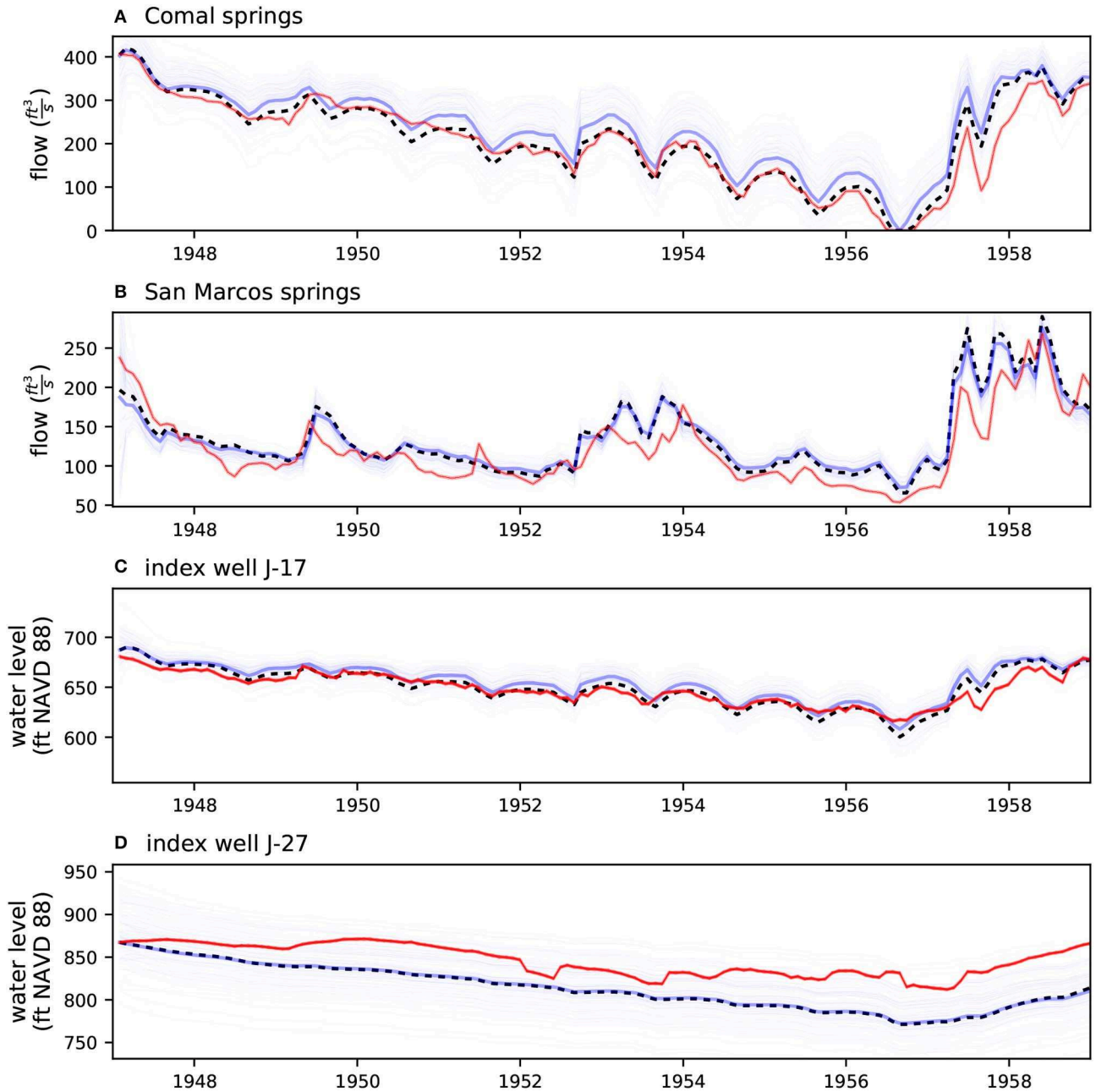


FIGURE 3 | Observed vs. simulated plots for the four locations of primary interest to groundwater resource managers for the scenario simulation. Red trace is the observed series, dashed black trace is the existing simulation results from Liu et al. (2017), thick blue trace is the *maximum a posteriori* simulation result, light blue traces are the posterior ensemble simulation results. The posterior ensemble brackets the observed series, however, some uncertainty remains, which we attribute to the scenario simulation forcing parameters.

the four locations of primary interest to groundwater resource managers during the history-matching simulation (**Figure 2**), as expected, due to these particular observed states being the dominant components of the likelihood function used in the PE analysis. The scenario simulation posterior ensemble (**Figure 3**) does not result in the same level of reproduction at the four locations. We attribute this to the inclusion of scenario-simulation-specific recharge and well-extraction uncertainty, expressed as parameters that only occur in the scenario simulation. That is, no matter how much the static properties are conditioned during history-matching simulation PE analysis, these scenario-only parameters remain at their prior uncertainty, and subsequently induce uncertainty in the scenario posterior simulated outputs.

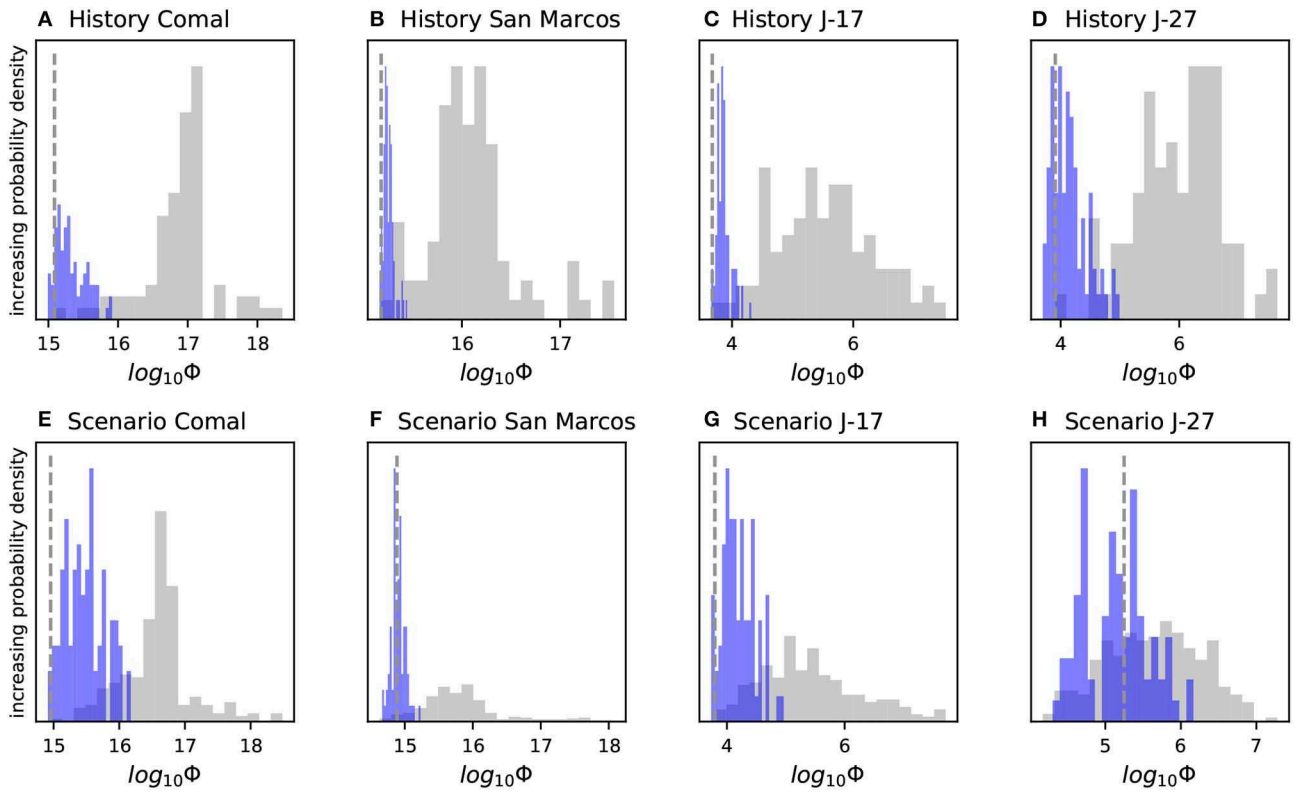


FIGURE 4 | Prior (gray) and posterior (blue) unweighted sum of squared residual (i.e., Φ) histograms for both history-matching and scenario simulations at the four locations of primary interest to groundwater resource managers. The vertical dashed gray line marks the Φ yielded by the existing model of Liu et al. (2017). Several posterior realizations better reproduce the observed series than the existing model; the history-matching simulation ensemble is more tightly clustered around the minimum Φ.

The fact that the posterior scenario simulation ensemble brackets the observed low spring-flow rates and low water levels at the springs and index wells of primary interest to groundwater resource managers (**Figure 3**) indicates that the combined UQ and PE analyses are likely to be robust at hindcasting (in a stochastic sense) the hydrologic response to drought at these four locations. This is an encouraging outcome and indicates that the automated workflow is functioning as expected. We attribute this success to the use of a high-dimensional parameter space (which helps to avoid under-estimation of uncertainty and limits the potential ill-effects of model error; Doherty and Christensen, 2011; White et al., 2014; Knowling et al., 2019), as well as the use of a likelihood function that was focused on outcomes of primary interest to groundwater resource managers (Doherty and Welter, 2010).

We also compared the residual $\mathcal{L}_2$ norm (Φ) at the four locations of primary interest to groundwater resource managers for both the history-matching and scenario simulations (**Figure 4**). In this light, we see that the PE analysis was able to reduce Φ for both the history-matching and scenario simulation ensembles, even though the scenario simulation outputs were not used in the PE analysis. We also note that for several realizations, the posterior Φ values are less than that of the existing Liu et al. (2017) models. The reduction in Φ across the ensemble under scenario conditions is attributable to the learning through PE about the static properties in the history-matching simulation and the subsequent transfer of these static properties to the scenario simulation.

An important aspect of PE is maintaining physically-plausible parameters (and corresponding simulation inputs). We have included several prior and posterior parameter realizations in the **Supplementary Material**. In general, the parameter changes resulting from the PE analysis are in agreement with the expected spatial and temporal patterns and are within the range of expectation.

The primary interest of groundwater resource managers is the number of low-flow/no-flow months at Comal springs during the scenario period. **Figure 5** shows the prior and posterior statistical distribution of simulated months with flow at Comal Springs less than $30\frac{ft^3}{s}$. The PE process has substantially reduced the uncertainty in this important simulated output. Specifically, the information in the state observations used for PE appears incompatible with prior realizations that yield more than 30 low-flow months. Or, put another way, parameter realizations that yield large values for Comal springs low-flow months during the scenario simulation do not fit the history-simulation observed states. Because of this incompatibility, the process of PE through history matching appears to be a valuable analysis to reduce uncertainty in the estimated Comal springs low-flow months. It is also important to note that while the posterior distributions

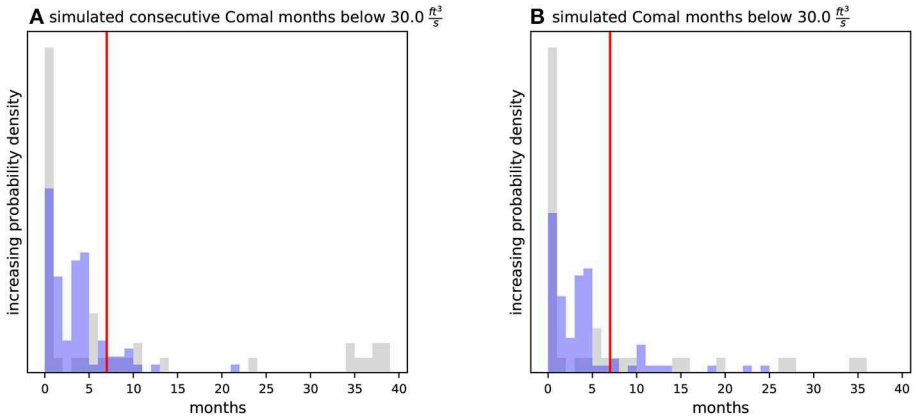


FIGURE 5 | Prior (gray) and posterior (blue) ensemble results for **(A)** the consecutive months below 30 $\frac{ft^3}{s}$ and **(B)** the total months below 30 $\frac{ft^3}{s}$ at Comal springs during the scenario simulation. The red vertical line marks the observed values (7 months for both). The PE analyses has reduced the number of extreme cases and has moved the ensemble closer to the observed values, however, considerable uncertainty remains. Note the X-axis scale was selected to focus on the posterior realizations; 37 prior realizations yielded greater than 40 consecutive months below 30 $\frac{ft^3}{s}$ and 44 prior realizations yielded greater than 40 total months below 30 $\frac{ft^3}{s}$; these prior realizations are not shown.

of both of these outputs bracket their respective observed values, both consecutive months and total months of low flow contain some posterior uncertainty and could range as high as 20 and 30 months, respectively.

5. DISCUSSION AND CONCLUSION

We have presented a demonstration of an approach to increase the reproducibility of UQ and PE for decision-support-scale groundwater modeling. This approach is predicated on the use of scripting to "drive" the modeling workflow. We recognize that not all environmental simulation practitioners will be proficient with scripting to the point that the approach we have demonstrated will be efficient. However, this approach offers many benefits, mostly toward increased transparency and reproducibility of the decision-support analyses—analyses that are typically at the center of the decision-making process. Furthermore, a script-based workflow affords increased efficiency when unforeseen factors arise that necessitate completing the analysis repeatedly. These "redos" are inevitable given the "ubiquity of errors" in computational science (e.g., Donoho et al., 2008) and occur, for example, when input errors are discovered or when the scope/purpose of the analysis changes). Greater reproducibility also makes the analysis more transferable to other projects and easier for third parties to evaluate or review the work (Kitzes et al., 2017).

To be clear, we are not stating that script-based analyses such as the one presented herein, will be free from input errors. While we have worked to implement the analyses herein as robustly and accurately as possible, the sheer number of operations and decisions required indicates that, in a statistical sense, there are faults or "bugs" in the script (and underlying modules) used to implement these analyses—the typical fault rate even in production-level software is between 15 and 50 faults per 1,000 lines (McConnell, 2004). However, in contrast to non-scripted modeling workflows, these "faults" can be identified and investigated by others practitioners long after the analysis has been completed—all the assumptions, decisions, and operations needed to implement our analysis are encoded transparently in the scripting workflow. This level of transparency and reproducibility has become a requirement in other fields—such as some omics cancer research—where the ramifications of decisions made in data processing can have life or death consequences (see e.g., Fienen and Bakker, 2016). Furthermore, as faults are discovered, they can be rectified programmatically in the script and the UQ and PE analyses can easily then be re-run, from beginning to end, without the complication of introducing new faults. In this way, while there is an initial "investment" to develop the scripting workflow, the returns on investment, as measured by efficiency and fidelity, are considerable.

The efficiency of the PE algorithm in PESTPP-IES has been shown to facilitate very high-dimensional (>300,000 parameters) history matching at a relatively low computational cost—the PE analysis required approximately 300 model evaluations, while the scenario prior and posterior Monte Carlo runs required approximately 100 model evaluations each. This efficiency allows practitioners to focus less on how model inputs are parameterized in the context of a computational trade-off and instead focus on expressing model input uncertainty as robustly as possible.

We realize that given the interest of groundwater resource managers in the hydrologic response to drought and the availability of state observations for PE during the scenario simulation, the scenario simulation could have also been subjected to PE (which could have easily been undertaken using our workflow and PESTPP-IES)—this would likely further reduce the posterior uncertainty in the outputs of primary interest to groundwater resource managers. However, in this study we are interested in evaluating the ability of UQ and PE to provide robust answers to management questions for which observations are not available (the more common use of environmental modeling).

DATA AVAILABILITY STATEMENT

We have placed all files and codes used for the analyses presented herein in an official U.S. Geological Survey Data Release available at https://doi.org/10.5066/P9AUZMI7 (White et al., 2020). The primary workflow driver script, `eaa.py`, has been annotated with extensive comments to guide readers. The project repository includes pre-compiled binaries for both MODFLOW-2005 and PESTPP-IES for Linux, PC and Mac. Additionally, the repository contains the versions of the FloPy and pyEMU modules that were used. The following software versions were used:

- python 3.7.3;
- numpy (Oliphant, 2006) 1.16.4;
- pandas (McKinney, 2012) 0.25.0;
- FloPy 3.2.13;
- pyEMU 0.8; and
- PESTPP-IES 4.2.15.

AUTHOR CONTRIBUTIONS

JWh, LF, and MF created the scripts and completed the analyses. JWh, MF, MK, and BH contributed to the development of pyEMU and FloPy that enabled the automated workflow herein and also contributed to the higher-level concepts around reproducibility through scripting. JWi provided the existing model files and also provided expert knowledge for defining the prior. JWh wrote the initial draft, however, all authors contributed to the writing of the manuscript.

FUNDING

The authors would like to acknowledge partial funding from the Edwards aquifer Authority under a Jointing Funding Agreement with the U.S. Geological Survey.

ACKNOWLEDGMENTS

The authors would like acknowledge all of the contributors and maintainers of FloPy, pyEMU, and PEST++. We would also like to acknowledge Nathan Pasley for making **Figure 1**.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feart.2020.00050/full#supplementary-material

The Supplementary Material included in this work includes:

- Parameter prior variance summary by group;
- Array-based multiplier parameterization summary for the first few realizations in the parameter ensemble for the active model area shown in figure 1 in main text.

REFERENCES

Anderson, M. P., Woessner, W. W., and Hunt, R. J. (2015). *Applied Groundwater Modeling, 2nd Edn*. San Diego, CA: Academic Press.

Bakker, M., Post, V., Langevin, C. D., Hughes, J. D., White, J. T., Starn, J. J., et al. (2016). Scripting modflow model development using python and flopy. *Groundwater* 54, 733–739. doi: 10.1111/gwat.12413

Barchard, K., and Pace, L. (2011). Preventing human error: the impact of data entry methods on data accuracy and statistical results. *Comput. Hum. Behav.* 27, 1834–1839. doi: 10.1016/j.chb.2011.04.004

Beven, K., and Binley, A. (1992). The future of distributed models: model calibration and uncertainty prediction. *Hydrol. Process.* 6, 279–298. doi: 10.1002/hyp.3360060305

Brakefield, L. K., White, J. T., Houston, N. A., and Thomas, J. V. (2015). *Updated Numerical Model With Uncertainty Assessment of 1950-56 Drought Conditions on Brackish-Water Movement Within the Edwards Aquifer, San Antonio, Texas*. U.S. Geological Survey Scientific Investigations Report 2015–5081. doi: 10.3133/sir20155081

Chen, Y., and Oliver, D. S. (2013). Levenberg–marquardt forms of the iterative ensemble smoother for efficient history matching and uncertainty quantification. *Comput. Geosci.* 17, 689–703. doi: 10.1007/s10596-013-9351-5

Chen, Y., and Oliver, D. S. (2016). Localization and regularization for iterative ensemble smoothers. *Comput. Geosci.* 21, 1–18. doi: 10.1007/s10596-016-9599-7

Doherty, J. (2003). Ground water model calibration using pilot points and regularization. *Ground Water* 41, 170–177. doi: 10.1111/j.1745-6584.2003.tb02580.x

Doherty, J. (2015a). *Calibration and Uncertainty Analysis for Complex Environmental Models - PEST: Complete Theory and What It Means for Modeling the Real World*. (Brisbane, QLD: Watermark Numerical Computing).

Doherty, J., and Christensen, S. (2011). Use of paired simple and complex models to reduce predictive bias and quantify uncertainty. *Water Resour. Res.* 47. doi: 10.1029/2011WR010763

Doherty, J., and Welter, D. (2010). A short exploration of structural noise. *Water Resour. Res.* 46. doi: 10.1029/2009WR008377

Doherty, J. E. (2015b). *PEST and Its Utility Support Software, Theory*. (Brisbane, QLD: Watermark Numerical Publishing).

Doherty, J. E., Hunt, R. J., and Tonkin, M. J. (2011). *Approaches to Highly Parameterized Inversion: A Guide to Using PEST for Model-Parameter and Predictive-Uncertainty Analysis*. Geological Survey Scientific Investigations Report 2010-5211.

Donoho, D. L., Maleki, A., Rahman, I. U., Shahram, M., and Stodden, V. (2008). Reproducible research in computational harmonic analysis. *Comput. Sci. Eng.* 11, 8–18. doi: 10.1109/MCSE.2009.15

Fienen, M. N., and Bakker, M. (2016). Hess opinions: Repeatable research: what hydrologists can learn from the duke cancer research scandal. *Hydrol. Earth Syst. Sci.* 20, 3739–3743. doi: 10.5194/hess-20-3739-2016

Fienen, M. N., and Hunt, R. J. (2015). High-throughput computing versus high-performance computing for groundwater applications. *Groundwater* 53, 180–184. doi: 10.1111/gwat.12320

Fisher, J. C. (2014). "*wrv: an R package for groundwater flow model construction, Wood River Valley Aquifer System, Idaho*," in *AGU Fall Meeting Abstracts* (San Francisco, CA).

Fisher, J. C., Bartolino, J. R., Wylie, A. H., Sukow, J., and McVay, M. (2016). *Groundwater-Flow Model for the Wood River Valley Aquifer System, South-Central Idaho*. in *U.S. Geological Survey Scientific Investigations Report 2016-5080* (Reston, VA), 84. doi: 10.3133/sir20165080

Goecks, J., Nekrutenko, A., and Taylor, J. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* 11:R86. doi: 10.1186/gb-2010-11-8-r86

Hanke, M. (1997). A regularizing levenberg-marquardt scheme, with applications to inverse groundwater filtration problems. *Inverse Prob.* 13:79. doi: 10.1088/0266-5611/13/1/007

Harbaugh, A. W. (2005). *MODFLOW-2005, the U.S. Geological Survey modular ground-water model – the Ground-Water Flow Process* (Reston, VA), Vol. 6. doi: 10.3133/tm6A16

Kitzes, J., Turek, D., and Deniz, F. (2017). *The Practice of Reproducible Research: Case Studies and Lessons From the Data-Intensive Sciences*. (Berkeley, CA: Univ of California Press).

Knowling, M. J., White, J. T., and Moore, C. R. (2019). Role of model parameterization in risk-based decision support: an empirical exploration. *Adv. Water Resour.* 128, 59–73. doi: 10.1016/j.advwatres.2019.04.010

Lindgren, R. J., Dutton, A. R., Hovorka, S. D., Worthington, S. R., and Painter, S. (2005). "Conceptualization and simulation of the Edwards aquifer, San Antonio region, Texas," in *Sinkholes and the Engineering and Environmental Impacts of Karst* (Reston, VA), 122–130. doi: 10.1061/40796(177)14

Liu, A., Troshanov, N., Winterle, J., Zhang, A., and Eason, S. (2017). *Updates to the Modflow Groundwater Model of the San Antonio Segment of the Edwards Aquifer*. in *Edwards Aquifer Authority Technical Report* (San Antonio, TX).

Liu, Z., Wang, S. P., and Meyer, D. (2019). Improving reproducibility in earth science research. *EOS* 100. doi: 10.1029/2019EO136216

McConnell, S. (2004). *Code Complete*. (Sebastopol, CA: Pearson Education).

McKenna, S. A., Akhriev, A., Ciaurri, D. E., and Zhuk, S. (2019). Efficient uncertainty quantification of reservoir properties for parameter estimation and production forecasting. *Math. Geosci.* 233–251. doi: 10.1007/s11004-019-09810-y

McKinney, W. (2012). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. O'Reilly Media.

Moore, C., and Doherty, J. E. (2005). Role of the calibration process in reducing model predictive error. *Water Resour. Res.* 41, 1–14. doi: 10.1029/2004WR003501

Oliphant, T. E. (2006). *Guide to NumPy*. Provo, UT.

Olsthoorn, T. (2010). "Improved groundwater modeling using an open and free environment called Mflab," in *NGWA 2010 Groundwater Summit* (Denver, CO).

Peng, R. D. (2011). Reproducible research in computational science. *Science* 334, 1226–1227. doi: 10.1126/science.1213847

Plesser, H. E. (2018). Reproducibility vs. replicability: a brief history of a confused terminology. *Front. Neuroinformat.* 11:76. doi: 10.3389/fninf.2017.00076

Puente, C. (1978). Method of estimating natural recharge to the edwards aquifer in the san antonio area, texas. *Sci. Investigat Rep*. doi: 10.3133/wri7810. [Epub ahead of print].

Sandve, G. K., Nekrutenko, A., Taylor, J., and Hovig, E. (2013). Ten simple rules for reproducible computational research. *Pub. Lib. Sci. Comput. Biol.* 9. doi: 10.1371/journal.pcbi.1003285

Stodden, V. (2010). "The scientific method in practice: reproducibility in the computational sciences," in *MIT Sloan Research Paper No. 4773-10*. 33. doi: 10.2139/ssrn.1550193

Tarantola, A. (2005). *Inverse Problem Theory and Methods for Model Parameter Estimation*. (Philadelphia, PA: SIAM). doi: 10.1137/1.9780898717921

Thain, D., Tannenbaum, T., and Livny, M. (2005). Distributed computing in practice: the condor experience. *Concur. Pract. Exp.* 17, 323–356. doi: 10.1002/cpe.938

White, J. T. (2018). A model-independent iterative ensemble smoother for efficient history-matching and uncertainty quantification in very high dimensions. *Environ. Model. Softw.* 109, 191–201. doi: 10.1016/j.envsoft.2018.06.009

White, J. T., Doherty, J. E., and Hughes, J. D. (2014). Quantifying the predictive consequences of model error with linear subspace analysis. *Water Resour. Res.* 50, 1152–1173. doi: 10.1002/2013WR014767

White, J. T., Fienen, M. N., and Doherty, J. E. (2016). A python framework for environmental model uncertainty analysis. *Environ. Model. Softw.* 85, 217–228. doi: 10.1016/j.envsoft.2016.08.017

White, J. T., Foster, L. K., Fienen, M. N., Knowling, M. J., Hemmings, B., and Winterle, J. R. (2020). *Data release for "towards reproducible environmental modeling for decision support: a worked example,"* (Reston, VA: U.S.Geological Survey Data Release).

Disclaimers: Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

This software has been approved for release by the U.S. Geological Survey (USGS). Although the software has been subjected to rigorous review, the USGS reserves the right to update the software as needed pursuant to further analysis and review. No warranty, expressed or implied, is made by the USGS or the U.S. Government as to the functionality of the software and related material nor shall the fact of release constitute any such warranty. Furthermore, the software is released on condition that neither the USGS nor the U.S. Government shall be held liable for any damages resulting from its authorized or unauthorized use.

Although these data have been processed successfully on a computer system at the U.S. Geological Survey (USGS), no warranty expressed or implied is made regarding the display or utility of the data for other purposes, nor on all computer systems, nor shall the act of distribution constitute any such warranty. The USGS or the U.S. Government shall not be held liable for improper or incorrect use of the data described and/or contained herein.

Conflict of Interest: MK and BH were employed by the GNS Science, a commercial entity owned by the New Zealand government.

The remaining 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.

frontiers in Earth Science

ORIGINAL RESEARCH
published: 27 May 2020
doi: 10.3389/feart.2020.00108

Efficient Ensemble-Based Stochastic Gradient Methods for Optimization Under Geological Uncertainty

Hoonyoung Jeong[1,2,3]*, *Alexander Y. Sun*[4], *Jonghyeon Jeon*[5], *Baehyun Min*[6] and *Daein Jeong*[7]

[1] Department of Energy Resources Engineering, Seoul National University, Seoul, South Korea, [2] Research Institute of Energy and Resources, Seoul National University, Seoul, South Korea, [3] Institute of Engineering Research, Seoul National University, Seoul, South Korea, [4] Bureau of Economic Geology, Jackson School of Geosciences, The University of Texas at Austin, Austin, TX, United States, [5] Department of Petroleum and Geosystems Engineering, Cockrell School of Engineering, The University of Texas at Austin, Austin, TX, United States, [6] Department of Climate and Energy Systems Engineering, Ewha Womans University, Seoul, South Korea, [7] Schlumberger Software Integrated Solutions, Tokyo, Japan

Ensemble-based stochastic gradient methods, such as the ensemble optimization (EnOpt) method, the simplex gradient (SG) method, and the stochastic simplex approximate gradient (StoSAG) method, approximate the gradient of an objective function using an ensemble of perturbed control vectors. These methods are increasingly used in solving reservoir optimization problems because they are not only easy to parallelize and couple with any simulator but also computationally more efficient than the conventional finite-difference method for gradient calculations. In this work, we show that EnOpt may fail to achieve sufficient improvement of the objective function when the differences between the objective function values of perturbed control variables and their ensemble mean are large. On the basis of the comparison of EnOpt and SG, we propose a hybrid gradient of EnOpt and SG to save on the computational cost of SG. We also suggest practical ways to reduce the computational cost of EnOpt and StoSAG by approximating the objective function values of unperturbed control variables using the values of perturbed ones. We first demonstrate the performance of our improved ensemble schemes using a benchmark problem. Results show that the proposed gradients saved about 30–50% of the computational cost of the same optimization by using EnOpt, SG, and StoSAG. As a real application, we consider pressure management in carbon storage reservoirs, for which brine extraction wells need to be optimally placed to reduce reservoir pressure buildup while maximizing the net present value. Results show that our improved schemes reduce the computational cost significantly.

Keywords: stochastic gradient, ensemble optimization, simplex gradient, stochastic simplex approximate gradient, hybrid simplex gradient, active pressure management

OPEN ACCESS

Edited by:
Liangping Li,
South Dakota School of Mines and Technology, United States

Reviewed by:
Xiaodong Luo,
NORCE Norwegian Research Centre, Norway
Devesh Kumar,
Chevron, United States

***Correspondence:**
Hoonyoung Jeong
hoonyoung.jeong@snu.ac.kr

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

Received: *24 December 2019*
Accepted: *24 March 2020*
Published: *27 May 2020*

Citation:
Jeong H, Sun AY, Jeon J, Min B and Jeong D (2020) Efficient Ensemble-Based Stochastic Gradient Methods for Optimization Under Geological Uncertainty. Front. Earth Sci. 8:108. doi: 10.3389/feart.2020.00108

INTRODUCTION

Since the ensemble Kalman filter was first introduced into the petroleum engineering (Lorentzen et al., 2001; Nævdal et al., 2002; Kim et al., 2018), many ensemble-based history matching methods have gained popularity because they are reduced rank methods (meaning less computational effort) and are relatively easy to implement, parallelize, and couple with any numerical simulator. Chen et al. (2009) first systematically applied the ensemble concept to optimization of well control

variables (e.g., well rates and bottom-hole pressures) to maximize the net present value in oil and gas fields. They named their scheme the ensemble optimization (EnOpt) method. Similar to the ensemble-based data assimilation methods, EnOpt can also be easily parallelized and coupled with any simulator.

Another strength of EnOpt is that EnOpt finds an optimal solution under geological uncertainty by maximizing the expectation of the objective function values of multiple models representing model uncertainties, whereas the conventional optimization methods typically require solving each model separately (Chen et al., 2009; van Essen et al., 2009) and optimization under uncertainty is non-trivial (Sun et al., 2013; Zhang et al., 2016). The idea of considering model uncertainties in optimization was also explored by van Essen et al. (2009). van Essen et al. (2009) named their method robust optimization, and the handling of model uncertainties in their method is essentially the same as that in EnOpt. However, EnOpt includes a specific way to compute the gradient, which is needed by all gradient-based optimization algorithms.

The gradient of EnOpt is determined on the basis of the cross covariance between randomly perturbed control variables (or decision variables) and the corresponding objective function values. Because this work is mainly concerned with the gradient approximation in various ensemble methods, hereafter, we will use EnOpt to refer to the gradient approximation in EnOpt where no confusion occurs. Traditional methods for gradient calculation include the finite-difference method (FDM) and adjoint-state method (Sun and Sun, 2015). FDM needs as many objective function evaluations as the product of the number of control variables and the number of ensemble members because FDM perturbs each control variable separately. The adjoint-state method typically requires derivation and solution of a dual problem of the original problem in the adjoint state space, which is not straightforward. In comparison, EnOpt only requires as many objective function evaluations as the number of ensemble members, because it computes the search directions by averaging the objective function anomalies resulting from simultaneous random perturbations of the control vector. Previous studies have shown that the EnOpt can efficiently and satisfactorily achieve improvement of the objective function, despite its low computational cost (Chen et al., 2009; Chen and Oliver, 2010, 2012). However, Fonseca et al. (2017) indicated that EnOpt may not produce satisfactory results for multiple geological models unless the variance in the ensemble models is sufficiently small. The first objective of our work is to show mathematically and experimentally why EnOpt may fail to produce satisfactory results when the variance of the ensemble models is not small.

EnOpt can be considered a variant of the simultaneous perturbation stochastic approximation (SPSA) method introduced by Spall (1992), and SPSA is appropriate for robust optimization because the computational cost of SPSA is significantly lower than that of FDM for a high-dimensional control vector. Even though accurate gradients are obtained using FDM at a high computational cost, gradient-based optimizations are likely to converge to local optima. Rather than spending considerable computational resources computing the gradients, it is more practical to find global optima by trying many initial solutions using the less accurate but more computationally efficient SPSA. SPSA computes the gradient of an objective function more efficiently than FDM does by perturbing control variables randomly and simultaneously (Spall, 1992, 1998). There are several variants of SPSA that can be used to compute the gradient of an objective function stochastically and quickly.

Bangerth et al. (2006) introduced the integer SPSA to solve an optimal well placement problem. Li et al. (2013) applied SPSA for joint optimization of well placement and controls under geological uncertainty. Li and Reynolds (2011) proposed a modification of the SPSA, which is called the stochastic Gaussian search direction (SGSD or G-SPSA). The original SPSA samples perturbations from a symmetric Bernoulli distribution, while SGSD and EnOpt generate perturbations from Gaussian distributions (Chen et al., 2009; Li and Reynolds, 2011). Do and Reynolds (2013) used the simplex gradient (SG) that has the perturbation coefficient of 1 in the formulation of SGSD.

However, Bangerth et al. (2006), Li et al. (2013), Li and Reynolds (2011), and Do and Reynolds (2013) applied the variants of SPSA to optimization of a single geologic model, which means that geological uncertainty was not considered. Fonseca et al. (2017) proposed an extension of SG, which is named the stochastic simplex approximate gradient (StoSAG), that improves the accuracy of the stochastic gradient by repeating multiple perturbations for each ensemble model. In this study, we propose practical ways to reduce the computational cost of EnOpt, SG, and StoSAG by approximating the objective function values of unperturbed control variables using those obtained for the perturbed ones. The proposed approaches reduce about 10% to 50% of the computational cost compared to EnOpt, SG, and StoSAG in our examples.

This paper is organized as follows. In the next section, we explain why EnOpt may fail when the variance of objective function values of the ensemble members is not small, by comparing the gradient approximation schemes in the original EnOpt and SG. Then we propose new hybrid schemes for further reducing computational costs in EnOpt, SG, and StoSAG. Finally, we demonstrate the efficacy of the different schemes using two examples, a test function that is popular for algorithm benchmarking and a well placement optimization problem for pressure management in geologic carbon storage reservoirs.

COMPARISON OF ENOPT AND SG

The steepest ascent or descent algorithm to maximize or minimize an objective function $J(\mathbf{u})$ is given as

$$\mathbf{u}_{k+1} = \mathbf{u}_k \pm \alpha_k \frac{\mathbf{d}_k}{\|\mathbf{d}_k\|_\infty}, \text{ for } k = 0, 1, \cdots \text{ until convergence,} \quad (1)$$

where $\mathbf{u}$ is the column vector of control variables; $\mathbf{u}_0$ is the initial guess; $\mathbf{d}$ is the search direction; α is the step size; and k is an iteration index. For convenience, all major notations used in this study are listed in the **Nomenclature** table attached at the end of the paper. In this problem, the objective function (J) is dependent only on $\mathbf{u}$. van Essen et al. (2009) replaced $J(\mathbf{u})$

with $J(\mathbf{m}, \mathbf{u})$ by adding another input ($\mathbf{m}$) to the scalar function, where $\mathbf{m}$ is a random vector generated from a known probability density function. In optimization of well placement and controls, for instance, $\mathbf{m}$ may represent uncertain rock properties arising from geologic heterogeneity. $J(\mathbf{m}, \mathbf{u})$ is dependent on both $\mathbf{m}$ and $\mathbf{u}$, but only $\mathbf{u}$ is a control vector. Because $J(\mathbf{m}, \mathbf{u})$ inherits the uncertainty of $\mathbf{m}$, van Essen et al. (2009) suggested to maximize the approximate expectation of $J(\mathbf{m}, \mathbf{u})$ for $\mathbf{m}_i$ ($i = 1, 2, 3, \ldots, N_e$) sampled from a given probability density function (Fonseca et al., 2017).

$$\max_{\mathbf{u}} \mathrm{E}_{\mathbf{m}}\left[J\left(\mathbf{m}, \mathbf{u}\right)\right] = \max_{\mathbf{u}} \frac{\mathbf{1}}{N_e} \sum_{i=1}^{N_e} J\left(\mathbf{m}_i, \mathbf{u}\right), \tag{2}$$

where $1/N_e \sum_{i=1}^{N_e} J\left(\mathbf{m}_i, \mathbf{u}\right)$ and $J(\mathbf{m}_i, \mathbf{u})$ are called the objective function and the J-function, respectively, to avoid confusion. The former is marginalized over all ensemble members, while the latter corresponds to a single ensemble member.

Ensemble-based gradients (EnOpt, SG, and StoSAG) generate the objective function anomalies by stochastically sampling perturbations for $\mathbf{u}$ from a multivariate Gaussian distribution with mean $\mathbf{u}_k$ and covariance matrix $\mathbf{C_u}$. The perturbed vector of $\mathbf{u}$ at the kth iteration is denoted by $\hat{\mathbf{u}}_k$. In the following, we explain situations that EnOpt may underperform. We then proceed to introduce several hybrid schemes that can significantly reduce the computational costs of the ensemble-based optimization in general while mitigating the underperformance of EnOpt.

SG is given by Do and Reynolds (2013) as.

$$\mathbf{d}_{k,SG} = \frac{1}{N_e} \sum_{i=1}^{N_e} \left(\hat{\mathbf{u}}_{k,i} - \mathbf{u}_k\right) \left(J\left(\mathbf{m}_i, \hat{\mathbf{u}}_{k,i}\right) - J\left(\mathbf{m}_i, \mathbf{u}_k\right)\right). \tag{3}$$

Equation (3) presented by Do and Reynolds (2013) was applied to a single model. However, once the objective function is replaced with Equation (2) in the formulation of Do and Reynolds (2013), Equation (3) can be applied to multiple models (or robust optimization). Equation (3) is identical to StoSAG with a single repetitive perturbation. Hereafter, we will use SG to refer to StoSAG with a single repetitive perturbation.

The search direction of EnOpt is given by Chen et al. (2009).

$$\mathbf{d}_{k,EnOpt} = \frac{1}{N_e - 1} \sum_{i=1}^{N_e} \left(\hat{\mathbf{u}}_{k,i} - \overline{\hat{\mathbf{u}}_k}\right) \left(J\left(\mathbf{m}_i, \hat{\mathbf{u}}_{k,i}\right) - \overline{J\left(\mathbf{m}, \hat{\mathbf{u}}_k\right)}\right), \tag{4}$$

$$\mathbf{u}_k \approx \overline{\hat{\mathbf{u}}_k} = \frac{1}{N_e} \sum_{i=1}^{N_e} \hat{\mathbf{u}}_{k,i}, \tag{5}$$

$$J\left(\mathbf{m}_i, \mathbf{u}_k\right) \approx \overline{J\left(\mathbf{m}, \hat{\mathbf{u}}_k\right)} = \frac{1}{N_e} \sum_{i=1}^{N_e} J\left(\mathbf{m}_i, \hat{\mathbf{u}}_{k,i}\right). \tag{6}$$

Compared to Equation (3), the unperturbed vector of control variables ($\mathbf{u}_k$) and the corresponding $J(\mathbf{m}_i, \mathbf{u}_k)$ are approximated by their corresponding ensemble means ($\overline{\hat{\mathbf{u}}_k}$ and $\overline{J\left(\mathbf{m}, \hat{\mathbf{u}}_k\right)}$) in Equation (4) that are further defined as in Equations (5) and (6) (Chen et al., 2009; Do and Reynolds, 2013; Fonseca et al., 2017). In Equation (4), EnOpt needs N_e J-function evaluations to compute a search direction. However, EnOpt requires additional N_e J-function evaluations to calculate the expectation $\mathrm{E}_{\mathbf{m}}[J(\mathbf{m}, \mathbf{u})]$ in Equation (2). In Equations (2) and (4), SG needs $2N_e$ J-function evaluations to compute a search direction. Thus, SG and EnOpt essentially take the same computational effort ($2N_e$) to compute a search direction and $\mathrm{E}_{\mathbf{m}}[J(\mathbf{m}, \mathbf{u})]$.

Do and Reynolds (2013) demonstrated that the performances of EnOpt and SG are almost identical for a single geological model. However, Fonseca et al. (2017) pointed out that EnOpt may produce unsatisfactory results because the two assumptions given in Equations (5) and (6) are invalid unless N_e is sufficiently large or the variance in the prior model for $\mathbf{m}$ is sufficiently small. Equation (5) is usually valid because perturbations for $\mathbf{u}_k$ should be sufficiently small to approximate $\nabla_{\mathbf{u}} \mathrm{E}_{\mathbf{m}}[J(\mathbf{m}, \mathbf{u})]$. Equation (6) is likely to be invalid if high variations in model parameters such as permeability and porosity cause large variations in J-function values. However, even though the variance of $\mathbf{m}$ is small, other variables in the J-function such as unit costs may still make large variations in the J-function values. The claim of Fonseca et al. (2017) can be expressed more quantitatively by simply manipulating Equation (4):

$$\begin{aligned}
\mathbf{d}_{k,EnOpt} &= \frac{1}{N_e - 1} \sum_{i=1}^{N_e} \left(\hat{\mathbf{u}}_{k,i} - \overline{\hat{\mathbf{u}}_k}\right) \left(J\left(\mathbf{m}_i, \hat{\mathbf{u}}_{k,i}\right)\right. \\
&\quad \left. - \overline{J\left(\mathbf{m}, \hat{\mathbf{u}}_k\right)}\right) \\
&= \frac{N_e}{N_e - 1} \frac{1}{N_e} \sum_{i=1}^{N_e} \left(\hat{\mathbf{u}}_{k,i} - \overline{\hat{\mathbf{u}}_k}\right) \left(J\left(\mathbf{m}_i, \hat{\mathbf{u}}_{k,i}\right) - J\left(\mathbf{m}_i, \mathbf{u}_k\right)\right. \\
&\quad \left. + J\left(\mathbf{m}_i, \mathbf{u}_k\right) - \overline{J\left(\mathbf{m}, \hat{\mathbf{u}}_k\right)}\right) \\
&= \frac{N_e}{N_e - 1} \left[\frac{1}{N_e} \sum_{i=1}^{N_e} \left(\hat{\mathbf{u}}_{k,i} - \overline{\hat{\mathbf{u}}_k}\right) \left(J\left(\mathbf{m}_i, \hat{\mathbf{u}}_{k,i}\right) - J\left(\mathbf{m}_i, \mathbf{u}_k\right)\right)\right. \\
&\quad \left. + \frac{1}{N_e} \sum_{i=1}^{N_e} \left(\hat{\mathbf{u}}_{k,i} - \overline{\hat{\mathbf{u}}_k}\right) \left(J\left(\mathbf{m}_i, \mathbf{u}_k\right) - \overline{J\left(\mathbf{m}, \hat{\mathbf{u}}_k\right)}\right)\right] \\
&= \frac{N_e}{N_e - 1} \left[\mathbf{d}_{k,SG} + \frac{1}{N_e} \sum_{i=1}^{N_e} \left(\hat{\mathbf{u}}_{k,i} - \overline{\hat{\mathbf{u}}_k}\right) \left(J\left(\mathbf{m}_i, \mathbf{u}_k\right)\right.\right. \\
&\quad \left.\left. - \overline{J\left(\mathbf{m}, \hat{\mathbf{u}}_k\right)}\right)\right].
\end{aligned} \tag{7}$$

From Equation (7), it follows that

$$\begin{aligned}
\mathbf{r}_k &= \frac{1}{N_e} \sum_{i=1}^{N_e} \left(\hat{\mathbf{u}}_{k,i} - \overline{\hat{\mathbf{u}}_k}\right) \left(J\left(\mathbf{m}_i, \mathbf{u}_k\right) - \overline{J\left(\mathbf{m}, \hat{\mathbf{u}}_k\right)}\right), \\
\mathbf{d}_{k,EnOpt} &= \frac{N_e}{N_e - 1} \left[\mathbf{d}_{k,SG} + \mathbf{r}_k\right], \\
\|\mathbf{r}_k\|_2 &= \left\| \frac{N_e - 1}{N_e} \mathbf{d}_{k,EnOpt} - \mathbf{d}_{k,SG} \right\|_2,
\end{aligned} \tag{8}$$

where $\mathbf{d}_{k,EnOpt}/||\mathbf{d}_{k,EnOpt}||_\infty \approx \mathbf{d}_{k,SG}/\mathbf{d}_{k,SG}||_\infty$ is true if $||\mathbf{r}_k||_2$ is sufficiently small compared to $\mathbf{d}_{k,SG}$. However, as $||\mathbf{r}_k||_2$ increases,

$\mathbf{d}_{k,EnOpt}$ becomes more inaccurate than $\mathbf{d}_{k,SG}$. In $\mathbf{r}_k$ in Equation. (8), because perturbations for $\mathbf{u}_k$ should be small enough to approximate a search direction accurately, $||\mathbf{r}_k||_2$ is significantly dependent on $J(\mathbf{m}_i, \mathbf{u}_k) - \overline{J(\mathbf{m}, \hat{\mathbf{u}}_k)}$. Thus, $||\mathbf{r}_k||_2$ is closely related to the variance in $J(\mathbf{m}, \mathbf{u}_k)$, which is $\sigma^2_{J(\mathbf{m},\mathbf{u}_k)}$ as given in Equation (9):

$$\sigma^2_{J(\mathbf{m},\mathbf{u}_k)} = \frac{1}{N_e - 1}\sum_{i=1}^{N_e}\left[J(\mathbf{m}_i, \mathbf{u}_k) - \overline{J(\mathbf{m}, \mathbf{u}_k)}\right]^2,$$

$$\overline{J(\mathbf{m}, \mathbf{u}_k)} = \frac{1}{N_e}\sum_{i=1}^{N_e} J(\mathbf{m}_i, \mathbf{u}_k). \tag{9}$$

The norm $||\mathbf{r}_k||_2$ is not exactly the same as the variance in $J(\mathbf{m}, \mathbf{u}_k)$, but $||\mathbf{r}_k||_2$ is expected to increase as $\sigma^2_{J(\mathbf{m},\mathbf{u}_k)}$ increases because $\overline{J(\mathbf{m}, \mathbf{u}_k)} \approx \overline{J(\mathbf{m}, \hat{\mathbf{u}}_k)}$. On the basis of the observation that the approximation of EnOpt becomes inaccurate as the variance in $J(\mathbf{m}_i, \mathbf{u}_k)$ increases, we propose a hybrid gradient of EnOpt and SG in the next section that first clusters ensemble members based on the variance of $J(\mathbf{m}_i, \hat{\mathbf{u}}_k)$ and then approximate $J(\mathbf{m}_i, \mathbf{u}_k)$ using $J(\mathbf{m}_i, \hat{\mathbf{u}}_k)$ within each cluster. Thus, the hybrid gradient is more computationally efficient than SG because the objective function for the unperturbed control vector in some ensemble members does not need to be evaluated. We also propose two additional practical measures that can save the computational cost of EnOpt and StoSAG.

COMPUTATIONAL COST REDUCTION OF ENOPT, SG, AND STOSAG

Here, we introduce three new formulations to save on the computational cost of EnOpt, SG, and StoSAG. The new formulations approximate the objective function values for unperturbed control variables using the objective function values for perturbed ones in the formulations of EnOpt, SG, and StoSAG. Thus, these new formulations require fewer J-function evaluations than that by the original ensemble-based gradients.

In EnOpt, the J-function for unperturbed control variables does not need to be evaluated for the search direction, but it needs to be evaluated for the objective function in Equation (2). However, the means of $J(\mathbf{m}_i, \mathbf{u}_k)$ and $J(\mathbf{m}_i, \hat{\mathbf{u}}_{k,i})$ are similar because the perturbations on $\mathbf{u}$ are small. Thus, $\mathrm{E}_{\mathbf{m}}[J(\mathbf{m}, \mathbf{u})]$ can be approximated using $J(\mathbf{m}_i, \hat{\mathbf{u}}_{k,i})$ as given in Equation (10):

$$\max_{\mathbf{u}} \mathbf{E}_{\mathbf{m}}[J(\mathbf{m}, \mathbf{u})] \approx \max_{\mathbf{u}} \frac{1}{N_e}\sum_{i=1}^{N_e} J(\mathbf{m}_i, \hat{\mathbf{u}}_{k,i}). \tag{10}$$

We call this the modified EnOpt (ModEnOpt), which uses Equation (10) instead of Equation (2) for approximating $\mathrm{E}_{\mathbf{m}}[J(\mathbf{m}, \mathbf{u})]$. ModEnOpt requires only half of the number of the J-function evaluations of EnOpt to approximate a search direction and calculate $\mathrm{E}_{\mathbf{m}}[J(\mathbf{m}, \mathbf{u})]$.

To save the computational cost of SG, we propose a hybrid gradient of EnOpt and SG, which is named the hybrid simplex gradient (HSG) method, on the basis of the observation that EnOpt provides satisfactory search directions if the variance in $J(\mathbf{m}, \mathbf{u}_k)$ is small. HSG clusters the ensemble members based on $J(\mathbf{m}_i, \hat{\mathbf{u}}_{k,i})$ and then uses the cluster mean instead of $J(\mathbf{m}_i, \mathbf{u}_k)$ for clusters that have more than one member as given in the first term of Equation (11), which is close to the search direction of EnOpt given in Equation (4). In the second term of Equation (11), $J(\mathbf{m}_i, \mathbf{u}_k)$ should be evaluated for the clusters that have only a single member, which is close to the search direction of SG given in Equation (3). The search direction of HSG is given by

$$\mathbf{d}_{k,HSG} = \frac{1}{N_e}\sum_{j=1}^{N_C}\left[\sum_{i\in C_j, N_{C_j}>1}\left(\hat{\mathbf{u}}_{k,i} - \overline{\hat{\mathbf{u}}_{kj}}\right)\left(J(\mathbf{m}_i, \hat{\mathbf{u}}_{k,i}) - \overline{J(\mathbf{m}_i, \hat{\mathbf{u}}_k)}_j\right)\right.$$
$$\left. + \sum_{i\in C_j, N_{C_j}=1}\left(\hat{\mathbf{u}}_{k,i} - \mathbf{u}_k\right)\left(J(\mathbf{m}_i, \hat{\mathbf{u}}_{k,i}) - J(\mathbf{m}_i, \mathbf{u}_k)\right)\right] \tag{11}$$

HSG uses a different approximation of $\mathrm{E}_{\mathbf{m}}[J(\mathbf{m}, \mathbf{u})]$ given in Equation (12), instead of Equation (2):

$$\max_{\mathbf{u}} \mathrm{E}_{\mathbf{m}}[J(\mathbf{m}, \mathbf{u})] \approx \max_{\mathbf{u}} \frac{1}{N_e}\sum_{j=1}^{N_C}\left[\sum_{i\in C_j, N_{C_j}>1} J(\mathbf{m}_i, \hat{\mathbf{u}}_{k,i})\right.$$
$$\left. + \sum_{i\in C_j, N_{C_j}=1} J(\mathbf{m}_i, \mathbf{u}_k)\right]. \tag{12}$$

In Equation (12), the J-function values for perturbed control variables are used to approximate $\mathrm{E}_{\mathbf{m}}[J(\mathbf{m}, \mathbf{u})]$, but the J-function values for unperturbed ones are used for clusters that have only a single member.

In Equation (11), the N_e ensemble members (models) are grouped based on $[J(\boldsymbol{m_i}, \hat{\mathbf{u}}_{\boldsymbol{k,i}})]$. However, determining the optimal number of clusters is still a challenging problem in data clustering (Jain, 2010). Furthermore, even though models are grouped to the optimal number of clusters, some groups might have significantly different $J(\mathbf{m}_i, \hat{\mathbf{u}}_{k,i})$ values because cluster algorithms group the models into the number of clusters unexceptionally regardless of how similar $J(\mathbf{m}_i, \hat{\mathbf{u}}_{k,i})$ values are in a group. For this reason, a mean of $J(\mathbf{m}_i, \hat{\mathbf{u}}_{k,i})$ in a group might not be properly representative of $J(\mathbf{m}_i, \mathbf{u}_k)$ of the group members. Thus, rather than trying to determine the optimal number of clusters, we use a predefined criterion to determine if models have similar $J(\mathbf{m}_i, \hat{\mathbf{u}}_{k,i})$. A standard deviation is the most common indicator of how dissimilar data are, but the standard deviation should be normalized in our clustering problem. For example, let us assume that there are two data sets (1, 2, 3) and (198, 200, 202). The standard deviation of (1, 2, 3) is smaller than that of (198, 200, 202), but (198, 200, 202) has relatively small differences in terms of magnitude compared to (1, 2, 3). In our clustering problem, group members should have relatively similar J-function values within each group. Thus, for the predefined criterion, we use the coefficient of variation shown

in Equation (13) instead of standard deviation to normalize the standard deviation.

$$\text{coefficient of variation (CV)} = \frac{\text{standard deviation}}{\text{mean}} \tag{13}$$

Thus, the N_e models are grouped so that the coefficient of variation in each cluster is smaller than a predefined value (CV_{HSG}). However, during clustering, a model should be assigned to a cluster such that the coefficient of variation becomes minimal. We introduce an algorithm to find groups that have small coefficient of variations that are lower than a predefined coefficient of variation. First, N_e models have random cluster indices where the initial number of groups (N_c) is the same to the number of models (N_e). Then whether the coefficient of variation of a group can be reduced by adding a model to the group is examined where the model makes the coefficient of variation of the group minimum. For example, let us assume that we try to select a cluster for a model between clusters A and B. The coefficients of variation of both A and B are smaller than those of CV_{HSG}. If the coefficients of variation of A and B (after including the model) are 0.001 and 0.002, respectively, then the model should be put in cluster A. This is repeated until the coefficient of variation of the group does not become smaller. Finding models that make the coefficient of variation of other groups smaller is repeated. Other details of the proposed algorithm are described in Algorithm 1. In HSG, we use Algorithm 1 to make the coefficients of variation of clusters smaller than CV_{HSG} and to drop the coefficients of variation of clusters. The procedure of clustering is given as follows:

The input of Algorithm 1 includes a predefined coefficient of variation, CV_{HSG}, and the J-function values for perturbed control variables. The number of clusters does not need to be inputted for Algorithm 1. The k-means clustering algorithm (MacQueen, 1967), which is one of the most commonly used clustering algorithms, cannot be used in this case because it requires the number of clusters to be specified and it tends to group members that are relatively close to each other. Determination of the CV_{HSG} value depends on how much computational cost of HSG is expected to be saved compared to SG. For example, if 70% of the computational cost is expected to be saved using HSG compared to SG, then CV_{HSG} is set to a number that makes the number of clusters 70% of the number of ensemble members. CV_{HSG} can be chosen based on the initial J-function values of an ensemble.

HSG is equivalent to ModEnOpt and SG for large and small CV_{HSG}, respectively, where the sum is divided by $N_e - 1$ in $\mathbf{d}_k$, $_{ModEnOpt}$, but this is canceled by $||\mathbf{d}_k$, $_{ModEnOpt}||_\infty$ in Equation (1). For large $||\mathbf{r}_k||_2$, a small CV_{HSG} should be used because the approximate gradient of ModEnOpt is inaccurate. HSG takes $N_e \sim 2N_e$ J-function evaluations where N_e and $2N_e$ correspond to the number of J-function evaluations of ModEnOpt and SG, respectively.

Algorithm 1 : Clustering using a predefined coefficient of variations for HSG (CV_{HSG})

Sort $1, \ldots, N_e$ randomly and assign the randomly sorted numbers to the cluster indices of N_e ensemble members (=*c_ind*) where the initial number of clusters (N_c) is the same as the number of models (N_e)
Mark *clustering_done* as false for N_e ensemble members
$j = 1$
While $j \leq N_c$
    Find ensemble members that belong to C_j and of which *clustering_done* is false
If the number of the ensemble members == 0
j++
    Else
        temp_cv = [inf, inf, …, inf] //size of *temp_cv* = N_e
For $m = 1$ to N_e
If m does not belong to C_j and clustering for m is not done
//*c_ind*[m] $\neq$ *c_ind*[j] and *clusetering_done*[m] == false
temp_cv[m] = coefficient of variation of perturbed objective function values of C_j members and m
    End
    For-Loop
Find an ensemble member corresponding to the minimum of *temp_cv*
If the minimum coefficient of variation $\leq CV_{HSG}$
Add the ensemble member to C_j
Else
Mark *clustering_done* as true for the ensemble member
j++
        End
    End
While-Loop

The search direction of StoSAG is given by Fonseca et al. (2017) as

$$\mathbf{d}_{k,StoSAG} = \frac{1}{N_e}\sum_{i=1}^{N_e}\left[\frac{1}{N_p}\sum_{j=1}^{N_p}\left(\hat{\mathbf{u}}_{k,i,j} - \mathbf{u}_k\right)\left(J\left(\mathbf{m}_i, \hat{\mathbf{u}}_{k,i,j}\right) - J\left(\mathbf{m}_i, \mathbf{u}_k\right)\right)\right]. \tag{14}$$

StoSAG repeats multiple perturbations (N_p) for each ensemble member, while SG takes a single perturbation for each ensemble member. StoSAG provides more accurate search directions than SG does because StoSAG takes more J-function evaluations to compute the search direction than SG. StoSAG needs $N_e(N_p + 1)$ J-function evaluations, and this is $(N_p + 1)/2$ times as many J-function evaluations as SG requires.

The computational cost of StoSAG can be reduced by approximating the J-function values of unperturbed control variables using the J-function values of perturbed ones as given below:

$$\mathbf{d}_{k,ModStoSAG} = \frac{1}{N_e}\sum_{i=1}^{N_e}\Big[\frac{1}{N_p}\sum_{j=1}^{N_p}\left(\hat{\mathbf{u}}_{k,i,j}-\mathbf{u}_k\right)\left(J\left(\mathbf{m}_i,\hat{\mathbf{u}}_{k,i,j}\right)-\overline{J\left(\mathbf{m}_i,\hat{\mathbf{u}}_{k,i}\right)}\right)\Big], \tag{15}$$

$$J\left(\mathbf{m}_i,\mathbf{u}_k\right) \approx \overline{J\left(\mathbf{m}_i,\hat{\mathbf{u}}_{k,i}\right)} = \frac{1}{N_p}\sum_{j=1}^{N_p} J\left(\mathbf{m}_i,\hat{\mathbf{u}}_{k,i,j}\right). \tag{16}$$

The modified StoSAG (ModStoSAG) uses different approximations of $\mathrm{E}_{\mathbf{m}}[J(\mathbf{m},\mathbf{u})]$ given in Equation (17) instead of Equation (2):

$$\max_{\mathbf{u}} \mathrm{E}_{\mathbf{m}}\left[J\left(\mathbf{m},\mathbf{u}\right)\right] \approx \max_{\mathbf{u}} \frac{1}{N_e N_p}\sum_{i=1}^{N_e}\sum_{j=1}^{N_p} J\left(\mathbf{m}_i,\hat{\mathbf{u}}_{k,i,j}\right). \tag{17}$$

ModStoSAG requires N_eN_p J-function evaluations, and it can save $1/(N_p+1)$ of the computational cost of StoSAG needed to compute a search direction. A summary of the ensemble-based stochastic gradient methods discussed thus far is presented in **Table 1**. ModEnOpt, HSG, and ModStoSAG use different objective functions, but the objective function in Equation (2) is calculated at initial and final steps for all the formulations to compare them in the next numerical examples.

NUMERICAL EXAMPLES

Rosenbrock Function

The performance of the six formulations given in **Table 1** is tested using the Rosenbrock (1960) function, which is widely used for benchmarking optimization solvers. The Rosenbrock function is given by

$$J\left(m_i,\mathbf{u}\right) = \sum_{j=1}^{N_\mathbf{u}}\left[\left(1-\mathbf{u}_{2j-1}\right)^2 + m_i\left(\mathbf{u}_{2j}-\mathbf{u}_{2j-1}{}^2\right)^2\right], \tag{18}$$

$$\min_{\mathbf{u}} \mathrm{E}_{\mathrm{m}}\left[J\left(m,\mathbf{u}\right)\right] = \min_{\mathbf{u}} \frac{1}{N_e}\sum_{i=1}^{N_e} J\left(m_i,\mathbf{u}\right), \tag{19}$$

where $\mathbf{u} = \left[u_1\ u_2\ \cdots\ u_{N_u}\right]^{\mathrm{T}}$ and $i = 1, 2, \ldots, N_e$. m_i is a constant (=100) in the original Rosenbrock function, but ensemble members have different m_i to mimic geological uncertainty where $m \sim N\left(100,\sigma_m^2\right)$. The number of ensemble members is 100 ($N_e = 100$), and the ensemble members have different m_i. The dimension of $\mathbf{u}$ is 50, and an initial solution is $\mathrm{u}_i = 2.0$ for $i = 1, 2, \ldots, N_u$. The objective function is $\frac{1}{N_e}\sum_{i=1}^{N_e} J\left(m_i,\mathbf{u}\right)$ in Equation (19), and 100 J-function evaluations are needed to calculate the objective function. Two values of σ_m (0.01 and 1.00) are used to generate two sets of $m_1, m_2, \ldots, m_{100}$ for the 100 ensemble members. A larger σ_m makes a larger variance in $J(m_i, \mathbf{u})$ and larger $||\mathbf{r}_k||_2$, which causes inaccurate search directions of EnOpt and ModEnOpt. Perturbations for $\mathbf{u}$ are generated using $N(0, \mathbf{C_u})$ where $\mathbf{C_u}$ is a diagonal matrix and the diagonal elements are 0.001^2. N_p is set to 3 for both StoSAG and ModStoSAG.

CV_{HSG} is set to a number that makes the number of clusters about 70 (=0.7 · 100). CV_{HSG} for $\sigma_m = 0.01$ and 1.00 are

TABLE 1 | Names, formulas, and number of objective function evaluations of six formulations.

Full name	Acronym	Approximation of gradient	Objective function	Number of J-function evaluations
Ensemble optimization	EnOpt	$\mathbf{d}_{k,EnOpt} = \frac{1}{N_e-1}\sum_{i=1}^{N_e}\left(\tilde{\mathbf{u}}_{k,i}-\overline{\tilde{\mathbf{u}}_k}\right)\left(J\left(\mathbf{m}_i,\tilde{\mathbf{u}}_{k,i}\right)-\overline{J\left(\mathbf{m}_i,\tilde{\mathbf{u}}_{k,i}\right)}\right)$	$\frac{1}{N_e}\sum_{i=1}^{N_e} J\left(\mathbf{m}_i,\mathbf{u}\right)$	$2N_e$
Modified ensemble optimization	ModEnOpt	$\mathbf{d}_{k,EnOpt} = \frac{1}{N_e-1}\sum_{i=1}^{N_e}\left(\hat{\mathbf{u}}_{k,i}-\overline{\hat{\mathbf{u}}_k}\right)\left(J\left(\mathbf{m}_i,\hat{\mathbf{u}}_{k,i}\right)-\overline{J\left(\mathbf{m}_i,\hat{\mathbf{u}}_{k,i}\right)}\right)$	$\frac{1}{N_e}\sum_{i=1}^{N_e}\sum_{j=1}^{1} J\left(\mathbf{m}_i,\hat{\mathbf{u}}_{k,i}\right)$	N_e
Simplex gradient	SG	$\mathbf{d}_{k,SG} = \frac{1}{N_e}\sum_{i=1}^{N_e}\left(\tilde{\mathbf{u}}_{k,i}-\mathbf{u}_k\right)\left(J\left(\mathbf{m}_i,\tilde{\mathbf{u}}_{k,i}\right)-J\left(\mathbf{m}_i,\mathbf{u}_k\right)\right)$	$\frac{1}{N_e}\sum_{i=1}^{N_e} J\left(\mathbf{m}_i,\mathbf{u}\right)$	$2N_e$
Hybrid simplex gradient	HSG	$\mathbf{d}_{k,HSG} = \frac{1}{N_e}\sum_{j=1}^{N_C}\left[\sum_{i\in C_j,N_{C_j}>1}\left(\tilde{\mathbf{u}}_{k,i}-\overline{\tilde{\mathbf{u}}_{k,j}}\right)\left(J\left(\mathbf{m}_i,\tilde{\mathbf{u}}_{k,i}\right)-\overline{J\left(\mathbf{m}_i,\tilde{\mathbf{u}}_{k,i}\right)}\right)+\sum_{i\in C_j,N_{C_j}=1}\left(\tilde{\mathbf{u}}_{k,i}-\mathbf{u}_k\right)\left(J\left(\mathbf{m}_i,\tilde{\mathbf{u}}_{k,i}\right)-J\left(\mathbf{m}_i,\mathbf{u}_k\right)\right)\right]$	$\frac{1}{N_e}\sum_{j=1}^{N_C}\left[\sum_{i\in C_j,N_{C_j}>1} J\left(\mathbf{m}_i,\tilde{\mathbf{u}}_{k,i}\right)+\sum_{i\in C_j,N_{C_j}=1} J\left(\mathbf{m}_i,\mathbf{u}_k\right)\right]$	N_e ~$2N_e$
Stochastic simplex approximate gradient	StoSAG	$\mathbf{d}_{k,StoSAG} = \frac{1}{N_e}\sum_{i=1}^{N_e}\left[\frac{1}{N_p}\sum_{j=1}^{N_p}\left(\hat{\mathbf{u}}_{k,i,j}-\mathbf{u}_k\right)\left(J\left(\mathbf{m}_i,\hat{\mathbf{u}}_{k,i,j}\right)-J\left(\mathbf{m}_i,\mathbf{u}_k\right)\right)\right]$	$\frac{1}{N_e}\sum_{i=1}^{N_e} J\left(\mathbf{m}_i,\mathbf{u}\right)$	$N_e(N_p+1)$
Modified stochastic simplex approximate gradient	ModStoSAG	$\mathbf{d}_{k,ModStoSAG} = \frac{1}{N_e}\sum_{i=1}^{N_e}\left[\frac{1}{N_p}\sum_{j=1}^{N_p}\left(\hat{\mathbf{u}}_{k,i,j}-\mathbf{u}_k\right)\left(J\left(\mathbf{m}_i,\hat{\mathbf{u}}_{k,i,j}\right)-\overline{J\left(\mathbf{m}_i,\hat{\mathbf{u}}_{k,i,j}\right)}\right)\right]$	$\frac{1}{N_eN_p}\sum_{i=1}^{N_e}\sum_{j=1}^{N_p} J\left(\mathbf{m}_i,\hat{\mathbf{u}}_{k,i,j}\right)$	N_eN_p

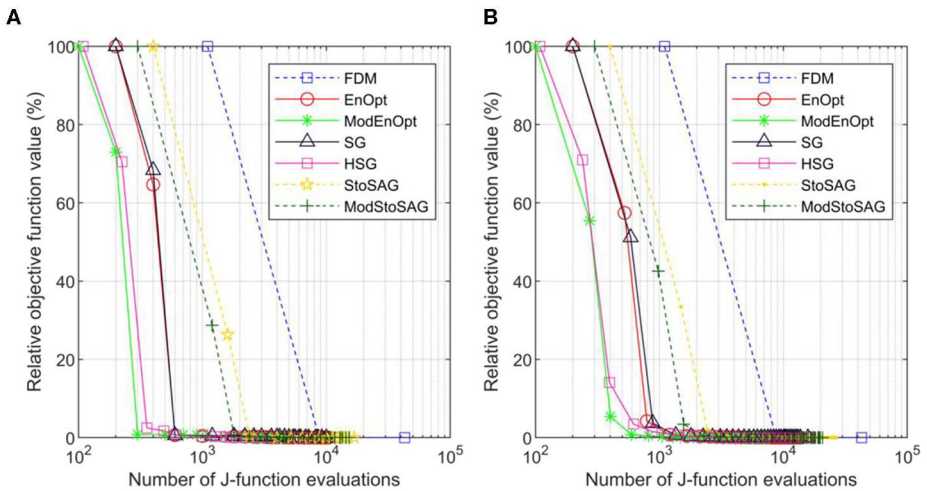


FIGURE 1 | Plot of numbers of *J*-function evaluations and relative objective function values of the seven gradient methods for $\sigma_m = 0.01$. **(A)** Relative objective function value is a percentage of the initial objective function value. Symbols represent iterations, and the *x*-axis is the total number of function evaluations. **(B)** Shows the average relative objective function values and the average total number of function evaluations of 100 optimization runs.

1E−5 and 5E−5, respectively, which were determined based on the *J*-function values for the initial solution ($u_i = 2.0$ for $i = 1, 2, \ldots, N_u$).

The performance of the stochastic gradients is compared to the gradient obtained using FDM (Sun and Sun, 2015). A search direction of FDM is computed using Equation (20), and $N_e(N_u + 1)$ function evaluations are needed.

$$\frac{\partial}{\partial u_j}\left(\sum_{i=1}^{N_e} J\left(m_i, \mathbf{u}\right)\right) = \frac{1}{\delta u_j}\begin{pmatrix} \sum_{i=1}^{N_e} J\left(m_i, \left[u_1 \ \cdots \ u_j + \delta u_j \ \cdots \ u_{N_u}\right]^T\right) \\ -\sum_{i=1}^{N_e} J\left(m_i, \left[u_1 \ \cdots \ u_j \ \cdots \ u_{N_u}\right]^T\right) \end{pmatrix}, j = 1, \ldots, N_u \tag{20}$$

The optimization problem is solved using the steepest descent method given in Equation (1). The optimization is terminated if either of the following two conditions is satisfied: (i) the relative increase of the objective function is <0.0001%, or (ii) the relative change of the norm of $\mathbf{u}_k - \mathbf{u}_{k+1}$ is <0.01%.

Figure 1 shows the optimization results for $\sigma_m = 0.01$. **Figures 1A,B** show the results of a single run and the average result of 100 runs, respectively. Because perturbations are stochastically generated for the stochastic gradients (EnOpt, ModEnOpt, SG, HSG, StoSAG, and ModStoSAG), the optimization runs using the stochastic gradients show the different numbers of *J*-function evaluations and relative objective function values for iterations. For this reason, in **Figure 1B**, the 100 optimization runs are repeated, and the relative objective function values and the total number of *J*-function evaluations are averaged for each iteration. In **Figure 1**, because σ_m (=0.01) is small, the variance in $J(m_i, \mathbf{u})$ is also small, and the small variance in $J(m_i, \mathbf{u})$ leads to small $||\mathbf{r}_k||_2$. Because EnOpt and ModEnOpt provide accurate search directions for small $||\mathbf{r}_k||_2$, EnOpt and ModEnOpt find satisfactory solutions as the other gradient methods do.

TABLE 2 | Numbers of function evaluations that are needed to achieve 5% of the relative objective function value in **Figure 1B**.

Method	Number of *J*-function evaluations	Method	Number of *J*-function evaluations
FDM	8,417	HSG	586
EnOpt	788	StoSAG	2,349
ModEnOpt	417	ModStoSAG	1,548
SG	876		

Table 2 shows the numbers of function evaluations that are needed to achieve 5% of the relative objective function value in **Figure 1B**. ModEnOpt, HSG, and ModStoSAG saved about 47, 33, and 34% of the function evaluations compared to EnOpt, SG, and StoSAG, respectively. In **Table 2**, ModEnOpt took the smallest number of function evaluations, and EnOpt needed a smaller number of function evaluations than SG did.

However, EnOpt and ModEnOpt do not achieve satisfactory reduction of the objective function value for $\sigma_m = 1.0$ as shown in **Figure 2A**. In **Figure 2B**, EnOpt and ModEnOpt show the abnormal relation between the number of *J*-function evaluations and relative objective function values because EnOpt and ModEnOpt do not reduce the objective function value sufficiently for all the 100 optimization runs. The large σ_m(=1.0) results in a large variance in $J(m_i, \mathbf{u})$ and, consequently, large $||\mathbf{r}_k||_2$. For this reason, EnOpt and ModEnOpt provide inaccurate search directions and unsatisfactory optimization results. The average angles between the search directions obtained using FDM and the stochastic gradients for $\sigma_m = 1.0$ are given in **Table 3**. The

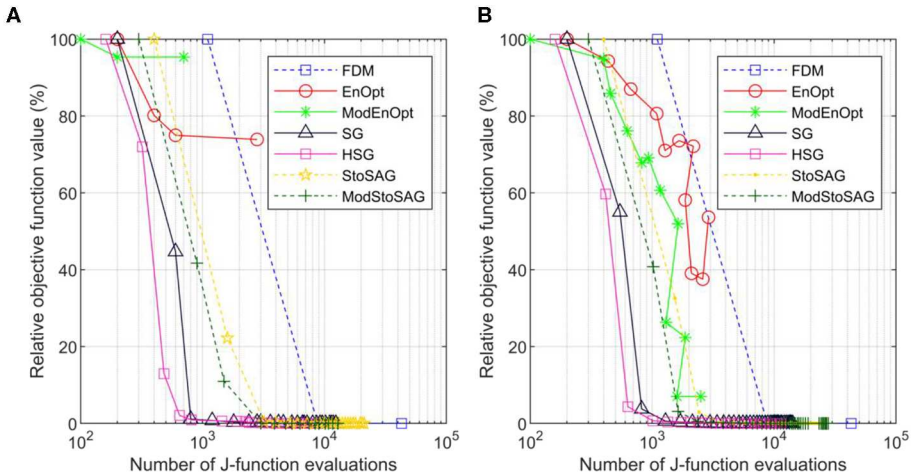


FIGURE 2 | Plot of numbers of *J*-function evaluations and relative objective function values of the seven gradient methods for $\sigma_m = 1$. **(A)** Relative objective function value is a percentage of the initial objective function value. Symbols represent iterations, and the *x*-axis is the total number of function evaluations. **(B)** Shows the average relative objective function values and the average total number of function evaluations of 100 optimization runs.

TABLE 3 | Average angles between the search directions obtained using FDM and the stochastic gradients for $\sigma_m = 1.0$.

Method	Average angle (degrees)	Method	Average angle (degrees)
EnOpt	83.98	ModEnOpt	81.25
SG	18.65	HSG	21.87
StoSAG	13.76	ModStoSAG	14.74

The angle calculations are repeated 100 times and averaged.

angle between two search direction vectors is calculated using Equation (21):

$$\theta = \cos^{-1}\left(\frac{\mathbf{u} \bullet \mathbf{v}}{|\mathbf{u}|\,|\mathbf{v}|}\right), \tag{21}$$

where **u** and **v** are vectors corresponding to search directions. The angle calculations are repeated 100 times and averaged. The search direction obtained using FDM is considered to be the most accurate, and a higher angle from the FDM search direction implies less accuracy. It can be seen from **Table 3** that SG, HSG, StoSAG, and ModStoSAG provide acceptable search directions, while EnOpt and ModEnOpt provide inaccurate search directions. The number of *J*-function evaluations for the same relative objective value increases from HSG, SG, ModStoSAG, and StoSAG, as shown in **Figure 2B**.

Optimization of Well Placement in a Carbon Storage Reservoir

Problem Formulation

As a realistic example, we now consider the optimization of brine extraction well placement in geological carbon sequestration (GCS) applications. CO_2 injection into saline aquifers or depleted oil and gas reservoirs necessarily leads to pore pressure increases. The pore pressure buildup not only can affect the CO_2 injectivity and storage performance but may also cause caprock damage, fault reactivation, induced seismicity, and leakage of brine and CO_2, posing severe problems to CO_2 long-term storage permanence and public safety (Birkholzer et al., 2012; Carroll et al., 2014; Cihan et al., 2015). Recently, active reservoir pressure management was proposed as a mitigation measure, which installs one or more brine extraction wells to reduce pressure buildup in the reservoir (Bergmo et al., 2011; Buscheck et al., 2011, 2012; Birkholzer et al., 2012; Cihan et al., 2015; Arena et al., 2017). The extraction, treatment, and disposal of the extracted brine, however, impose additional expenses to GCS operators (Cihan et al., 2015). Thus, the placement and control of brine extraction wells need to be optimized to improve the economic feasibility of GCS projects. Cihan et al. (2015) optimized well placement and controls of brine extraction wells in different geological models to minimize extracted brine volume and keep pressure buildups under critical thresholds for potentially activating fault leakage and/or fault slippage. In their study, Cihan et al. (2015) adopted a constrained differential evolution algorithm, which is a heuristic stochastic evolution algorithm similar to the genetic algorithm. Here, we demonstrate the use of the more efficient ensemble-based gradient algorithms. Optimal placement of a brine extraction well is sought in multiple geological models (i.e., geologic uncertainty) to maximize the objective function, which is expressed in the form of the net present value. Performance of the six formulations given in **Table 1** is compared.

The *J*-function for a single model, $\mathbf{m}_i$, is given by

$$J(\mathbf{m}_i, \mathbf{u}) = \sum_{n=1}^{N_t}\left[\frac{\Delta t_n}{(1+b)^{\frac{\Delta t_n}{365}}}\left\{\sum_{j=1}^{N_{inj}}\left(r_{ci}\cdot\overline{q_{ci,j}^n} - f_{ciq}\left(\overline{q_{ci,j}^n}\right)\right)\right.\right. \tag{22}$$

$$- \sum_{k=1}^{N_{ext}} \left(c_{be} \cdot \overline{q^n_{be,k}} - c_{ce} \cdot \overline{q^n_{ce,k}} \right) - \sum_{l=1}^{N_{leak}} \left(c_{bl} \cdot \overline{q^n_{bl,l}} - c_{cl} \cdot \overline{q^n_{cl,l}} \right) \Bigg\} \Bigg]$$

$$f_{ciq}\left(\overline{q^n_{ci,j}}\right) = \begin{cases} \text{if } \overline{q^n_{ci,j}} < q^n_{ciq},\ c_{ciq}\left(q^n_{ciq} - \overline{q^n_{ci,j}}\right) \\ \text{otherwise, } 0 \end{cases} \tag{23}$$

where the control vector $\mathbf{u}$ is a two-dimensional column vector including I and J indices of brine extraction wells. In words, the objective function in Equation (22) can be described as the tax credit for injected CO_2 minus the penalty for unfulfilled CO_2 injection, the cost of brine extraction wells, and the damage cost related to brine and CO_2 leakage.

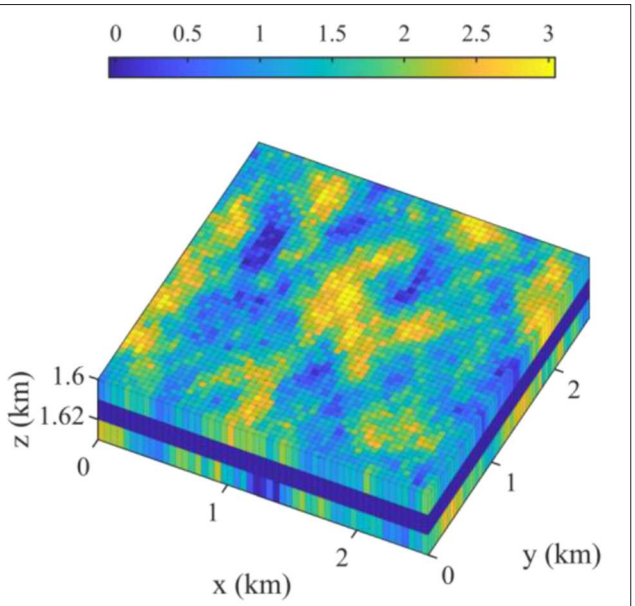


FIGURE 3 | Three-dimensional view of $\log_{10}$ permeability (md). The z-axis is exaggerated by a factor of 20.

The brine extraction wells are vertically perforated in a CO_2 injection zone, and I and J indices are rounded off to integers during iterations. Δt_n is the size (days) of the nth time step in the reservoir simulation, and b is the annual discount rate. N_{inj}, N_{ext}, and N_{leak} are the numbers of CO_2 injection, brine extraction, and leaky wells, respectively. $\overline{q^n_{ci,j}}$, $\overline{q^n_{be,k}}$, and $\overline{q^n_{ce,k}}$ represent the average CO_2 injection rate at the jth CO_2 injector, the average brine extraction rate at the kth brine extractor, and the average CO_2 extraction rate at the kth brine extractor for Δt_n, respectively.

TABLE 4 | Cost factors and unit costs for brine extraction.

Cost factor	Unit cost
Tax credit for CO_2 injection (r_{ci})	\$50/ton
Quota of CO_2 injection (q_{ciq})	2.5 tons/day
Penalty for quota of CO_2 injection (c_{ciq})	\$100/ton/day
Brine treatment (c_{be})	\$10/ton
CO_2 reinjection (c_{ce})	\$50/ton
Brine leakage treatment (c_{bl})	\$10/ton
CO_2 leakage (c_{cl})	\$1,000/ton

TABLE 5 | Statistical parameters used to generate the geological model.

Parameter	Value
Mean of porosity (fraction)	0.2
Standard deviation of porosity (fraction)	0.05
Mean of $\log_{10}$ horizontal permeability (md)	2
Standard deviation of $\log_{10}$ horizontal permeability (md)	0.6
Correlation coefficient between porosity and $\log_{10}$ horizontal permeability	0.7
Correlated direction of porosity and $\log_{10}$ horizontal permeability	North–south
Correlation lengths (major, minor, vertical) of porosity and $\log_{10}$ horizontal permeability (m)	500, 300, 10

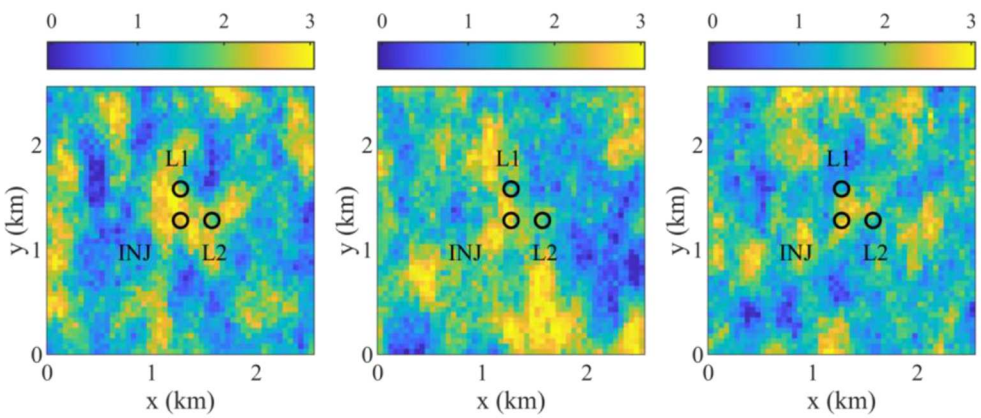


FIGURE 4 | Top view of $\log_{10}$ permeability (md) of 3 of 20 geological models, and the locations of an injector (INJ) and two leaky abandoned wells (L1 and L2).

One of the main risks related to GCS is leakage from abandoned wells—if abandoned wells are leaky, then brine and CO_2 can migrate to an overlying formation along the leaky wells (Birkholzer et al., 2012; Sun et al., 2013). Because the risk related to leakage must be minimized for safe long-term CO_2 storage, abandoned wells are assumed to be leaky for conservative estimation. In Equation (20), $\overline{q^n_{bl,l}}$ and $\overline{q^n_{cl,l}}$ are the brine leakage rate at the lth leaky abandoned well and the CO_2 leakage rate at the lth leaky abandoned well for the nth time step. If brine extraction wells are placed near the leaky wells, the brine and CO_2 leakage amount decreases because the installed brine extraction wells reduce the pressure buildup at the leaky wells; on the flip side, the brine extraction wells also need to be shut in early if they are placed near CO_2 injectors because they have early CO_2 breakthroughs. The brine extraction wells should be placed at locations that minimize the leakage costs at the leaky wells while postponing CO_2 breakthrough at the brine extraction wells as late as possible. r_{ci} is the credit for CO_2 storage, which is often provided by the government subsidy or driven by the carbon trading market (Jahangiri and Zhang, 2012; Allen et al., 2017). In Equation (23), q^n_{ciq} is the minimum required CO_2 injection rate for Δt_n, and a penalty is imposed when the injection rate cannot be met, $\overline{q^n_{ci,j}} < q^n_{ciq}$, in which case the upstream capturing facility needs to find an alternative means for temporary storage. In Equation (20), c_{be}, c_{ce}, c_{bl}, and c_{cl} are the unit costs of brine treatment, CO_2 reinjection, brine leakage, and CO_2 leakage, respectively.

Overall, the brine extraction well placement optimization is complex, especially when geologic uncertainty is involved. The ensemble gradient methods are well-suited to solve such problems because of their efficiency and the ability to incorporate geologic uncertainty. As a demonstration, we consider the three-dimensional reservoir model shown in **Figure 3**, which is 2.55 km (x-axis) by 2.55 km (y-axis) by 30 m (z-axis), and the dimensions of a grid block are 50 m by 50 m by 10 m. The vertical structure consists of three formations, which are named the above zone, caprock, and injection zone. CO_2 is injected at 30 tons/day at the center of the injection zone for 5 years, and the maximum bottom-hole pressure is 20,000 kPa. Brine is extracted at an extraction well at 60 m^3/day, and it is shut in if the ratio of produced CO_2 volume to produced brine volume is >100. The caprock blocks the flow between the above and injection

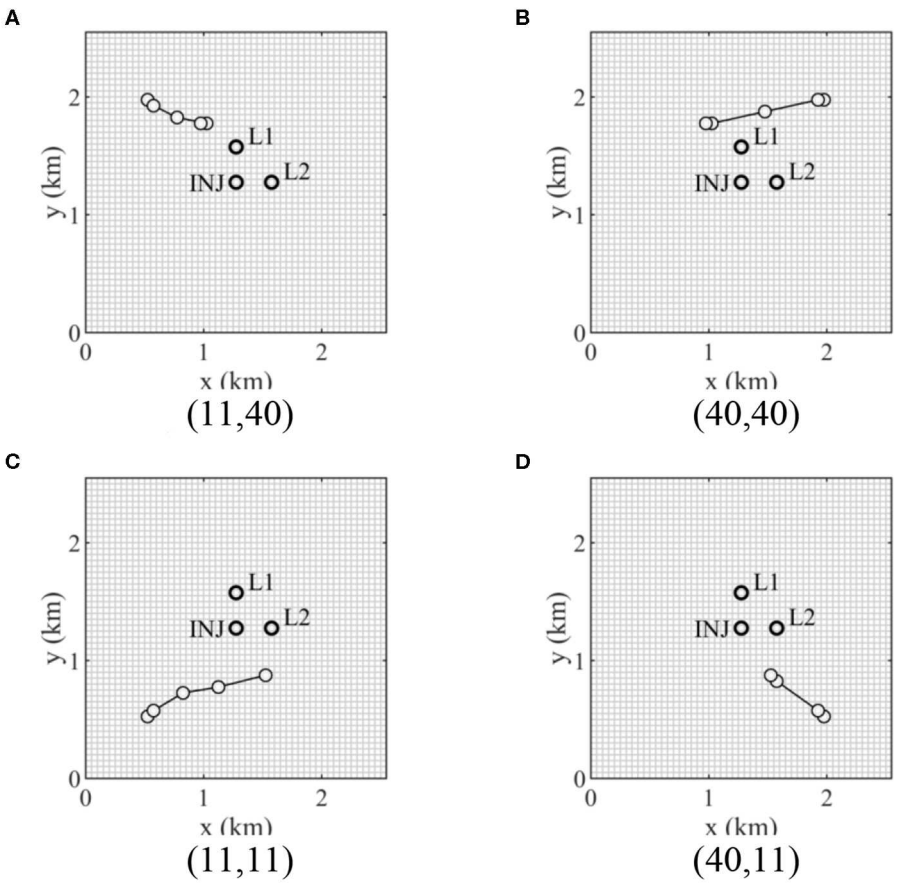


FIGURE 5 | Iteration steps of well placement of a brine extraction well using HSG for four initial solutions (I and J indices). INJ and L1 and L2 represent the CO_2 injector and two leaky abandoned wells, respectively. **(A)** 1st initial solution (11,40). **(B)** 2nd initial solution (40,40). **(C)** 3rd initial solution (11,11). **(D)** 4th initial solution (40,11).

zones, but brine and CO_2 can vertically flow up along two leaky abandoned wells where the leaky wells are located 300 m north and east of the injector in **Figure 4**. The unit cost data needed for calculating the J-function in Equations (22) and (23) are given in **Table 4**.

Heterogeneous porosity and $\log_{10}$ permeability fields are generated using the sequential Gaussian simulation and the sequential Gaussian co-simulation modules of SGeMS (Remy et al., 2011). Porosity and $\log_{10}$ permeability follow normal distributions, and their statistical parameters are given in **Table 5**. **Figure 4** shows three realizations of log10 permeability out of a total of 20 geological models. The porosity and permeability of the caprock zone are assumed deterministic and are assigned values of 0.1 and 0.001 md, respectively. The flow simulation is conducted using a compositional multiphase reservoir simulator, CMG-GEM (CMG, 2015). The two leaky abandoned wells are described using the local grid refinement of CMG-GEM, and the porosity and the vertical permeability of the leaky wells are set to 0.2 and 1,000 md.

The convergence criteria are as follows: (i) the relative increase of the objective function is <0.1%, or (ii) the relative change of the norm of $\mathbf{u}_k - \mathbf{u}_{k+1}$ is <1%.

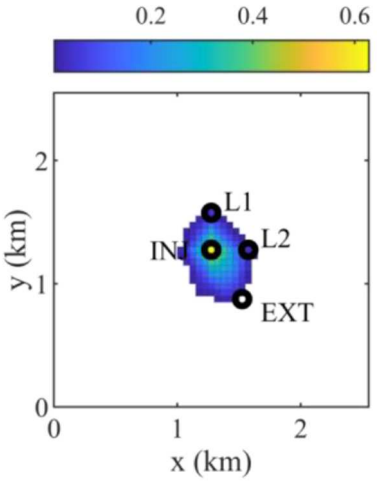


FIGURE 6 | Average CO_2 saturation map of 20 geological models after CO_2 is injected for 5 years when placing a brine extraction well at the optimal location ($I = 31$ and $J = 18$). INJ, EXT, and L1 and L2 represent the CO_2 injector, the brine extraction well, and two leaky abandoned wells, respectively.

RESULTS AND DISCUSSION

The objective of this optimization problem is to find the optimal location of a brine extraction well that maximizes the mean of the J-function values given in Equation (22) of the 20 geological models. As described in Equation (22), a solution for a brine extraction well is a two-dimensional vector including I and J indices of the brine extraction well. Thus, in this example, the vector of a search direction has two elements for the I and J indices of a brine extraction well, and the I and J indices of a solution are updated at the same time using the search direction. The optimal solution is sought using the six stochastic gradients, and then the performance of the six stochastic gradients is compared. In HSG, CV_{HSG} is set to a number that makes the number of clusters about 80% of the number of ensemble members. $N_p = 2$ is used for StoSAG and ModStoSAG.

Figure 5 shows the brine extraction well locations at the iterations obtained using HSG for four different initial solutions. In general, initial solutions for optimization are sampled in a prior distribution, but the initial solutions located at the four corners in **Figure 5** are selected because optimal solutions are expected to be located between CO_2 plume and the boundary. The four different initial solutions are fixed to compare the performance of the stochastic gradients. As described in the previous section, a brine extraction well should be placed as close to the CO_2 injector and the leaky wells as possible to mitigate the reservoir pressure buildup, as well as CO_2 and brine leakage amounts. However, the brine extraction well is shut in early if the brine extraction well is placed in the extent of the CO_2 plume, which changes for different brine extraction well locations because the CO_2 migration is affected by the reservoir pressure drawdown caused by the brine extraction well. Furthermore, the reservoir pressure buildup and drawdown, the CO_2 and brine leakage amounts, and the CO_2 plume extent are significantly affected by the heterogeneity of rock permeability in the 20 geological models. As shown in **Figure 6**, a brine extraction well should be placed out of the CO_2 plume extent and as close to the CO_2 injector and the leaky wells. (11, 11) and (40, 11) shown in **Figures 5C,D** are converged to the same solution, which is (31, 18) shown in **Figure 6**.

Figure 7 shows the numbers of simulation runs and objective function values of the six stochastic gradient methods for four initial solutions. In **Figure 7**, the number of simulation runs is how many times the flow simulator is conducted until an iteration is finished. For example, if the suite of the 20 models is simulated five times until an iteration ends, then the number of simulation runs is 100.

EnOpt and ModEnOpt do not achieve satisfactory objective function values compared to the others. This implies that $\overline{\|r_k\|_2}$ in Equation (8) or the variance in $J(m_i, \mathbf{u})$ is too large for EnOpt and ModEnOpt to provide acceptable accuracy of search directions. In **Figure 7**, for the same number of simulation runs, HSG and ModStoSAG reach higher objective function values than SG and StoSAG do.

CONCLUSION

Ensemble-based optimization algorithms are widely used for reducing the computational costs of optimization, especially when the forward problem requires a significant amount of time to run. In this work, we theoretically and experimentally showed when EnOpt may fail to achieve satisfactory performance. If $\overline{\|r_k\|_2}$ in Equation (8) or the variance in $J(m_i, \mathbf{u})$ is large, EnOpt produces unsatisfactory optimization results because the search direction of EnOpt is inaccurate as shown in the Rosenbrock

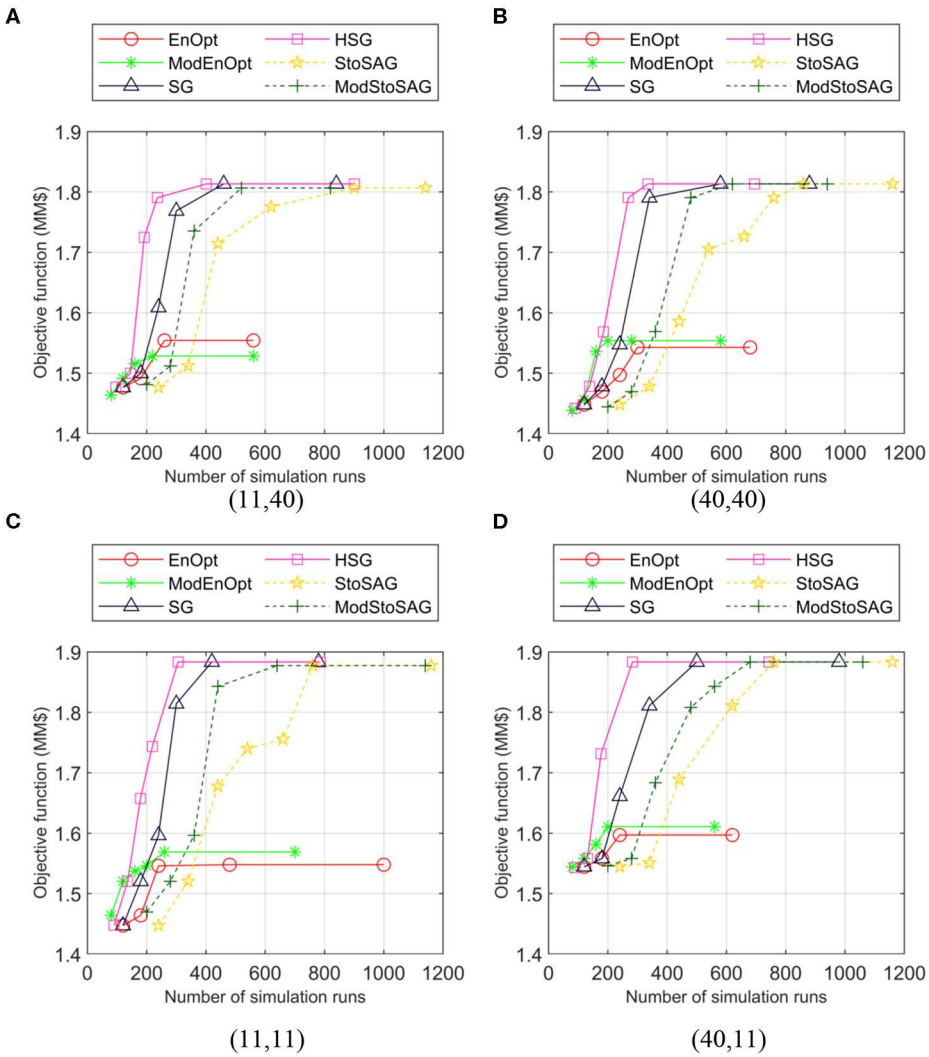


FIGURE 7 | Plot of numbers of simulation runs and objective function values of the six stochastic gradient methods for four initial solutions (*I* and *J* locations). Symbols represent iterations, and the *x*-axis is the total number of simulation runs. **(A)** 1st initial solution (11,40). **(B)** 2nd initial solution (40,40). **(C)** 3rd initial solution (11,11). **(D)** 4th initial solution (40,11).

function example. We also introduced hybrid schemes to reduce the computational costs of EnOpt, SG, and StoSAG. For the benchmark example and the geological carbon sequestration example, the computational costs of EnOpt, SG, and StoSAG can be significantly reduced by replacing the *J*-function values for the unperturbed control variables with those for the perturbed ones. The ensemble-based optimization schemes proposed in this study are generic and can be readily used on other types of problems involving computationally expensive forward simulations or optimization under uncertainty.

DATA AVAILABILITY STATEMENT

The datasets generated for this study will not be made publicly available. It is a confidential data set. Requests to access the datasets should be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

HJ was a main author to make research results and wrote this manuscript. HJ had technical discussions with AS, JJ, BM, and DJ.

FUNDING

HJ was supported by the National Research Foundation of Korea (NRF) under grant number 2018R1C1B5045260, and the Korea Institute of Geoscience and Mineral Resources(KIGAM) and the Ministry of Science, ICT and Future Planning of Korea under grant number GP2020-006. AS was supported by the U.S. Department of Energy, National Energy Technology Laboratory (NETL) under grant number DE-FE0026515. BM was supported by the National Research Foundation of Korea (NRF) under grant numbers 2018R1A6A1A08025520 and 2019R1C1C1002574.

REFERENCES

Allen, R., Nilsen, H. M., Andersen, O., and Lie, K.-A. (2017). On obtaining optimal well rates and placement for CO2 storage. *Comput. Geosci.* 21, 1403–1422. doi: 10.1007/s10596-017-9631-6

Arena, J. T., Jain, J. C., Lopano, C. L., Hakala, J. A., Bartholomew, T. V., Mauter, M. S., et al. (2017). Management and dewatering of brines extracted from geologic carbon storage sites. *Int. J. Greenh. Gas Control* 63, 194–214. doi: 10.1016/j.ijggc.2017.03.032

Bangerth, W., Klie, H., Wheeler, M. F., Stoffa, P. L., and Sen, M. K. (2006). On optimization algorithms for the reservoir oil well placement problem. *Comput. Geosci.* 10, 303–319. doi: 10.1007/s10596-006-9025-7

Bergmo, P. E. S., Grimstad, A.-A., and Lindeberg, E. (2011). Simultaneous CO_2 injection and water production to optimise aquifer storage capacity. *Int. J. Greenh. Gas Control* 5, 555–564. doi: 10.1016/j.ijggc.2010.09.002

Birkholzer, J. T., Cihan, A., and Zhou, Q. (2012). Impact-driven pressure management via targeted brine extraction—conceptual studies of CO_2 storage in saline formations. *Int. J. Greenh. Gas Control* 7, 168–180. doi: 10.1016/j.ijggc.2012.01.001

Buscheck, T. A., Sun, Y., Chen, M., Hao, Y., Wolery, T. J., Bourcier, W. L., et al. (2012). Active CO_2 reservoir management for carbon storage: Analysis of operational strategies to relieve pressure buildup and improve injectivity. *Int. J. Greenh. Gas Control* 6, 230–245. doi: 10.1016/j.ijggc.2011.11.007

Buscheck, T. A., Sun, Y., Hao, Y., Wolery, T. J., Bourcier, W., Tompson, A. F. B., et al. (2011). Combining brine extraction, desalination, and residual-brine reinjection with CO_2 storage in saline formations: Implications for pressure management, capacity, and risk mitigation. *Energy Procedia* 4, 4283–4290. doi: 10.1016/j.egypro.2011.02.378

Carroll, S. A., Keating, E., Mansoor, K., Dai, Z., Sun, Y., Trainor-Guitton, W., et al. (2014). Key factors for determining groundwater impacts due to leakage from geologic carbon sequestration reservoirs. *Int. J. Greenh. Gas Control* 29, 153–168. doi: 10.1016/j.ijggc.2014.07.007

Chen, Y., and Oliver, D. S. (2010). Ensemble-based closed-loop optimization applied to Brugge field. *SPE Reserv. Eval. Eng.* 13, 56–71. doi: 10.2118/118926-PA

Chen, Y., and Oliver, D. S. (2012). Localization of ensemble-based control-setting updates for production optimization. *SPE J.* 17, 122–136. doi: 10.2118/125042-PA

Chen, Y., Oliver, D. S., and Zhang, D. (2009). Efficient ensemble-based closed-loop production optimization. *SPE J.* 14, 634–645. doi: 10.2118/112873-PA

Cihan, A., Birkholzer, J. T., and Bianchi, M. (2015). Optimal well placement and brine extraction for pressure management during CO_2 sequestration. *Int. J. Greenh. Gas Control* 42, 175–187. doi: 10.1016/j.ijggc.2015.07.025

CMG (2015). *User's Guide GEM*. Calgary, AB: Computer Modelling Group Ltd.

Do, S. T., and Reynolds, A. C. (2013). Theoretical connections between optimization algorithms based on an approximate gradient. *Comput. Geosci.* 17, 959–973. doi: 10.1007/s10596-013-9368-9

Fonseca, R. R.-M., Chen, B., Jansen, J. D., and Reynolds, A. (2017). A stochastic simplex approximate gradient (StoSAG) for optimization under uncertainty. *Int. J. Numer. Methods Eng.* 109, 1756–1776. doi: 10.1002/nme.5342

Jahangiri, H. R., and Zhang, D. (2012). Ensemble based co-optimization of carbon dioxide sequestration and enhanced oil recovery. *Int. J. Greenh. Gas Control* 8, 22–33. doi: 10.1016/j.ijggc.2012.01.013

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognit. Lett.* 31, 651–666. doi: 10.1016/j.patrec.2009.09.011

Kim, S., Min, B., Lee, K., and Jeong, H. (2018). Integration of an iterative update of sparse geologic dictionaries with ES-MDA for history matching of channelized reservoirs. *Geofluids* 2018, 1–21. doi: 10.1155/2018/1532868

Li, G., and Reynolds, A. C. (2011). Uncertainty quantification of reservoir performance predictions using a stochastic optimization algorithm. *Comput. Geosci.* 15, 451–462. doi: 10.1007/s10596-010-9214-2

Li, L., Jafarpour, B., and Mohammad-Khaninezhad, M. R. (2013). A simultaneous perturbation stochastic approximation algorithm for coupled well placement and control optimization under geologic uncertainty. *Comput. Geosci.* 17, 167–188. doi: 10.1007/s10596-012-9323-1

Lorentzen, R. J., Fjelde, K. K., Frøyen, J., Lage, A. C. V. M., Nævdal, G., and Vefring, E. H. (2001). "Underbalanced and low-head drilling operations: real time interpretation of measured data and operational support," in *SPE Annual Technical Conference and Exhibition* (New Orleans, LA: Society of Petroleum Engineers).

MacQueen, J. (1967). "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Berkeley, CA).

Nævdal, G., Trond, M., and Vefring, E. H. (2002). "Near-well reservoir monitoring through ensemble kalman Fflter," in *SPE/DOE Improved Oil Recovery Symposium* (Tulsa, OK: Society of Petroleum Engineers).

Remy, N., Boucher, A., and Wu, J. (2011). *Applied Geostatistics With SGeMS: A User's Guide*. Cambridge University Press. Available online at: https://books.google.com/books/about/Applied_Geostatistics_with_SGeMS.html?id=eSxstwAACAAJ&pgis=1 (accessed August 14, 2015).

Rosenbrock, H. H. (1960). An automatic method for finding the greatest or least value of a function. *Comput. J.* 3, 175–184. doi: 10.1093/comjnl/3.3.175

Spall, J. C. (1992). Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Trans. Automat. Contr.* 37, 332–341. doi: 10.1109/9.119632

Spall, J. C. (1998). Implementation of the simultaneous perturbation algorithm for stochastic optimization. *IEEE Trans. Aerosp. Electron. Syst.* 34, 817–823. doi: 10.1109/7.705889

Sun, A. Y., Nicot, J.-P., and Zhang, X. (2013). Optimal design of pressure-based, leakage detection monitoring networks for geologic carbon sequestration repositories. *Int. J. Greenh. Gas Control* 19, 251–261. doi: 10.1016/j.ijggc.2013.09.005

Sun, N.-Z., and Sun, A. (2015). *Model Calibration and Parameter Estimation: For Environmental and Water Resource Systems*. New York, NY: Springer.

van Essen, G., Zandvliet, M., van den Hof, P., Bosgra, O., and Jansen, J.-D. (2009). Robust waterflooding optimization of multiple geological scenarios. *SPE J.* 14, 202–210. doi: 10.2118/102913-PA

Zhang, X., Sun, A. Y., and Duncan, I. J. (2016). Shale gas wastewater management under uncertainty. *J. Environ. Manage.* 165, 188–198. doi: 10.1016/j.jenvman.2015.09.038

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.

NOMENCLATURE

$\mathbf{u}$	Control vector
$\mathbf{d}$	Search direction
α	Step size
$\mathbf{m}$	Reservoir model parameters
$J(\mathbf{m}, \mathbf{u})$	Objective function of a single model and a control vector
N_e	Number of ensemble members
$\mathbf{C_u}$	Covariance matrix of perturbations for a control vector
$\hat{\mathbf{u}}_k$	Matrix of perturbed control column vectors for N_e ensemble members
$\hat{\mathbf{u}}_{k,i}$	Perturbed control vector of the ith ensemble member at the kth iteration
$\overline{\hat{\mathbf{u}}_{k,j}}$	Mean of control vectors in the jth cluster
C_j	jth cluster
N_p	Number of perturbations for StoSAG and ModStoSAG
$\hat{\mathbf{u}}_{k,i,j}$	jth perturbed control vector of the ith ensemble member at the kth iteration

frontiers in Earth Science

ORIGINAL RESEARCH
published: 16 June 2020
doi: 10.3389/feart.2020.00202

Conditioning Multi-Gaussian Groundwater Flow Parameters to Transient Hydraulic Head and Flowrate Data With Iterative Ensemble Smoothers: A Synthetic Case Study

Dan-Thuy Lam[1]*, *Jaouher Kerrou*[1], *Philippe Renard*[1], *Hakim Benabderrahmane*[2] and *Pierre Perrochet*[1]

[1] Centre for Hydrogeology and Geothermics (CHYN), University of Neuchâtel, Neuchâtel, Switzerland, [2] French National Radioactive Waste Management Agency (ANDRA), Châtenay-Malabry, France

OPEN ACCESS

Edited by:
Teng Xu,
Hohai University, China

Reviewed by:
Xiaodong Luo,
Norwegian Research Institute (NORCE), Norway
Andrea Zanini,
University of Parma, Italy

***Correspondence:**
Dan-Thuy Lam
dan-thuy.lam@outlook.fr

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

Received: 03 February 2020
Accepted: 18 May 2020
Published: 16 June 2020

Citation:
Lam D-T, Kerrou J, Renard P, Benabderrahmane H and Perrochet P (2020) Conditioning Multi-Gaussian Groundwater Flow Parameters to Transient Hydraulic Head and Flowrate Data With Iterative Ensemble Smoothers: A Synthetic Case Study. Front. Earth Sci. 8:202. doi: 10.3389/feart.2020.00202

Over the last decade, data assimilation methods based on the ensemble Kalman filter (EnKF) have been particularly explored in various geoscience fields to solve inverse problems. Although this type of ensemble methods can handle high-dimensional systems, they assume that the errors coming from whether the observations or the numerical model are multivariate Gaussian. To handle existing non-linearities between the observations and the variables to estimate, iterative methods have been proposed. In this paper, we investigate the feasibility of using the ensemble smoother and two iterative variants for the calibration of a synthetic 2D groundwater model inspired by a real nuclear storage problem in France. Using the same set of sparse and transient flow data, we compare the results of each method when employing them to condition an ensemble of multi-Gaussian groundwater flow parameter fields. In particular, we explore the benefit of transforming the state observations to improve the parameter identification performed by one of the two iterative algorithms tested. Despite the favorable case of a multi-Gaussian parameter distribution addressed, we show the importance of defining an ensemble size of at least 200 to obtain sufficiently accurate parameter and uncertainty estimates for the groundwater flow inverse problem considered.

Keywords: inverse problem, transient groundwater flow, parameter identification, iterative ensemble smoother, data assimilation, uncertainty

1. INTRODUCTION

Since the ensemble Kalman filter (EnKF) (Evensen, 1994) has been introduced as a computationally efficient Monte Carlo approximation of the Kalman filter (Kalman, 1960; Anderson, 2003), ensemble methods for data assimilation have been widely used for high-dimensional estimation problems in geosciences (Evensen, 2009b). In all these methods, an initial ensemble of realizations which should capture the initial uncertainty of the state or parameter variables of interest is first generated. Then, thanks to the assimilation of available uncertain observations, an updated

ensemble of realizations that are conditioned by the observations is obtained. However, a main limitation is that ensemble Kalman methods assume multivariate Gaussian error statistics in the distributions involved in the computation of the update. As a result, departures from this multi-Gaussian assumption can lead to an important loss of optimality in the estimated ensemble mean and variance.

EnKF is for example extensively applied in meteorology in order to estimate the current state of the atmosphere in real time (Anderson, 2009). In such atmospheric applications, newly obtained observations of the atmosphere are assimilated sequentially in order to update the initial conditions of weather predictions models. In reservoir modeling, ensemble methods have also become a standard tool with the use, more particularly, of smoother algorithms for inverse modeling (Evensen, 2018). Still with the aim of improving model forecasts, time series of state observations collected during the production of a reservoir are with smoother methods processed all simultaneously in order to update the static parameters of reservoir simulation models.

EnKF has also been applied in surface and groundwater hydrology for the estimation of both the parameter and state variables of a system (Moradkhani et al., 2005; Hendricks Franssen and Kinzelbach, 2008). In particular, it has been shown that the increased degree of freedom introduced by the larger number of unknowns can make the estimation of EnKF particularly unstable, especially in the presence of non-linear dynamics (Moradkhani et al., 2005). When the problem is non-linear, such joint estimation can also result in inconsistent predicted data after the update and physical inconsistencies between the updated states and parameters (Gu and Oliver, 2007; Chen et al., 2009). These issues have particularly motivated the development of iterative EnKF methods based on the iterative minimization of a cost function for each iteration of the standard EnKF (Gu and Oliver, 2007; Emerick and Reynolds, 2012b).

Because of the need of restarting the dynamic model multiple times in the context of non-linear parameter estimation with EnKF, the simultaneous assimilation of all the data set in the ensemble smoother method (van Leeuwen and Evensen, 1996) has been considered a suitable alternative to EnKF in reservoir applications. Instead of having to update the variables at each assimilation time step, the ensemble smoother can process all the data of the time series in one single update step. Similarly to the iterative EnKF, successive updates can also be applied using iterative forms of the ensemble smoother in order to improve the data fit in non-linear problems (Chen and Oliver, 2012; Emerick and Reynolds, 2012a; Luo et al., 2015). (Chen and Oliver, 2012; Emerick and Reynolds, 2012a; Luo et al., 2015).

This paper focuses on the performance of two existing iterative forms of ensemble smoother for a synthetic groundwater flow application. Although the model is synthetic, it is inspired by the real hydraulic perturbation observed at the Andra's Meuse/Haute-Marne site since the construction and operation of the Underground Research Laboratory (Benabderrahmane et al., 2014; Kerrou et al., 2017). In particular, the problem considers a transient flow induced by one vertical shaft, and transient observations at points located in a restricted region of the model. Note that these situations of sparse and unevenly distributed data are very common in real groundwater modeling studies and makes the problem of subsurface characterization by inverse modeling more difficult.

The main objective of this synthetic application is to assess the benefit of assimilating different types of flow data, namely hydraulic heads and flow rates, for the identification of multi-Gaussian log hydraulic conductivity (log K) fields. In particular, the effects of increasing the ensemble size on the accuracy of the mean estimate and its associated uncertainty captured by the ensemble will be compared for both methods. In addition, similarly to the work of Schoniger et al. (2012) which introduced the benefit of using a normal-score transform on the state variables prior to updating multi-Gaussian log K fields with EnKF, we will also assess in this study the benefit of using a similar transformation approach but in the specific context of one of the tested smoother algorithms. Indeed, it has been shown that such state transformations could improve the accuracy of the updates computed by EnKF thanks to a pseudolinearization between the multi-Gaussian parameter field and the transformed states variables (Schoniger et al., 2012). As a matter of fact, the normal-score transform approach has already been applied in the context of iterative ensemble smoothing in the recent work of Li et al. (2018). However, in that study, the proposed approach specifically aimed at addressing the problem of identifying non-Gaussian parameter fields. Hence the approach required transforming not only the state variables but also the parameters in order to perform the update in both the transformed parameter and data spaces. The main motivation for the normal-score transform in that particular context was to get closer to the assumption of multi-Gaussian variables which underlie ensemble Kalman methods. However, unlike the normal-score transform approach introduced by Schoniger et al. (2012), one main drawback of transforming both the parameter and state variables is that it can actually increase the non-linearity between those variables (Zhou et al., 2011).

Hereinafater, we first present in section 2 the ensemble smoother and the two iterative ensemble smoother considered in this study: LM-EnRML (Chen and Oliver, 2013) and ES-MDA (Emerick and Reynolds, 2012a). Both are the main iterative variants currently used for inverse modeling in reservoir applications (Evensen, 2018). The synthetic case including the model set up and the generation of the initial ensemble are presented in section 3. The performance used to analyze our results on the synthetic case are presented in section 4. Finally, the results are discussed in section 5.

2. GENERAL BACKGROUND ON THE ENSEMBLE SMOOTHER AND ITERATIVE

The ensemble smoother (ES) introduced by van Leeuwen and Evensen (1996) is an extension of the ensemble Kalman filter. Both are similar in that a set of N realizations $\{m_1^{pr}, ..., m_N^{pr}\}$ is used to represent a presumed multi-Gaussian distribution, and is updated by the assimilation of measurements in order to form a new conditional distribution. When using either ES or EnKF for the inverse modeling of parameters based on observations

of state variables, each conditioned parameter realization m_i^{post} of the ensemble is calculated from the unconditioned realization m_i^{pr} according to the following equation

$$m_i^{post} = m_i^{pr} + K(d_{obs,i} - g(m_i^{pr})) \quad (1)$$

with

$$K = C_{MD}^{pr}(C_{DD}^{pr} + C_{err})^{-1}$$

The matrix K represents an approximation of the so-called "Kalman gain" in the Kalman filter update equation, derived so as to minimize the error covariance of the posterior estimate. It is here computed based on approximations from the ensemble of the cross-covariance matrix between the vector of prior parameters and the vector of predicted data, noted C_{MD}^{pr}, the auto-covariance matrix of predicted data C_{DD}^{pr}, and the covariance matrix of observed data measurement errors C_{err}. By evaluating the relative uncertainty of the measurements and prior estimate, the Kalman gain weights the contribution of each conditioning observation relatively to the prior estimate m_i^{pr} for the computation of the update. More precisely, it weights the contribution from each component of the mismatch between the vector of perturbed observations $d_{obs,i}$, i.e., the observations corrupted with noise $z_{obs,i} \sim N(0, C_{err})$, and the corresponding vector of predicted states $g(m_i^{pr})$ using the forward operator g.

Unlike EnKF however, ES does not assimilate the data sequentially in time. Instead, it assimilates all the available observations simultaneously in a single conditioning step. Hence the prediction step in ES prior to the single update will be longer than each recursive one in EnKF since the ensemble of prior realizations need to be forwarded in time until the time of the last conditioning observation. Evensen and van Leeuwen (2002) showed that when the prior realizations are multi-Gaussian and the forward model is linear, ES and EnKF at the last data assimilation will give the same result. In this special case, they will converge to the exact solution in the Bayesian sense as the ensemble size increases to infinity (hence the subscripts *pr* and *post* used in the previous equations to denote the unconditioned and conditioned realizations respectively). In non-linear cases, EnKF has been shown to outperform ES (Crestani et al., 2013). Indeed, the sequential processing of fewer data in EnKF effectively allows the computation of smaller updates than the single global update of ES. This fact particularly allows EnKF to better match the measurements than ES in non-linear problems.

Even so, if the whole data set for the parameter estimation is already acquired, the assimilation with ES of the whole set of data in a global update step may seem more convenient to implement. Indeed, the additional computations of intermediate conditional ensembles over time with EnKF can be avoided. For non-linear problems, iterative versions of the ES have been especially developed in order to improve the insufficient data match obtained with ES. Like ES, these iterative variants assimilate the complete data set during the conditioning step. However, the assimilation is performed multiple times on the same data set in order to reach the final solution. In the following sections, we introduce two existing iterative ensemble smoother algorithms which are particularly used in reservoir applications (Evensen, 2018), namely the simplified version of the Levenberg-Marquardt Ensemble Randomized Maximum Likelihood (LM-EnRML) of Chen and Oliver (2013) and the Ensemble smoother with Multiple Data Assimilation (ES-MDA) of Emerick and Reynolds (2012a).

2.1. Levenberg–Marquardt Ensemble Randomized Likelihood (LM-EnRML)

LM-EnRML is an iterative ensemble smoother based on a modified form of the Levenberg-Marquardt algorithm (Chen and Oliver, 2013). By modifying the Hessian term, LM-EnRML avoids the explicit computation of the sensitivity matrix of the predicted data to the model parameters using the ensemble at each iteration as in the original LM-EnRML formulation. This allows LM-EnRML to reduce the numerical instability usually observed with the original method for large-scale problems where the ensemble size becomes smaller than the number of parameters to estimate (Chen and Oliver, 2013). For our study, we will use a simplified version of LM-EnRML referred as LM-EnRML (approx.) in Chen and Oliver (2013) which neglects the contribution to the update of the mismatch between the updated and the prior realization. Indeed, Chen and Oliver (2013) showed that using this simpler variant did not significantly affect the results obtained with an ensemble of 104 realizations for a large-scale estimation problem involving 165000 parameters and 4000 observations.

Assuming a prior multi-Gaussian distribution of realizations m_i^{pr} with $i = 1, ..., N$, the algorithm aims to generate a posterior ensemble of N realizations m_i that each individually minimizes an objective function

$$\begin{aligned} O(m_i) = & \frac{1}{2}(g(m_i) - d_{obs,i})^T C_{err}^{-1}(g(m_i) - d_{obs,i}) \\ & + \frac{1}{2}(m_i - m_i^{pr})^T C_M^{-1}(m_i - m_i^{pr}) \end{aligned} \quad (2)$$

which measures the distance between m_i and the realization m_i^{pr} sampled from a prior distribution and the distance between the noisy observations $d_{obs,i}$ and the corresponding vector of predictions, noted $g(m_i)$, which results from the application of the forward operator g to m_i.

Each minimization of the ensemble of objective functions is performed iteratively, so that for each ensemble member, the updated realization at iteration $k+1$ is computed using the results of the previous iteration k as follows

$$\begin{aligned} m_i^{k+1} = & m_i^k - C_{sc}^{1/2}\Delta m_k^e V_D^{P_D} W_D^{P_D}((1+\lambda_k)I_{P_D} + {W_D^{P_D}}^2)^{-1} \\ & {U_D^{P_D}}^T C_{err}^{-1/2}(g(m_i^k) - d_{obs,i}^k) \end{aligned} \quad (3)$$

where C_{err} is the covariance of measurement errors, C_{sc}, a scaling matrix for the model parameters. Although C_{sc} is typically defined diagonal with its diagonal elements equal to the variance of the prior distribution in the general form of LM-EnRML, C_{sc} can simply be the identity matrix in the approximate version LM-EnRML as it allows the algorithm to converge more

quickly (Chen and Oliver, 2013). Δm^e represents an ensemble of deviations from the mean of the parameters vectors, computed as

$$\Delta m_k^e = -C_{sc}^{-1/2}(m_k^e - \overline{m_k^e})/\sqrt{N-1} \quad (4)$$

with m^e the ensemble of N parameter vectors and $\overline{m^e}$ the ensemble mean. As for the terms $U_D^{P_D}$, $W_D^{P_D}$, and $V_D^{P_D}$, they result from the truncated singular value decomposition, based on a number P_D of singular values, of the ensemble of deviations from the mean of the vectors of predictions, noted Δd^e, so that $\Delta d^e = U_D^{P_D} W_D^{P_D} V_D^{PD^T}$. This ensemble of predicted data deviations is calculated as

$$\Delta d_k^e = -C_D^{-1/2}(d_k^e - \overline{d_k^e})/\sqrt{N-1} \quad (5)$$

with d^e the ensemble of N prediction vectors and $\overline{d^e}$ the ensemble mean. Finally, λ is a damping parameter, also known as the Levenberg-Marquardt regularization parameter (Chen and Oliver, 2013), and is adjusted by the algorithm. This parameter is critical for the convergence of the algorithm as it affects the search direction and length of each update step. The LM-EnRML algorithm as proposed by Chen and Oliver (2013) adjusts λ after each parameter update computation according to whether the updated parameters lead to a decrease or increase of the objective function. If the objective function is decreased, the update is accepted and the parameter search continues with a decreased λ. Otherwise, λ is increased until an update that decreases the objective function is found. LM-EnRML will stop iterating either after a maximum number of iterations allowed is reached, or the relative decrease of the objective function or the magnitude of the realization update falls below a defined threshold. To allow a convergence toward a sufficiently low value of the objective function, we set for our synthetic problem the initial value of λ to 100 and defined a factor for either decreasing or increasing the damping parameter equal to 4. Indeed, after testing different initial values of λ, we found that lower values such as 1×10^{-3} did not allow a significant change of the results in terms of convergence and final data match. However, we noticed that values larger than 1×10^5 could result to very small changes in the objective function and consequently to the termination of the algorithm before a good data match could be reached.

2.2. Ensemble Smoother With Multiple Data Assimilation (ES-MDA)

Thanks to its simple formulation, ES-MDA of Emerick and Reynolds (2012a) is perhaps the most used iterative form of the ensemble smoother in geoscience applications. The standard algorithm consists simply in repeating a predefined number of times the standard ensemble smoother (ES) (Emerick and Reynolds, 2012a). However, unlike in the ES update, the covariance of the measurement errors in ES-MDA is inflated so that each realization m_i of the ensemble of size N at iteration $k+1$ is updated as follows

$$m_i^{k+1} = m_i^k + C_{MD}^k(C_{DD}^k + \alpha_{k+1}C_{err})^{-1}(d_{obs,i}^k - g(m_i^k)) \quad (6)$$

where $d_{obs,i}^k = d_{obs} + \sqrt{\alpha_{k+1}}C_{err}^{1/2}z_{d,i}$ with $z_{d,i} \sim N(0, I_d)$.

The purpose of inflating the Gaussian noise, sampled at every iteration, via the coefficient $\alpha > 1$ is to limit the confidence given to the data as they will be assimilated multiple times (Emerick and Reynolds, 2012a). In so doing, the parameter covariance reduction which occurs after each data assimilation is also limited. The inflation factors used to inflate the covariance matrix of the measurement errors need to satisfy the following condition

$$\sum_{k=1}^{N_a} \frac{1}{\alpha_k} = 1 \quad (7)$$

where N_a is the number of times we repeat the data assimilation. Indeed, this condition has been derived in order to make the single update by ES and multiple data assimilation by ES-MDA equivalent for the linear Gaussian case. For all our tests using ES-MDA, we set for simplicity the inflation coefficients equal to the predefined number of assimilations, as varying them in a decreased order do not lead to a significant improvement of the data match (Emerick, 2016).

ES-MDA effectively improves the data fit obtained by ES in the non-linear case because ES is in fact equivalent to one single iteration of the Gauss-Newton procedure to minimize the objective function (2) when using a full step length and an average sensitivity matrix calculated from the ensemble (Emerick and Reynolds, 2012a; Le et al., 2016). The motivation for applying ES-MDA in non-linear problems is that it would be comparable to several Gauss-Newton iterations with an average sensitivity matrix which is updated after each new data assimilation. By calculating smaller updates than one single potentially large ES update, ES-MDA is expected to lead to better results than ES.

The quality of the final data fit achieved with ES-MDA will particularly depend on the predefined number of data assimilations. The standard ES-MDA algorithm is therefore not an optimized procedure as it requires some amount of trial before finding a number of iterations which allows an acceptable match. Although we did not consider them in this synthetic study, implementations of ES-MDA which allow to adapt the inflation coefficients and the number of iterations as the history match proceeds have been proposed (Emerick, 2016; Le et al., 2016).

2.3. Normal-Score Transform of State Variables With ES-MDA

Considering that the forward model g can be non-linear due to the physical process being modeled and/or the influence of imposed boundary conditions on the predicted states, the assumption of multi-Gaussian dependence among state variables in ensemble Kalman methods is generally not justified in subsurface flow modeling even in the case of multi-Gaussian log hydraulic conductivity fields (Schoniger et al., 2012). In addition to the linearization around the ensemble mean introduced by the use of an ensemble gradient, the derivation of the ES update equation (6) particularly involves a linearization of the forward model around the local estimate m_i^k (Luo et al., 2015; Evensen, 2018). Hence ES-MDA, which is based on the repeated application of ES using inflated measurement errors, also applies such a local linearization at each update step (Evensen, 2018).

Because of this aforementioned local linearization in the update step, we are here interested in assessing the effects of a normal-score transform in order to apply the ES-MDA update to locally Gaussian state distributions. In a previous study, (Schoniger et al., 2012) observed an improved performance of EnKF for the identification of multi-Gaussian log hydraulic conductivity fields when assimilating normal-score transformed state variables such as hydraulic heads. They attributed this improvement to an "implicit pseudolinearization" of the relationship between the multi-Gaussian log hydraulic fields and the transformed predicted data which benefited the "linear updating step" of EnKF. In addition to this statement, we conjecture here that it is more specifically the local linearization highlighted by Evensen (2018), rather than the linearization around the ensemble mean introduced by the approximated gradient, which will benefit from such a normal-score transform performed locally at the data points. The "pseudolinearization" observed by Schoniger et al. (2012) would then allow a better approximation of the forward model from the local linearization applied at each update step, thereby yielding more accurate updates.

The effects of the ES-MDA update in the transformed space will be tested as follows

$$m_i^{k+1} = m_i^k + C_{M\widehat{D}}^k(\widehat{C_{DD}^k} + \alpha_{k+1}\widehat{C_{err}})^{-1}(\widehat{d_{obs,i}^k} - \widehat{g(m_i^k)}) \quad (8)$$

where

$$d_{obs,i}^k = d_{obs} + \sqrt{\alpha_{k+1}}C_{err}^{1/2}z_{d,i}$$

with

$$z_{d,i} \sim N(0, I_d)$$

where the hat indicates either transformed variables or covariances calculated based on transformed variables. A normal-score transform (NST) function denoted ψ_k, for each data type, is calculated based on the time series of state variables predicted at each one of the k observation locations (**Figure 1**) which we perturb with noise beforehand. Indeed, we perturb the predicted data using the same inflated measurement errors added to the observations d_{obs} in (8). This NST function maps the p-quantile of the ensemble distribution of original perturbed predicted values at one location to the p-quantile of a standard normal distribution. In order to evaluate properly the data mismatch in the transformed space, the predicted data $g(m_i^k)$ without added noise and the perturbed observations $d_{obs,i}^k$ are transformed using the same NST functions defined depending on the location of the data. Note that $C_{M\widehat{D}}$ is the cross-covariance between the parameters and the normal-score transform of the predicted data without added measurement errors.

To transform the term $C_{DD}^k + \alpha_{k+1}C_{err}$, Schoniger et al. (2012) suggested to calculate the covariance using directly the transformed perturbed predicted data from which we built the NST functions. In this way, the terms C_{DD}^k and C_{err} don't need to be transformed separately. However, unlike in EnKF where the inversion of the matrix $C_{DD}^k + \alpha_{k+1}C_{err}$ is fast thanks to the generally small number of data at each observation time (Liu and Oliver, 2005), the assimilation of a large data set here with ES-MDA will require the use of the subspace inversion procedure of Evensen (2004). Consequently, the calculation of the transformed terms $\widehat{C_{DD}^k}$ and $\widehat{C_{err}}$ will be needed in any case to compute the pseudo-inverse of $\widehat{C_{DD}^k + \alpha_{k+1}C_{err}}$ in (8). In the end, $\widehat{C_{DD}^k}$ can be calculated simply on the basis of the normal-score transform of the predicted data without added measurement errors, while the transformed measurement error covariance $\widehat{C_{err}}$ will be determined so as to respect the following ratio

$$\widehat{C_{err}}(C_{err})^{-1} = \widehat{C_{D'D',diag}}(C_{D'D',diag})^{-1} \quad (9)$$

where C_{err} is the untransformed measurement error covariance which is often assumed diagonal, $\widehat{C_{D'D',diag}}$ is the diagonal matrix constructed from the diagonal elements of the covariance of the transformed perturbed predicted data, and $C_{D'D',diag}$ is the untransformed equivalent. It is worth pointing out that transforming properly the measurement error covariance is particularly critical in this proposed application of ES-MDA in the transformed space as the data mismatch reduction with ES-MDA will rely on the inflation of $\widehat{C_{err}}$. Moreover, to better taking into account the data of different orders of magnitude in the update, Emerick (2016) suggested using the measurement error covariance to rescale the predicted data before calculating the truncated singular value decomposition (SVD) in the subspace inversion procedure.

3. A SYNTHETIC INVERSE PROBLEM INSPIRED BY THE ANDRA'S SITE

A synthetic case inspired by the hydraulic situation encountered at the Andra's Meuse/Haute-Marne site during the construction and the operation of the Underground Research Laboratory (Benabderrahmane et al., 2014; Kerrou et al., 2017) was created. The model represents a two-dimensional vertical cross-section of a multilayered aquifer system and is inspired by previous modeling studies on the Andra's site (Bourgeat et al., 2004; Deman et al., 2015). The model was designed to mimic the hydraulic behavior of the Oxfordian limestone multi-layered aquifer above the Callovo-Oxfordian clay host formation which is not included in the model. It will be used to analyze the performance of the ensemble smoother and the two iterative ensemble smoother algorithms presented previously for the identification of multi-Gaussian parameter fields. We present hereinafter the inverse problem set up, the assumptions for the application of the methods, and the performance criteria.

3.1. Model Set Up

The synthetic model is two-dimensional over a vertical domain of 5,000 m along the west-east direction and 500m in depth. In order to solve the groundwater flow equation, the domain is discretized by 50 × 500 square elements of 10m wide. Along the top boundary, constant heads are imposed. The head values vary from 267.5m on the west side to 284m on the east side according to a small hydraulic gradient observed in the shallow aquifers which we did not represent in our model. At depth, between 150

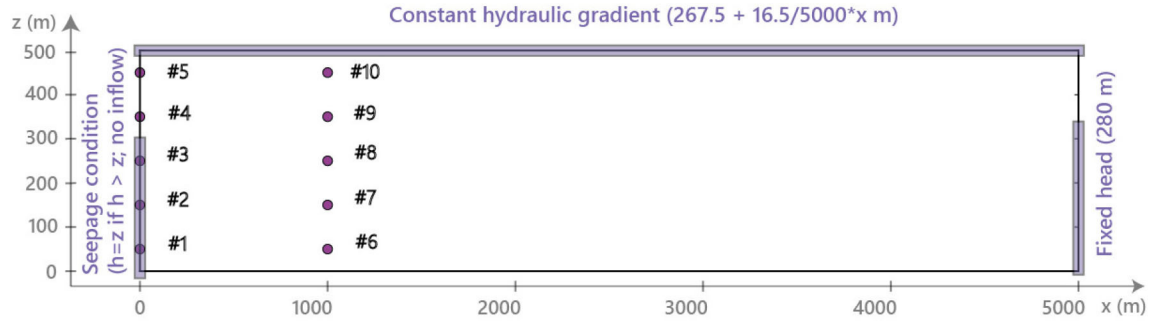


FIGURE 1 | The two-dimensional vertical domain modeled of 5,000 m length and 500m depth. The thin rectangles indicates the location of nodes with a boundary condition applied. The dots indicates the head observation nodes. Not shown here, the seepage boundary is divided into 5 zones where the produced flowrates are measured. Reprinted from Lam et al. (2020) with permission.

and 500m, a constant head of 280m is imposed on the east side to maintain a continuous east-west flow in the system (**Figure 1**). To mimic the hydraulic perturbation created by one vertical shaft to the underground research laboratory at the Andra's site, a subsequent transient flow simulation is calculated by adding from time zero a seepage condition over the nodes corresponding to the first 300 m of the western boundary starting from the bottom (**Figure 1**). The nodes over the remaining 200m toward the surface do not contribute to the production of water as they model the isolated upper part of the shaft as in the real case. Since the produced flowrates at the shaft are unknown prior to the computed head solution at each time step, using a seepage boundary condition instead of a pumping well with a prescribed extraction rate seems indeed more appropriate here. Note that in exploratory simulations, a sensitivity analysis was carried out on different levels of the finite element mesh refinement and the solver parameters, as well as on various types of flow boundary conditions and hydraulic property values. This was in order to ensure robust numerical solutions and optimized CPU time, and also to avoid bias from boundary effects.

For simplicity, the groundwater flow equation was solved numerically under saturated conditions. All the groundwater flow simulations in the framework of this study were performed with the simulator GroundWater (GW) (Cornaton, 2014). This numerical code uses the standard Galerkin Finite Element and the Control Volume Finite Element methods, and has been validated on the basis of a series of standard benchmarks by comparison with analytical solutions as well as with commercial numerical simulators.

3.2. Synthetic Data Set

Since we intend in this synthetic study to apply ensemble Kalman methods to an initial ensemble of multi-Gaussian parameter realizations, a reference field of values of decimal logarithm of hydraulic conductivity (log K) with mean -5 and variance 0.49 was generated with a multi-Gaussian simulation technique (**Figure 2**). To generate the synthetic data set of the inverse problem, a steady state groundwater flow was first simulated prior to the activation of the seepage boundary condition. The computed heads were then used as initial conditions for a subsequent transient flow simulation after the activation of the seepage condition at time zero. A constant specific storage value of 10^{-6} m^{-1} was assumed over the whole domain as they will not be considered in the inverse modeling. In the end, the data set for the inversion is composed of simulated heads collected every 1,200 s from $t = 0$ to $t = 43,200s$, hence at 37 time steps, at the ten observation points shown in **Figure 1**, and of the flowrates produced every 300s in five defined zones of nodes spread along the seepage boundary during the first 6,000 s of the time series. As a result, the data set will be composed of at most 570 observations in our experiments. Note that the final time step of the flow simulation was chosen so as to capture approximatively the drawdown before its stabilization, similarly to the situation observed at the Andra's site in Meuse/Haute Marne (Kerrou et al., 2017).

3.3. Initial Ensemble of Parameters and Assumptions for the Update Step

In this synthetic case, an initial multi-Gaussian distribution of the variables to condition was considered to respect the underlying multivariate Gaussian assumption of ensemble Kalman methods. The initial ensemble is composed of multi-Gaussian realizations of mean -5 of log hydraulic conductivity (log K) values generated using a fast Fourier transform method. This mean corresponds to the one used for the reference field as we assume this value to be known in our case. Although different variogram models could have been considered for the generation of the initial ensemble, Wen and Chen (2005) showed that the covariance model was not critically important to reproduce the main heterogeneity features from the assimilation of flow data with an ensemble Kalman method. Interestingly, Jafarpour and Khodabakhshi (2011) concluded in a study which addressed variogram uncertainty that the direct estimation of the variogram model parameters with an ensemble Kalman method was made difficult due to an insufficient strength of the linear correlation between the flow data and those parameters. Ultimately, we chose to simply assume a known variogram and hence used the same variogram model as the one used for the reference to describe the spatial variability of the ensemble of log K fields, i.e., an exponential variogram of variance 0.49 and

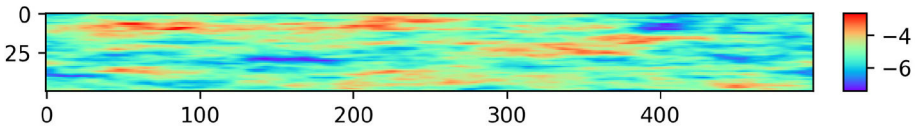


FIGURE 2 | Reference field of decimal logarithm of hydraulic conductivities (m/s).

correlation lengths of 120 gridblocks in the horizontal direction and 10 gridblocks in the vertical direction. We assumed longer ranges in the horizontal direction to mimic the presence of the elongated porous horizons "HP" observed in the real field case. In the vertical direction, we assumed the existence of five horizons, hence a correlation length was set to one fifth of the vertical dimension of the model. The realizations generated are considered quite heterogeneous since the correlation lengths are shorter with respect to the domain size, namely about one fifth of each dimension.

One practical issue of ensemble methods is that the finite number of members, usually no more than $O(100)$ in applications (Wen and Chen, 2005; Gillijns et al., 2006; Anderson, 2009; Evensen, 2009a), can cause spurious covariances between widely separated components of the parameter vector and between components of the vector of parameters and vector of predicted data. Since these long-range spurious correlations are non-physical, they can bring an undesirable response in the update computed. In general, they cause an underestimation of the ensemble variance which can potentially lead to "filter divergence" (Evensen, 2009a). This term refers to the situation where the assimilation of new measurements stops being effective because the spread of the ensemble has overly reduced or has "collapsed" to take them into account.

Considering that the observations are spatially restricted to the modeled domain (**Figure 1**) and that the hydraulic perturbation from the shaft is local relatively to the horizontal extension of the model, such long range correlations will be inevitable in this synthetic case. Therefore, it is necessary to try to filter them out as much as possible before computing the update. To this end, we will use a simple approach which consists in "localizing" the Kalman gain matrix so that only the parameters located within a certain distance of an observation will be influenced by this observation during the update (Chen and Oliver, 2017). We will multiply the Kalman gain element-wisely with a "localization matrix" of the same size, i.e., number of parameters by number of data, which each entry will be a factor between 0 and 1 calculated by the correlation function defined by Gaspari and Cohn (1999)

$$\rho = \begin{cases} -\frac{1}{4}(\frac{\delta}{L})^5 + \frac{1}{2}(\frac{\delta}{L})^4 + \frac{5}{8}(\frac{\delta}{L})^3 - \frac{5}{3}(\frac{\delta}{L})^2 + 1, \text{ if } 0 \leq \delta \leq L \\ \frac{1}{12}(\frac{\delta}{L})^5 + \frac{1}{2}(\frac{\delta}{L})^4 + \frac{5}{8}(\frac{\delta}{L})^3 - \frac{5}{3}(\frac{\delta}{L})^2 - 5(\frac{\delta}{L}) + 4 - \frac{2}{3}(\frac{\delta}{L})^{-1}, \\ \qquad \text{if } L \leq \delta \leq 2L \\ 0, \text{ if } \delta > 2L \end{cases} \tag{10}$$

where δ denotes the distance between each couple of parameter and data variable and L is a predefined "critical length" beyond which the contribution of the data to the parameter update will be negligible. In this way, long-range spurious correlations will be partly removed from the Kalman gain and the performance of the ensemble method will be improved. To avoid potential discontinuities in the localized parameter update between the two columns of data points (**Figure 1**), we set L to 1,600 m, hence over a little larger distance than the distance separating the points horizontally, in all tests. In this synthetic case, this distance is consistent with the domain of the model that is identifiable given our data set, namely the region influenced by the observed hydraulic perturbation.

The ensemble smoother, LM-EnRML and ES-MDA all consider the vector of observations in their update equation as a random vector with the addition of a random noise vector sampled from the measurement error covariance C_{err}. For all tests performed, we assumed a diagonal covariance matrix of independent measurement errors of $0.05m^2$ for every head observation. When flowrate data were assimilated in addition to the hydraulic heads, C_{err} also included independent flowrate measurement errors set equal to the square of 20% of the flowrate value. For the application of ES-MDA with transformed state variables, C_{err} was transformed as described in section 2.3.

4. PERFORMANCE CRITERIA

Ideally, the application of ensemble Kalman methods for our parameter estimation problem aims to satisfy the following two criteria: (1) to reproduce the dynamic observations of state variables with the final ensemble of conditioned realizations, and (2) to obtain a final ensemble of conditional realizations which variations around the mean correctly quantify the uncertainty. As we assumed in this synthetic case no model errors or errors in the choice of the prior distribution used to sample the parameter space, meeting both these criteria should ensure that the ensemble-based uncertainty in the model predictions will be well-represented.

To assess the quality of the fit between the simulated $g(m_i)$ and observed data d_{obs} using either LM-EnRML or ES-MDA with or without transformed data, we will consider the evolution of the sum of squared errors

$$D(m_i) = (g(m_i) - d_{obs})^T (g(m_i) - d_{obs}) \tag{11}$$

Note that this sum of squared errors does not correspond to the data mismatch term of the objective function which is actually minimized by each algorithm considered (cf. section 2). Indeed, d_{obs} here denotes the vector of the original observations even in the case of ES-MDA when considering transformed data. In

this manner, we will be able to more easily compare the data mismatch results obtained using LM-EnRML, ES-MDA with or without transformed data. Note also that in this defined sum of squared residuals, no normalization by the observation error is considered. Therefore, when assimilating both heads and flowrates, this sum will be dominated by the head data.

In addition, to compare the efficiency of ES-MDA when using transformed vs. untransformed data, we will also compute for each case the data mismatch term of the different cost functions being minimized at each iteration. Indeed, Evensen (2018) noted that each updated estimate m_i^{k+1} of ES-MDA actually corresponds to an estimation of the minimum of a cost function written as

$$\begin{aligned} D(m_i^{k+1}) =& \frac{1}{2}(g(m_i^{k+1}) - d_{obs} - \sqrt{\alpha_k} z_{d,i})^T (\alpha_k C_{err})^{-1} \\ & \times (g(m_i^{k+1}) - d_{obs} - \sqrt{\alpha_k} z_{d,i}) \\ & + \frac{1}{2}(m_i^{k+1} - m_i^k)^T (C_M^k)^{-1} (m_i^{k+1} - m_i^k) \end{aligned} \tag{12}$$

where $m_i^{k=1} = m_i^{pr}$ and $C_M^{k=1} = C_M^{pr}$ with C_M^{pr} the prior model covariance approximated from the ensemble around the prior ensemble mean $m^{pr} = \overline{m_i^{pr}}$.

Once an acceptable data match has been reached during the conditioning procedure, the uncertainty based on the variability of the final ensemble of conditioned realizations can be assessed. In practice, ensemble-based Kalman methods are known to particularly overestimate the uncertainty reduction when they are not applied in sufficiently optimal conditions. In particular, the use of a finite ensemble inevitably introduces sampling errors in the approximated covariances which can lead to an overly reduced variance of the conditioned ensemble. Moreover, the solution space spanned by the ensemble when computing the update will likely be under-sampled (Evensen, 2009a) and the estimated uncertainty will not be reliable. To reduce these sampling errors, a natural solution is to increase the ensemble size. However, because the Monte Carlo sampling errors decrease proportionally to $1/\sqrt{N}$, with N the ensemble size, the improved performance of the ensemble method can come at a significant computational cost. For example, Chen and Zhang (2006) noted that an ensemble size of 1,000 allowed EnKF to obtain an accurate uncertainty estimation for their synthetic case, but ultimately concluded that an ensemble size of 200 was sufficient to achieve results with both accuracy and efficiency. When applying iterative ensemble smoothers, the choice of the ensemble size will be particularly constrained by the time needed to run all the forward predictions of the ensemble during one iteration. Consequently, larger ensemble sizes than $O(100)$ are generally not affordable if the available computational resources are not sufficient to efficiently parallelize the prediction step of every ensemble member. For these practical reasons, we will discuss the performance of the different ensemble methods based on tests involving ensemble sizes of the order of $O(10^2)$ at most.

To assess the accuracy of the uncertainty captured by the final conditioned ensemble, a common approach consists in comparing the error between the ensemble mean of log K realizations and the known reference to the ensemble mean error that was obtained by the ensemble method, also called the "ensemble spread" (Houtekamer and Mitchell, 1998). For the first error mentioned, we will simply compute the root-mean-squared error (RMSE) as follows

$$RMSE = \sqrt{\frac{1}{N_m} \sum_{i=1}^{N_m} (m_{true}^i - m_e^i)^2} \tag{13}$$

where N_m is the number of gridblocks, m_{true} is the reference log K field, and m_e stands for the mean of the ensemble of log hydraulic conductivity fields.

The ensemble spread, here noted S_{ens}, corresponds to an average uncertainty on the log K property calculated from the ensemble

$$S_{ens} = \sqrt{\frac{1}{N_m} \sum_{i=1}^{N_m} \sigma_{ens,i}^2} \tag{14}$$

where $\sigma_{ens,i}^2$ corresponds to the variance estimated from the ensemble of log K realizations at one gridblock i. For our uncertainty analysis, we will systematically evaluate these average errors by considering the gridblocks of the model located within a certain distance from the seepage boundary which includes all the data points. In particular, we will set this distance equal to the "critical length" used for the localization of the update (cf. section 3.3). In this way, the comparison between the RMSE and the ensemble spread will be based on the set of parameters which are the most informed by the observations and hence which uncertainty will decrease the most during the assimilation.

One goal of this analysis will be to illustrate how the spread between ensemble members updated using either LM-EnRML, ES-MDA with or without transformed data is representative of the difference between the ensemble mean and the reference (RMSE) depending on the ensemble size. Indeed, as the size of the ensemble increases, the discrepancy between the final conditioned ensemble and the RMSE should reduce thanks to the reduced sampling errors. However as previously mentioned, a trade-off between accuracy of the uncertainty and computational efficiency will need to be found. In the end, if the ensemble spread is not overly underestimating the RMSE, we will consider that the ensemble method has performed correctly given its intrinsic limitations. In addition, although the inverse problem is ill-posed, i.e., there are more parameters to infer than data to inform them in a unique way, we expect for this synthetic case that the information contained in the data set (i.e., hydraulic heads and flowrates) is sufficient in order to yield after the data assimilation estimates that are closer to the reference than initially in the most updated region of parameters. Therefore, the performance of the smoother will also be assessed in its ability to decrease the RMSE value.

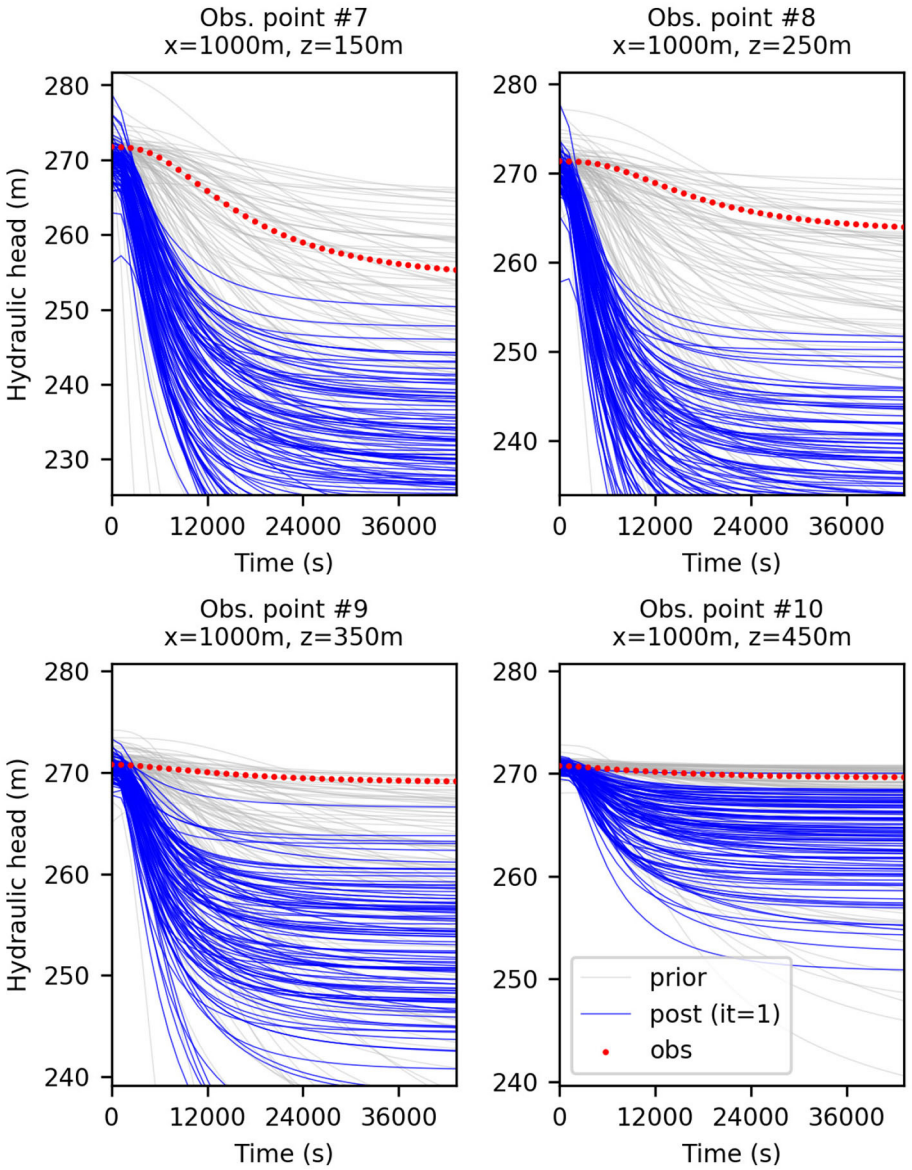


FIGURE 3 | Predicted hydraulic head at four different locations, before (in gray) and after (in blue) the assimilation with ES of the head data without normal-score transform. The red dots are the observed data; The ensemble size is 100.

5. RESULTS AND DISCUSSION

5.1. Ensemble Smoother and Benefit of Data Transformation

As expected in a non-linear case, the application of a non-iterative ensemble method such as the ensemble smoother (ES) results in an insufficient match between the final ensemble of predicted data, i.e., computed from the ensemble of updated realizations, and the observations. **Figure 3** shows for three locations at different depths 1 km away from the producing shaft the simulated heads before and after the update of 100 realizations with ES. By using the same initial ensemble, the evolution of the data mismatch obtained from the assimilation of the head data only and both the head and flowrate data simultaneously are shown respectively in **Figures 4A,C**. Because the represented data mismatch does not include the inflated measurement error in the observed data and hence the normalization by the error variance, as defined in section 4, the additional contribution to the mismatch from the flowrates in **Figure 4C** is not visible. Even so, compared to **Figure 4A**), we can observe that the assimilation of the flowrates in addition to the heads allows the data mismatch to decrease thanks to a better match of the head data.

The strategy of transforming the data prior to each update of ES-MDA, as described in section 2.3, was also tested here with ES using the same initial ensemble. Indeed, ES is equivalent

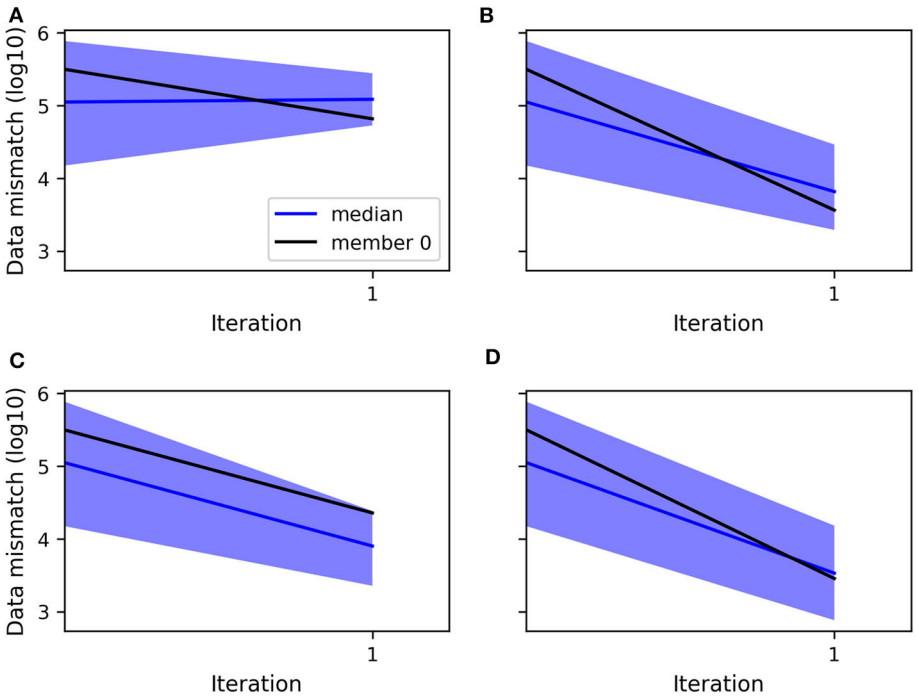


FIGURE 4 | Evolution of the decimal logarithm of the data mismatch when assimilating with ES: **(A)** the head data, **(B)** the transformed head data, **(C)** both the head and flowrate data, **(D)** both the transformed head and transformed flowrate data. The blue curve is the median, the blue area corresponds to the region between the percentiles P5 and P95 of the ensemble of values, and the black curve is the values computed for the ensemble member no. 0.

to ES-MDA with only one iteration and hence applies once a local linearization of the model around the initial parameter estimate to compute its single update. As shown in **Figures 4B,D**, whether only the head data are assimilated or both the head and flowrate data, the data mismatch is reduced in both cases with a slightly lower mismatch in the case mentioned last. Therefore, compared to the case where only untransformed head are assimilated (**Figure 4A**), the data transformation seems to have a positive impact on the decreased data mismatch which, we remind, considers the original data even for the transformed case. The improved history match can be observed for example by comparing **Figure 3** with **Figure 5**. In particular, we note that when data transformation is used, the ensemble of predicted values better tracks the observations, most certainly as a result of a more accurate update of each ensemble member. Indeed, the comparison of the RMSE and the ensemble spread in **Table 1** for each one of the cases considered in **Figure 4** shows that the assimilation of the transformed heads instead of the original data results in a more accurate estimation of the ensemble mean and spread thanks to a decrease of the RMSE which reduces the discrepancy with the reduced ensemble spread. As for the assimilation of both the head and flowrate data simultaneously, we can see that compared to when the data are transformed, the assimilation is less efficient with a less reduced ensemble spread and a RMSE that has almost not reduced. Based on both the evolution of the RMSE and the "coverage", which we defined as the percentage of true values of the reference captured by the final ensemble spread in the most updated region of the model, we can conclude from **Table 1** that transforming the data clearly benefits the performance of ES. However, we note that the assimilation of both the transformed head and flowrate data with ES does not allow to further reduce the RMSE compared to when using the transformed heads only, although the ensemble spread is further reduced.

5.2. Comparing the Accuracy of LM-EnRML and ES-MDA Estimates

Although ES-MDA and the LM-EnRML were developed with the aim of improving the data match obtained with ES in the non-linear case, the iterative updates will likely cause a more important underestimation of the uncertainty than with ES. Indeed, at each iteration, the sampling errors in the approximated covariances of the Kalman gain will affect the estimates of the method, and more particularly the ensemble uncertainty. Seeing as how the Monte Carlo sampling errors decrease according to $1/\sqrt{N}$ where N is the ensemble size (Evensen, 2009a), we tested for our synthetic case the effect of increasing N on the performance of ES-MDA and LM-EnRML when only assimilating the head data. **Table 2** shows the evolution of the RMSE and the ensemble spread when applying ES-MDA with 4 arbitrarily defined iterations. Whether we consider an ensemble of 100, 200, or 400, hence of the order of $O(10^2)$ to remain computationally efficient, we note that starting from the first update, the ensemble spread measuring the variability between the ensemble members systematically underestimates the RMSE which measures the error between the reference and the ensemble

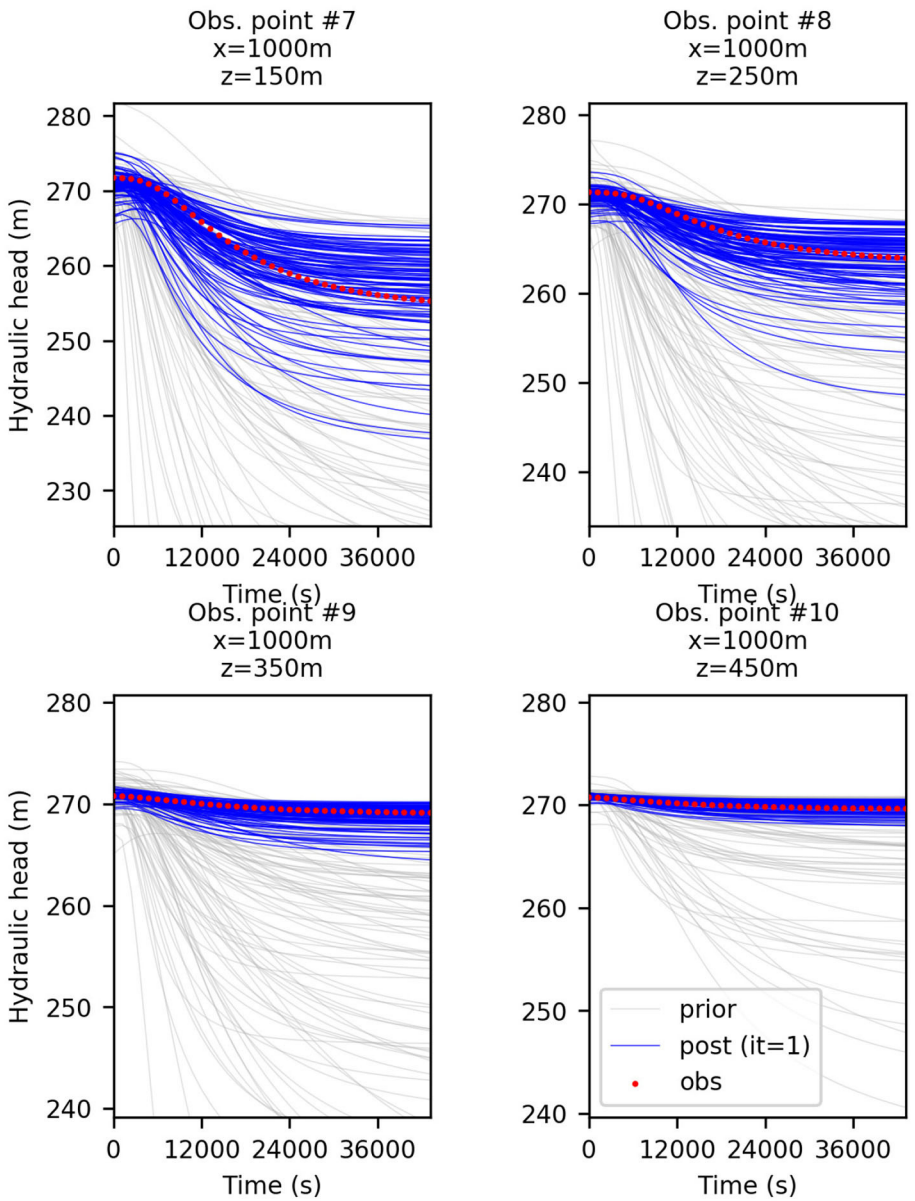


FIGURE 5 | Predicted hydraulic head at four different locations, before (in gray) and after (in blue) the assimilation with ES of normal-score transformed head data. The red dots are the observed data; The ensemble size is 100.

mean. This discrepancy with the RMSE further increases with each iteration as the spread keeps decreasing while the RMSE remains higher. Although we do not show it here, not localizing the update resulted in a larger discrepancy between the RMSE and ensemble spread for the different ensemble sizes tested.

Nevertheless, we note that the larger the ensemble size, the more reduced the gap between the RMSE and the ensemble spread becomes. As expected, the estimated uncertainty gains in accuracy with the ensemble size. We note however that the most significant improvement occurs when increasing the ensemble from 100 to 200 realizations for both algorithms (**Tables 2**, **3**). Indeed, as indicated for ES-MDA in **Table 2**, the coverage, defined as the proportion of reference values that are within the ensemble range in the most updated region of the model, has more than doubled with a final value of 77% after the same number of iterations. In a similar manner, the coverage also doubled with LM-EnRML (**Table 3**). Moreover, when using an ensemble size of 100, we can see that the evolution of the RMSE is particularly unstable for both LM-EnRML and ES-MDA. The ensemble means estimated with LM-EnRML even lead to a final RMSE value which is much higher than initially. In contrast, an ensemble size of 200 allows LM-EnRML to provide a more accurate estimation of the true uncertainty than with ES-MDA. The results shown in **Table 3** were all calculated based on the

TABLE 1 | Influence of different types of data assimilated (transformed or not) on the RMSE, ensemble spread, and coverage calculated with ES.

Assimilated data	RMSE	Ens. spread	Coverage (%)
h	0.72–1.00	0.70–0.42	57
h, q	0.72–0.69	0.70–0.53	93
transf. h	0.72–0.58	0.70–0.44	91
transf. h, transf. q	0.72–0.59	0.70–0.41	88

The ensemble size was 100 for all the tests shown in this table.

TABLE 2 | Effect of increasing the ensemble size on the RMSE, ensemble spread, and coverage after 4 predefined ES-MDA iterations.

Ens. size	Assimilated data	RMSE	Ens. spread	Coverage (%)
100	h	0.72–0.71	0.70–0.19	37
200	h	0.71–0.66	0.70–0.32	77
400	h	0.72–0.57	0.70–0.40	96

TABLE 3 | Effect of increasing the ensemble size on the RMSE, ensemble spread, and coverage with LM-EnRML at iteration 8.

Ens. size	Assimilated data	RMSE	Ens. spread	Coverage (%)
100	h	0.72–0.88	0.70–0.19	49
200	h	0.70–0.60	0.70–0.34	93
400	h	0.72–0.66	0.70–0.39	94

estimates of LM-EnRML obtained after 8 iterations when the reduction of the objective function seemed to have decreased to a sufficiently low data mismatch value as shown in **Figure 6A**. We note also from **Figure 6**, however, that it seems to take more iterations for LM-EnRML to converge to the same level of data match achieved by ES-MDA using 6 predefined iterations.

From all these tests, it is clear that the larger the size, the more accurate the uncertainty estimated by LM-EnRML or ES-MDA will be. However, as can be observed in **Table 2**, this improvement in the estimates comes at an increased computational cost. For our synthetic problem, we note that an ensemble of 400 realizations would still not allow the ensemble to represent correctly the true error on the ensemble mean. Since increasing the ensemble size does not seem to be a reasonable option computationally, it seems that the application of such ensemble methods necessarily requires to achieve a balance between accuracy and computational efficiency. This computational efficiency is of even more importance for large-scale problems since the biggest constraint would often be the long flow simulation time required to run the prediction step for each ensemble member. Hence the larger the ensemble, the longer it will take to complete all the predictions of the ensemble before being able to start the update step. Indeed, we remind that the update of each member is based on the same ensemble-based Kalman gain matrix. It is worth mentioning that this matter of efficiency also concerns the case where the ensemble of predictions steps is parallelized. Indeed, it would still require to have enough resources to efficiently run the whole ensemble of forward simulations simultaneously after each update step.

From **Tables 2**, **3**, it is clear that the improvement on both the ensemble mean and spread is very significant when increasing the size from 100 to 200 with either LM-EnRML and ES-MDA. However, the differences in the estimated ensemble mean fields shown in **Figure 7** suggest that the two methods converge toward different solutions. This is confirmed by the fact that when increasing the ensemble size from 200 to 400, the accuracy of both the error and uncertainty estimated with ES-MDA keeps increasing whereas for LM-EnMRL, no visible improvement is observed as if an optimal ensemble size has been reached around 200.

Although LM-EnRML provides the more accurate estimation of the uncertainty compared to ES-MDA for any of the tested ensemble sizes, ES-MDA seems to converge faster, as shown in **Figure 6**. Indeed, it takes more iterations with LM-EnRML to reach the same mismatch value which corresponds to a good match as obtained with ES-MDA using 6 iterations. In the end, although the estimates of ES-MDA definitely improve with an ensemble size of 400, we will consider that using an ensemble of size 200 with either ES-MDA or LM-EnRML in our case is acceptable in order to achieve both accuracy and efficiency.

5.3. Assimilating Both Hydraulic Head and Flowrate Data With LM-EnRML and ES-MDA

As shown in **Table 4**, the assimilation of both head and flowrate data with LM-EnRML yields estimates of similar accuracy to when only the heads are assimilated. Neither the larger number of observations nor the higher "weights" of the flowrates in the objective function, given their much lower magnitudes, seem to have affected the minimization process. **Figure 8A** shows in fact that the mismatch associated to the flowrate data in the objective function, i.e., the sum of squared residuals weighted by the inverse of the measurement variance, is mostly reduced after the first iteration so that ultimately the remaining of the optimization deals with the reduction of the mismatch associated to the head data (**Figure 8B**). Although the convergence of LM-EnRML seems effective when assimilating different data types, we note that the accuracy of the estimates does not improve with the additional information coming from the assimilated flowrates. This lack of improvement could be related to possible over-corrections of the parameters during the first iteration as often observed with gradient-based methods. One way to reduce such over-corrections would be to artificially increase the measurement errors. We did not try that however since the parameter estimates obtained seem sufficiently accurate in terms of the estimated uncertainty.

For the assimilation of both data types, we first note that ES-MDA converged more rapidly, i.e., in 8 iterations, to the same level of data match achieved with LM-EnRML after 16 iterations (**Figures 8A**, **9A**). For brevity, we do not show the associated good match of the head and flowrate observations obtained with

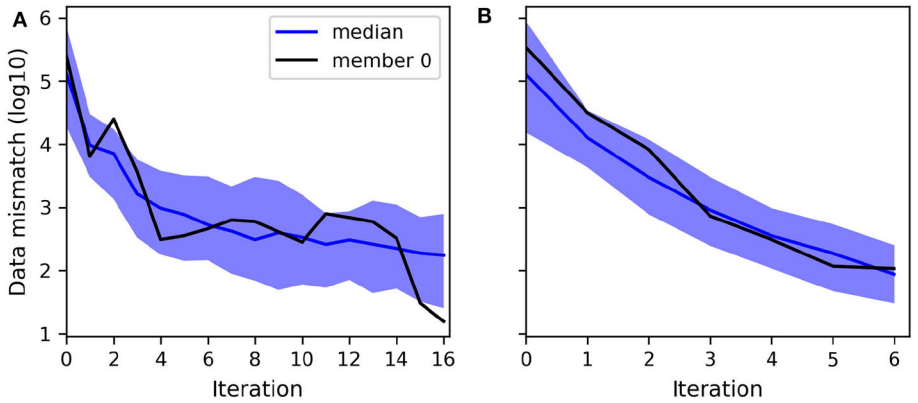


FIGURE 6 | Evolution of the decimal logarithm of the data mismatch when assimilating head data with: **(A)** LM-EnRML, **(B)** ES-MDA (×6). The ensemble size is 200.

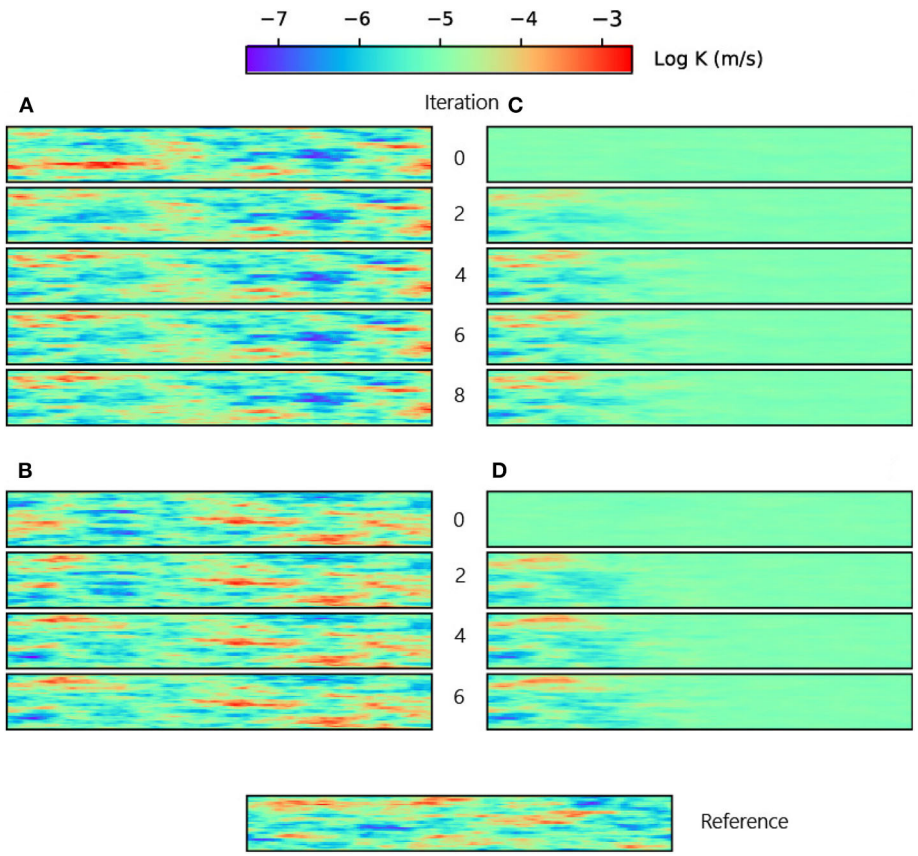


FIGURE 7 | Two initial different realizations of log hydraulic conductivity being updated, respectively, with LM-EnRML **(A)** and ES-MDA **(B)** throughout the iterations. Panels **(C,D)** show the corresponding ensemble mean fields at each iteration. The ensemble size is 200.

both methods. However, unlike previously with LM-EnRML, the mean and variance estimates obtained with ES-MDA are less accurate than those obtained by the assimilation of head data only (**Table 4**). This illustrates that the different data types are not properly taken into account with ES-MDA during the multiple assimilations compared to LM-EnRML.

As done previously with ES in order to mitigate the effects of non-Gaussianity on the performance of the method, we applied a normal-score transform on the data so as to obtain ensemble distributions of predicted variables which are locally Gaussian, for each data type, before performing each update step of ES-MDA in the transformed data space. As shown in **Figure 10**,

the proposed state transformation allows ES-MDA to achieve an acceptable match of the head and flowrate observations after 8 iterations. **Table 4** shows the estimates obtained after assimilating only the transformed heads resulted in a slight improvement. The coverage has slightly increased, which indicates that the estimated uncertainty, although still underestimated, has improved a little. A much more significant improvement, however, can be observed when both the head and flowrate data are assimilated in the transformed space. Compared to when no transformation is applied, the assimilation of both transformed data types is much more efficient given the larger variability of the ensemble and the smaller RMSE.

TABLE 4 | Effect of assimilating the flowrate data in addition to the hydraulic heads on the RMSE, ensemble spread, and coverage with LM-EnRML and ES-MDA with or without transformed data.

Algorithm	Assimilated data	RMSE	Ensemble spread	Coverage (%)
LM-EnRML	h	0.72–0.60	0.70–0.34	93
LM-EnRML	h, q	0.71–0.60	0.69–0.34	93
ES-MDA(×6)	h	0.72–0.65	0.69–0.29	72
ES-MDA(×8)	h, q	0.72–0.70	0.70–0.20	58
ES-MDA(×8)	transf. h	0.72–0.66	0.69–0.31	80
ES-MDA(×8)	transf. h, q	0.70–0.63	0.70–0.30	82

The ensemble size was 200 for all the tests shown in this table. The corresponding number of iterations was chosen so as to obtain an acceptable and comparable data match for all the results.

Figure 9 shows that when both data types are transformed, the decrease of the cost functions minimized for each iteration of ES-MDA (Evensen, 2018) is more stable than when no transformation is applied. In the case without transformation, the larger range of the initial ensemble of values reflects the differences of magnitude of the mismatch associated to the flowrate data on the one hand, and the one associated to the heads on the other hand. As shown in **Figure 11A**, the distribution is spread around two distinct modes. Hence the less stable reduction observed is probably related to the contribution of the flowrate data which dominates the cost function to

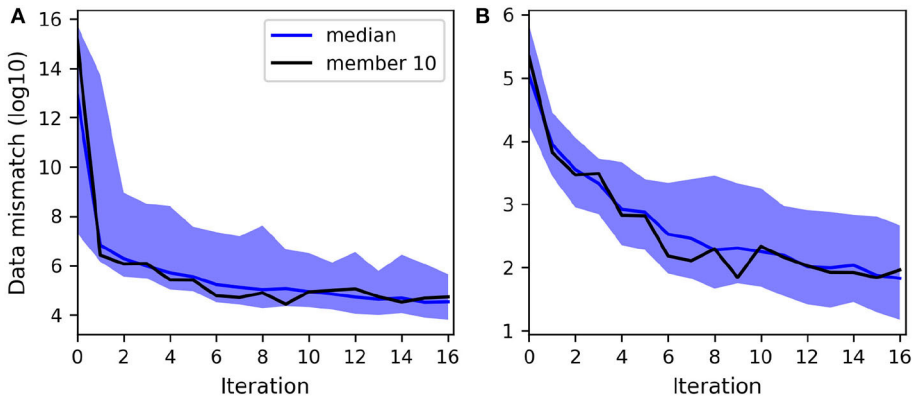


FIGURE 8 | Evolution of the decimal logarithm of the data mismatch when assimilating both head and flowrate data with LM-EnRML: **(A)** With the normalization of the residuals by their variance (as minimized by the objective function). **(B)** Without normalization, hence the values reflect the head data mismatch reduction mainly. The ensemble size is 200.

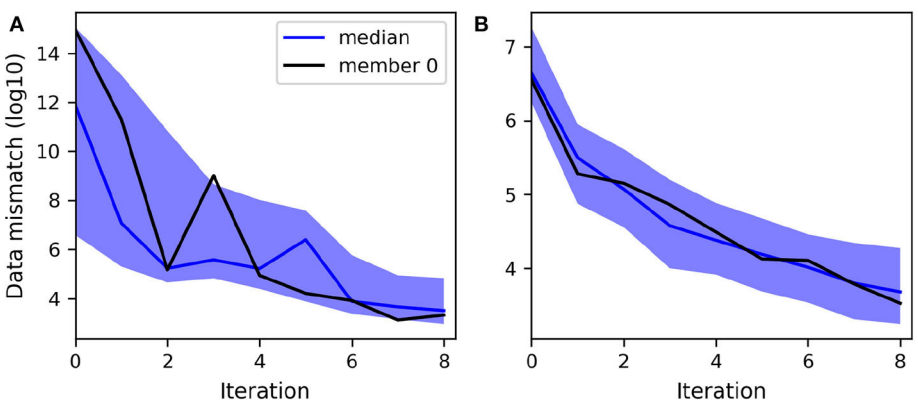


FIGURE 9 | Evolution of the decimal logarithm of the data mismatch term of the cost function minimized at each iteration when assimilating both head and flowrate data with ES-MDA: **(A)** Without transformed data. **(B)** With transformed data. The ensemble size is 200.

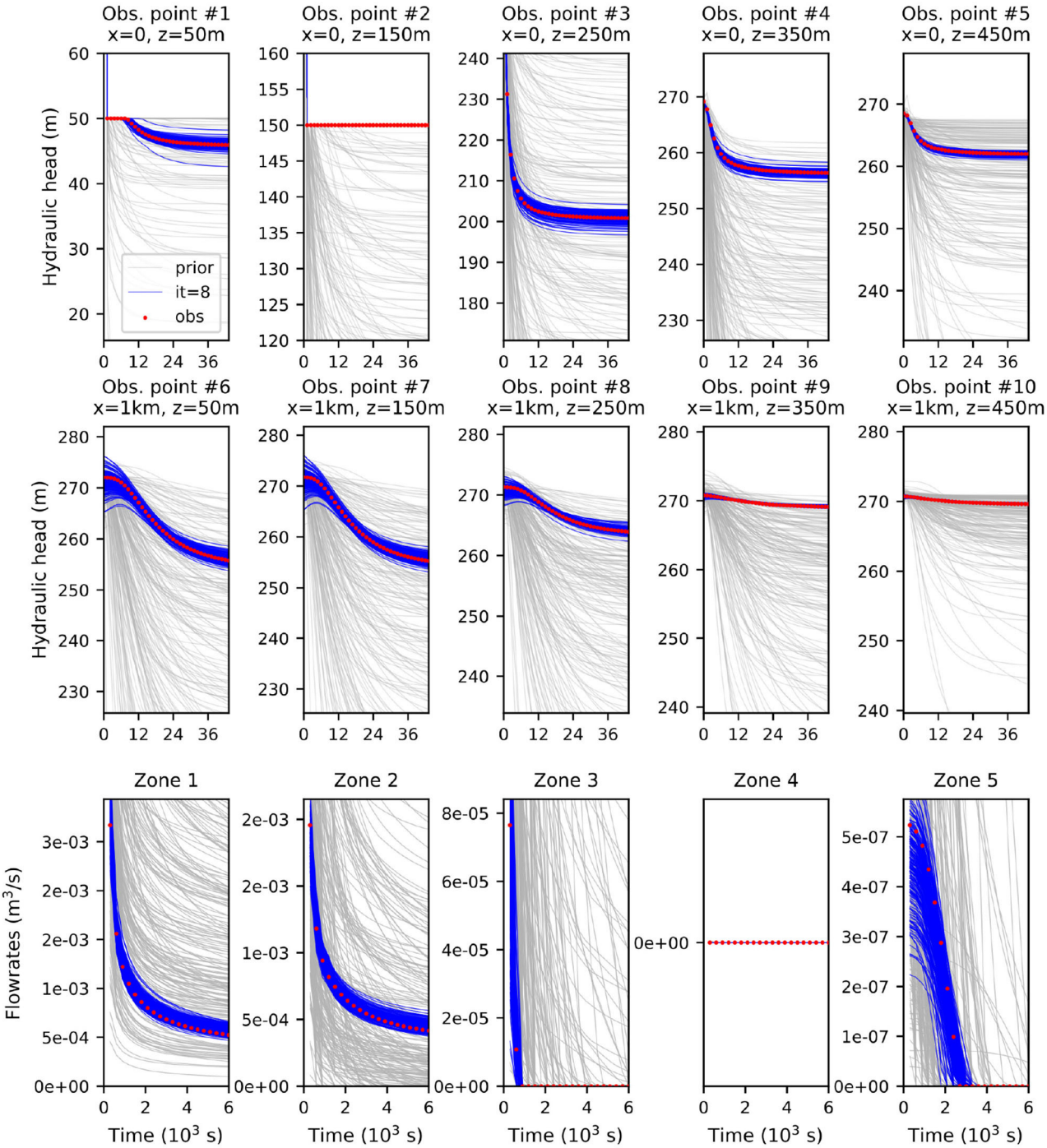


FIGURE 10 | Predicted hydraulic head and flowrate data at the observation locations, before (in gray) and after (in blue) the assimilation of both the transformed hydraulic heads and transformed flowrate data with ES-MDA. The red dots are the observed data; the ensemble size is 200.

minimize because of their much more smaller measurement errors. In contrast, the initial ensemble of mismatch values resulting from the transformed data (**Figure 9B**) is much less spread and resembles a Gaussian distribution (**Figure 11**). By making Gaussian the distribution of all mismatch values defining the first cost function to be minimized with ES-MDA, the normal-score transform applied to each data type seems to have benefited the data assimilation with ES-MDA. Since a new cost function is minimized after each ES-MDA iteration, we suppose that the initial Gaussian distribution of all the data mismatch values resulting from the transformation contributes to make the minimization more efficient by allowing to take equally

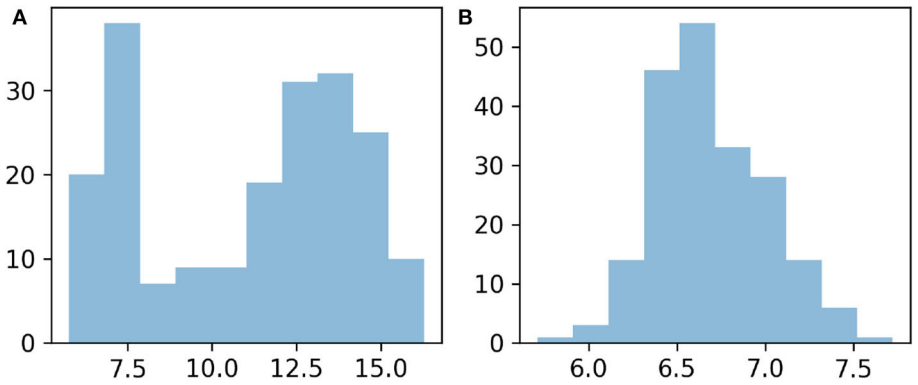


FIGURE 11 | Distribution of the decimal logarithm of the sum of weighted squared errors prior to the assimilation of head and flowrate data with ES-MDA: **(A)** Without transformation. **(B)** With transformation.

account of each data type in the assimilation from the start of the assimilation. However, as discussed by Schoniger et al. (2012), it is important to note that such local transformation of the state variables would not have led to a successful application of ES-MDA if the multivariate dependence of the state variables was not sufficiently Gaussian in the first place. Indeed, since the proposed transformation only affects the marginal distribution, it is essential that the multivariate structure of the state variables is near-Gaussian in order to really observe an improvement of the performance of the ensemble Kalman method. Crestani et al. (2013) illustrated this point by observing that similar local transformation performed on concentration data can deteriorate the performance of EnKF because of an insufficient dependence between the multi-Gaussian log hydraulic conductivity parameters and the concentrations.

It could be argued that equal contributions from both the heads and flowrate data to the first cost function could also have been obtained in the case without transformation by decreasing the measurement variance of the head measurements. However, the use of very small values of measurement errors can cause problems during the data assimilation as it may lead to severe underestimations of the parameter uncertainty due to the overconfidence in the assimilated data.

6. CONCLUSION

In this synthetic study, two iterative forms of ensemble smoother, LM-EnRML (Chen and Oliver, 2013) and ES-MDA (Emerick and Reynolds, 2012a), were applied and compared. They were both used to condition an ensemble of multi-Gaussian log hydraulic conductivity fields to transient hydraulic data. Given the non-linear dynamics of the groundwater flow problem inspired by a real hydraulic situation, the results show the necessity of using an iterative instead of non-iterative ensemble smoother in order to obtain an ensemble of hydraulic conductivity fields which all match properly the data.

Despite the multi-Gaussian log K distribution considered, the repeated assimilations of the data set have highlighted the known tendency of ensemble Kalman methods to underestimate the true error due to the finite ensemble size. A trade-off between accuracy of the estimation and computational efficiency hence needs to be found when applying such methods. Our uncertainty analysis based on the comparison between the ensemble spread and the true error, here accessible in this synthetic study, clearly indicated the existence of a threshold ensemble size below which the updated mean and variance are not reliable. For our case, we found that the final estimates using an ensemble of 100 were not acceptable whereas increasing the ensemble to a still reasonable size of 200 improved significantly the accuracy of the updated mean and spread of the ensemble for both algorithms. Because determining this threshold ensemble size a priori is not possible, it seems all the more important before tackling large-scale applications to try to estimate this threshold size on a smaller-sized problem but which will take into account the main characteristics of the real inverse problem.

Although LM-EnRML outperforms ES-MDA in terms of the accuracy of the estimated uncertainty when using an ensemble size of 200, the performance of ES-MDA seems to improve steadily with the ensemble size while LM-EnRML does not. In particular, the RMSE of the ES-MDA estimates keeps decreasing. This particularly underlines that both algorithms do not converge to the same solution. It is not so surprising considering that LM-EnRML was derived so as to minimize an objective function using a gradient-based approach while ES-MDA was derived to minimize the variance of the error (Chen and Oliver, 2013; Evensen, 2018).

The benefit of transforming the state variables with ES-MDA was mainly observed when assimilating the hydraulic head and flowrate data simultaneously. Indeed, the normal-score transform of the state variables allowed a normalization of the magnitudes of all data types to values drawn from a standard normal distribution. Consequently, both data types could be more equally taken into account during the data assimilation as shown by the more stable reduction of the cost functions compared to the application of ES-MDA without transformation. However, as commented in previous studies (Schoniger et al., 2012; Crestani et al., 2013), the applicability of such state transformation could be questioned in cases where

the multi-Gaussian dependence of the state variables is not as strong as in this multi-Gaussian log K case, either as a result of the non-linearity of the model and/or because of an initial non-multi-Gaussian parameter distribution.

Although this paper focused on the performance of iterative ensemble smoothers in the specific case of multi-Gaussian distributions, their application to condition non-Gaussian distributions is of special interest for many real field applications. For a case such as the one of the ANDRA where the built groundwater model is very high-dimensional, the observed efficiency of iterative ensemble smoothers to decrease the data mismatch in the multi-Gaussian case is very interesting. However, as summarized by Zhou et al. (2014), one main issue when conditioning directly non-multi-Gaussian distributions using ensemble Kalman methods is that the consistency of the initial geological structures is lost. A relevant future perspective for the case of the ANDRA will hence to consider an appropriate parameterization for the application of iterative ensemble smoother methods to non-Gaussian heterogeneous fields such as when generated with truncated Gaussian or multiple-point statistics simulation techniques.

DATA AVAILABILITY STATEMENT

A repository with the data set used in this paper is published on Zenodo (https://doi.org/10.5281/zenodo.3878820).

AUTHOR CONTRIBUTIONS

D-TL implemented, tested, and analyzed the methods presented and wrote the article. JK helped in the set up of the synthetic groundwater flow model and supervised the work. PR gave critical feedback on the intermediary results which helped improving the final manuscript. HB gave important information of the real problem which inspired the addressed synthetic case. PP commented on the progress of the project and manuscript. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

The authors thank the University of Neuchâtel for funding this work and for providing the necessary computational resources.

REFERENCES

Anderson, J. L. (2003). An ensemble adjustment Kalman filter for data assimilation. *Mthly. Weather Rev.* 129, 2894–2903. doi: 10.1175/1520-0493(2001)129<2884:AEAKFF>2.0.CO;2

Anderson, J. L. (2009). Ensemble Kalman filters for large geophysical applications: adaptive algorithms for improving ensemble filter performance. *IEEE Control Syst.* 29, 66–82. doi: 10.1109/MCS.2009.932222

Benabderrahmane, H., Kerrou, J., Tacher, L., Deman, G., and Perrochet, P. (2014). Modelling of predictive hydraulic impacts of a potential radioactive waste geological repository on the Meuse/Haute-Marne multilayered aquifer system (France). *J. Appl. Math. Phys.* 02, 1085–1090. doi: 10.4236/jamp.2014.212125

Bourgeat, A., Kern, M., Schumacher, S., and Talandier, J. (2004). The COUPLEX test cases: nuclear waste disposal simulation. *Comput. Geosci.* 8, 83–98. doi: 10.1023/B:COMG.0000035073.03009.5d

Chen, Y., and Oliver, D. S. (2012). Ensemble randomized maximum likelihood method as an iterative ensemble smoother. *Math. Geosci.* 44, 1–26. doi: 10.1007/s11004-011-9376-z

Chen, Y., and Oliver, D. S. (2013). Levenberg-Marquardt forms of the iterative ensemble smoother for efficient history matching and uncertainty quantification. *Comput. Geosci.* 17, 689–703. doi: 10.1007/s10596-013-9351-5

Chen, Y., and Oliver, D. S. (2017). Localization and regularization for iterative ensemble smoothers. *Comput. Geosci.* 21, 13–30. doi: 10.1007/s10596-016-9599-7

Chen, Y., Oliver, D. S., and Zhang, D. (2009). Data assimilation for nonlinear problems by ensemble Kalman filter with reparameterization. *J. Petrol. Sci. Eng.* 66, 1–14. doi: 10.1016/j.petrol.2008.12.002

Chen, Y., and Zhang, D. (2006). Data assimilation for transient flow in geologic formations via ensemble Kalman filter. *Adv. Water Resour.* 29, 1107–1122. doi: 10.1016/j.advwatres.2005.09.007

Cornaton, F. (2014). *GroundWater (GW) A 3-D GroundWater and SurfaceWater Flow, Mass Transport and Heat Transfer Finite Element Simulator*. Technical report. University of Neuchâtel.

Crestani, E., Camporese, M., Baú, D., and Salandin, P. (2013). Ensemble Kalman filter versus ensemble smoother for assessing hydraulic conductivity via tracer test data assimilation. *Hydrol. Earth Syst. Sci.* 17, 1517–1531. doi: 10.5194/hess-17-1517-2013

Deman, G., Konakli, K., Sudret, B., Kerrou, J., Perrochet, P., and Benabderrahmane, H. (2015). Using sparse polynomial chaos expansions for the global sensitivity analysis of groundwater lifetime expectancy in a multi-layered hydrogeological model. *Reliabil. Eng. Syst. Saf.* 147, 156–169. doi: 10.1016/j.ress.2015.11.005

Emerick, A. A. (2016). Analysis of the performance of ensemble-based assimilation of production and seismic data. *J. Petrol. Sci. Eng.* 139, 219–239. doi: 10.1016/j.petrol.2016.01.029

Emerick, A. A., and Reynolds, A. C. (2012a). Ensemble smoother with multiple data assimilation. *Comput. Geosci.* 55, 3–15. doi: 10.1016/j.cageo.2012.03.011

Emerick, A. A., and Reynolds, A. C. (2012b). History matching time-lapse seismic data using the ensemble Kalman filter with multiple data assimilations. *Comput. Geosci.* 16, 639–659. doi: 10.1007/s10596-012-9275-5

Evensen, G. (1994). Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.* 99:10143. doi: 10.1029/94JC00572

Evensen, G. (2004). Sampling strategies and square root analysis schemes for the EnKF. *Ocean Dyn.* 54, 539–560. doi: 10.1007/s10236-004-0099-2

Evensen, G. (2009a). *Data Assimilation*. Berlin; Heidelberg: Springer.

Evensen, G. (2009b). The ensemble Kalman filter for combined state and parameter estimation: Monte Carlo techniques for data assimilation in large systems. *IEEE Control Syst.* 29, 83–104. doi: 10.1109/MCS.2009.932223

Evensen, G. (2018). Analysis of iterative ensemble smoothers for solving inverse problems. *Comput. Geosci.* 22, 885–908. doi: 10.1007/s10596-018-9731-y

Evensen, G., and van Leeuwen, P. J. (2002). An ensemble Kalman smoother for nonlinear dynamics. *Mthly. Weather Rev.* 128, 1852–1867. doi: 10.1175/1520-0493(2000)128<1852:AEKSFN>2.0.CO;2

Gaspari, G., and Cohn, S. E. (1999). Construction of correlation functions in two and three dimensions. *Q. J. R. Meteorol. Soc.* 125, 723–757. doi: 10.1002/qj.49712555417

Gillijns, S., Mendoza, O., Chandrasekar, J., De Moor, B., Bernstein, D., and Ridley, A. (2006). "What is the ensemble Kalman filter and how well does it work?" in *American Control Conference, 2006* (IEEE), 4448–4453. doi: 10.1109/ACC.2006.1657419

Gu, Y., and Oliver, D. S. (2007). An iterative ensemble Kalman filter for multiphase fluid flow data assimilation. *SPE J.* 12, 438–446. doi: 10.2118/108438-PA

Hendricks Franssen, H. J., and Kinzelbach, W. (2008). Real-time groundwater flow modeling with the Ensemble Kalman Filter: Joint estimation of states and parameters and the filter inbreeding problem. *Water Resour. Res.* 44:W09408. doi: 10.1029/2007WR006505

Houtekamer, P. L., and Mitchell, H. L. (1998). Data assimilation using an ensemble Kalman filter technique. *Mthly. Weather Rev*. 126, 796–811. doi: 10.1175/1520-0493(1998)126<0796:DAUAEK>2.0.CO;2

Jafarpour, B., and Khodabakhshi, M. (2011). A probability conditioning method (PCM) for nonlinear flow data integration into multipoint statistical facies simulation. *Math. Geosci*. 43, 133–164. doi: 10.1007/s11004-011-9316-y

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *J. Basic Eng*. 82:35. doi: 10.1115/1.3662552

Kerrou, J., Deman, G., Tacher, L., Benabderrahmane, H., and Perrochet, P. (2017). Numerical and polynomial modelling to assess environmental and hydraulic impacts of the future geological radwaste repository in Meuse site (France). *Environ. Modell. Softw*. 97, 157–170. doi: 10.1016/j.envsoft.2017.07.018

Lam, D.-T., Renard, P., Straubhaar, J., and Kerrou, J. (2020). Multiresolution approach to condition categorical multiple-point realizations to dynamic data with iterative ensemble smoothing. *Water Resour. Res*. 56:e2019WR025875. doi: 10.1029/2019WR025875

Le, D. H., Emerick, A. A., and Reynolds, A. C. (2016). An adaptive ensemble smoother with multiple data assimilation for assisted history matching. *SPE J*. 21, 2195–2207. doi: 10.2118/173214-PA

Li, L., Stetler, L., Cao, Z., and Davis, A. (2018). An iterative normal-score ensemble smoother for dealing with non-Gaussianity in data assimilation. *J. Hydrol*. 567, 759–766. doi: 10.1016/j.jhydrol.2018.01.038

Liu, N., and Oliver, D. S. (2005). Ensemble Kalman filter for automatic history matching of geologic facies. *J. Petrol. Sci. Eng*. 47, 147–161. doi: 10.1016/j.petrol.2005.03.006

Luo, X., Stordal, A. S., Lorentzen, R. J., and Nævdal, G. (2015). Iterative ensemble smoother as an approximate solution to a regularized minimum-average-cost problem: theory and applications. *SPE J*. 20, 962–982. doi: 10.2118/176023-PA

Moradkhani, H., Sorooshian, S., Gupta, H. V., and Houser, P. R. (2005). Dual state-parameter estimation of hydrological models using ensemble Kalman filter. *Adv. Water Resour*. 28, 135–147. doi: 10.1016/j.advwatres.2004.09.002

Schoniger, A., Nowak, W., and Hendricks Franssen, H. J. (2012). Parameter estimation by ensemble Kalman filters with transformed data: approach and application to hydraulic tomography. *Water Resour. Res*. 48, 1–18. doi: 10.1029/2011WR010462

van Leeuwen, P. J., and Evensen, G. (1996). Data assimilation and inverse methods in terms of a probabilistic formulation. *Mthly. Weather Rev*. 124, 2898–2913. doi: 10.1175/1520-0493(1996)124<2898:DAAIMI>2.0.CO;2

Wen, X. H., and Chen, W. (2005). "Real-time reservoir model updating using ensemble Kalman filter," in *SPE Reservoir Simulation Symposium* (Houston, TX), 1–14. doi: 10.2118/92991-MS

Zhou, H., Gómez-Hernández, J. J., Hendricks Franssen, H. J., and Li, L. (2011). An approach to handling non-Gaussianity of parameters and state variables in ensemble Kalman filtering. *Adv. Water Resour*. 34, 844–864. doi: 10.1016/j.advwatres.2011.04.014

Zhou, H., Gómez-Hernández, J. J., and Li, L. (2014). Inverse methods in hydrogeology: evolution and recent trends. *Adv. Water Resour*. 63, 22–37. doi: 10.1016/j.advwatres.2013.10.014

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.

frontiers in Environmental Science

ORIGINAL RESEARCH
published: 25 September 2020
doi: 10.3389/fenvs.2020.00142

Null-Space Monte Carlo Particle Backtracking to Identify Groundwater Tetrachloroethylene Sources

Loris Colombo, Luca Alberti, Pietro Mazzon and Matteo Antelmi*

Dipartimento di Ingegneria Civile e Ambientale, Politecnico di Milano, Milan, Italy

Groundwater in most urban areas around the globe is often contaminated by toxic substances. Among the various sources of contamination, industries cause the heaviest impact when toxic compounds are released underground, mainly through leaking tanks or pipelines. Some contaminants (typically chlorinated hydrocarbons) tend to persist within the underground and are hard to biodegrade. As a result, substances that leaked decades ago are still impacting groundwater. Milano and its surroundings (Functional Urban Area) is a good example of an area that has been hosting industries of all dimensions for over a century, many of them contributing to groundwater contamination from chlorinated hydrocarbons. While the position of the biggest industrial facilities is well-known, many smaller sources are hard to identify in many cases where direct surveys have not been undertaken. Furthermore, the overlapping effects of big, small, known, and unknown sources of groundwater contamination make it challenging to identify the contribution of each. In order to identify the contribution of several point sources responsible for tetrachloroethylene contamination in public water supply wells, a numerical model (MODFLOW-2005) has been implemented and calibrated using PEST in the northwestern portion of the Milano Functional Urban Area. In contaminant transport modeling, the deterministic approach is still favored over the stochastic approach because of the simplicity of its application. Nevertheless, the latter is considered by the authors as the most suitable for dealing with problems characterized by high uncertainty, such as hydrogeological parameter distributions. Adopting a Null-Space Monte Carlo analysis, 400 different sets of hydraulic conductivity fields were randomly generated of which only 336 were selected using an objective function threshold. Subsequently, particle backtracking was performed for each of the accepted hydraulic conductivity fields, by placing particles in a contaminated well. The number of particle passages is considered as being proportional to the contribution of each unknown point source to the tetrachloroethylene contamination identified in the target well. The study provides a methodology to help public authorities to locate the "more probable than not" area responsible for the tetrachloroethylene contamination detected in groundwater and to focus environmental investigations in specific sectors of Milano.

Keywords: particle tracking, Null-Space Monte Carlo, Stochastic MODPATH, groundwater pollution, inverse modeling, uncertainty prediction, PEST

OPEN ACCESS

Edited by:
Teng Xu,
Hohai University, China

Reviewed by:
Ty Ferre,
University of Arizona, United States
Christian Moeck,
Swiss Federal Institute of Aquatic Science and Technology, Switzerland
John Molson,
Laval University, Canada

****Correspondence:***
Loris Colombo
loris.colombo@polimi.it

Specialty section:
This article was submitted to Environmental Informatics and Remote Sensing, a section of the journal Frontiers in Environmental Science

Received: *24 April 2020*
Accepted: *27 July 2020*
Published: *25 September 2020*

Citation:
Colombo L, Alberti L, Mazzon P and Antelmi M (2020) Null-Space Monte Carlo Particle Backtracking to Identify Groundwater Tetrachloroethylene Sources. Front. Environ. Sci. 8:142. doi: 10.3389/fenvs.2020.00142

1. INTRODUCTION

In urban areas impacted by historical industrialization, the main problem of groundwater contamination is related to chlorinated hydrocarbons (CHCs) (Menichetti and Doni, 2017; La Vigna et al., 2019). Such contamination is typically associated with point sources (PS) or discrete zones, formerly linked to petrochemical plants, refineries, automotive, dry-cleaning, or metal degreasing operations. However, due to the intense industrial and urban development during the last 60 years, it is very difficult to apply the "Polluter Pay Principle" (PPP) (De Sadeleer, 2014; Schwartz, 2018; Covucci, 2019; Milon, 2019). Pollution plumes pose challenges in terms of characterization of contaminant migration pathways, identification of the source, and the polluter. Due to the high uncertainty linked to the exact position of the source, it is necessary to identify the PS with a stochastic approach in order to apply the "more probable than not" principle. Jurisprudence has frequently highlighted the fact that public authorities must identify the operator liable for damage to the environment caused by pollution (D.Lgs 152/2006, Environmental Ministry of Italy, 2006). Particularly, article 239 recalls principles set out by the EU including the PPP. From Directive 2004/35/EC (2004), there is the stem that supports the obligation of imposing the costs of preventive and remediation measures on the operator whose actions or omissions have caused the environmental damage. In the Lombardy Region (Italy), one of the most urbanized and industrialized areas in Europe, nearly 3,000 potential brownfields are present in the contaminated site regional database (AGISCO, ARPA Lombardia, 2019). In such a context, there are many obstacles for remediation as suggested by Alderuccio et al. (2019) and Barilari et al. (2020). The PPP is quite different in many countries: for example, in China (Zahar, 2018), the PPP has been introduced into environmental law, but it seems to be difficult to apply because of the ownership of the National Government. In Spain, the PPP is mixed to the concept of "pay as you throw" (PAYT), which is the method that most directly relates user charges to contributions to environmental sustainability (Chamizo-González et al., 2018). A number of developing countries have recently extended this principle to make it an obligation of the state to compensate the victims of environmental harm (Luppi et al., 2012). In order to apply in Italy both administrative and penal proceedings, a detailed historical analysis of the activities conducted at a site, in conjunction with the link between substances, materials used in production cycles and relationship with the owner, creates the basis for environmental forensics. The link, always "causal," requires strong supporting "evidence"; in this sense, the criterion of "more probable than not" can be applied. This is a fundamental step for regional authorities that, without identifying those responsible for contamination, have to take charge of the remediation processes. In order to reconstruct such difficult historical backgrounds in highly urbanized areas (i.e., in the Lombardy Region), new scientific and robust methodologies are required in order to identify the suspected sources of contaminant. Among others, compound-specific isotope analysis (Hunkeler et al., 2008; Shouakar-Stash et al., 2009; Alberti et al., 2017), integral pumping tests (Bauer et al., 2004; Alberti et al., 2011), inverse groundwater modeling (Carrera et al., 2005; Tonkin and Doherty, 2009; Alberti et al., 2018; Moeck et al., 2020), and their combination are the most promising techniques. The modeling methodologies have to take into account uncertainties within an "acceptable level," which depends on the effects and costs of possible contamination and on the complexity of the hydrogeological systems (Frind and Molson, 2018). Uncertainties are given by heterogeneity in the aquifer units, in the regional flow system, in the recharge, and in the conceptual model setup (Rojas et al., 2008). The classical method to identify sources is particle tracking (PT), which involves flow system simulation modeling combined with advective transport of particles along flow lines. For contamination in highly urbanized areas, where a simple termination point (i.e., end-point) analysis is not feasible due to the multiple unknown source release times, the combination of particle tracking with the Null-Space Monte Carlo approach (PT+NSMC) is able to delineate the contribution of each PS under geological uncertainty associating a probability to each of them. It thus enables the public authorities to focus their investigations and to effectively apply the "Polluter Pay Principle".

2. MATERIALS AND METHODS

2.1. Study Area and Hydrogeology

The methodology has been developed within the AMIIGA Project (Interreg Central Europe Grant N. CE32), where inverse transport modeling was one of the tools used to assess the unknown sources in the northwestern part of the Milano Functional Urban Area (FUA). The area of 157 km^2 covers 12 municipalities with high urbanization density (about 4,000 inhabitants per km^2) and a large presence of industrial sites. The area is historically affected by many chlorinated hydrocarbon plumes originating from the northern outer border of the Milano municipality (Giovanardi, 1979; Segre, 1987; Provincia di Milano, 1992); over the last 40 years, the contamination has migrated into Milano city because of the intake area induced by the water supply wells, which provoked a coalescence of several plumes and causing a deterioration of water quality. The evidence is that many plumes that originated in the early 70*s* in the vicinity of Milano are still affecting the aquifers that supply water to the Milano aqueduct system due to the presence of secondary sources and the retardation process of transport. Many chloride compounds were commonly used for degreasing processes by metallurgical, chemical, galvanizing, dry cleaning, and other industrial plants. Today, the main solvents observed in the water supply wells are tetrachloroethylene (PCE), trichloroethylene (TCE), and thichloromethane (TCM) whose concentrations exceed the CCT (Contamination Concentration Threshold) of Italian Law (D. Lgs 152/2006, Environmental Ministry of Italy, 2006, D. Lgs 31/2001, Environmental Ministry of Italy, 2001). The problems owing to the presence of these substances in waters are (1) the high risk of carcinogenicity in humans and (2) the measures that water managers have to adopt in order to control and remove organic chlorinated compounds from drinking water.

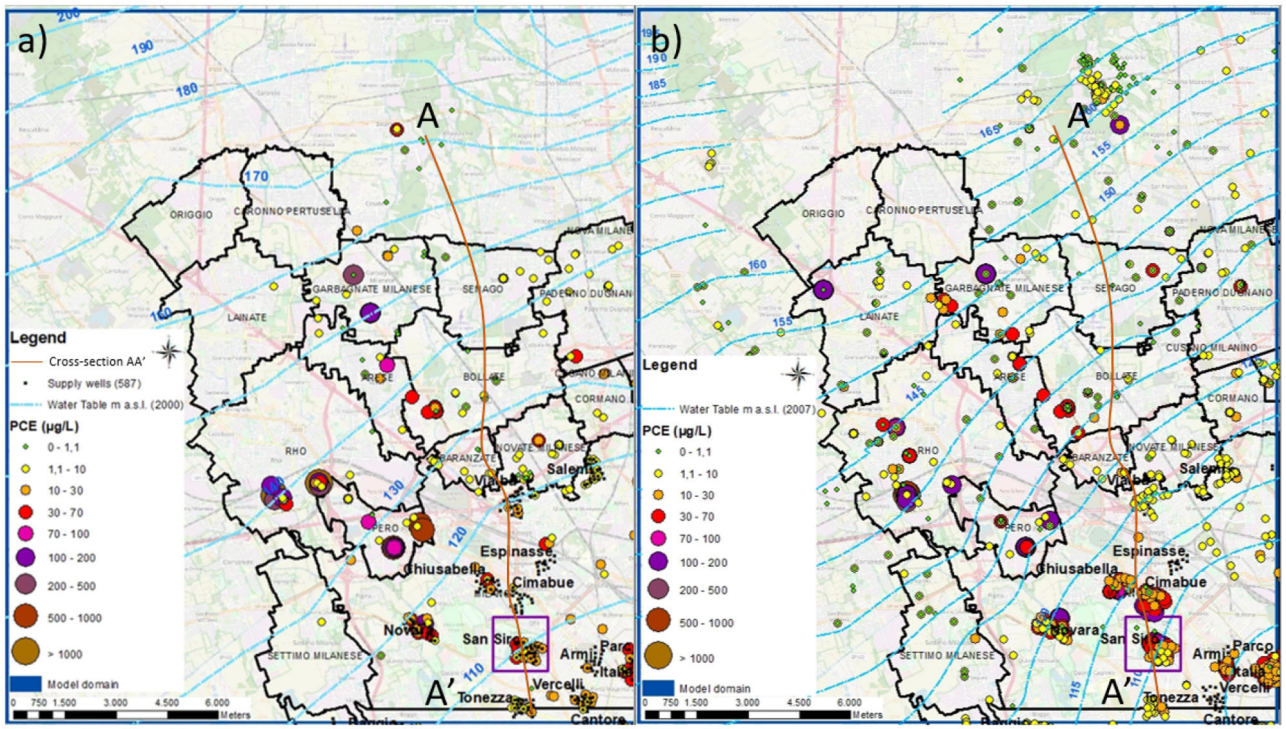


FIGURE 1 | Study area: FUA municipalities with black contours, PCE contamination sampled in **(a)** 2000 and **(b)** 2007 with the associated piezometric map. The cross-section AA′ is described in the **Supplementary Figure 1**.

Figure 1 shows the evolution of PCE contamination from 2000 (a) to 2007 (b) in the water supply wells in Milano (in some wells the detected values are higher than 200 μg/l, as represented by violet and brown circles). Furthermore, the urban texture has significantly changed during the last 15 years [the change in the use of soil (from 2000 to 2015) is provided in the **Supplementary Figure 2**. Some green areas are becoming industrialized and vice versa. The continuous change of use of soil (Masetti et al., 2008) can be one of the causes affecting the groundwater contamination (Pollicino et al., 2019). It is therefore very difficult to reconstruct the suspected areas and consequently to attribute the responsibility of contamination observed in the water supply wells. The Lombardy Region aquifer classification (Regione Lombardia and ENI Divisione Agip, 2002) considers the presence of four different hydrostratigraphic units named A, B, C, and D (a vertical cross-section is shown in the **Supplementary Figure 1**) that originated due to the overlay of plio-pleistocenic alluvial sediments that filled the Neogene Po plain fore-deep, reaching a maximum thickness of approximately 500 m (Bini, 1987). The main aquifers affected by the contamination (A and B) have alluvial and glacio-fluvial origins, and they have a high transmissivity. The unconfined Aquifer A, has a sandy-gravel texture, and the underlying semi-confined aquifer (Aquifer B) has a fine sand composition, interrupted by clay lenses that subdivide it in several small aquifers (i.e., multilayered aquifer). The two aquifers are undifferentiated in the northern part of the study area, whereas they are hydraulically separated by a clay layer of about a meter in thickness in the Milano area. The historical piezometric data collected for the two aquifers show that Aquifer A presents a hydraulic head, higher than in Aquifer B, of about 1 to 1.5 m, and there is therefore a natural downwelling of flow from the shallow aquifer where the clay aquitard shows discontinuities. More details about the conceptual model can be found in previous works (Cavallin et al., 1983; Beretta et al., 1992, 2004; Alberti et al., 2016; Colombo et al., 2019).

2.2. Methodology

The model used to carry out the stochastic simulations in this work is briefly described in this section; a more detailed description is provided by Colombo et al. (2019). Nine model layers with variable thickness (**Figure 2**) are used to represent the two most contaminated geological units (A and B) in the groundwater model (MODFLOW-2005, Harbaugh, 2005), using Groundwater Vistas as the graphical interface (Rumbaugh and Rumbaugh, 2011): Layer 1 represents Aquifer A, layer 2 represents the aquitard, and layers 3–9 represent the so-called multilayered Aquifer B. The model also represents rivers and sinks (i.e., pumping wells), and the horizontal discretization is a 100x100 m grid. Boundary conditions are chosen to be CHs (Constant Heads) and are extrapolated from previous modeling studies (Alberti et al., 2016). The main river (Olona) is implemented as a Cauchy boundary condition. The exchange between superficial water and groundwater is quite low and influences a zone of a few hundred meters around the riverbed. The flow rate depends on the low conductance term defined

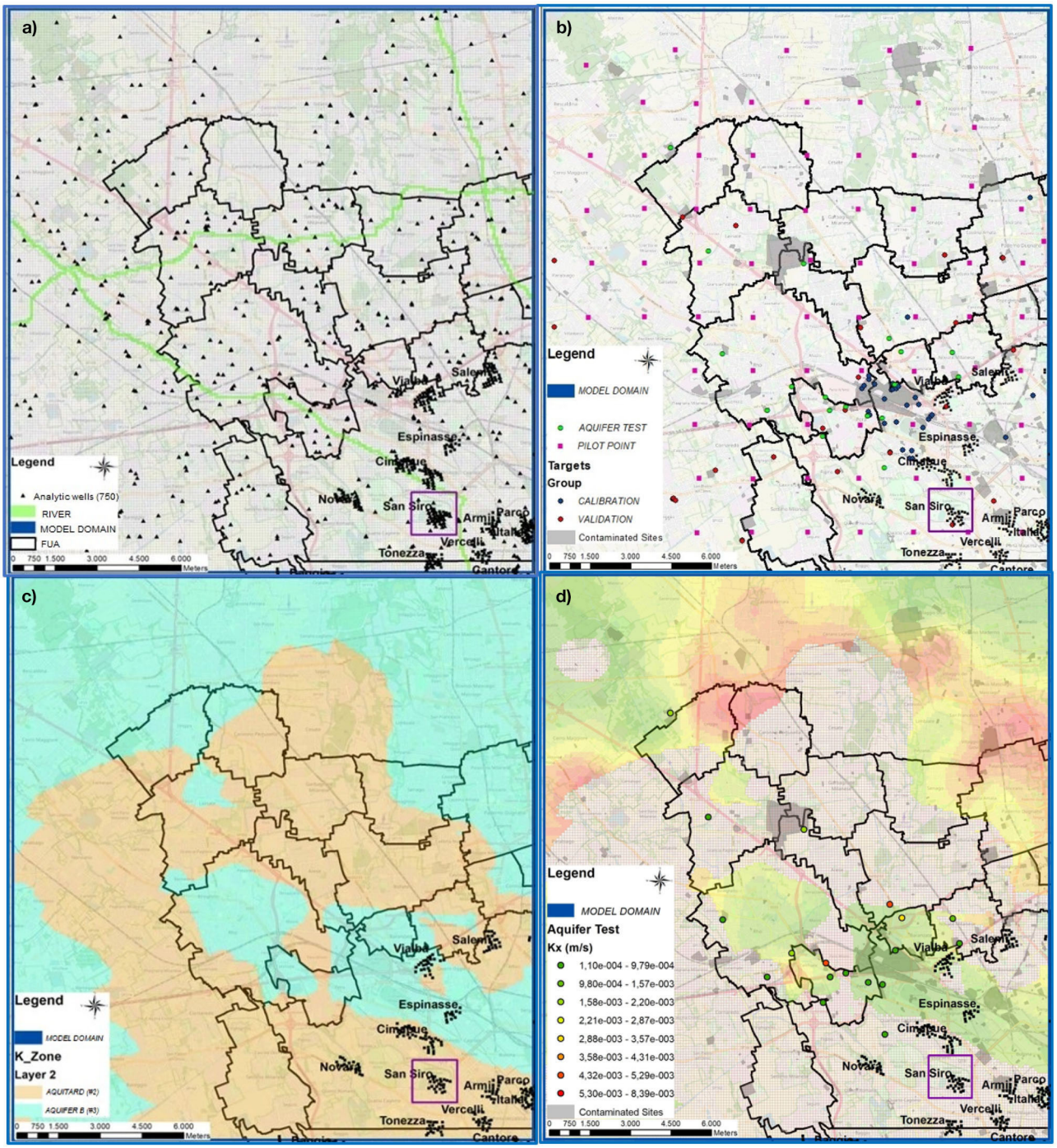


FIGURE 2 | Groundwater model: **(a)** boundary conditions, grid, and internal condition; **(b)** Pilot Point (PP) positions and targets used for calibration (blue) and validation (red); **(c)** example of different zones used in calibration process for Layer 2 and **(d)** calibrated hydraulic conductivity distribution for Layer 2 in the deterministic process. The violet squared area is the San Siro pumping station, where the particles are located for the backtracking analysis.

by hydraulic conductivity and the riverbed thickness (Alberti et al., 2007). An artificial canal used for irrigation purposes is also represented using a Cauchy boundary condition, even if its exchanges with the aquifer are quite low because of its concrete structure. Over the rest of model top, the amount of vertical recharge has been considered uniform in areas with similar land use (divided in urban, agricultural, and green) and estimated through the Thornthwaite method (Thornthwaite, 1948). In the

urbanized areas, the contribution of natural recharge is almost null, but a low groundwater recharge is nevertheless assigned in order to consider water network losses (Alberti et al., 2016). More than 750 pumping wells are included in the model (Colombo et al., 2019) of which 200 wells are drinking water wells affected by CHCs contamination.

The model has been calibrated in a steady state, as described in Colombo et al. (2019), using head values from 63 piezometers using the inverse parameter estimation code PEST (Doherty, 2010). The spatial variations in hydraulic conductivity are estimated by using a pilot point (PP) approach for each of the hydrogeology units (A and B). An anisotropy ratio between the horizontal direction ($K_{xx} = K_{yy}$) and the vertical direction ($K_{zz} = 0.1K_{xx}$) is applied. Four different zones, where the PPs are able to interpolate (using Kriging with an exponential variogram), the K-values have been set (zone 1 for Aquifer A, zone 3 for Aquifer B, zone 2–4 respectively for aquitard and lenses in Aquifer B). As suggested by Doherty (2015) and Moeck et al. (2020), using a high number of PPs, the variogram assumes less importance and only an estimation is required. During the process of inverse calibration, PEST is able to modify the PP values in order to minimize the objective target function (i.e., residual sum of squares expressed in m^2, $PHI = \sum[h_{obs} - h_{sim}]^2$, where h_{obs} is the observed head and h_{sim} is the computed head) in subsequent iterations. Expert knowledge in this framework has been used to assign input values to the PPs representing aquifer tests performed in the study area (i.e., characterization of contamination site and new hydraulic barrier installations). For these PPs, as the hydraulic conductivity values have been estimated by a test (i.e., pumping tests), the boundary range is narrower compared to the others (+/- 0.5 times the magnitude of the value) in order to constrain the PEST estimation in a "real geological" interval. PPs are distributed differently in each zone (83 in zone 1, 46 in zone 2, 117 in zone 3, and 71 in zone 4). The density of the PPs is proportional to the head observation number, as suggested by Christensen and Doherty (2008), and they are uniformly distributed within the domain except for the aquifer test points. In general, the PP density has to take into account the possibility of exploring the heterogeneity, which is needed to have a good fit with observations; on the other hand, the more PPs are considered, the more complex is the inverse parameter estimation with its higher computational time. After this step, a singular value decomposition (SVD), in conjunction with a Tikhonov regularization, is applied in order to have a stable inverse process and to consider prior information (which includes the expert knowledge) into the model to obtain a reasonable geological structure. The subsequent step is to combine PPs with a NSMC analysis (Tonkin and Doherty, 2009) to provide spatial variations in the K-field. A total of 400 different parameter sets are generated through a random log-normal distribution based on the provided upper and lower parameter range (**Table 1**).

The parameter sets are modified in the null-space projections and are adapted from the base model. In other words, PEST decomposes the parameter space into two perpendicular sub-spaces (solution and null-space). The first space is the representation of parameter combinations that can be estimated on the current field used for the calibration. The parts not "explained" in the solution space are spanned into the null-space (i.e., the second space). Each generated random parameter is therefore projected into the calibration null-space and the solution-space component is removed, as it is replaced with the calibrated parameter field from the base model. As the groundwater model is not linear, the stochastic parameter field needs to be re-calibrated. One option is a two-iteration method calibration. The final objective target function is then compared to the desired objective function to keep only the best calibrated K-distributions. Alberti et al. (2018) considered an arbitrary threshold value of the objective function corresponding to the objective function values of the initial deterministic calibrated model, while Moeck et al. (2020) considered both the expected measurement error and structural model error (to avoid overfitting).

TABLE 1 | Hydraulic conductivity in m/s (initial and bounding values) of the parameter used under NSMC (for the geological zones description see **Supplementary Figure 1**).

Geological zone	Initial level	Lower limit	Upper limit
1	10^{-4}	10^{-5}	10^{-2}
2	10^{-6}	10^{-8}	10^{-6}
3	10^{-5}	10^{-5}	5×10^{-3}
4	10^{-6}	10^{-8}	10^{-6}

In this analysis, following Alberti et al. (2018), the threshold has been set to 35 m^2, corresponding to the objective function value of the "native" (primary) calibrated model. From the total set of 400 random models, 64 were discarded and 336 accepted for the further analyses (as presented in **Supplementary Figure 3**). Generally, the results in terms of probability in the study area are not different (in **Supplementary Figure 4** shows the differences, while in **Supplementary Figure 5** is shown the particle frequency distribution in Layer 2). The scheme of the adopted methodology is represented in **Figure 3**.

In the literature, several examples of the NSMC method with different uses have been presented: from calibration purposes (Tonkin and Doherty, 2009; Doherty, 2015) to diffuse pollution assessment (Alberti et al., 2018) and from assessing aquifer pathways (Moeck et al., 2020) to testing barrier effectiveness (Formentin et al., 2019).

2.3. Integration of Particle Tracking and NSMC

To assess the most likely source area of the PCE contamination in the San Siro pumping wells in Milano and to identify the most likely occurring flow path, the particle backtracking technique has been applied by using the code MODPATH v.5 (Pollock, 1994). For all accepted hydraulic conductivity fields, 3D backward particle tracking has been carried out from most contaminated pumping well (well N°500), where the detected concentrations of PCE are historically higher than the threshold limit for drinking water set by the Environmental Ministry of Italy (2001). **Figure 4** provides details about the area of the

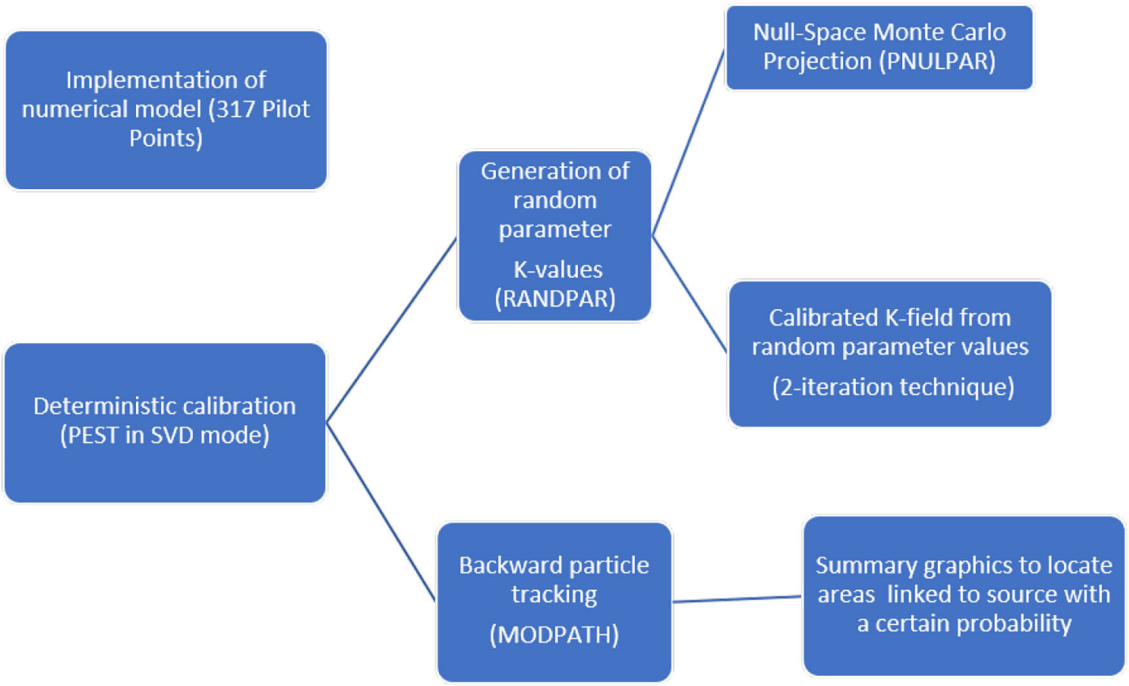


FIGURE 3 | Conceptual flowchart for the PT+NSMC methodology.

San Siro pumping station. In general, the wells of the Milano pumping stations are screened in Aquifer B, which is mostly protected by a clay lenses of variable thickness (between 1 and 3 m) whose extension is shown in **Figure 2d**. **Figure 4** shows the cross-sections obtained from the pumping station well log-stratigraphy, where pumping well N°500 is screened in three different layers. The particles have been computed based on a "screen-position" hypothesis: since wells can screen more than one layer and so the actual depth of the contamination is uncertain, it has been assumed that the number of particles can be distributed in different modeled layers. Then, for well N°500, a total of three particles are positioned basing on the screened levels in permeable layers (Layer 5,7, and 9), as the contaminant is considered diluted along the screened length of the well.

3. RESULTS AND DISCUSSION

Figure 5 shows the values of the objective function before and after the K-field re-calibration (paragraph 2.2). It can be seen that the randomly generated K-fields are generally "not calibrated" as generated by a "classical" Monte Carlo process: they have higher objective functions than the posterior calibration (399 models are not calibrated compared to the objective function of the "native" deterministic model during the first random extraction).

3.1. Spatial Distribution of K-field Realization

During the calibration process, the hydraulic conductivity (K) random distributions are estimated over the available prior knowledge of the geological setting. **Figure 6** shows the variability of the K pattern in some of the randomly generated realizations. The areas where the K-values are higher generally remain the same spatially. For example, in the northern part of the model, the extension of the high permeability zone (in white) changes the shape but remains located in the upper part of the area. This area is affected by limited information (only a few targets are available), and for this reason the uncertainty has a strong influence on the realizations. On the other hand, in the central area of the model (near the San Siro pumping station), the density of information (target measurements) is higher and the realizations are less influenced by the lack of observations. At these locations, the information contained in the head measurements is able to constrain the parameter estimation process.

3.2. Mapping the Backward NSMC Particle Tracking: the Probability Associated With the Source Areas of Contamination

Similar to Alberti et al. (2018) and Moeck et al. (2020), a computation of the number of particles crossing the model cells has been carried out, considering all different accepted simulations. **Figure 7a** represents the particle frequency distribution for all calibrated K-fields and for all model layers. For areas delineated with high pathline density (i.e., higher particle passage frequency), it is possible to delineate the most likely area that can contain one or more PS and it allows to apply the criterion "more probable than not." It is important to remember that, also in those areas where the frequency is low, the possibility to find a source area is still not negligible.

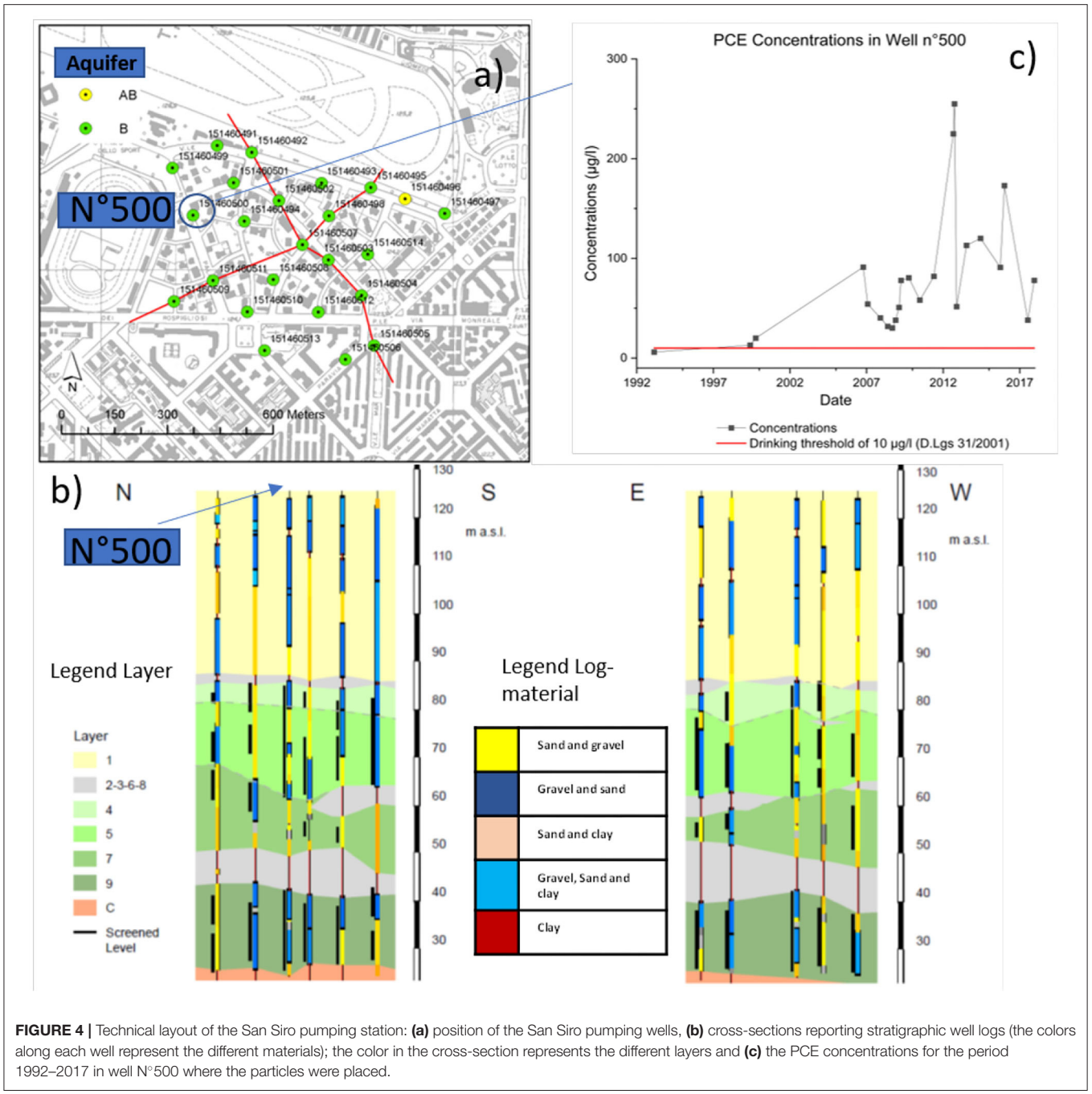


FIGURE 4 | Technical layout of the San Siro pumping station: **(a)** position of the San Siro pumping wells, **(b)** cross-sections reporting stratigraphic well logs (the colors along each well represent the different materials); the color in the cross-section represents the different layers and **(c)** the PCE concentrations for the period 1992–2017 in well N°500 where the particles were placed.

Most of the water pumped from the well is collected from a central narrow line whose frequency is higher than 10%. Superimposing the AGISCO database (from ARPA Lombardia, 2019) with these results (**Figure 7**), it is possible to underline the contaminated sites that are more probable to be responsible for the contamination detected in well N°500. The single pathline is comparable to the path with the highest pathline density. In this case, a single pathline can therefore be used as an initial screening tool.

Figure 7a shows that there are few contaminated sites (Provincia di Milano, 1992, 1994) falling in the area having a higher probability (in yellow), whereas the majority of them fall in green areas having a low probability. Considering a probability higher than 10%, two industrial districts (represented in red and blue) can be considered the most probable source of the San Siro PCE contamination. Nevertheless, even if the backtracking PT assigns the same probability to those districts, the distance from well N°500 has to be considered. Indeed,

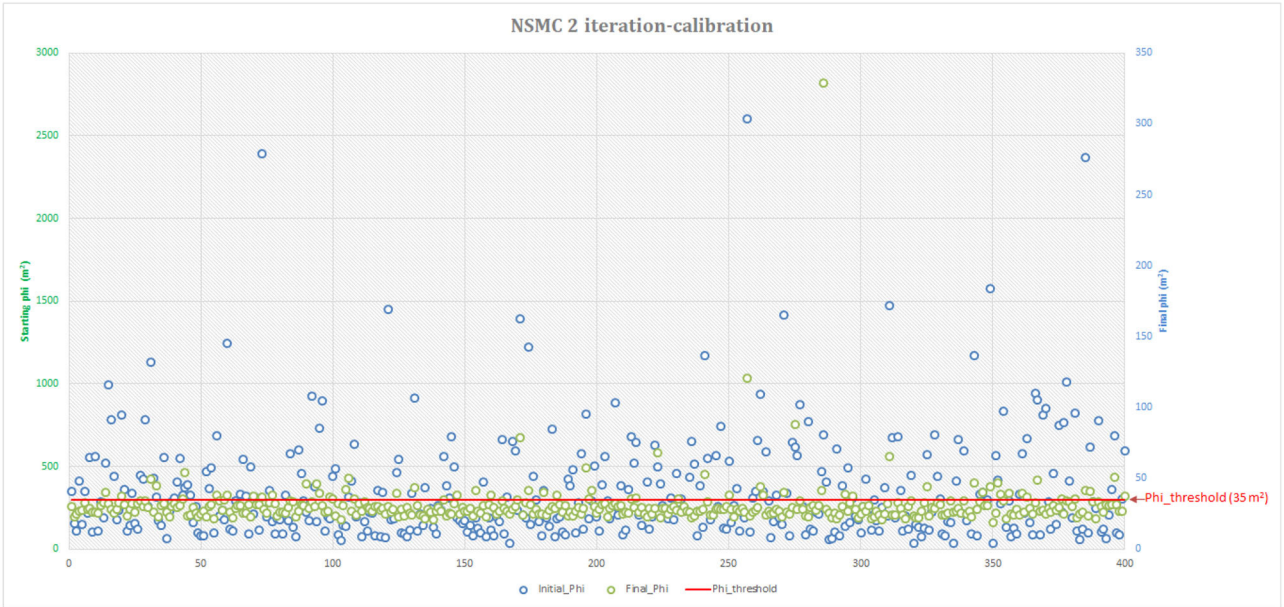


FIGURE 5 | Comparison between the initial objective function (RANDPAR) and the final objective function obtained with the two-iteration process (SVD). The red line represents the threshold objective function limit (35 m^2) used to retain the models with a good misfit to track particles.

chlorinated hydrocarbon plumes can be very long as natural attenuation needs anoxic aquifer conditions. However, the further the contaminated target is from the potential source, the less probable for that source to be responsible (McGuire et al., 2004). For this reason, as the red rectangle is closer to well N°500 in the San Siro pumping wells (approximately 6 km) compared to the blue rectangle (approximately 12 km), it can be stated that the latter has a globally lower probability of being responsible.

In order to better identify the contaminated sites and to clarify the contamination paths, two different maps have been prepared separating the two aquifers: **Figure 7b** shows the probability based only on particles flowing in Aquifer A, whereas **Figure 7c** represents the probability only in Layer 5, which can be considered representative of Aquifer B because most of the particles flow within this hydrogeological unit.

The following analysis of the particle paths in each aquifer was applied in order to narrow the areas directly impacted by the contamination:

- In Aquifer A, the area with higher frequency (more than 15%) is more evident, which was used to restrict the analyses and then to find the potential responsible site.
- In Aquifer B, the probability is significantly less than 10%. This map has the advantage to show that the withdrawals of the Novara pumping station are high enough to affect the flow direction in Aquifer B, acting as a "hydraulic barrier", protecting the San Siro pumping station.

3.3. Analyses of the Cross-Sections and Discussion

In order to understand how the contamination can move and reach the screens of well N°500, it is also useful to analyze, through some cross-sections, the probability distribution of particles in the vertical direction. The represented particle frequencies have been computed summing the number of passages for each cell (the model coordinates are row and column) and dividing it by the total sum in each column (**Figure 8**): for example, in the cell coordinate (155;116), 29% of total particles crossing this specific column (whose total is indicated with $\sum$ in **Figure 8**) pass in Layer 1, whereas 2% pass in Layer 2, representing the shallow aquifer A and the aquitard, respectively. To present the advantages of this analysis, two different cross-sections (**Figure 7a**) have been drawn near the red squared area. To complete the analysis, a third cross-section is presented in the Supplementary Material (**Supplementary Figure 7**).

Longitudinal cross-section B (**Figure 8**): this is oriented along the main flow direction (NW-SE) and stretches for 7 km, starting from the red square to the San Siro pumping station. This representation allows to show in detail the vertical migration of the contamination along the main flow direction. Starting from the screens of well N°500 (cell 189;151) and backtracking the particle pathlines from there, **Figure 8** shows that the particles flow only in Aquifer B (Layer from 5 to 9), but they are suddenly (cell 182;144) able to partially reach the shallower layers. Continuing northward, the number of particles in Layer 9 and 7 progressively decreases (cell 166;127), and, simultaneously, they increase in the more permeable upper Layers 5 and 1; due to the presence of a high number of clay lenses, Layers 2–4 are not significantly interested by the particles. As soon as a clay layer becomes discontinuous or disappears (from cell 154;114), the particle frequencies also increase in those layers, and similar percentages are consequently present in Layers 1 to 5, corresponding to industrial sites 4 and 5. These results

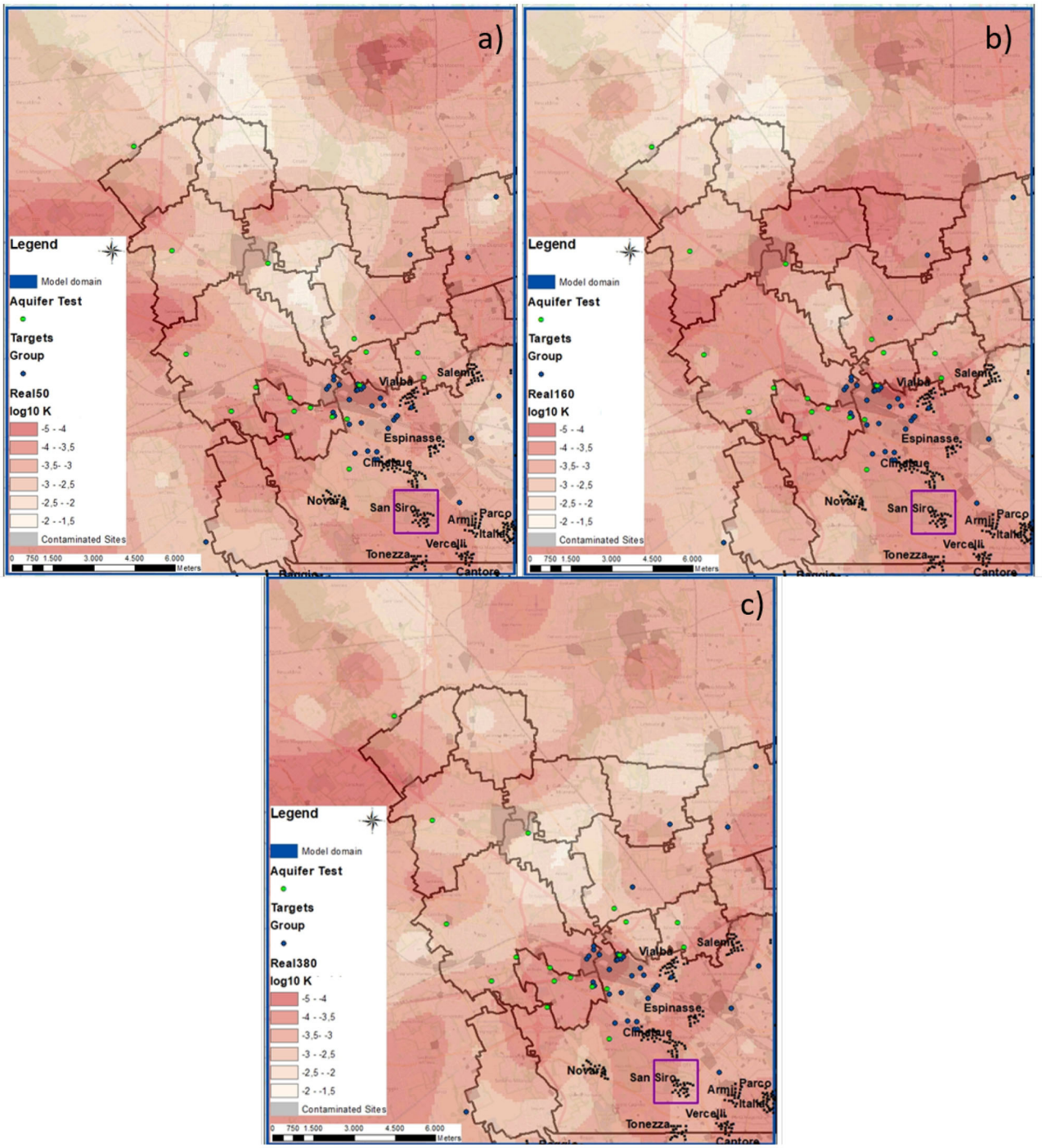


FIGURE 6 | K distribution for Aquifer A for three randomly selected parameter realizations **(a–c)**. The K values are shown as log10 values, where white and red areas show higher and lower K, respectively. Blue dots represent the head targets and green dots represent the aquifer tests.

show that contamination released in Aquifer A, 7 km from San Siro, is able to reach the upper part of the screen at well N°500 (corresponding to Layer 5). In contrast, in order to contaminate the deeper screens (Layer 7 and 9), PCE would have to be mainly released directly in the deeper part of Aquifer B, as few particles are able to cross the clay strata underneath Layer 5.

Transverse cross-section A (**Figure 9**): this is orthogonal to the main flow direction and stretches for 1.5 km. The capital letter B indicates the crossing point with the longitudinal

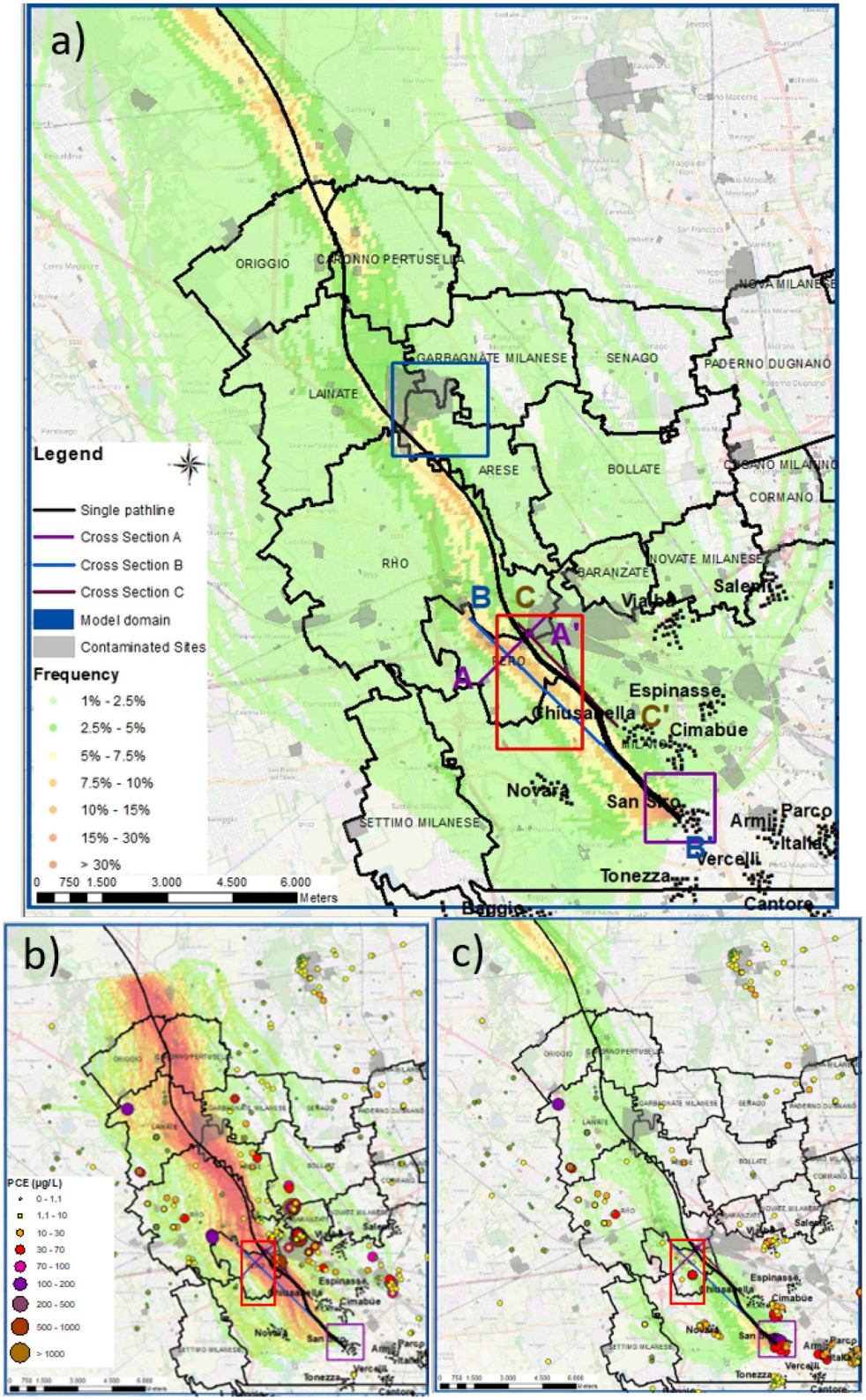
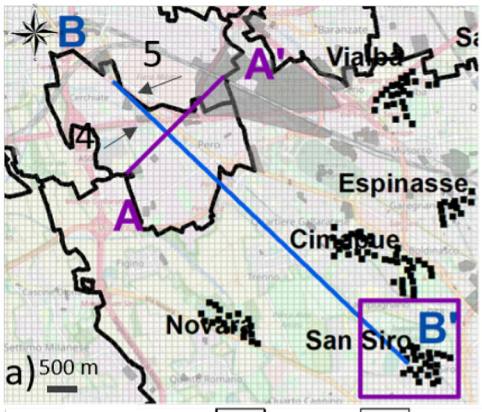


FIGURE 7 | Particle frequency distribution of the ensemble backtracked pathlines for **(a)** all layers, **(b)** Aquifer A, and **(c)** Layer 5, representing the more affected sector of Aquifer B. Particle tracking has been carried out for all 336 accepted and re-calibrated model realizations. To represent the particle passage frequency, a threshold value of 1% was set in all maps. Black lines represent the pathlines resulting from the native deterministic model.



b)

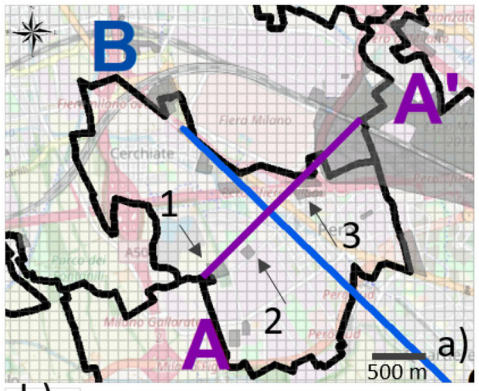
B					5			4			A																		
	Lyr	143;103	144;104	145;105	146;106	147;107	148;108	149;109	150;110	151;111	152;112	153;113	154;114	154;115	155;116	156;117	157;118	158;119	159;120	160;121	161;122	162;123	163;124	164;125	165;126	166;127
		90	106	99	118	135	135	140	106	107	104	105	107	108	108	102	94	86	81	80	70	66	73	74	77	72
	1	24%	20%	16%	24%	21%	22%	20%	20%	24%	29%	28%	29%	30%	29%	35%	37%	35%	36%	39%	40%	38%	37%	41%	35%	33%
	2	10%	7%	3%	6%	9%	6%	8%	0%	0%	0%	0%	0%	0%	2%	0%	1%	1%	1%	0%	0%	0%	0%	0%	1%	0%
	3	10%	12%	7%	6%	11%	8%	12%	7%	4%	2%	2%	2%	2%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	0%
	4	8%	11%	7%	6%	8%	7%	9%	7%	9%	8%	9%	7%	7%	2%	2%	2%	0%	0%	4%	4%	2%	0%	1%	4%	0%
	5	24%	25%	37%	35%	29%	35%	32%	34%	30%	31%	33%	32%	32%	27%	28%	32%	35%	36%	34%	39%	45%	44%	41%	38%	44%
	6	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	1%	0%	1%	0%	1%	1%	0%	0%	0%	0%	0%	0%	0%
	7	20%	21%	25%	20%	19%	18%	15%	24%	23%	22%	22%	24%	24%	28%	27%	26%	27%	25%	23%	16%	14%	18%	14%	17%	18%
	8	3%	4%	4%	3%	4%	4%	4%	7%	7%	7%	5%	4%	4%	6%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	9	0%	0%	0%	0%	0%	1%	1%	3%	3%	2%	2%	2%	2%	6%	6%	2%	1%	1%	1%	1%	2%	1%	4%	4%	4%

																								Well N°500	
167;128	168;129	169;130	170;131	171;132	172;133	173;134	174;135	175;136	176;137	177;138	177;139	178;140	179;141	180;142	181;143	182;144	183;145	184;146	185;147	186;148	187;149	188;150	189;151	B'
71	65	78	78	80	72	74	78	63	73	60	58	56	64	79	74	82	95	118	141	179	285	533	1008	Σ
32%	34%	31%	28%	26%	22%	16%	21%	25%	25%	17%	10%	11%	13%	14%	15%	12%	7%	3%	3%	1%	0%	0%	0%	965
0%	2%	3%	3%	1%	0%	0%	0%	0%	1%	0%	0%	0%	0%	1%	0%	0%	0%	0%	1%	0%	0%	0%	0%	77
0%	2%	3%	9%	9%	7%	4%	6%	3%	4%	7%	2%	2%	3%	3%	1%	4%	4%	4%	1%	2%	1%	0%	0%	163
0%	0%	0%	0%	1%	1%	0%	1%	2%	0%	0%	0%	0%	0%	1%	0%	0%	0%	2%	0%	1%	0%	0%	0%	134
46%	37%	32%	31%	34%	40%	43%	41%	38%	34%	35%	36%	34%	38%	33%	36%	39%	42%	42%	41%	43%	43%	44%	33%	2182
0%	2%	0%	0%	0%	0%	0%	1%	0%	3%	2%	0%	2%	0%	3%	0%	0%	0%	0%	1%	0%	0%	0%	0%	15
15%	20%	23%	22%	23%	26%	24%	15%	16%	16%	20%	28%	30%	30%	28%	27%	23%	26%	27%	26%	26%	27%	27%	33%	1485
0%	0%	1%	0%	0%	0%	1%	1%	0%	1%	0%	2%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	78
6%	5%	8%	8%	6%	3%	11%	13%	16%	15%	20%	22%	21%	17%	18%	20%	22%	20%	21%	27%	27%	29%	29%	33%	908

FIGURE 8 | (a) Cross-section B, oriented NW-SE (BB') with the blue arrow indicating the flow direction: line one shows the cell coordinates, line two shows the sum of particles passing in each vertical column and then the percentage for each layer, and the last column represents the number of particles for each layer; **(b)** zoom of **Figure 7**, showing the area and two different industrial sites (4 and 5). The hydrogeological cross-section is shown in the **Supplementary Figure 6B**.

cross-section. **Figure 9** shows that at the red squared area, a large part of particles traverse Layer 1 and 5 where the particle sum is 239 and 212 respectively. Nevertheless, some differences can be noticed moving to the right and left of the B crossing point. On the right side, moving toward the industrial site 2, frequencies are quite similar from Layer 1 to 5 as displayed by the cross-section B in the cell (149;109). On the left side, toward the industrial site 3, frequencies are decreasing in Layer 2–4, indicating that here clay lenses are able to better separate Aquifer A from Aquifer B. Summarizing the results, it is possible to affirm that a PCE release in the shallow aquifer at site 2, 4, and 5, compared to sites 1 and 3, has a higher probability to be responsible for the contamination detected in well N°500 (mainly for screens positioned in Layer 5, as shown in **Figure 8**). On the other hand, for the contamination detected in screens located in Layers 7 and 9, a direct release (i.e., dry well) in the deeper part of Aquifer B should be hypothesized. Results show how the stochastic approach, combined with backtracking PT, is able to display the probability of the advective pathways followed by the contaminant.

Stochastic particle tracking (PT+NSMC) is an innovative tool that can be helpful in the search for contamination source areas and plume monitoring. Its simplicity and versatility allows several advantages such as the ability to perform multiple analyses based on the same results, indeed two possible outcomes have been presented: (1) plan map (x,y) of the particle passage frequency in each cell of the model domain, showing all the possible pathlines that the contaminant can follow starting from a specific location, reached by the contamination, considering the uncertainty related to the hydraulic conductivity. The main advantage of the maps is that they allow us to consider a global frequency (i.e., counts all the particles belonging to a specific layer or group of layers) in order to locate the more likely area to be investigated by the installation of new piezometers or the characterization of brownfield areas. On the contrary, these maps are not able to link the vertical passage of the contamination (i.e., it would be necessary to produce as many maps as the number of layers); (2) vertical cross-sections representing the distribution of the particles cell by cell as a function of the layer (z), able to represent the impact of the contamination in each model layer and to give a 3D



b)

A		1			2			B			3								A'
	Lyr	158;105	157;106	156;107	155;108	154;109	153;110	152;111	151;112	150;113	149;114	148;115	147;116	146;117	145;118	144;119	143;120	142;121	
		8	6	12	27	52	98	101	103	102	71	39	54	29	34	13	10	8	∑
	1	25%	0%	0%	11%	29%	18%	28%	25%	31%	39%	31%	50%	55%	56%	46%	50%	25%	239
	2	0%	0%	0%	0%	12%	10%	0%	0%	0%	0%	0%	6%	17%	12%	8%	10%	13%	31
	3	0%	0%	0%	4%	8%	10%	3%	5%	3%	0%	0%	0%	0%	0%	0%	0%	0%	26
	4	0%	0%	0%	7%	2%	7%	8%	7%	9%	6%	0%	0%	0%	0%	0%	0%	0%	38
	5	25%	33%	50%	33%	19%	20%	35%	36%	29%	30%	33%	26%	21%	6%	23%	10%	13%	212
	6	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	1
	7	25%	17%	42%	30%	17%	23%	17%	19%	20%	18%	26%	17%	3%	18%	15%	10%	13%	148
	8	0%	17%	0%	11%	8%	6%	6%	7%	8%	4%	5%	0%	0%	6%	8%	10%	25%	46
	9	25%	33%	8%	4%	6%	4%	4%	1%	0%	1%	5%	2%	3%	3%	0%	10%	13%	26

FIGURE 9 | (a) Cross-section A, oriented SW-NE (AA'): line one shows the cell coordinate, line two the sum of particles passing in each vertical column and then the frequency for each layer, and the last column the count of particles for each layer; **(b)** zoom of **Figure 7**, showing the area and three different industrial sites (1, 2, and 3). The hydrogeological cross-section is shown in the **Supplementary Figure 6A**.

vision if orthogonal cross-sections are used. The advantage of these analyses is the calculation of the number of particles per vertical column and therefore the frequency for each layer (i.e., accounting for the particles passing through a specific cell) in order to show the contamination impact relative to a single cell. However, the vertical cross-section is representative of a specific area and as many different cross-sections as the number of different contaminated sites should be made (i.e., a reconstruction in a greater detail of the contamination paths). The two analyses are complimentary and can be used together to identify both potential contamination sources (i.e., higher frequency in the plan map) in order to find appropriate areas to install monitoring wells and suitable positions for the well screens (i.e., in the more contaminated layer computed in the vertical cross-section). However, this methodology shows some limitations that can be further improved. For example, a preliminary linear uncertainty analysis (Moore and Doherty, 2005; James et al., 2009; Dausman et al., 2010; Moeck et al., 2015, 2020) can be carried out in order to identify the most important parameters linked to the uncertainty and likely to be used in the NSMC procedure: for example, porosity, well abstractions, the vertical recharge, and the boundary conditions can be considered in the analyses. In this paper, only the hydraulic conductivity has been considered as uncertain because concerning the porosity, previous analyses showed that the model is not sensitive to this parameter (Colombo et al., 2019); well abstraction data have been provided by water managers so are considered to be only slightly uncertain. To deal with BC's, the choice has been to use a large model domain in order to set them far away from the area of interest and therefore would have a low influence. Since the results obtained by the deterministic pathline and the higher pathline density area obtained with NSMC are coherent (**Figure 7**), it is possible that the flow path is mainly controlled by the BCs and therefore the subsurface heterogeneity might be less important. For this reason, the influence of the BCs it will be further investigated in the future. In addition, as reported by Chow et al. (2016), particle tracking provides an approximation of groundwater flow paths and results can differ between simulators because of velocity approximations which are used for computation of the particle tracks. In the model presented here, a regular grid has been used due to the domain dimensions, and the choice was to apply the option "pass through weak sinks" in order to not terminate any particle and as a factor of safety from an environmental point of view. Thanks to the unstructured grid of the MODFLOW-USG version (Panday et al., 2013), it also would be possible to reduce the particle tracking numerical error using a refined grid where wells fall in the area of interest. Another possible improvement of the proposed methodology could consider the combination of stochastic results with contaminant concentrations detected in the monitoring network. Overlapping this information would exclude some areas removing the pathlines passing through clean wells, following the methodology presented in Formentin et al. (2019). Finally, in dense urban contexts, the results are not

able to guarantee a unique liable area for the contamination (i.e., the presence of many contaminated sites along the main water flow direction as in cross-section B). Therefore, it can be coupled with other analyses to further reduce the list of potential responsible sites. For example, some statistics about the general behavior of chlorinated solvent plumes can be considered, like those presented in McGuire et al. (2004). Considering a sample of 45 chlorinated solvent sites (managed through monitored natural attenuation), 96% of the plumes were shorter than approximately 1.5 km and over a sample of 25 sites 68% of the remediation time-frame was less than 30 years. According to these statistics, a more distant potential source (e.g., the blue squared area in **Figure 7**) is less likely to be responsible and therefore the spatial range likely to contain the contamination source can be limited, weighting the probability on the target distance or the contaminant travel time. These ideas could be a future improvement of the proposed methodology.

4. CONCLUSIONS

The northwestern area in the Milano Functional Urban Area is affected by a strong presence of chlorinated hydrocarbons, originating from several contaminated sites historically present in the territory. Over the last years, this contamination has posed important safety problems in the intake area of the water supply wells of Milano. For this reason, the remediation of groundwater for drinking water purposes has become a major problem for water managers; on one hand, the water treatment procedures became more onerous both in terms of management and new plants (like new active carbon filter installations), but, on the other hand, the cost of the raw water clean-up has to be imposed on the citizens within the water bill costs. In this situation, the Polluter Pay Principle becomes crucial: its application to groundwater contamination due to the "more probable than not" responsible source is able to help charge the cost of site remediation and (mainly) the treatment costs of the water extracted for public use. The support of a groundwater model is very useful to study the contamination pathways, but, especially in the presence of multiple possible sources, the deterministic simulation fails as the uncertainties strongly influence the results both in terms of groundwater flow and in terms of contaminant transport. In this context, new methodologies and models have been developed (Alberti et al., 2018; Moeck et al., 2020). One innovative methodology is to combine particle backtracking with a constrained Monte Carlo approach (i.e., considering the uncertainties linked to the hydraulic conductivity). This combination has been developed throughout this paper to identify PCE point sources in urban areas. Starting from one of the pumping wells most contaminated by PCE in the San Siro pumping station, a particle backtracking stochastic analysis has been developed. Considering all the accepted K-field realizations (336), all the contaminated sites within the non-zero probability areas must be considered as potential sources for this pollution. However, in such a densely industrialized zone, this result is not able to narrow the area where the actual source of the contamination can be identified. Then, a further analysis has been developed in order to: (1) narrow the suspected areas considering the higher particle density, recognizing as possible sources only five sites among several others present upstream of the detected contamination; (2) study the particle paths in greater detail dividing the map results into two different hydrogeological units (Aquifer A and Aquifer B) using a high vertical discretization of the model (nine layers); and (3) provide vertical cross-sections of the particle passage frequency for selected suspected areas. In conclusion, particle tracking with Monte Carlo can be used as a method to identify the "more probable than not" areas potentially responsible for contamination detected in sensitive targets. For this reason, the method can be very helpful to the public authorities as a decision support tool to prepare a ranking list of possible potential responsible sites for contamination and to design a suitable well monitoring network. In densely industrialized districts, the proposed procedure is not able to identify just one source site with certainty and it needs to be coupled together with other analyses to provide strong evidence of liability. This method provides a new basis for future modelers, where the geological uncertainty plays a fundamental role for a reasonable and responsible assessment of the polluters.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: Data about contaminated sites are confidential due to ongoing legal procedures. Requests to access these datasets should be directed to Regione Lombardia through the online form ambiente_clima@pec.regione.lombardia.it.

AUTHOR CONTRIBUTIONS

LC, PM, and LA developed the model and carried out the analyses. LC and PM contributed to the development of Excel Macros and scripts that enabled the automated analysis of counting particles herein. LC, PM, and LA also contributed to the higher-level workflow. LA provided expert knowledge for defining the conceptual model. LC, PM, and MA wrote the initial draft. However, all authors contributed to the writing of the manuscript.

FUNDING

The authors would like to acknowledge funding from the Interreg Central Europe Project (CE No. 32 AMIIGA - Integrated Approach to Management of Groundwater Quality in Functional Urban Areas).

ACKNOWLEDGMENTS

The authors would like to acknowledge Regione Lombardia for the data availability on the contaminated sites (information and shapefile-AGISCO DB) and contaminant concentration data (i.e., PCE concentration). We would also like to acknowledge Eng. Giovanni Formentin and John Doherty for supporting the

methodology and scripts. We thank the Milano public water manager (Metropolitana Milanese S.p.A.) for the information and data about the Milano aqueduct wells. Finally, we would like to acknowledge the reviewers that help us to improve the text and the quality of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fenvs.2020.00142/full#supplementary-material

REFERENCES

Alberti, L., Azzellino, A., Colombo, L., and Lombi, S. (2016). Use of cluster analysis to identify tetrachloroethylene pollution hotspots for the transport numerical model implementation in urban functional area of milan, italy. *Int. Multidiscipl. Sci. GeoConfer. Surv. Geol. Min. Ecol. Manage.* 1, 723–729. doi: 10.5593/SGEM2016/B11/S02.091

Alberti, L., Brogioli, G., Formentin, G., Marangoni, T., and Masetti, M. (2007). "Experimental studies and numerical modeling of surface water-groundwater interaction in a semi-disconnected system," in *IAH Congress, Groundwater and Ecosystems* (Lisbon).

Alberti, L., Colombo, L., and Formentin, G. (2018). Null-space monte carlo particle tracking to assess groundwater pce (tetrachloroethene) diffuse pollution in North-Eastern Milan functional urban area. *Sci. Total Environ.* 621, 326–339. doi: 10.1016/j.scitotenv.2017.11.253

Alberti, L., Lombi, S., and Zanini, A. (2011). Identification de sources d'hydrocarbures aliphatiques chlorés dans une zone résidentielle en italie par la méthode d'essai de pompage intégral. *Hydrogeol. J.* 1:76. doi: 10.1007/s10040-011-0742-1

Alberti, L., Marchesi, M., Trefiletti, P., and Aravena, R. (2017). Compound-specific isotope analysis (csia) application for source apportionment and natural attenuation assessment of chlorinated benzenes. *Water* 9:872. doi: 10.3390/w9110872

Alderuccio, M., Cantarella, L., Ostoich, M., Ciuffi, P., Gattolin, M., Tomiato, L., et al. (2019). Identification of the operator responsible for the remediation of a contaminated site. *Environ. For.* 20, 339–358. doi: 10.1080/15275922.2019.1657521

ARPA Lombardia (2019). *Anagrafe e gestione integrata dei siti contaminati* (agisco).

Barilari, A., Quiroz Londoño, M., Paris, M., Lima, M., and Massone, H. (2020). Groundwater contamination from point sources. a hazard index to protect water supply wells in intermediate cities. *Groundwater Sustain. Dev.* 10:100363. doi: 10.1016/j.gsd.2020.100363

Bauer, S., Bayer-Raich, M., Holder, T., Kolesar, C., Müller, D., and Ptak, T. (2004). Quantification of groundwater contamination in an urban area using integral pumping tests. *J. Contamin. Hydrol.* 75, 183–213. doi: 10.1016/j.jconhyd.2004.06.002

Beretta, G. P., Avanzini, M., and Pagotto, A. (2004). Managing groundwater rise: experimental results and modelling of water pumping from a quarry lake in milan urban area (italy). *Environ. Geol.* 45, 600–608. doi: 10.1007/s00254-003-0918-7

Beretta, G. P., Civita, M., and Francani, V. (1992). *Idrogeologia per il disinquinamento delle acque sotterranee: tecniche per lo studio e la progettazione degli interventi di prevenzione, controllo, bonifica e recupero*. Milano: Pitagora Editrice Bologna.

Bini, A. (1987). *L'apparato glaciale wurmiano di Como* (Ph.D. thesis) Tesi di Dottorato di Ricerca, Università di Milano, Milan, Italy.

Carrera, J., Alcolea, A., Medina, A., Hidalgo, J., and Slooten, L. J. (2005). Inverse problem in hydrogeology. *Hydrogeol. J.* 13, 206–222. doi: 10.1007/s10040-004-0404-7

Cavallin, A., Francani, V., and Mazzarella, S. (1983). *Studio idrogeologico della pianura compresa fra adda e ticino*. Milano: Costruzioni.

Chamizo-González, J., Cano-Montero, E. I., and Muñoz-Colomina, C. I. (2018). Does funding of waste services follow the polluter pays principle? The case of Spain. *J. Clean. Prod.* 183, 1054–1063. doi: 10.1016/j.jclepro.2018.02.225

Chow, R., Frind, M. E., Frind, E. O., Jones, J. P., Sousa, M. R., Rudolph, D. L., et al. (2016). Delineating baseflow contribution areas for streams–a model and methods comparison. *J. Contamin. Hydrol.* 195, 11–22. doi: 10.1016/j.jconhyd.2016.11.001

Christensen, S., and Doherty, J. (2008). Predictive error dependencies when using pilot points and singular value decomposition in groundwater model calibration. *Adv. Water Resour.* 31, 674–700. doi: 10.1016/j.advwatres.2008.01.003

Colombo, L., Alberti, L., Mazzon, P., and Formentin, G. (2019). Transient flow and transport modelling of an historical chc source in north-west milano. *Water* 11:1745. doi: 10.3390/w11091745

Covucci, D. (2019). The "polluter pays principle": environmental liabilities and scientific evidence under the italian law system. *Acque Sotterranee Ital. J. Groundwater* 8, 69–72. doi: 10.7343/as-2019-427

Dausman, A. M., Doherty, J., Langevin, C. D., and Sukop, M. C. (2010). Quantifying data worth toward reducing predictive uncertainty. *Groundwater* 48, 729–740. doi: 10.1111/j.1745-6584.2010.00679.x

De Sadeleer, N. (2014). *Polluter Pays Principle*. Oxford: OUP.

Directive 2004/35/EC (2004). *Member States Have Until the End of April 2007 to Transpose This Directive Into Domestic Law*.

Doherty, J. (2010). Methodologies and software for pest-based model predictive uncertainty analysis. *Watermark Numer. Comput.* 1–33.

Doherty, J. (2015). Calibration and uncertainty analysis for complex environmental models. *Watermark Numer. Comput.* 1–224.

Environmental Ministry of Italy (2001). *Legislative Decree no. 31, Approving the Directive 98/83/ce Based on the Drinkable Water*. Gazzetta Ufficiale.

Environmental Ministry of Italy (2006). *Legislative Decree no. 152 Approving the Code on the Environment*.

Formentin, G., Terrenghi, J., Vitiello, M., and Francioli, A. (2019). Evaluation of the performance of a hydraulic barrier by the null space monte carlo method. *Acque Sotterranee Ital. J. Groundwater*. 8, 55–61. doi: 10.7343/as-2019-420

Frind, E. O., and Molson, J. W. (2018). Issues and options in the delineation of well capture zones under uncertainty. *Groundwater* 56, 366–376. doi: 10.1111/gwat.12644

Giovanardi, A. (1979). Pollution by organic chloride compounds of the aquifers of the region of milan [contaminazione da composti organoclorurati delle falde idriche del territorio milanese]. *Riv. Ital. d'Igiene* 39, 323–344.

Harbaugh, A. W. (2005). *MODFLOW-2005, the US Geological Survey Modular Ground-Water Model: The Ground-Water Flow Process*. Reston, VA: US Department of the Interior; US Geological Survey.

Hunkeler, D., Meckenstock, R. U., Lollar, B. S., Schmidt, T. C., and Wilson, J. T. (2008). *A Guide for Assessing Biodegradation and Source Identification of Organic Ground Water Contaminants Using Compound Specific Isotope Analysis (CSIA)*. Ada, OK: USEPA Publication.

James, S. C., Doherty, J. E., and Eddebbarh, A.-A. (2009). Practical postcalibration uncertainty analysis: Yucca mountain, nevada. *Groundwater* 47, 851–869. doi: 10.1111/j.1745-6584.2009.00626.x

La Vigna, F., Sbarbati, C., Bonfà, I., Martelli, S., Ticconi, L., Aleotti, L., et al. (2019). First survey on the occurrence of chlorinated solvents in groundwater of Eastern sector of rome. *Rend Lincei Sci. Fisiche Nat.* 30, 297–306. doi: 10.1007/s12210-019-00790-z

Luppi, B., Parisi, F., and Rajagopalan, S. (2012). The rise and fall of the polluter-pays principle in developing countries. *Int. Rev. Law Econ.* 32, 135–144. doi: 10.1016/j.irle.2011.10.002

Masetti, M., Poli, S., Sterlacchini, S., Beretta, G. P., and Facchi, A. (2008). Spatial and statistical assessment of factors influencing nitrate contamination in groundwater. *J. Environ. Manage.* 86, 272–281.

McGuire, T. M., Newell, C. J., Looney, B. B., Vangelas, K. M., and Sink, C. H. (2004). Historical analysis of monitored natural attenuation: a survey of 191 chlorinated solvent sites and 45 solvent plumes. *Remediat. J.* 15, 99–112. doi: 10.1002/rem.20036

Menichetti, S., and Doni, A. (2017). Organohalogen diffuse contamination in firenze and prato groundwater bodies. investigative monitoring and definition of background values. *Acque Sotterranee Ital. J. Groundwater* 6, 47–63. doi: 10.7343/as-2017-260

Milon, J. W. (2019). The polluter pays principle and everglades restoration. *J. Environ. Stud. Sci.* 9, 67–81. doi: 10.1007%2Fs13412-018-0529-y

Moeck, C., Hunkeler, D., and Brunner, P. (2015). Tutorials as a flexible alternative to guis: an example for advanced model calibration using pilot points. *Environ. Model. Softw.* 66, 78–86. doi: 10.1016/j.envsoft.2014.12.018

Moeck, C., Molson, J., and Schirmer, M. (2020). Pathline density distributions in a null-space monte carlo approach to assess groundwater pathways. *Groundwater* 58, 189–207. doi: 10.1111/gwat.12900

Moore, C., and Doherty, J. (2005). Role of the calibration process in reducing model predictive error. *Water Resour. Res.* 41:W05020. doi: 10.1029/2004WR003501

Panday, S., Langevin, C. D., Niswonger, R. G., Ibaraki, M., and Hughes, J. D. (2013). *Modflow–Usg Version 1: An Unstructured Grid Version of Modflow for Simulating Groundwater Flow and Tightly Coupled Processes Using a Control Volume Finite-Difference Formulation*. Technical report, US Geological Survey.

Pollicino, L. C., Masetti, M., Stevenazzi, S., Colombo, L., and Alberti, L. (2019). Spatial statistical assessment of groundwater pce (tetrachloroethylene) diffuse contamination in urban areas. *Water* 11:1211. doi: 10.3390/w11061211

Pollock, D. W. (1994). *User's Guide for MODPATH/MODPATH-PLOT, Version 3: A Particle Tracking Post-processing Package for MODFLOW, the US: Geological Survey Finite-difference Ground-water Flow Model*.

Provincia di Milano (1992). *Indagini sulla presenza di composti organo-alogenati nelle acque di falda della Provincia di Milano*. Technical report, Sistema Informativo Falda, Milano, Italy.

Provincia di Milano (1994). *Acqua sotto i piedi*. Technical report, Sistema Informativo Falda, Milano, Italy.

Regione Lombardia and ENI Divisione Agip (2002). *Geologia degli acquiferi padani della regione lombardia*.

Rojas, R., Feyen, L., and Dassargues, A. (2008). Conceptual model uncertainty in groundwater modeling: combining generalized likelihood uncertainty estimation and bayesian model averaging. *Water Resour. Res.* 44:W12418. doi: 10.1029/2008WR006908

Rumbaugh, J., and Rumbaugh, D. (2011). *Guide to Using Groundwater Vistas, Version 6*. New York, NY: Environmental Simulations.

Schwartz, P. (2018). "The polluter-pays principle," in *Elgar Encyclopedia of Environmental Law* (Cheltenham, UK: Edward Elgar Publishing Limited), 260–271.

Segre, M. (1987). *Organic Chlorine Compounds Pollution of Milan Groundwater: Two Cases*. Acqua Aria, 1.

Shouakar-Stash, O., Frape, S., Aravena, R., Gargini, A., Pasini, M., and Drimmie, R. (2009). Analysis of compound-specific chlorine stable isotopes of vinyl chloride by continuous flow–isotope ratio mass spectrometry (fc–irms). *Environ. For.* 10, 299–306. doi: 10.1080/15275920903347628

Thornthwaite, C. W. (1948). An approach toward a rational classification of climate. *Geogr. Rev.* 38, 55–94. doi: 10.2307/210739

Tonkin, M. and Doherty, J. (2009). Calibration-constrained monte carlo analysis of highly parameterized models using subspace techniques. *Water Resour. Res.* 45:W00B10. doi: 10.1029/2007WR006678

Zahar, A. (2018). Implementation of the polluter pays principle in China. *Rev. Eur. Compar. Int. Environ. Law*. 27, 293–305. doi: 10.1111/reel.12242

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.

frontiers in Earth Science

METHODS
published: 04 November 2020
doi: 10.3389/feart.2020.565707

Integration of Soft Data Into Geostatistical Simulation of Categorical Variables

Steven F. Carle[1]* *and Graham E. Fogg*[2]

[1]Lawrence Livermore National Laboratory, Livermore, CA, United States, [2]University of California, Davis, Davis, CA, United States

Uncertain or indirect "soft" data, such as geologic interpretation, driller's logs, geophysical logs or imaging, offer potential constraints or "soft conditioning" to stochastic models of discrete categorical subsurface variables in hydrogeology such as hydrofacies. Previous bivariate geostatistical simulation algorithms have not fully addressed the impact of data uncertainty in formulation of the (co) kriging equations and the objective function in simulated annealing (or quenching). This paper introduces the geostatistical simulation code **tsim-s**, which accounts for categorical data uncertainty through a data "hardness" parameter. In generating geostatistical realizations with **tsim-s**, the uncertainty inherent to soft conditioning is factored into both 1) the data declustering and spatial correlation functions in cokriging and 2) the acceptance probability for change of category in simulated quenching. The degree or sensitivity to which soft data conditions a realization as a function of hardness can be quantified by mapping category probabilities derived from multiple realizations. In addition to point or borehole data, arrays of data (e.g., as derived from a depth-dependency function, probability map, or "prior realization") can be used as soft conditioning. The **tsim-s** algorithm provides a theoretically sound and general framework for integrating datasets of variable location, resolution, and uncertainty into geostatistical simulation of categorical variables. A practical example shows how **tsim-s** is capable of generating a large-scale three-dimensional simulation including curvilinear features.

Keywords: geostatistics, hydrogeology, conditioning, cokriging, simulation, soft, prior, inversion

OPEN ACCESS

Edited by:
Liangping Li,
South Dakota School of Mines and Technology, United States

Reviewed by:
Hélio Correia Da Silva Jhunior,
Universidade de São Paulo São Carlos, Brazil
Yunwei Tang,
Chinese Academy of Sciences, China

***Correspondence:**
Steven F. Carle
carle1@llnl.gov

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

Received: *26 May 2020*
Accepted: *14 September 2020*
Published: *04 November 2020*

Citation:
Carle SF and Fogg GE (2020) Integration of Soft Data Into Geostatistical Simulation of Categorical Variables. Front. Earth Sci. 8:565707. doi: 10.3389/feart.2020.565707

1 INTRODUCTION

In many hydrogeological modeling applications, much of the available characterization data for categorical variables, such as lithology, texture, or hydrofacies, are uncertain or less than 100% accurate. In some geostatistical applications, the categorical data such as soil texture are treated as 100% accurate or "hard" data despite the fact these data are uncertain for various reasons (Carle, 1996; Burow et al., 1997; Carle et al., 1998; Weissmann et al., 1999). In other applications, uncertain or indirect "soft" data such as geophysical imaging are available but found difficult to apply as "soft conditioning" to geostatistical simulations of categorical variables (Falivene et al., 2007; Koch et al., 2014). Categorization of hydrogeological variables often has uncertainty attributable to sample quality, geologic interpretation, or indirectness of measurement (e.g., geophysical logs, cone penetrometer data). For example, so-called "driller's logs" or lithologic descriptions by well drillers based on interpretations of drilling conditions, cuttings, and limited core samples are, understandably, uncertain (Oatfield and Czarnecki, 1989; Smith, 2002; Dumedah and Schuurman,

2008; Arihood, 2009; Tsai and Elshall, 2013). Yet driller's logs may provide the most detailed characterization information available for many local site to basin-scale hydrogeological modeling applications (Ezzedine et al., 1999; Weissmann et al., 1999; Weissmann and Fogg, 1999; Carle et al., 2006; Fleckenstein et al., 2006; Elshall et al., 2013). Alternatively, hydrogeological categories can be inferred from geophysical measurements or imaging, but the resulting inferences are inherently uncertain. Paradis et al. (2015) found that even after application of sophisticated machine learning techniques to cone penetrometer test and soil moisture and resistivity probe data, classification errors persist in hydrofacies identification. The resolution of geophysical images varies as a result of petrophysical relationships and, therefore, provides indirect contraints on lithology, texture, or hydrofacies, as further discussed in application of three-dimensional resistivity maps derived from airborne electromagnetic surveys to characterization of spatial occurrences of sand-clay textures (Koch et al., 2014; Hoyer et al., 2015) and electrical resistance tomographs to determine spatial distributions of alluvial or fluvial hydrofacies (Carle et al., 1999; Carle and Ramirez, 1999; Hermans and Irving, 2017). It is a well-accepted fact in the Earth sciences that uncertainty is common to subsurface data.

A general framework for integrating soft data into categorical geostatistical simulation should be useful to stochastic subsurface characterization and inversion of discrete heterogeneity within hydrogeologic systems. One, two, or three-dimensional (3-D) spatial information derived from geophysical imaging or logging (e.g., seismic, electrical) or hydrogeologic interpretation (e.g., cross-sections or calibrated flow models) could be treated as a type of soft data available for conditioning geostatistical simulation of categorical variables. In stochastic inversion, a general framework for integrating prior information could be used, for example, to selectively manipulate discrete heterogeneity structure, such as the interconnectivity of permeable units or the continuity of impermeable units (Carle and Ramirez, 1999; Carle et al., 1999; Aines et al., 2002; Wainwright et al., 2014). A "Markov-Bayes" approach has been proposed to transform the soft data to prior probability distributions (Zhu, 1991; Deutsch and Journel, 1998), but this approach has been deemed intractable because of nonlinear relationships and large volumetric scale of the soft data (Deutsch and Wen, 2000).

Categorical or "indicator" bivariate geostatistical approaches have long been recognized as offering realistic and practical means for assessing the impact of subsurface heterogeneity on field- and basin-scale flow and transport processes (Poeter and Townsend, 1994; McKenna and Poeter, 1995; Poeter and McKenna, 1995; Tsang et al., 1996). As part of the **T-ProGS** software package (Carle, 1999; Carle, 2007), the **tsim** code was developed to take advantage of the interpretability of transition probability statistic to ensure that spatial cross-correlations and juxtapositional tendencies of hydrogeologic units or hydrofacies are fully considered in categorical stochastic simulation (Carle, 1996; Carle, 1997; Carle and Fogg, 1996; Carle and Fogg, 1997). The **tsim** code was modified from the variogram-based **sisim** code (Deutsch and Journel, 1992; Deutsch and Journel, 1998) in three main ways: 1) formulation of the estimate of the local probability of occurence of a discrete category by a cokriging system of equations instead of multiple indicator kriging equations (Carle and Fogg, 1996), 2) addition of a simulated quenching step to improve match of simulated and modeled simulated variability (Carle, 1997), and 3) addition of an option to vary the local direction of anisotropy direction (Carle, 1996; Carle, 1999; Carle, 2007). The latter modification enables simulation of curvilinear features such as variable stratigraphic dip, major direction of anistropy, and sinuous meandering of fluvial facies (Carle et al., 1998; Tompson et al., 1999; Carle et al., 2006; Green et al., 2010; Engdahl et al., 2012). These modifications were made to improve the ability of geostatistical methods to simulate realistic three-dimensional alluvial or fluvial hydrofacies architecture that influences site-scale flow and transport behavior (Fogg, 1986; Tompson et al., 1999; Fogg et al., 2000; Labolle and Fogg, 2001). Another development of the **T-ProGS** simulation environment is that it leverages the geologic reality that the spatial variability of many hydro- or geo-facies systems can be characterized by a continuous-lag Markov chain, which provides an intuitive yet statistically rigorous framework for developing geologically realistic models with a minimal number of parameters (Vistelius, 1949; Krumbein and Dacey, 1969; Harbaugh and Bonham-Carter, 1970; Agterberg, 1974; Doveton, 1994; Carle and Fogg, 1997). However, the user of **tsim** can choose to implement other 3-D models of the transition probability, if so desired (Carle, 1999; Carle, 2007).

The **tsim** code and its associated transition probability-based categorical geostatistical methodologies have subsequently been found to be useful to characterization and modeling of a variety of heterogeneous hydrogeological systems (Carle, 2000; Ritzi, 2000; Lu and Zhang, 2002; Lu et al., 2002; Bohling and Dubois, 2003; Zhang and Fogg, 2003; Carle et al., 2004; James, 2004; Troldborg, 2004; Weissmann et al., 2004; McDonald et al., 2005; Ye and Khaleel, 2008; Janza, 2009; Sakaki et al., 2009; Alberto, 2010; Engdahl et al., 2010b; Janza, 2009; Doherty and Christensen, 2011; Papapetrou and Theodossiou, 2012; Purkis et al., 2012; Guastaldi et al., 2014; He et al., 2015; Song et al., 2015; Weissmann et al., 2015; Krage et al., 2016; Zhu et al., 2016a; Meirovitz et al., 2017; Muskus and Falta, 2018; Erdal et al., 2019; Sun et al., 2019; Wu et al., 2019; Arshadi et al., 2020). Furthermore, **tsim** has been applied to the study of a range of problems and processes involving subsurface heterogeneity including groundwater recharge or river flow loss (Tompson et al., 1999; Izbicki, 2002; Fleckenstein et al., 2006; Frei et al., 2009; Engdahl et al., 2010a, Pryshlak et al., 2015; Ganot et al., 2018; Maples et al., 2019, Maples et al., 2020), integration of geophysical data or imaging (Carle and Ramirez, 1999; Carle et al., 1999; Zhu et al., 2016b), aquifer or pore system interconnectivity, percolation, and preferential flow (Fogg et al., 2000; Proce et al., 2004; Harter, 2005; Knudby et al., 2006; Bianchi et al., 2011; Huang et al., 2012), risk analysis (Maxwell et al., 2000; Maxwell et al., 2008), heterogeneity effects on groundwater flow (Lu et al., 2001; Jones et al., 2002; Phillips et al., 2007; Traum et al., 2014; Bianchi, 2017; Liao et al., 2020), heterogeneity effects on contaminant transport (Pawloski et al., 2001; Tompson et al.,

2002; Hu et al., 2003; Maxwell et al., 2003; Zhang and Fogg, 2003; Pozdniakov et al., 2005; Zhang et al., 2007; Sivakumar et al., 2005a; Sivakumar et al., 2005b; Carle et al., 2006; Maji and Sudicky, 2008; Sun et al., 2008; Ye et al., 2009; Cooper et al., 2010; Baidariko and Pozdniakov, 2011; Zhang and Meerschaert, 2011; Pozdniakov et al., 2012; Zhang et al., 2013; Glinskii et al., 2014; Liu et al., 2014; Zhang et al., 2014; Beisman et al., 2015; Bianchi et al., 2015; Lu et al., 2015; Maghrebi et al., 2015; Siirila-Woodburn and Maxwell, 2015; Siirila-Woodburn et al., 2015; Mi et al., 2016; Bianchi and Pedretti, 2017; Giraldo et al., 2017; Soltanian et al., 2017a; Teramoto et al., 2017; Chen et al., 2018; Guo et al., 2019b; Guo et al., 2019c; Vincent Henri and Harter, 2019; Guo et al., 2020), efficacy of remediation (Labolle and Fogg, 2001; Lee, 2004; Misut, 2014; Abriola et al., 2019; Guo et al., 2019a), effects of diffusion, dispersion, or fractionation on groundwater tracers (Labolle and Fogg, 2001; Weissmann et al., 2002; Labolle et al., 2006; Labolle et al., 2008; Green et al., 2010; Green et al., 2014; Engdahl et al., 2012; Yin et al., 2020), sequestration of carbon dioxide (Hovorka et al., 2001; Doughty and Pruess, 2004; Ramirez et al., 2006; Ramirez et al., 2010; Deng et al., 2012; Espinet et al., 2013; Sun et al., 2013; Yang et al., 2013; Carroll et al., 2014; Mansoor et al., 2014; Mukhopadhyay et al., 2015; Bianchi et al., 2016; Kitanidis, 2016; Soltanian et al., 2016; Trainor-Guitton et al., 2016; Amooie et al., 2017; Soltanian et al., 2017b; Damico et al., 2018; Buscheck et al., 2019; Yang et al., 2019; Yang et al., 2020), stochastic inversion of hydrofacies spatial distributions, hydraulic properties, or transport behavior (Aines et al., 2002; Jones et al., 2003; Harp et al., 2008; Bohling and Butler, 2010; Harp and Vesselinov, 2010; Blessent et al., 2011; Espinet and Shoemaker, 2013; Berg and Illman, 2015; Wang et al., 2017; Lee et al., 2018; Song et al., 2019), spatial variability of reactive mineral assemblages (Carle et al., 2002; Deng et al., 2010), analysis of contaminant plumes (Reed et al., 2004; Maji et al., 2006; Maji and Sudicky, 2008), assessment of nitrate contamination, reduction, and removal (Carle et al., 2004; Carle et al., 2006; Hansen et al., 2014; Sawyer, 2015; Wallace et al., 2020), permeability structure within fractured rock (Park et al., 2004; Blessent et al., 2011; Blessent, 2013), 3-D modeling of ore-grade distributions (Fisher et al., 2005), probabilistic well location (Stevick et al., 2005), non-point source contamination (Zhang et al., 2006; Refsgaard et al., 2014; Zhang et al., 2018), upscaling of flow and transport parameters (Dai et al., 2007; Fleckenstein and Fogg, 2008; Bakshevskaia and Pozdniakov, 2016), geotechnical engineering (Beretta and Felletti, 2007; Felletti and Berretta, 2009; Zetterlund et al., 2011; Grasmick et al., 2020), assessment of nuclear waste disposal (Back and Sundberg, 2007), effects of subsurface heterogeneity on remote sensing (Eslinger et al., 2007), wellhead protection and contamination vulnerability (Burow et al., 2008; Heywood, 2013; Yager and Heywood, 2014; Theodossiou and Fotopoulou, 2015), physical and chemical heterogeneity in streambeds and the hyporheic zone (Schornberg et al., 2010; Faulkner et al., 2012; Zhou et al., 2014; Pryshlak et al., 2015; Singh et al., 2018; Pescimoro et al., 2019; Liu et al., 2020), effects of micro-topography on surface-subsurface exchange (Frei et al., 2010), analysis of transport at the macrodispersion experiment site (Bianchi et al., 2011; Zheng et al., 2011; Bianchi and Zheng, 2016; Pedretti and Bianchi, 2018; Yin et al., 2020), 3-D soil texture (Haugen et al., 2011; Roig-Silva et al., 2012; Li et al., 2014b), characterization of groundwater ecosystems (Larned, 2012), effects of petroleum reservoir heterogeneity (Purkis et al., 2012; Kwon et al., 2017), geologic units of the Swiss Jura (Sartore, 2013), groundwater hydrology of fens (Sampath et al., 2015; Sampath et al., 2016), potential for liquefaction (Munter et al., 2016; Munter et al., 2017; Boulanger et al., 2019), coupled surface and subsurface flow (Blessent et al., 2017; Erdal et al., 2019), and geomechanical modeling of land subsidence (Zhu et al., 2020). Given the usefulness of the **tsim** algorithm, which was originally designed for categorical stochastic simulation with conditioning by hard data only, an improved capability to assimilate conditioning from soft data or prior information of variable quality is expected to be useful to hydrogeological and related subsurface applications and research.

In this paper, a simple theoretical framework is developed for incorporating uncertain, indirect, or soft categorical data into categorical geostatistical simulation. Geostatistical realizations will honor or be conditional to both hard and soft data. The theory considers that soft data should not be treated the same as hard data in formulating (co)kriging equations and objective functions in simulated annealing (or quenching) as implemented in the original categorical stochastic simulation codes using bivariate spatial statistics such as **isim3d** (Gomez-Hernandez and Srivastava, 1990), **tsim** (Carle 1996; Carle et al., 1998), **sisim** and **anneal** (Deutsch and Journel, 1998), and **iksim** (Ying, 2000).

A new version of **tsim**, called **tsim-s**, has been coded to enable incorporation of soft categorical data or prior information. The **tsim-s** algorithm was originally conceived to enable **tsim** to produce and perturb stochastic realizations for Monte Carlo Markov chain inversion (Aines et al., 2002; Carle, 2003; Glaser et al., 2004). The new capabilities in **tsim-s** have more general applicability and flexibility to handle conditioning data of variable quality or uncertainty. **tsim-s** will be made available by request as the open-source fortran code **tsim** has been distributed in the past. The development of the equations necessary for implementation of the **tsim-s** algorithm are included in this paper to fully document the methods and to facilitate coding of the **tsim** algorithms in higher-level languages such as R (Sartore, 2013; Sartore et al., 2016). As will be seen in the equations, the computational overhead for **tsim-s** is not signifcantly different from **tsim** because the only modifications are to the entries in the cokriging matrices and the parameters of the quenching objective function.

This paper provides the theory behind the **tsim-s** algorithms and results from example applications. The paper first reviews transition probability-based indicator geostatistical theory implemented in the **tsim** algorithms. Next, the paper derives the equations used for implementing the new soft data capabilities to account for uncertainty in categorical variables using the "hardness" concept previously introduced for continuous variables (Deutsch and Wen, 2000). Cokriging equations and simulated quenching objective functions are re-

ormulated to account for soft data using the hardness concept. Example applications are given for hard and soft conditioning at boreholes and from prior conditioning by an array of data such as another realization or a depth-dependent uncertainty function. Results of a 3-D simulation by Carle et al. (2006) are included to demonstrate that **tsim-s** can be used to produce large-scale stochastic realizations useful for investigation of flow and transport processes in hydrogeologic systems. These examples are provided to show the flexibility of the algorithm and by no means cover all potential applications, a topic that is beyond the scope of this paper. Further discussion is added to clarify the capabilities and limitations of the **tsim** and **tsim-s** algorithms in relationship to variogram-based and multi-point geostatistical methods and the current understanding of how the simulation algorithms can and should be implemented in a geological context.

2 MATERIALS AND METHODS

2.1 Transition Probability-Based Indicator Geostatistics

Transition probability-based indicator geostatistics is a categorical geostatistical approach where the transition probability bivariate statistic is used to analyze spatial variability and formulate cokriging equations (Carle and Fogg, 1996). The transition probability approach enables consideration of spatial cross correlations (e.g., how different facies tend to locate in space relative to each other) and facilitates a Markov chain modeling framework that can be linked to geologic interpretation (Krumbein and Dacey, 1969; Miall, 1973; Carle and Fogg, 1997). Indicator variogram-based geostatistical approaches do not fully consider spatial cross-correlations and rely on data-intensive empirical curve fitting for model development (Deutsch and Journel, 1992; Deutsch and Journel, 1998; Goovaerts, 1996; Goovaerts, 1997).

In a categorical geostatistical approach, an indicator variable is defined with respect to mutually exclusive or discrete categorical variables (e.g., lithofacies, hydrofacies) by

$$I_k(\mathbf{x}) = \begin{cases} 1, \text{ if category } k \text{ occurs at location } \mathbf{x} \\ 0, \text{ otherwise} \end{cases} \qquad k = 1, \ldots, K, \tag{1}$$

where $\mathbf{x}$ is location, and K is the number of categories. The probability that category k occurs at $\mathbf{x}$ is equivalent to the expected value of the indicator variable:

$$\Pr\{I_k(\mathbf{x}) = 1\} = E\{I_k(\mathbf{x})\}. \tag{2}$$

In transition probability-based indicator geostatistics (Carle and Fogg, 1996), the transition probability bivariate spatial statistic is used to quantitatively describe spatial variability of the discrete categorical variables, which we will generally refer to as "facies." Assuming second-order stationarity, the transition probability $t_{jk}(\mathbf{h})$ is defined as a conditional probability that depends on a lag separation vector $\mathbf{h}$ by

$$t_{jk}(\mathbf{h}) = \Pr\{k \text{ occurs at } \mathbf{x} + \mathbf{h} \,|\, j \text{ occurs at } \mathbf{x}\}. \tag{3}$$

Applying Bayes theorem and **Eqs. 1 and 3** is formulated with respect to indicator variables by

$$t_{jk}(\mathbf{h}) = \frac{\Pr\{j \text{ occurs at } \mathbf{x} \text{ and } k \text{ occurs at } \mathbf{x} + \mathbf{h}\}}{\Pr\{j \text{ occurs at } \mathbf{x}\}} = \frac{\Pr\{I_j(\mathbf{x}) = 1 \text{ and } I_k(\mathbf{x} + \mathbf{h}) = 1\}}{\Pr\{I_j(\mathbf{x}) = 1\}}. \tag{4}$$

The transition probability entries, $t_{jk}(\mathbf{h})$, form the transition probability matrix, $\mathbf{T}(\mathbf{h})$, as

$$\mathbf{T}(\mathbf{h}) = \begin{bmatrix} t_{11}(\mathbf{h}) & \cdots & t_{1K}(\mathbf{h}) \\ \vdots & \ddots & \vdots \\ t_{K1}(\mathbf{h}) & \cdots & t_{KK}(\mathbf{h}) \end{bmatrix}. \tag{5}$$

Other bivariate statistics, such as the indicator (cross-) variogram or covariance can be used to implement indicator geostatistical techniques. However, the transition probability has several advantages:

- The transition probability is defined as a conditional probability, which facilitates the connection of statistical measures to geologic interpretation of facies architecture (Miall, 1973; Carle et al., 1998).
- The geologically observable and interpretable parameters of proportions, mean length, and juxtapositional tendencies can be used to develop Markov chain models (Carle and Fogg, 1996; Carle and Fogg, 1997).
- Non-symmetric juxtapositional tendencies can be considered (Carle and Fogg, 1996).
- Three-dimensional (3-D) transition probability models of spatial variability are readily developed from 1-D Markov chains along principal stratigraphic directions (Carle and Fogg, 1997).
- Continuous-lag Markov chains have been found suitable for 3-D modeling of vertical and lateral spatial transitioning among geo- or hydro-facies (Carle 1996; Carle and Fogg, 1996; Carle and Fogg, 1997; Carle et al., 1998; Fogg et al., 1998; Zhang and Fogg, 2003; Proce et al., 2004; Ye and Khaleel, 2008; Engdahl et al., 2010a; Bianchi et al., 2011; Pozdniakov et al., 2012; Purkis et al., 2012; Bakshevskaia and Pozdniakov, 2016; Krage et al., 2016; Sartore et al., 2016; Zhu et al., 2016a; Meirovitz et al., 2017; Guo et al., 2019b).

2.2 Simulation with Hard Data Only

The **tsim** computer code is used to generate geostatistical "realizations" of categorical variables such as lithology, soil texture, or hydrofacies. The realizations generated by **tsim** consist of a rectangular block of regularly-spaced grid cells. The conditional simulation algorithm consists of two steps:

(1) cokriging-based "sequential indicator simulation" (SIS), and
(2) simulated quenching.

summarized below and described in further detail by Carle (1996), Carle et al. (1998), and Deutsch and Journel (1998).

2.2.1 Sequential Indicator Simulation Step

The algorithm and code used by the SIS step in **tsim** is modified from the **sisim** code (Deutsch and Journel, 1992; Deutsch and Journel, 1998). The SIS step in **tsim** traces a random path through every grid cell in the realization. At each grid cell **tsim** uses cokriging (instead of repeated kriging steps as in **sisim**) to estimate conditional probabilities that a facies occurs at a grid cell given surrounding conditioning data, which are typically facies occurrences located at nearby grid cells. Initially, hard data are the only conditioning information. In the process of completing the simulation, nearby simulated values serve as hard conditioning data for the future cokriging estimates along the random path. Based on the cokriging estimates of the conditional probability that a facies occurs at a particular grid cell given facies occurrences at other nearby cells, a uniformly-distributed random number is used to select the category that occurs at a grid cell in the realization. This process continues one grid cell at a time until all cells have been reached by the random path.

The indicator cokriging estimate at a location, $\mathbf{x}_0$, is formulated as a weighted sum by

$$\Pr\left\{k \text{ occurs at } \mathbf{x}_0 \mid i_j(\mathbf{x}_\alpha);\ \alpha = 1,\ldots,N;\ j = 1,\ldots,K\right\} \approx \sum_{\alpha=1}^{N}\sum_{j=1}^{K} i_j(\mathbf{x}_\alpha) w_{jk,\alpha}, \quad (6)$$

where $i_j(\mathbf{x}_\alpha)$ are indicator data values, N is the number of data, K is the number of categories, and $w_{jk,\alpha}$ represent weighting coefficients. The weighting coefficients $w_{jk,\alpha}$ are computed by the transition probability-based cokriging system of equations (Carle, 1996; Carle and Fogg, 1996)

$$\begin{bmatrix} \mathbf{T}(\mathbf{x}_1-\mathbf{x}_1) & \cdots & \mathbf{T}(\mathbf{x}_N-\mathbf{x}_1) \\ \vdots & \ddots & \vdots \\ \mathbf{T}(\mathbf{x}_1-\mathbf{x}_N) & \cdots & \mathbf{T}(\mathbf{x}_N-\mathbf{x}_N) \end{bmatrix} \begin{bmatrix} \mathbf{W}_1 \\ \vdots \\ \mathbf{W}_N \end{bmatrix} = \begin{bmatrix} \mathbf{T}(\mathbf{x}_0-\mathbf{x}_1) \\ \vdots \\ \mathbf{T}(\mathbf{x}_0-\mathbf{x}_N) \end{bmatrix}, \quad (7)$$

where N is the number of data and

$$\mathbf{W}_n = \begin{bmatrix} w_{11,n} & \cdots & w_{1K,n} \\ \vdots & \ddots & \vdots \\ w_{K1,n} & \cdots & w_{KK,n} \end{bmatrix} \qquad \text{for } n = 1,\ldots,N. \quad (8)$$

The indicator cokriging estimate is only an approximation of the conditional probability on the left side of **Eq. 6**. The SIS step provides the "initial configuration" for the next step in **tsim**, simulated quenching (Carle, 1997; Carle et al., 1998).

2.2.2 Simulated Quenching Step

The SIS step alone does not ensure that the realization will honor the model of spatial variability including all spatial auto-correlations [$t_{jk}(\mathbf{h})$ for $j = k$] and cross-correlations [$t_{jk}(\mathbf{h})$ for $j \neq k$]. The simulated quenching step is used to improve the match between modeled and simulated spatial variability by attempting to minimize an objective function, **O,** defined by

$$\mathbf{O} = \sum_{l=1}^{M}\sum_{j=1}^{K}\sum_{k=1}^{K}\left[t_{jk}(\mathbf{h}_l)_{SIM} - t_{jk}(\mathbf{h}_l)_{MOD}\right]^2, \quad (9)$$

where $\mathbf{h}_l$ denote $l = 1,\ldots,M$ specified lag vectors and "SIM" and "MOD" distinguish simulated and modeled transition probabilities, respectively (Aarts and Korst, 1989; Deutsch and Journel, 1992; Deutsch and Journel, 1998; Deutsch and Cockerham, 1994; Carle, 1997; Deutsch and Journel, 1998). Simulated quenching is implemented by cycling through every grid cell of the realization several times along a random path and querying whether a change in facies will decrease **O**; if so, the category is changed. Conditioning of hard data is maintained during quenching by not allowing changes of categories at conditioning locations. The quenching step continues until a specified number of iterations through the every grid cell is reached or **O** is reduced below a specified minimum threshold (Carle et al., 1998; Carle, 1999, Carle, 2007).

Simulated quenching is the "zero temperature" form of simulated annealing, where an "annealing schedule" determines a probability of acceptance for changes that increase **O** to avoid high-valued local minima in the solution space for **O** (Deutsch and Journel, 1992; Carle, 1997). The main advantages of using simulated quenching over annealing are 1) the difficulty of designing and implementing an annealing schedule is avoided and 2) quenching is much faster. In **tsim**, the cokriging-based SIS step avoids high-valued local minima of **O** by providing a spatially-correlated initial configuration prior to quenching. The quenching step simply modifies existing spatial structures in the initial configuration to be consistent with the transition probability model.

2.3 Soft Data Conditioning in Categorical Geostatistical Simulation

Two concepts are presented here to enable location-specific soft data conditioning for mutually exclusive categories:

- "prior probability," which assigns probabilities between zero and one to each category (Deutsch and Journel, 1998; Ying, 2000), and
- "hardness," which assigns one uncertainty measure to account for overall uncertainty of the data at a given location (Deutsch and Wen, 2000).

A common practical situation is that the data are categorized (e.g., as textures in driller's logs or by interpreted hydrofacies) but known to be uncertain (e.g., because the driller's logs or geological interpretations are not 100% accurate). In this situation, the latter approach is more straightforward to apply, although both approaches can be applied simultaneously.

2.3.1 Prior Probabilities

Uncertainty in the indicator data or the "softness" of categorical variables can be accounted for in **tsim** by assigning prior probabilities between zero and unity to the indicator values, which Weissmann and Fogg (1999) implemented in application use of driller's logs. This prior probability approach has also been implemented in integration of geophysical imaging to condition stochastic simulations of hydrofacies architecture (Carle et al., 1999; Carle and Ramirez, 1999; Hermans and Irving, 2017). However, the prior probability

approach can be problematic in many practical situations for several reasons:

- The most readily available categorical data (e.g., from driller's logs or geological interpretations) are discretely categorized and not presented in the form of prior probabilities.
- The practitioner may find difficulty, tedium, and confusion in assigning multiple prior probabilities between zero and unity to the different facies categories over hundreds, thousands, or more data points.
- Uncertainty in indicator data is not accounted for in the cokriging **Eq. 7** and objective function **Eq. 9** used the SIS and simulated quenching steps, respectively, of **tsim**.

The latter reason is a persistent theoretical shortcoming of past and current bivariate geostatistical methods, wherein the (co) kriging or simulated annealing equations are weighting the soft data in the same manner as the hard data. On this topic, Deutsch and Wen (2000) stated that "a significant problem with kriging-based approaches (to stochastic simulation) is that there is no convenient way to handle the fact that the soft data have locally variable precision (or accuracy)" and, as a result, proposed a simulated-annealing approach. However, we believe there is a convenient way to handle soft data in both cokriging and simulated quenching of categorical variables.

2.3.2 Categorical Data with Hardness

We introduce a simple alternative approach to integration of uncertain or soft categorical data into stochastic simulation by extending the concept of data hardness to categorical variables, as previously proposed for continuous variables (Deutsch and Wen, 2000). This approach requires assignment of a hardness value ranging between zero and unity to the set of facies (or indicator) probabilities given for each data location. On the extremes, a hardness value of 1.0 represents hard data, and 0.0 represents data that provide no additional information. To incorporate hardness into transition probability-based indicator geostatistics, a soft datum is assumed to consist of a weighted sum of both *certain* and *uncertain* information, with weights that sum to unity. The certain portion is a set of indicator values represented in binary form (e.g., the presence or absence of a certain lithology) or as prior probabilities, and the uncertain portion is, in effect, a state of no useful information.

Assuming stationarity, the condition of complete uncertainty for the expected value of the indicator variable, $I_k(\mathbf{x})$, for a facies at location $\mathbf{x}$ is the marginal probability or proportion, p_k, such that

$$E\{I_k(\mathbf{x})\} = p_k. \tag{10}$$

The soft indicator value, denoted by $\tilde{i}_k(x)$, consists of a weighted sum of a hard indicator value $i_k(\mathbf{x})$ and the marginal probability, p_k, according to

$$\tilde{i}_k(\mathbf{x}) = \alpha(\mathbf{x}) i_k(\mathbf{x}) + \beta(\mathbf{x}) p_k,$$

where the weights, $\alpha(\mathbf{x})$ and $\beta(\mathbf{x})$, indicate hardness and softness, respectively, at location $\mathbf{x}$. The hardness and softness weights are complementary to each other as

$$\alpha(\mathbf{x}) + \beta(\mathbf{x}) = 1. \tag{11}$$

Values of hardness or softness weights are assumed to depend only on location and not on individual categories. A soft indicator variable $\tilde{I}_k(\mathbf{x})$ is defined with respect to a hard indicator variable $I_k(\mathbf{x})$ defined in **Eq. 1** by

$$\tilde{I}_k(\mathbf{x}) = \alpha(\mathbf{x}) I_k(\mathbf{x}) + \beta(\mathbf{x}) p_k. \tag{12}$$

In practice, a single hardness value is assigned to the set of indicator values (i.e., facies probabilities). Compared to the original **tsim** code, hardness values are the only additional conditioning data information needed to implement the soft data approach described herein for **tsim-s**.

2.4 Transition Probabilities and Cokriging with Soft Data

The transition probability values used in the cokriging **Eq. 7** were originally formulated under the assumption of hard data (Carle and Fogg, 1996). To incorporate soft data, transition probability values in **Eq. 7** will need to be modified to reflect the uncertainty of the data used to formulate the cokriging estimate. For example, if a datum has zero hardness [$\alpha(\mathbf{x}) = 0$ or $\beta(\mathbf{x}) = 1$] and, therefore, provides no additional information, that datum should not impact the cokriging estimate. In particular, the left hand side matrix of **Eq. 7**, which accounts for the "declustering" of the data, and the right hand side of **Eq. 7**, which accounts for the statistical closeness (spatial correlation) of the data with respect to the estimation location (Isaaks and Srivastava, 1989; Deutsch and Journel, 1992, Deutsch and Journel, 1998), should be modified to account for data uncertainty.

To account for soft data, the transition probability entries in **Eq. 7** must be modified to account for decreased spatial correlation of soft data relative to hard data. This decrease in spatial correlation is derived below using the transition probability as the bivariate spatial statistic. Substituting **Eq. 2** into **Eq. 4**, the transition probability is defined with respect to hard indicator variables by

$$t_{jk}(\mathbf{h}) = \frac{E\{I_j(\mathbf{x}) I_k(\mathbf{x}+\mathbf{h})\}}{E\{I_j(\mathbf{x})\}}. \tag{13}$$

Substituting soft indicator values as defined by **Eq. 12** into **Eq. 13**, a "soft transition probability" $\tilde{t}_{jk}(\mathbf{h})$ is formulated by

$$\tilde{t}_{jk}(\mathbf{h}) = \frac{E\{[\alpha(\mathbf{x}) I_j(\mathbf{x}) + \beta(\mathbf{x}) p_j][\alpha(\mathbf{x}+\mathbf{h}) I_k(\mathbf{x}+\mathbf{h}) + \beta(\mathbf{x}+\mathbf{h}) p_k]\}}{E\{[\alpha(\mathbf{x}) I_j(\mathbf{x}) + \beta(\mathbf{x}) p_j]\}}. \tag{14}$$

Expanding (as shown step-by-step in the **Appendix**), assuming stationarity, applying **Eqs. 10 and 11**, and combining terms, **Eq. 14** reduces to

$$\tilde{t}_{jk}(\mathbf{h}) = \frac{E\{\alpha(\mathbf{x})\alpha(\mathbf{x}+\mathbf{h})I_j(\mathbf{x})I_k(\mathbf{x}+\mathbf{h}) + [\beta(\mathbf{x}+\mathbf{h}) + \beta(\mathbf{x})\alpha(\mathbf{x}+\mathbf{h})]p_j p_k\}}{p_j}. \tag{15}$$

Assuming that hardness values are independent and again applying **Eq. 10, Eq. 15** reduces to

$$\tilde{t}_{jk}(\mathbf{h}) = \alpha(\mathbf{x})\alpha(\mathbf{x}+\mathbf{h})\frac{E\{I_j(\mathbf{x})I_k(\mathbf{x}+\mathbf{h})\}}{E\{I_j(\mathbf{x})\}} + \frac{[1-\alpha(\mathbf{x}+\mathbf{h}) + [1-\alpha(\mathbf{x})]\alpha(\mathbf{x}+\mathbf{h})]p_j p_k}{p_j}. \tag{16}$$

Applying **Eq. 13** and simplifying the right hand side, **Eq. 16** reduces to

$$\tilde{t}_{jk}(\mathbf{h}) = \alpha(\mathbf{x})\alpha(\mathbf{x}+\mathbf{h})t_{jk}(\mathbf{h}) + [1-\alpha(\mathbf{x})\alpha(\mathbf{x}+\mathbf{h})]p_k. \tag{17}$$

According to **Eq. 17**, the soft transition probability is a weighted sum of the transition probability $t_{jk}(\mathbf{h})$ and the marginal probability p_k. The weight for $t_{jk}(\mathbf{h})$ is the product of the hardness values $\alpha(\mathbf{x})$ and $\alpha(\mathbf{x}+\mathbf{h})$ at the two datum locations, and the weight for p_k is the complement to the weight for $t_{jk}(\mathbf{h})$.

To consider soft data, the transition probability-based indicator cokriging equations are simply modified by substituting $\tilde{\mathbf{T}}(\mathbf{h})$ for $\mathbf{T}(\mathbf{h})$ in **Eq. 7** as follows

$$\begin{bmatrix} \tilde{\mathbf{T}}(\mathbf{x}_1-\mathbf{x}_1) & \cdots & \tilde{\mathbf{T}}(\mathbf{x}_N-\mathbf{x}_1) \\ \vdots & \ddots & \vdots \\ \tilde{\mathbf{T}}(\mathbf{x}_1-\mathbf{x}_N) & \cdots & \tilde{\mathbf{T}}(\mathbf{x}_N-\mathbf{x}_N) \end{bmatrix} \begin{bmatrix} \mathbf{W}_1 \\ \vdots \\ \mathbf{W}_N \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{T}}(\mathbf{x}_0-\mathbf{x}_1) \\ \vdots \\ \tilde{\mathbf{T}}(\mathbf{x}_0-\mathbf{x}_N) \end{bmatrix},$$

where,

$$\tilde{\mathbf{T}}(\mathbf{x}_m-\mathbf{x}_n) = \begin{bmatrix} \tilde{t}_{11}(\mathbf{x}_m-\mathbf{x}_n) & \cdots & \tilde{t}_{1K}(\mathbf{x}_m-\mathbf{x}_n) \\ \vdots & \ddots & \vdots \\ \tilde{t}_{K1}(\mathbf{x}_m-\mathbf{x}_n) & \cdots & \tilde{t}_{KK}(\mathbf{x}_m-\mathbf{x}_n) \end{bmatrix}.$$

If $\alpha(\mathbf{x}_m) = 1$ and $\alpha(\mathbf{x}_n) = 1$, the soft transition probability matrix $\tilde{\mathbf{T}}(\mathbf{x}_m-\mathbf{x}_n)$ is identical to the hard transition probability matrix $\mathbf{T}(\mathbf{x}_m-\mathbf{x}_n)$ as defined by **Eq. 5**. If either $\alpha(\mathbf{x}_m) = 0$ or $\alpha(\mathbf{x}_n) = 0$,

$$\tilde{\mathbf{T}}(\mathbf{x}_m-\mathbf{x}_n) = \begin{bmatrix} p_1 & \cdots & p_K \\ \vdots & & \vdots \\ p_1 & \cdots & p_K \end{bmatrix} \qquad k = 1,\ldots,K,$$

where the entries in each column k are the proportions p_k. Assuming stationarity and ergodicity,

$$\lim_{h_\phi \to \infty} = t_{jk}(h_\phi) = p_k \tag{18}$$

for lags h_ϕ in any direction ϕ. **Eq. 18** indicates that for large lags (beyond the range of spatial correlation), the transition probabilities converge on the marginal probabilities. Thus, the matrix values for $\tilde{\mathbf{T}}(\mathbf{x}_m-\mathbf{x}_n)$ when $\alpha(\mathbf{x}_m) = 0$ and $\alpha(\mathbf{x}_m) = 0$ are identical to the case for hard data where the lag $(\mathbf{x}_m-\mathbf{x}_n)$ is large enough such the two data are not spatially correlated.

Importantly, the right hand side entries, $\tilde{\mathbf{T}}(\mathbf{x}_0-\mathbf{x}_n)$ for $n = 1,\ldots,N$, should be formulated assuming $\alpha(\mathbf{x}_0) = 1$ because:

- Cokriging is used to estimate the probability that a category exists at location $\mathbf{x}_0$ assuming that $I_k(\mathbf{x}_0)$ for $k = 1,\ldots,K$ are hard indicator variables, as defined in **Eq. 1**.
- The estimation location, $\mathbf{x}_0$, is known.

However, if the location of $\mathbf{x}_0$ is not certain, the hardness value $\alpha(\mathbf{x}_0)$ could be used to account for estimation location uncertainty.

2.5 Simulated Quenching with Soft Data

2.5.1 Use of Acceptance Probability

In **tsim**, the simulated quenching step enforces hard conditioning by not allowing any changes in categories at grid cell locations with hard data. At grid cell locations with no data, categories are changed along the random path wherever **O** can be reduced. One can view the quenching algorithm in terms of a bi-modal acceptance probability for changing the category on the basis of reducing **O**: 0.0 if the category is determined by data and 1.0 if the category is not determined by data.

In **tsim-s**, categories are also queried for change at each grid cell, but with a lesser probability of acceptance at grid cell locations containing soft data as compared to cells with no data. The probability of accepting a change of categories that reduces **O** is set at $\beta(\mathbf{x})$; corresponding to **Eq. 11**, the probability of rejecting the change is $\alpha(\mathbf{x})$. Thus, if hardness equals unity $[\alpha(\mathbf{x}) = 1]$, categories are not allowed to be changed at location **x**. For soft data with hardness less than unity $[\alpha(\mathbf{x}) < 1]$, changes in categories that reduce **O** are accepted with a probability of $\beta(\mathbf{x})$ along the random path implemented by the quenching algorithm until **O** is sufficiently reduced in **Eq. 9**. This approach enables the simulated quenching step to impart the most change in the realizations at the locations where data are least certain, and the least change where data are most certain. This algorithm is, in effect, a location-dependent simulated annealing schedule where the acceptance probability is proportional to the softness of the data.

2.5.2 Use of the Joint Probability

In the example discussed later in Section 3.1, one category (*gravel*) has a very low proportion of 0.006. In application of tsim, low-proportion facies could be problematic in the simulated quenching step because of the small amount of sample statistics for matching the measured and modeled transition probabilities. Alternatively, the objective function can be re-formulated with respect to the joint probability, $p_{jk}(\mathbf{h})$

$$\begin{aligned} p_{jk}(\mathbf{h}) &= \Pr\{k \text{ occurs at } \mathbf{x}+\mathbf{h} \text{ and } j \text{ occurs at } \mathbf{x}\} \\ &= E\{I_j(\mathbf{x})I_k(\mathbf{x}+\mathbf{h})\} \end{aligned} \tag{19}$$

In **tsim-s** an option is available for using the joint probability defined in **Eq. 19** to formulate the simulated quenching objective function **O** as

$$\mathbf{O} = \sum_{l=1}^{M}\sum_{j=1}^{K}\sum_{k=1}^{K}\left[p_{jk}(\mathbf{h}_l)_{SIM} - p_{jk}(\mathbf{h}_l)_{MOD}\right]^2, \tag{20}$$

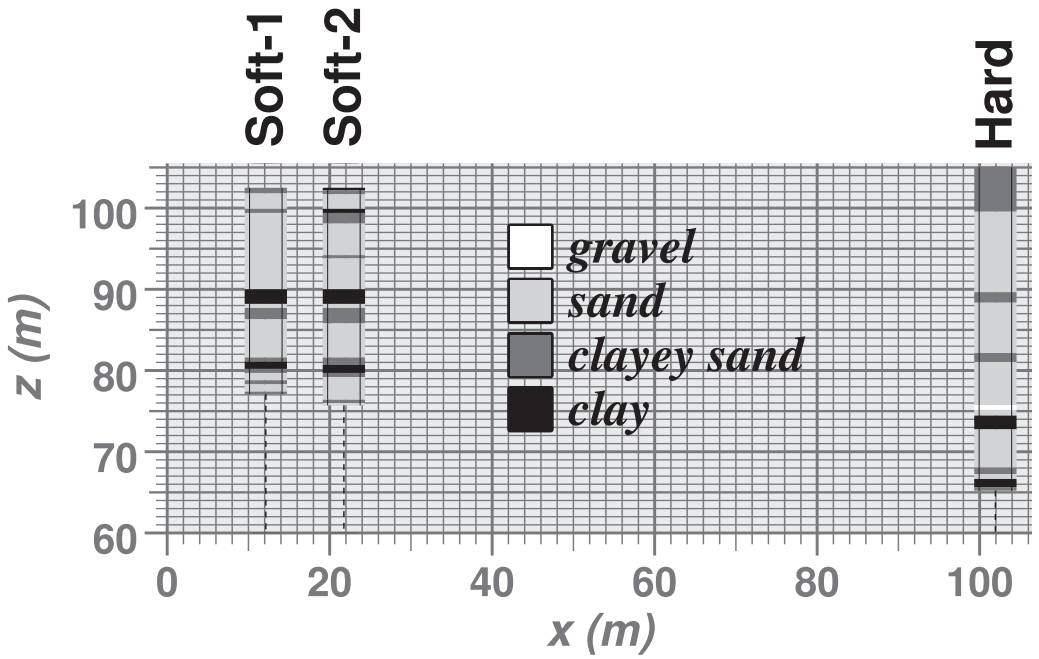


FIGURE 1 | Lithologic data within a geologic cross-section, Savannah River Site, South Carolina. Data labeled "Soft-1" and "Soft-2″ are inferred from resistivity logs, and data marked "Hard" are obtained from core descriptions.

for the lags $l = 1, \ldots, M$ included in the quenching process. The joint probability formulation of **O** using **Eq. 20** de-emphasizes matching of measured and modeled transition probabilities for low-proportion categories compared to **Eq. 9**.

3 RESULTS-APPLICATIONS TO SAVANNAH RIVER SITE AND LAGAS BASIN

3.1 Initial Characterization of Savannah River Site

The shallow subsurface beneath the Savannah River Site (SRS) in South Carolina consists of Tertiary siliciclastic sediments deposited in shoreline and nearshore depositional environments (Aadland et al., 1995; Falls et al., 1997). Characterization of heterogeneity at SRS is of interest to improve understanding of vadose zone, groundwater flow, and contaminant migration processes (Miller et al., 2000). A primary concern is characterization of vertical and lateral extent of clay lenses within the sand-dominated flow and transport regime. A two-dimensional (2-D) analysis (for lateral x and vertical z directions) is performed for the following SRS example, with the goal of generating 2-D realizations of lithofacies heterogeneity conditioned by both hard and soft data.

Figure 1 shows lithologic and geophysical log data from SRS treated as hard and soft data, respectively, to condition realizations generated by **tsim-s**. Four texturally-based lithofacies are distinguished, with proportions in parenthesis: *gravel* (0.006), *sand* (0.735), *clayey sand* (0.156), and *clay* (0.103). The hard data (labeled "Hard") are derived from core descriptions in the borehole data, and the soft data (labeled "Soft-1" and"Soft-2") are inferred from resistivity logs from two boreholes. In some sedimentary environments, resistivity log data are not necessarily highly correlated with texturally-derived facies (Burow et al., 1997).

The lithologic data for two boreholes in the vicinity of the cross-section were used to calculate matrices of transition probability measurements with dependence on vertical lag shown by circles in **Figure 2**. The transition probability measurements associated with the gravel category are somewhat erratic because of the very low proportion of gravel. As a practical matter, this sort of measurement variability caused by data sparseness should not be unduly fitted in the transition probability modeling process. A Markov chain model, shown by the solid lines in **Figure 2**, was fitted the calculated vertical transition probabilities through use of a matrix exponential function

$$\mathbf{T}(h_z) = \exp\mathbf{R}_z h_z = \exp\begin{bmatrix} -2.233 & 2.205 & 0.028 & 0.000 \\ 0.013 & -0.256 & 0.105 & 0.138 \\ 0.018 & 0.571 & -0.952 & 0.362 \\ 0.000 & 0.846 & 0.692 & -1.538 \end{bmatrix} h_z \tag{21}$$

where the rate coefficients in $\mathbf{R}_z$ are given in units of m^{-1}. As shown by Agterberg (1974) and Carle and Fogg (1997), the entries in $\mathbf{T}(h_z)$ for **Eq. 21** are computed by an eigenvalue decomposition, so that each entry, $t_{jk}(h_z)$, is a weighted sum of the k (column) category proportion and three exponential functions:

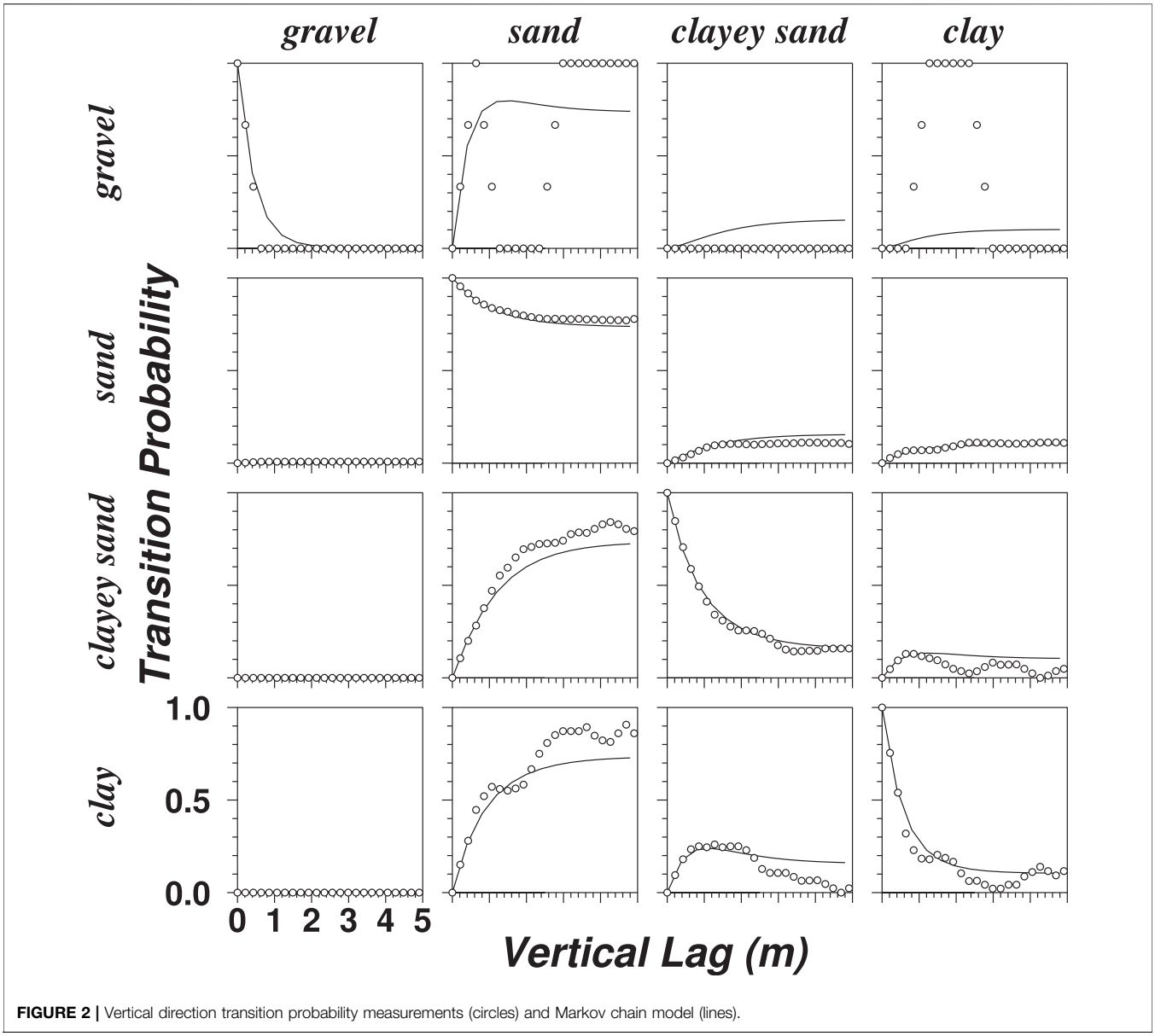


FIGURE 2 | Vertical direction transition probability measurements (circles) and Markov chain model (lines).

$$\mathbf{T}(h_z) = \begin{bmatrix} 0.006 & 0.735 & 0.156 & 0.103 \\ 0.006 & 0.735 & 0.156 & 0.103 \\ 0.006 & 0.735 & 0.156 & 0.103 \\ 0.006 & 0.735 & 0.156 & 0.103 \end{bmatrix}$$
$$+ \begin{bmatrix} 0.9832 & -1.1807 & -0.0662 & 0.2637 \\ -0.0072 & 0.0086 & 0.0048 & -0.0019 \\ -0.0180 & 0.0216 & 0.0012 & -0.0048 \\ 0.0260 & -0.0313 & -0.0018 & 0.0070 \end{bmatrix} \exp\left[-2.250\mathrm{m}^{-1} h_z \right]$$
$$+ \begin{bmatrix} 0.0117 & 0.0573 & 0.2117 & -0.2807 \\ 0.0019 & 0.0095 & 0.0351 & -0.0465 \\ 0.0116 & 0.0568 & 0.2100 & -0.2784 \\ -0.0321 & -0.1568 & -0.5800 & 0.7688 \end{bmatrix} \exp\left[-1.840\mathrm{m}^{-1} h_z \right]$$
$$+ \begin{bmatrix} -0.0004 & 0.3881 & -0.3018 & -0.0860 \\ -0.0003 & 0.2466 & -0.1917 & -0.0546 \\ 0.0008 & -0.8137 & 0.6326 & 0.1802 \\ 0.0006 & -0.5472 & 0.4255 & 0.1212 \end{bmatrix} \exp\left[-0.890\mathrm{m}^{-1} h_z \right]$$

For the lateral x direction, Markov chain transition probability models were developed for the SRS example based on prior geological estimates of lithofacies mean lengths and juxtapositional tendencies indicated by geologic cross-sections. Such geological information can be converted into lateral transition rates (e.g., $\mathbf{R}_x$) to enable development of 2- and 3-D Markov chain models and demonstrated by Carle and Fogg (1997), Carle et al. (1998), and Weissmann et al. (1999).

3.2 Geostatistical Simulation with Soft Data at SRS

3.2.1 Effect of Soft Data

Figure 3 compares **(A)** six realizations generated with hard data only with **(B)** six realizations generated with hard and soft data. In **(A)** and **(B)**, the hard data are honored on the right side of each realization where the solid vertical line is shown. In **(B)**, the soft

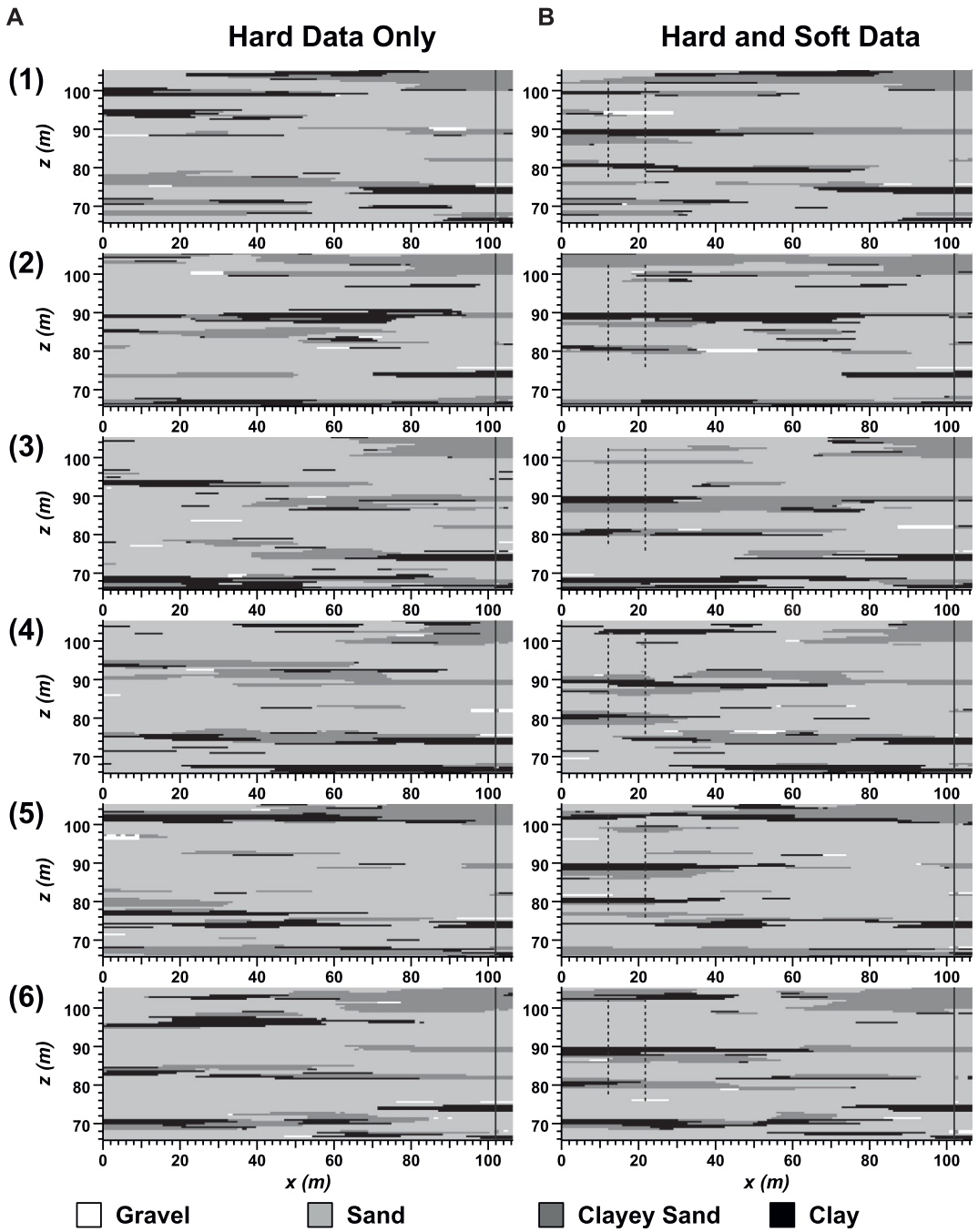


FIGURE 3 | Six lithofacies realizations generated for cases with **(A)** hard data only and **(B)** hard and soft data. Hard data locations are indicated by solid line at right, and soft data locations are indicated by dashed lines at left in **(B)**.

data are added as conditioning with hardness = 0.80 at the locations shown by dashed vertical lines toward the left side of each realization. With careful examination near soft data locations, the soft data impart a strong yet inexact influence on the lithofacies occurrences. For example, the occurrences of clay between about z = 87–90 m in both soft data boreholes produces a persistent clay layer in all six realizations, although the fit to the soft data values is not always exact. Near the top of the soft data shown in **Figure 1** at z = 98–102.5 m, more *clay* and *clayey sand* is indicated on the right. The realizations in **Figure 3** honor these data, but to a lesser extent than between z = 87–90 m where lateral correlation of the soft data is stronger. This comparison is a simple example of how **tsim-s** can be used to further constrain the realizations with soft data as compared to

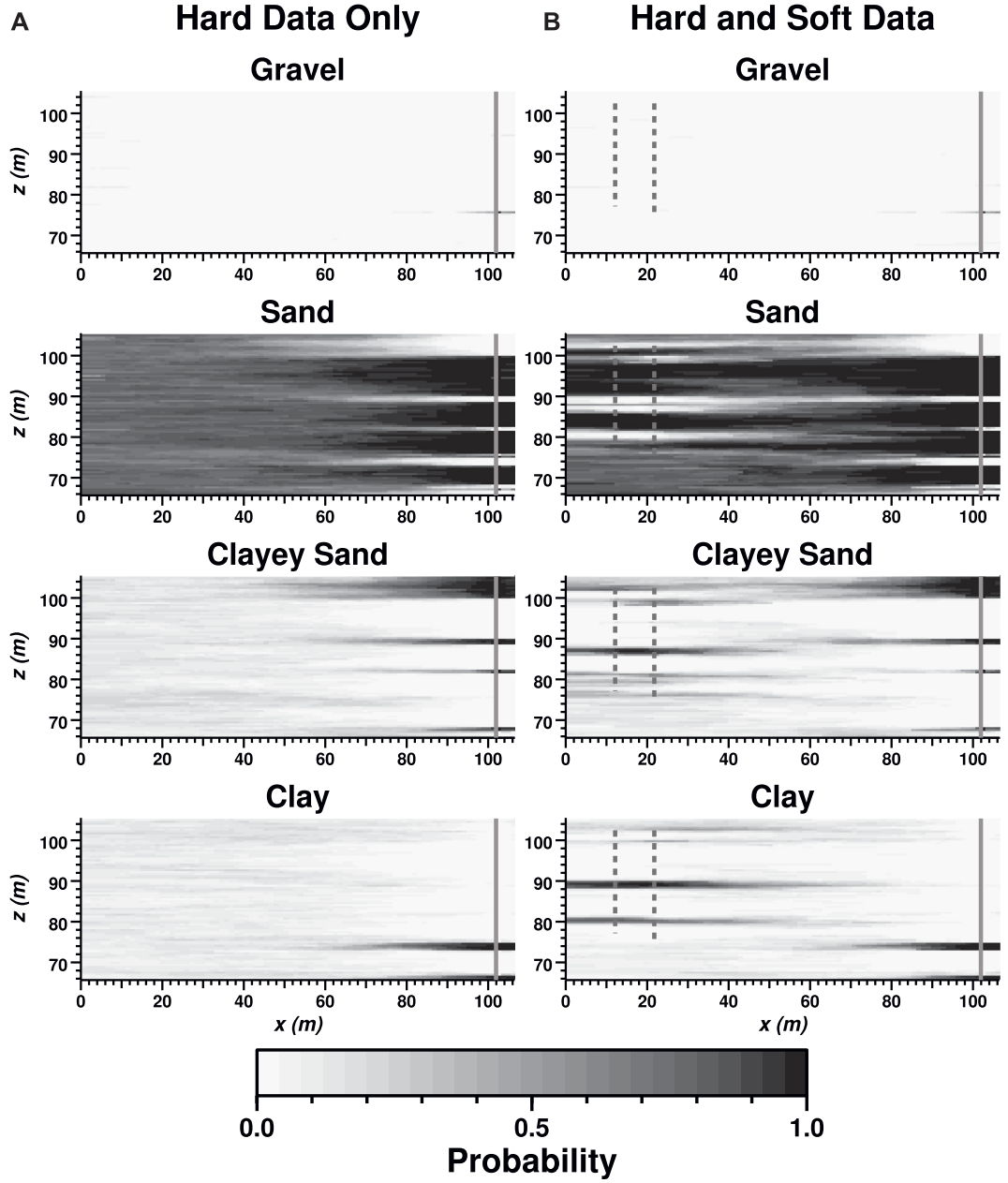


FIGURE 4 | Maps of probability of occurrence for each lithofacies, based on mean indicator value of 100 realizations, for cases with **(A)** hard data only and **(B)** hard and soft data (soft data hardness = 0.8).

conditioning with only hard data, as could otherwise be implemented with **tsim**.

Another way to visualize the impact of the soft data is to average the indicator values over many realizations to produce a "probability map" for each lithofacies. **Figure 4** shows probability maps for each lithofacies derived from 100 realizations with and without soft data conditioning. In case (A) with hard data only, the lithofacies probabilities approach marginal probabilities (proportions) toward the left portion of the realizations. In case (B) with hard and soft data, the soft data further constrain the probability structure toward the left side of the realizations. However, less-refined "gray areas" remain between the borehole data because of the limited lateral correlation of the lithofacies units. Overall, the soft conditioning at 0.80 hardness imparts a strong influence on the realizations.

Through the hardness parameter, the degree of influence by the soft data can be controlled as needed in application of **tsim-s**. **Figure 5** shows probability maps for the *sand*, *clayey sand*, and *clay* lithofacies where hardness values for the soft data are reduced to 0.5 and 0.2. Reducing the hardness produces less contrast or

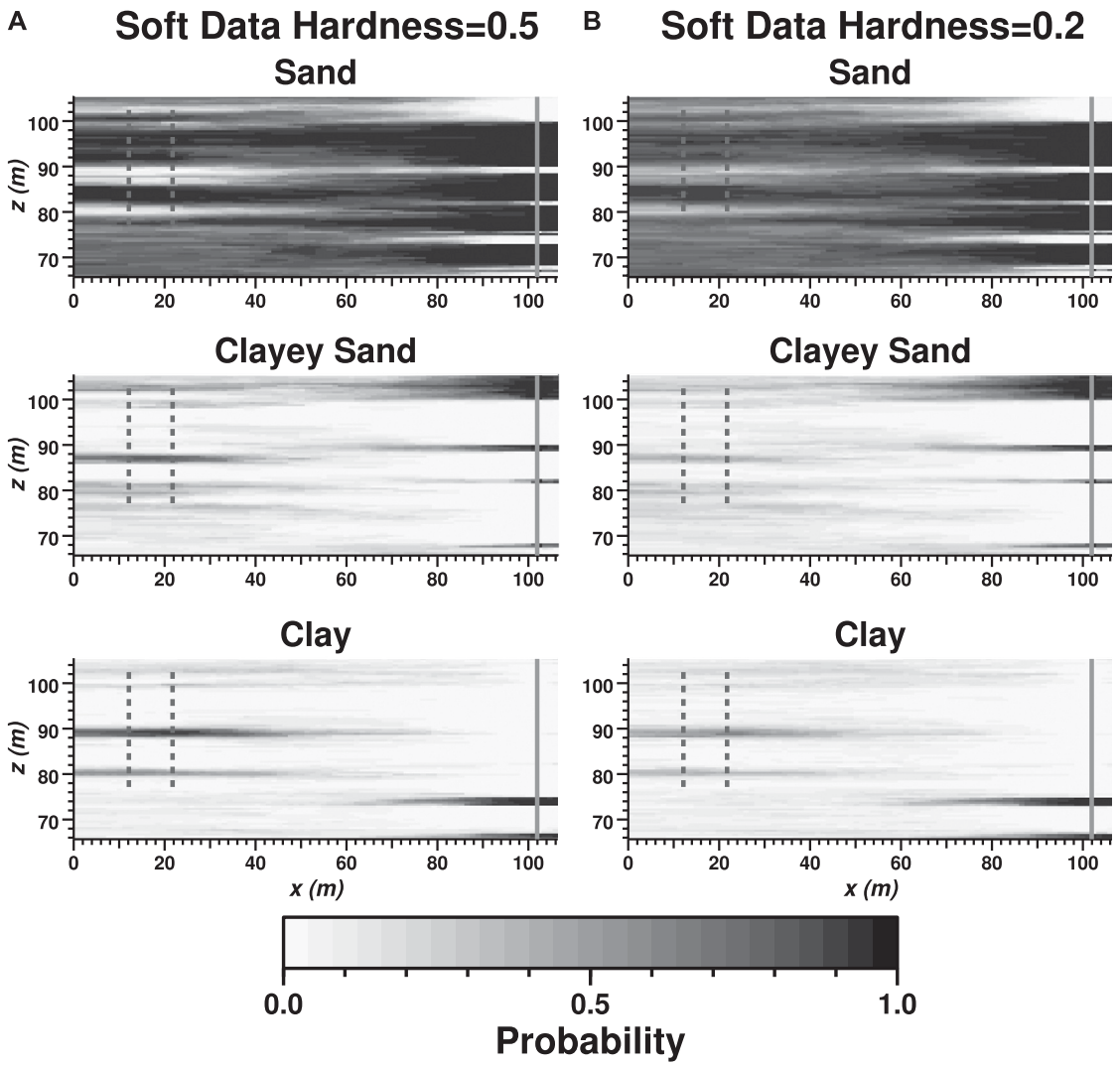


FIGURE 5 | Maps of probability of occurrence for sand, clayey sand, and clay lithofacies, based on mean indicator value of 100 realizations, for cases of hard and soft data with **(A)** soft data hardness = 0.5, and **(B)** soft data hardness = 0.2.

more gray areas in the probability map, which reflects the increased uncertainty of the soft data.

3.2.2 Use of Realizations as Soft Conditioning

There is increasing interest in use of stochastic inversion approaches to modify heterogeneity structures within geostatistical realizations to be more consistent with geophysical or hydraulic testing data (Aines et al., 2002; Harp et al., 2008; Wainwright et al., 2014; Berg and Illman, 2015; Wang et al., 2017). To implement these approaches, there can be a need to modify an initial heterogeneity structure in a controlled or incremental manner.

Another application of the soft data capability in **tsim-s** is to use all or part of a realization as soft conditioning to exert control on the modification of heterogeneity structures from one realization to the next. One realization (or any available field of categorical values) can be used as prior information for conditioning of a new realization, which makes several new capabilities available in **tsim-s**:

- The degree of correlation between a series of realizations or can be controlled.
- The degree of variation from one realization to the next can be controlled at different locations within each realization.
- By exerting control on the difference between one realization and the next, Monte Carlo Markov chain algorithms can be implemented as a Bayesian inverse approach to optimization of local heterogeneity structure (Aines et al., 2002; Wainwright et al., 2014; Wang et al., 2017).

In practice, we refer to a realization used for soft conditioning as the "prior realization" and a subsequent realization that is produced as the "posterior realization." Use of prior realizations

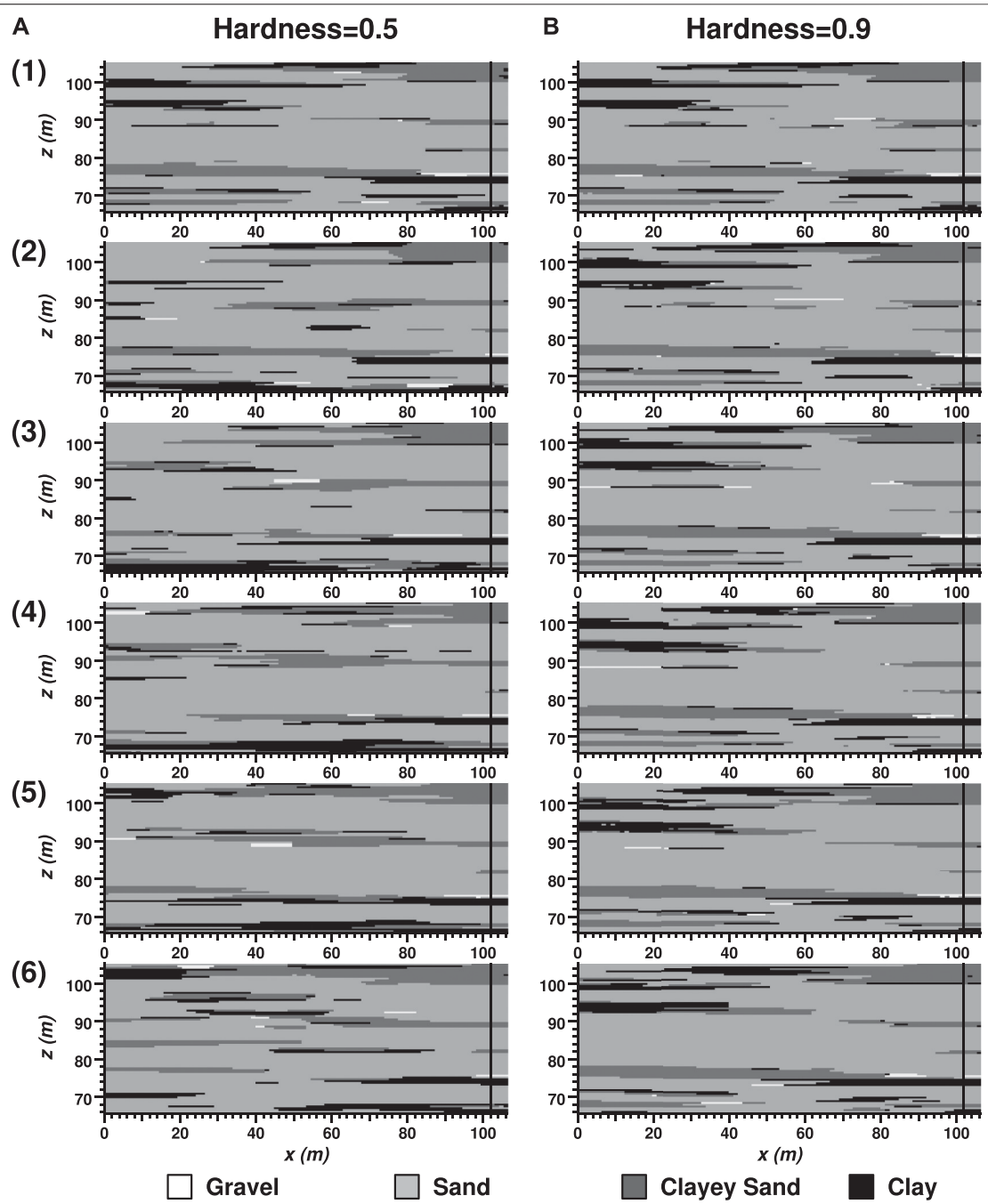


FIGURE 6 | Six successive realizations where the previous realization serves as soft conditioning, with hardness set for case **(A)** at 0.5 and case **(B)** at 0.9.

as soft conditioning in the SIS step is implemented by adding only one additional soft datum to each cokriging estimate **Eq. 6**; that additional soft datum consists of the indicator values from the grid cell of the prior realization corresponding to the cokriging estimation location for the posterior realization. More soft data can be used, but one prior realization soft datum at the cokriging estimation location itself provides sufficient conditioning for generating correlated realizations without adding much more computational burden. The degree of correlation between the prior and posterior realizations is controlled by setting the hardness values, which may vary with location.

Figure 6 shows a sequence of six realizations where each posterior realization is soft-conditioned to the prior realization for cases of (A) hardness = 0.5 and (B) hardness = 0.9. Because the degree of hardness controls the degree of similarity (or rate of change) between one realization and the next, the realizations in case (A) are less similar (or more different) from one realization

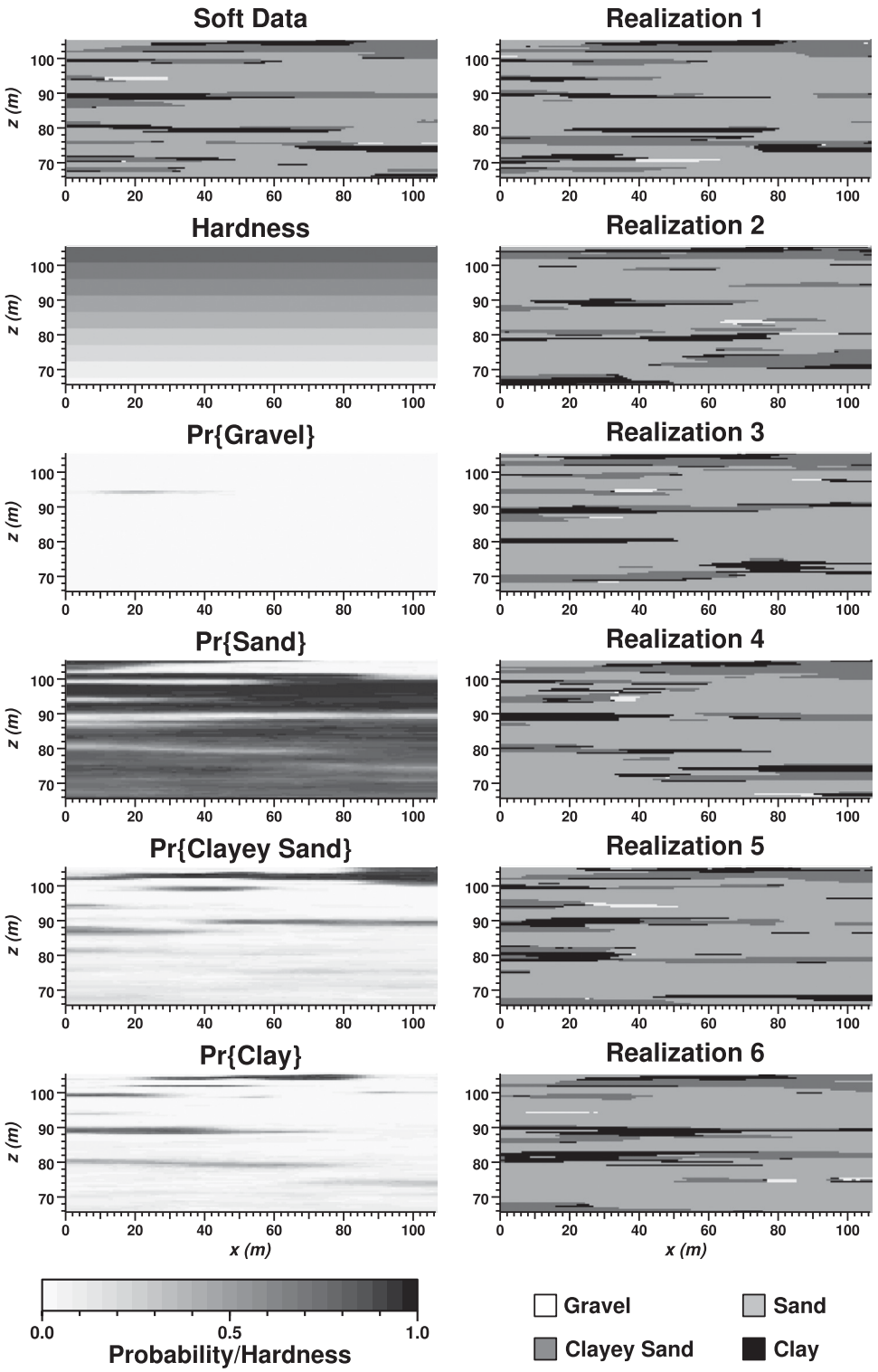


FIGURE 7 | Example using a realization as soft data **(upper left)** with hardness varying with depth **(left, second from top)**. Probability maps for gravel, sand, clayey sand, and clay facies **(lower left)** are derived from 100 realizations. The first six soft-conditioned realizations are shown at right.

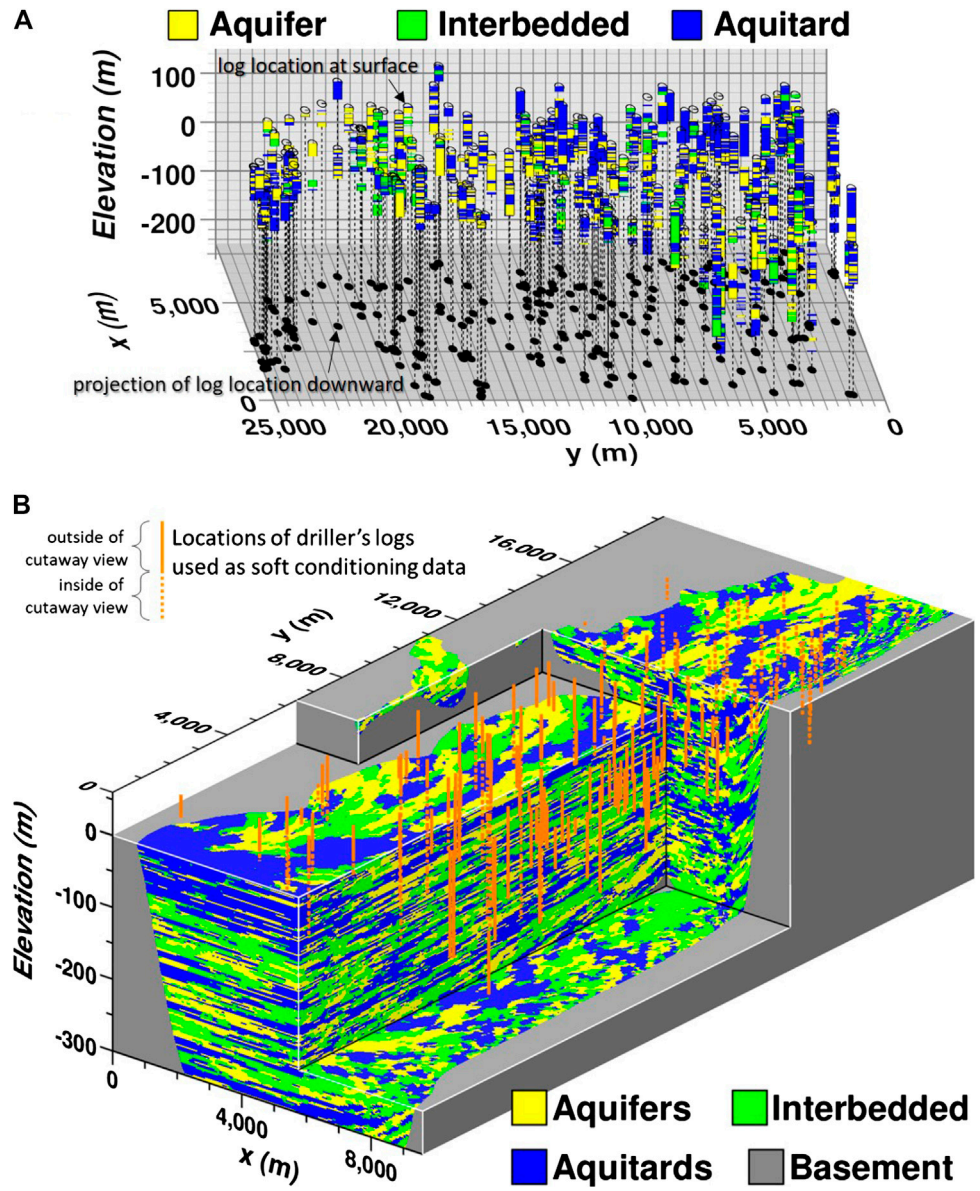


FIGURE 8 | (A) Hydrofacies interpretations of driller logs and **(B)** example 3-D simulation of aquifer system heterogeneity using **tsim-s** showing locations of conditioning data and variable dip and flow directions of anisotropy. Modified from Carle et al. (2006).

to the next. Comparing the left side (away from the hard conditioning) of realizations 1 and 6, similarities have largely disappeared for case (A) but remain for case (B). Thus, the higher rate of change (lesser hardness) shortens the memory of heterogeneity patterns within a sequence of realizations soft-conditioned by prior realizations.

Figure 7 shows an example using the same realization as soft data (upper left) with hardness varying with depth from 0.8 at top to 0.0 at bottom. The probability maps for *gravel*, *sand*, *clayey sand*, and *clay* facies, derived from 100 realizations generated by **tsim-s**, show increasing uncertainty with depth. In particular, the location of individual *clayey sand* and *clay* lenses in the soft data become less distinct with depth in the probability maps. Likewise, the sand probabilities become less distinct with depth and approach the proportion of 0.735. The influence of a single *gravel* lens in the soft data is evident in the *gravel* probability map. This example illustrates how 2or 3-D geophysical images or geological interpretations of categorical data might be used to soft-condition geostatistical realizations with consideration of variable spatial resolution such as decreasing resolution with increasing depth.

3.3 Application to Large-Scale Simulation

3.3.1 Computational Aspects

Like **tsim**, **tsim-s** is readily extended to large-scale three-dimensional (3-D) applications. From a computational standpoint, the main difference between running **tsim** and

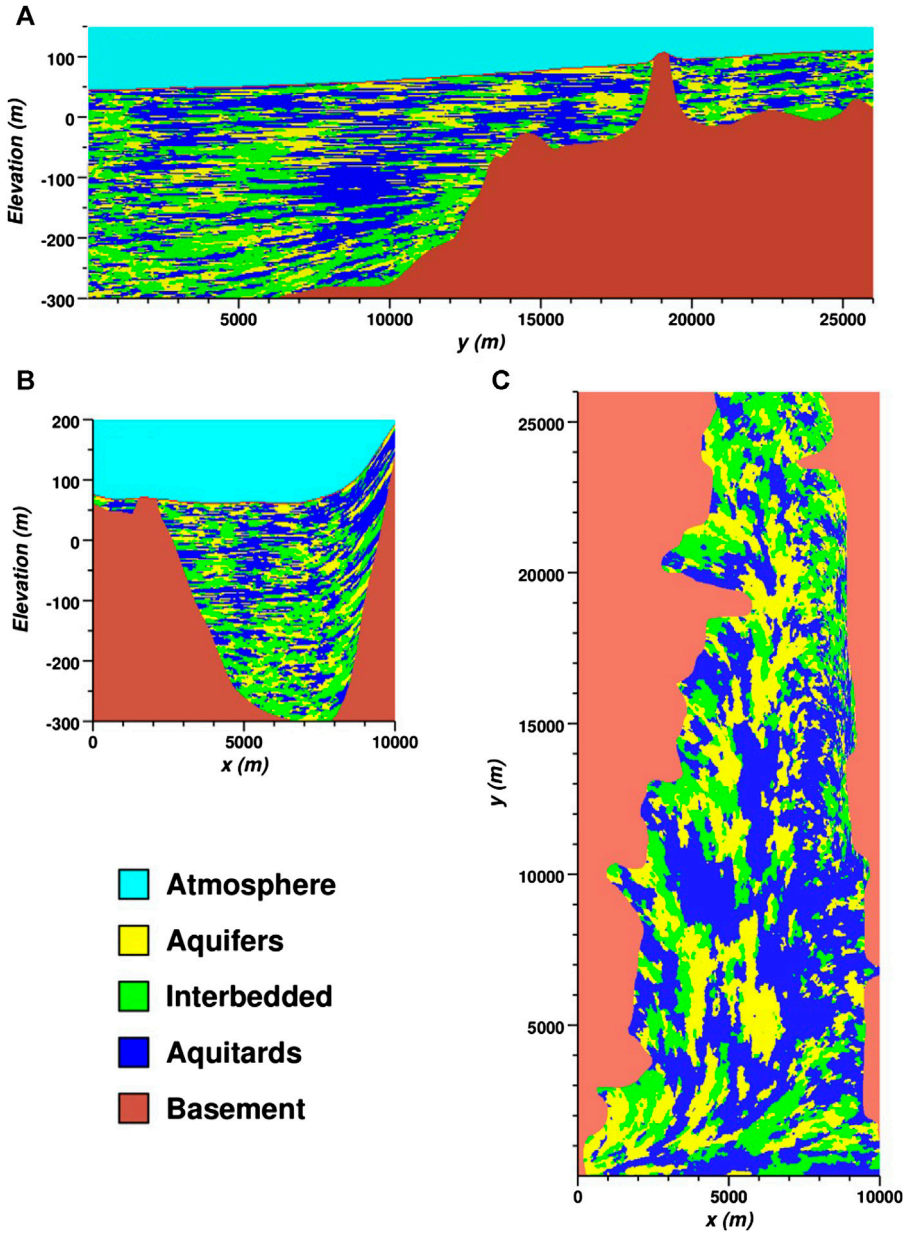


FIGURE 9 | Cross-section through the 3-D simulation shown in **Figure 8** throught **(A)** longitudinal **(B)** lateral, and **(C)** horizontal planes.

tsim-s will be in memory use. For very large grid cell counts (tens of millions or more), the memory use in **tsim** is mostly taken up by integer arrays storing the grid cell category and anisotropy direction information (if used). In the late 1990s when memory was quite restricted compared to the year 2020, a 45-million cell 3-D realization was produced on a computer with 48 megabytes (not gigabytes!) of memory by modifying **tsim** to use a 1-byte integer format for the grid cell array and implementing an analytical function to define variable stratigraphic directions (Carle, 1996; Carle et al., 1998; Carle, 1999; Tompson et al., 1999). The open-source nature of the **T-ProGS** package of codes enables the user to make similar modifications to conserve memory. Depending on the application, **tsim-s** will require a factor of as much as ten or more times the memory use as **tsim**, but computational time **tsim-s** will not be significantly higher because the corresponding arrays for the cokriging equations and quenching objective function are of the same dimensions under similar model parameter settings. Considering that computer memory and computational speed are orders of magnitude higher today and into the future as compared to the late 1990s, we do not anticipate large-scale simulation of categorical heterogeneity with **tsim-s** to be highly constrained by the current or future computational technology.

3.3.2 Llagas Basin Example

As discussed in the introduction, the use of driller's logs in categorical geostatistical simulation is of interest in hydrogeology, yet presents some uncertainty in how to treat driller's logs as conditioning data. As a demonstration of the use of driller's logs and **tsim-s** to large-scale simulation, **Figure 8A** shows an interpreted version of a driller's log dataset, and **Figure 8B** shows a 3-D simulation of hydrostratigraphic architecture for the Llagas groundwater subbasin south of San Jose, California (Carle et al., 2004; Carle et al., 2006).

In this application, the driller's log data were categorized into aquifer, interbedded, and aquitard hydrofacies, where the interbedded category represents relatively thin interlayers of aquifer and aquitard materials. The addition of the interbedded hydrofacies addresses unresolvable fine-scale heterogeneity of coarse- and fine-grained textural classifications and adds flexibility to development of the Markov chain model. If only two hydrofacies, aquifer and aquitard, had been distinguished, the 3-D Markov chain model would consist of only four parameters - proportion and mean length in the three principal depositional directions for one of the hydrofacies; the remaining transition probabilities are all determined by probability law (Carle and Fogg, 1996). A two-category characterization of heterogeneity presents no distinct advantage in the Markov chain spatial variability modeling framework; the spatial variability model is equivalent to a two-category indicator variogram-based approach modeled by an exponential variogram. With three categories, the number of parameters for the 3-D Markov chain is raised to 14, allowing for development of more complexity in the heterogeneity structure including asymmetry such as fining-upward and outward tendencies (Miall, 1973; Carle and Fogg, 1997; Fogg et al., 1998).

This simulation was generated by **tsim-s** with soft conditioning data divided into three sets of driller's logs to which hardness levels were set to 0.3, 0.7, and 1.0 based on data quality. Two separate **tsim-s** simulations of 162,500,000 and 117,000,000 cells were generated for the upper and lower portions of the final realization of hydrofacies architecture to address differences in the spatial structure of deeper and shallower alluvial hydrofacies evident in the driller's logs. Carle et al. (2006) provides further detail on the process of selection of the hardness parameter in the practical hydrogeological situation of using driller's logs for conditioning data. This example confirms that large-scale 3-D stochastic simulation using **tsim-s** is feasible. The stochastic analysis was further applied to investigate permeability heterogeneity effects on nitrate transport from agricultural sources toward municipal wells (Carle et al., 2004, Carle et al., 2006).

4 DISCUSSION

The discussion focuses on capabilities and limitations of **t-sim** and **tsim-s** with attention to the current literature on comparison and evaluation of geostatistical methods for subsurface characterization of categorical variables.

4.1 Curvilinear Features

Both **tsim** and **tsim-s** have the capability to produce curvilinear features by specifying azimuthal and dip angles local to each grid cell in an *a priori* manner (Carle, 1999; Carle, 2007) that can be deterministic (Tompson et al., 1999; Carle et al., 2006) or stochastic (Carle et al., 1998). These angles may be derived or inferred from prior geological knowledge or geologically reasonable interpretation of the depositional or stratigraphic architecture, surface mapping of the deep soil horizons, interpretation of seismic or surface geophysical data, or stochastic modeling (e.g., by modeling variation in the major axis of deposition due to meandering by a gaussian random field). Local anisotropy directions are implemented in **tsim** and **tsim-s** in a manner similar to "local anistropy kriging" (te Stroet and Snepvangers, 2005). This is not a coordinate transformation approach, as applied to the variogram-based **isim3D** (Gomez-Hernandez and Srivastava, 1990).

The Llagas basin example application of **tsim-s** uses the local anisotropy direction option to impart variable angles of dip and principle direction of deposition into the geostatistical realizations. **Figure 9** shows vertical and horizontal slices through the 3-D Llagas basin realization to better reveal the nature of the curvilinear features. These include variable dip angles evident in longitudinal and transverse-plane cross-sections (A) and (B) and variable direction in the major axis of deposition in the horizontal-plane cross-section (C).

The simulated aquifer/interbedded facies architecture for the Llagas basin example does not show pronounced sinuousity, which is consistent with how channel belt deposits were conceptualized in this alluvial depositional setting by the California Department of Water Resources (1981). This is in contrast to the "true" or "training image" concept of continuous "channels" worming their way through homogeneous low-pemeability media, a common argument posed for replacing bivariate statistical methods with multi-point statistical (**mps**) methods (Strebelle, 2000; Caers, 2001; Krishnan and Journel, 2003; Feyen and Caers, 2006; Ronayne et al., 2008; Li et al., 2014a; Li et al., 2015; Mariethoz and Caers, 2015; Zovi et al., 2017; Ramgraber et al., 2020). Such surficially-based conceptual models gloss over fundamental geologic concepts showing how depositional processes produce amalgamations of channel and adjacent sediments that are broader, less sinuous, and more variable in lateral extent as compared to an active fluvial channel viewed on the Earth's surface (Galloway and Hobday, 1996; Miall, 2013). It is a well-known fact in interpretation of borehole data that sedimentary features on the surface are not necessarily preserved in the subsurface (Smith, 2002).

4.2 Methods Comparison

The **T-ProGS** software package has been used somewhat frequently for comparison of methods for geostatistical simulation of categorical variables (Carle, 1996; Carle, 2000; Lee et al., 2007; Yong et al., 2009; Bianchi et al., 2011; dell'Arciprete et al., 2012; Ranjineh Khojasteh, 2013; Kessler et al., 2013; Guastaldi et al., 2014; Serrano et al., 2014; Hoyer et al., 2015; Damico et al., 2018). Bianchi et al. (2011) and Ranjineh Khojasteh (2013) present rigorous comparisons of

tsim and **sisim**, showing that **sisim** does not honor the 3-D model of spatial variability, which was one of the original motivations for developing **T-ProGS** (Carle, 1996; Carle, 1999; Carle, 2000).

We further discuss methods comparison below because **tsim-s** carries forward several capabilities of **tsim** that appear not to be recognized in the methods comparison literature. Flawed methods comparison causes a trickle-down effect of selective references that propogate misleading appraisals of the capabilities of the available methods for stochastic simulation of categorical variables.

In a methods comparsion of **sisim**, **tsim**, and **mps**, dell'Arciprete et al. (2012) did not consider variable anistropy directions in either conceptualization or parameterization of their analysis of spatial variablity and applications of **tsim** despite obvious dipping structures in their data. They chose not to apply Markov chain modeling concepts in the structural framework of a stratigraphic coordinate system relevant to the hydrofacies of a sedimentary depositional system (Carle et al., 1998; Tompson et al., 1999; Carle et al., 2006). The example of a **tsim** realization shown in **Figure 4** of dell'Arciprete et al. (2012) displays completely random-looking and geologically implausible spatial structuring inconsistent with the data and geologic setting.

An example of a trickle-down effect is how dell'Arcipreti et al. (2012) becomes a main reference in He et al. (2017) for promoting application of **mps**, then (He et al., 2017) becomes a main reference in Langousis et al. (2018) for criticizing the 3-D Markov chain modeling approach of Carle and Fogg (1997). To "reveal" limitations of Markov chain models, Langousis et al. (2018) execute a "simple test-case" of a 2-D dipping layer with a coordinate system anchored in the vertical and horizontal directions of their statistical analysis. Neither dell'Arciprete et al. (2012) nor Langousis et al. (2018) appeared to grasp the concept that bivariate geostatistical analysis and simulation should be applied in a geologically-based coordinate system, as demonstrated for at least 30 years (e.g., Gomez-Hernandez and Srivastava, 1990). The geological realism of the **mps** application by He et al. (2017) stands wide open to geological criticism too, with a **mps**-generated "real world buried valley system" showing unrealistic topography and isolated occurrences of the valley fill buried beneath pre-valley-fill strata. Geostatistical analysis should recognize and make adjustments to account for geological slopes, directions, and depositional ordering and not be strictly anchored in a rectilinear cooordinate system pertinent only to data locations.

Another pitfall of methods comparison is use of oversimplified example applications. In Kessler et al. (2013), **tsim** was compared to **mps** in application to stochastic simulation of sand lenses within clayey till. This application of **tsim** adhered to an oversimplified two-category Markov chain to model "a complex type of heterogeneity" exhibiting a bi-modal distribution in the size of sand lenses in 2-D training and reference images. A simple way to address this facies size distribution complexity would be to define two types of sand lenses (perhaps "small" and "large") and model a three-category system. Instead of adhering to a textural basis for defining categories (e.g., "sand" and "clay"), a more geological approach is to interpret hydrofacies appropriate to the geometric framework of the depositional system (Fogg et al., 1998).

The impact of methods comparison studies can get more convoluted by selective referencing of the literature, particularly on the topic of curvilinear features. For example, in promoting **mps**, Barfod et al. (2018) dissmiss applicability of **tsim** by stating that "...**T-ProGS** also has difficulties in reconstructing curvilinear geological features" without any reference to previous work in which the variable anisotropy direction capability of **tsim** was implemented. Barfod et al. (2018) refer to the oversimplified two-category sand-clay Markov chain analysis of Kessler et al. (2013) as "revealing a sub-optimal pattern reproduction, in comparison to other simulation tools such as multiple-point statistics..." Linde et al. (2015) in reviewing "variogram based models" state that "transition probability techniques such as **T-ProGS** ... cannot properly produce curvilinear features..." This is after referencing only dell'Arciprete et al. (2012), Falivene et al. (2007), Lee et al. (2007), and Reffsgaard et al., 2014) as **T-ProGS** applications, all four of which did not employ variable anisotropy direction capability in **tsim**. The misconception that all bivariate statistical methods for stochastic simulation of categorical variables lack the ability to produce curvilinear features appears to derive from selective referencing and comparison of **mps** to only variogram-based methods (Strebelle, 2002; Caers, 2001; Krishnan and Journel, 2003; Feyen and Caers, 2006; Linde et al., 2015; Barfod et al., 2018) irrespective of previous hydrogeologic studies (Carle et al., 1998; Tompson et al., 1999; Carle et al., 2006; Green et al., 2010; Engdahl et al., 2012) and the **T-ProGS** user manual (Carle, 1999; Carle, 2007).

If **tsim** or **tsim-s** is producing spatial structure that falls well short of the intended model for represention of the spatial heterogeneity, the conceptualization, implementation, or utilization of the capabilities of **tsim** or **tsim-s** may be at fault. Methods comparison studies in geostatistics should fully investigate the capabilities of each method including both the geological and statistical conceptual underpinnings before making sweeping judgments. The introduction of this paper provides references to varied applications of **T-ProGS** that should be useful for methods comparison and capabilities assessment.

4.3 Method Limitations

All models have limitations. The **T-ProGS** package was originally conceived in 1996 (Carle, 1996; Carle, 1997; Carle and Fogg, 1996; Carle and Fogg, 1997) to improve or add to the capabilities of the then state-of-the art variogram-based geostatistical methods of the Geostatistical Software Library (Deutsch and Journel, 1992) and subsequently analyze pumping test data in a highly heterogeneous alluvial aquifer system (Carle, 1996; Lee et al., 2007). It is easy to argue that natural systems are geometrically inconsistent with certain geostatistical representations or exhibit far more complexity than any bivariate geostatistical model could ever characterize. Different practitioners will be more comfortable with different methods or levels of complexity in the statistical or bio/chemo/hydrogeological components of their models. There will always be an open question as to what levels of complexity are necessary to sufficiently analyze the subsurface processes of interest.

The stationarity assumption in regard to the spatial continuity model and category proportions can be limiting. Early applications of **tsim** recognized this issue and incorporated "nonstationary" qualities into the geostatistics through data conditioning, geologically-based zones based on stratigraphic analysis or sedimentary environments, and spatially variable angles of anistropy (Carle, 1996; Carle, 2000; Carle et al., 1998; Tompson et al., 1999; Weissmann and Fogg, 1999). There continue to be applications of **tsim** that recognize and address nonstationarity (Weissmann et al., 2004; Traum et al., 2014; Weissmann et al., 2015; Meirovitz et al., 2017; Zhu et al., 2016a; Zhu et al., 2016b; Zhang et al., 2018; Liao et al., 2020; Maples et al., 2020). As discussed previously for application of **tsim-s**, Carle et al. (2006) employed two-zones of differing spatial statistics and apply variable anisotropy directions in their application of **tsim-s**. To address geological realism, adding a modicum of geological insight can be more effective than adding more statistical complexity.

The model of spatial variability is another limitation to any geostatistical approach. A methods limitations analysis by Langousis et al. (2018) criticizes the interpretive framework of transition probability-based Markov chain model development as "based on unverified/untested simplifying assumptions" and "ad-hoc manipulations." Langousis et al. (2018) further contend that.

> . . .stochastic modeling of actual geologies using the [T-ProGS] approach of Carle and Fogg (1997), is characterized by simplifying assumptions and theoretical limitations, with the simulated random fields exhibiting statistical structures that strongly depend on the problem under consideration and the modeling assumptions made, leading to increased epistemic uncertainties in the obtained results.

We offer some perspectives on assessing the limitations of **T-ProGS** that apply to **tsim-s** as well:

- Use of geological concepts in categorical geostatistical simulation may involve subjectivity, which can be viewed by some as either a strength or a limitation to reducing uncertainty.
- The implication that injection of subjective geologic interpretation increases epistemic uncertainties in stochastic modeling would appear to expose a lack of understanding of subsurface geology and it's role in modeling the subsurface.
- As referenced in the introduction of this paper, many applications in hydrogeology and related fields have found the Markov chain modeling framework to be useful to characterization of bivariate spatial statistical cross-relationships (i.e., juxtapositional tendencies), which can be related in an interpretive manner to geological concepts such as Walther's Law in the stratigraphic context of sedimentary depositional environments (Leeder, 1982; Doveton, 1994).
- As evident in the **T-ProGS** manual (Carle, 1999; Carle, 2007), the Markov chain is not actually required to run **tsim** (or **tsim-s**). So one could investigate epistemic aspects of other transition probability or indicator covariance models using **tsim**, **tsim-s**, or the various variogram-based simulation methods, if so desired. However, such methods comparison exercises, which can certainly be expanded to all stochastic models, will not prove that a spatial Markov chain is not useful to modeling subsurface hydrofacies heterogeneity when applied in the appropriate geologic context.

In every subsurface heterogeneity modeling project we have encountered, the heterogeneity contains both deterministic and stochastic aspects that should be treated differently. For example, conventional geologic stratigraphic analysis is quite effective for identifying and mapping the major formations, depositional systems, and the bounding unconformities and structural discontinuities (e.g., faults). Accordingly, such features can typically be treated deterministically and then used as the basis for appropriate geologic zoning of the system into quasi-stationary subdomains. The stochastic aspect of subsurface characterization typically lies within those geologically defined subdomains. If one, however, lumps together both the deterministic and stochastic aspects of the heterogeneity and calls on the stochastic geostatistical algorithm to sort out those spatial patterns, one only invites naive mischaracterization of the heterogeneity that produces unnecessary uncertainty and unrealistic results. An effective way to reduce or moderate uncertainty is to recognize and separate out deterministic and stochastic parts of the problem.

4.4 Stochastic Methods Evaluation

Stochastic methods evaluation in hydrogeology should strive to determine appropriate levels of complexity necessary for gaining insight to 3-D subsurface flow and transport processes at scales relevant to measurement diagnostics used in decision-making. An example of a process-oriented methods comparison, Damico et al. (2018) compared dynamics and trapping metrics for 3-D carbon-dioxide plume simulations using subsurface representations of heterogeneity derived from **tsim** and a more rigorous model for representation of complex features in fluvial architecture. They concluded:

> . . .in the context of representing plume dynamics and residual trapping within fluvial deposits, and within the scope of the parameters used here, the simpler geostatistical model of braided fluvial deposits appears to give an adequate representation of the smaller scale heterogeneity. The depositional- and geometric-based benchmark models represented more features of the fluvial architecture, including variability in the dip of cross sets, variability in the geometry and orientation of unit bars, and the occurrence of channel fills. Depending on context, representing those features may be quite important to understanding some multiphase flow processes in aquifers and reservoirs. However, the simpler geostatistical model (**tsim**) is able to capture the important aspects of fluvial architecture within the

context of understanding the general effect of smaller scale heterogeneity on residual trapping of CO_2 in geosequestration reservoirs, within the scope of the parameters used here.

The worthiness of a stochastic methods to subsurface applications does not necessarily depend on statistical rigor or geological detail, it also depends on the perception of relative value, including both benefits and costs, in the application (Ginn, 2004). The relative value of a model is not necessarily in its complexity, given that calibration of more complex models may erode rather than enhance predictive ability (Doherty and Christensen, 2011).

This paper is offering an approach to address uncertainty in data conditioning of categorical geostatistical models. It would be quite straightforward to add more statistical complexity to transition probability or hardness concepts. However, in our nearly 25-year experience with using transition probability-based geostatistics, we find simpler and more interpretetable geologically-based tools to be quite useful in the study of the effects of various scales and types of heterogeneity on subsurface flow and transport processes.

5 CONCLUSIONS

Many Earth science applications would benefit from increased ability to incorporate "soft" (uncertain or indirect) data to further constrain subsurface models of heterogeneity. In categorical geostatistical simulation applications, often abundant soft data on lithology or hydrofacies (e.g., geophysical logs and imaging, geological interpretations, driller's logs, etc.) offer opportunity for imposing increased or relaxed model constraint.

A soft data capability has been incorporated into the categorical geostatistical simulation code **tsim-s**. Soft data for categorical variables are input either as indicator values or prior probabilities, and a "hardness" value accounts for uncertainty in the data. This approach is particularly conducive to soft data that is already categorical, such as texture inferred from driller's logs, hydrofacies interpretations, or electrofacies based on resistivity cutoffs. In generating realizations with **tsim-s**, the impact of uncertainty in the soft data is factored into formulation of both the cokriging and simulated quenching geostatistical simulation steps. The extent to which the realizations honor the soft data is balanced by the values of hardness, the model of spatial variability, and the values of other nearby hard and soft data.

The degree to which soft data reduces variability in simulation outputs can be quantified by mapping facies probabilities derived by averaging indicator values from many realizations. Example applications in this paper using different values and spatial distributions of hardness illustrate how the impact of data uncertainty can be controlled in the stochastic realizations. Such control will be useful for assimilating different data sets of variable resolution and accuracy. The soft conditioning can be arrays of data, including "prior realizations," to incrementally adjust or evolve the spatial heterogeneity structure of the realizations. The ability to manipulate localized heterogeneity structure or rate of change in a sequence of realizations should be useful for flow and transport model calibration, inverse approaches, or sensitivity analysis. The **tsim-s** algorithm is ammenable to large-scale 3-D simulation including curvilinear features.

Overall, the **tsim-s** code more rigorously integrates data uncertainty and prior information into the categorical stochastic simulation algorithm as compared to previous indicator-based geostatistical simulation codes including its direct predecessor **tsim**. However, users and evaluators of bivariate geostatistical models should become familiarized with capabilities, limitations, and varied uses of **T-ProGS** or other geostatistical software packages before applying or evaluating **tsim** or **tsim-s**.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: must be obtained by permission. Requests to access these datasets should be directed to carle1@llnl.gov.

AUTHOR CONTRIBUTIONS

SC developed code, executed applications, and wrote first draft of article. GF aided in theoretical development, outside application, tie-in to other research, and assisted SC in revision of the manuscript.

ACKNOWLEDGMENTS

The authors thank Roger Aines, John Nitao, and Abe Ramirez at the Lawrence Livermore National Laboratory for their input and support toward developing tsim-s, Rolf Aadland and other SRS staff for supplying data and hydrogeological cross-sections, and Yong Zhang at University of California Davis for a preliminary geostatistical analysis of the SRS data. This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344.

REFERENCES

Aadland, R. K., Gellici, J. A., and Thayer, P. A. (1995). "Hydrogeologic framework of west-central South Carolina," in *Water resources division report*. Columbia, US: South Carolina Department of Water Resources, Vol. 5.

Aarts, E., and Korst, J. (1989). *Simulated annealing and Boltzmann machines*. New York, NY: John Wiley and Sons.

Abriola, L. M., Cápiro, N. L., Christ, J. A., Chu, L., Miller, E. L., and Pennell, K. D. (2019). *Final report, project ER-2311*. Alexandria, VA: Strategic Environmental Research and Development Program.

Agterberg, F. P. (1974). *Geomathematics*. Amsterdam, NY: Elsevier Scientific Publishing Company.

Aines, R., Nitao, J., Newmark, R., Carle, S., Ramirez, A., Harris, D., et al. (2002). UCRL-ID-148221. *The stochastic engine initiative: improving prediction of behavior in geologic environments we cannot directly observe*. Livermore, CA: Lawrence Livermore National Laboratory.

Alberto, M. C. (2010). Heterogeneidades geológicas e o gerenciamento de áreas contaminadas em local situado na Interface da Serra do Mar com a Planície Aluvionar do Rio Cubatão (Cubatão/SP). Tese (doutorado). Brazil: Universidade Estadual Paulista, Instituto de Geocióncias e Cióncias Exatas.

Amooie, M. A., Soltanian, M. R., Xiong, F., Dai, Z., and Moortgat, J. (2017). Mixing and spreading of multiphase fluids in heterogeneous bimodal porous media. *Geomech. Geophys. Geo-energ. Geo-resour.* 3 (3), 225–244. doi:10.1007/s40948-017-0060-8

Arihood, L. D. (2009). *Processing, analysis, and general evaluation of well-driller records for estimating hydrogeologic parameters of the glacial sediments in a ground-water flow model of the Lake Michigan Basin*. Scientific Investigations Report 2008-5184. Reston, VA: US Geological Survey, 26.

Arshadi, M., De Paolis Kaluza, M. C., Miller, E. L., and Abriola, L. M. (2020). Subsurface source zone characterization and uncertainty quantification using discriminative random fields. *Water Resour. Res.* 56 (3), e2019WR026481. doi:10.1029/2019wr026481

Back, P. E., and Sundberg, J. (2007). *Thermal site descriptive model. A strategy for the model development during site investigations-version 2 (No. SKB-R–07-42)*. Stockholm, SE: Swedish Nuclear Fuel and Waste Management Co., 1402–3091.

Baidariko, E. A., and Pozdniakov, S. P. (2011). Simulation of liquid waste buoyancy in a deep heterogeneous aquifer. *Water Resour.* 38 (7), 972–981. doi:10.1134/s0097807811070037

Bakshevskaia, V. A., and Pozdniakov, S. P. (2016). Simulation of hydraulic heterogeneity and upscaling permeability and dispersivity in sandy-clay formations. *Math. Geosci.* 48 (1), 45–64. doi:10.1007/s11004-015-9590-1

Barfod, A. S., Moller, I., Christiansen, A. V., Hoyer, A. S., Hoffimann, J., Straubhaar, J., et al. (2018). Hydrostratigraphic modeling using multiple-point geostatistics and airborne transient electromagnetic methods. *Hydrol. Earth Syst. Sci.* 22, 3351–3373. doi:10.5194/hess-22-3351-2018

Beisman, J. J., Maxwell, R. M., Navarre-Sitchler, A. K., Steefel, C. I., and Molins, S. (2015). ParCrunchFlow: an efficient, parallel reactive transport simulation tool for physically and chemically heterogeneous saturated subsurface environments. *Comput. Geosci.* 19 (2), 403–422. doi:10.1007/s10596-015-9475-x

Berg, S. J., and Illman, W. A. (2015). Comparison of hydraulic tomography with traditional methods at a highly heterogeneous site. *Groundwater.* 53 (1), 71–89. doi:10.1111/gwat.12159

Berretta, G. P., and Felletti, F. (2007). Previsione della presenza di massi nello scavo di gallerie in depositi glaciali: approccio geostatistico mediante T-ProGS (transition probability geostatistics). *Geoing. Ambient. e Mineraria.* 2, 5–21. doi:10.1016/j.enggeo.2009.06.006

Bianchi, M. (2017). Validity of flowmeter data in heterogeneous alluvial aquifers. *Adv. Water Resour.* 102, 29–44. doi:10.1016/j.advwatres.2017.01.003

Bianchi, M., Kearsey, T., and Kingdon, A. (2015). Integrating deterministic lithostratigraphic models in stochastic realizations of subsurface heterogeneity. Impact on predictions of lithology, hydraulic heads and groundwater fluxes. *J. Hydrol.* 531 (3), 557–573. doi:10.1016/j.jhydrol.2015.10.072

Bianchi, M., and Pedretti, D. (2017). Geological entropy and solute transport in heterogeneous porous media. *Water Resour. Res.* 53 (6), 4691–4708. doi:10.1002/2016wr020195

Bianchi, M., Zhang, L., and Birkholzer, J. T. (2016). Combining multiple lower-fidelity models for emulating complex model responses for CCS environmental risk assessment. *Int. J. Greenh. Gas Control.* 46, 248–258. doi:10.1016/j.ijggc.2016.01.009

Bianchi, M., and Zheng, C. (2016). A lithofacies approach for modeling non-Fickian solute transport in a heterogeneous alluvial aquifer. *Water Resour. Res.* 52 (1), 552–565. doi:10.1002/2015WR018186

Bianchi, M., Zheng, C., Wilson, C., Tick, G. R., Liu, G., and Gorelick, S. M. (2011). Spatial connectivity in a highly heterogeneous aquifer: from cores to preferential flow paths. *Water Resour. Res.* 47 (5), W05524. doi:10.1029/2009WR008966

Blessent, D. (2013). Stochastic fractured rock facies for groundwater flow modeling. *Dyna.* 80 (182), 88–94.

Blessent, D., Barco, J., Temgoua, A. G. T., and Echeverrri-Ramirez, O. (2017). Coupled surface and subsurface flow modeling of natural hillslopes in the Aburrá Valley (Medellín, Colombia). *Hydrogeol. J.* 25 (2), 331–345. doi:10.1007/s10040-016-1482-z

Blessent, D., Therrien, R., and Lemieux, J. M. (2011). Inverse modeling of hydraulic tests in fractured crystalline rock based on a transition probability geostatistical approach. *Water Resour. Res.* 47 (12), 88-94. doi:10.1029/2011wr011037

Bohling, G. C., and Butler, J. J., Jr. (2010). Inherent limitations of hydraulic tomography. *Groundwater.* 48 (6), 809–824. doi:10.1111/j.1745-6584.2010.00757.x

Bohling, G. C., and Dubois, M. K. (2003). *An integrated application of neural network and Markov chain techniques to the prediction of lithofacies from well logs, Kansas Geological Survey Open-File Report 2003-50*. Lawrence, KS: Kansas Geological Survey, 6.

Boulanger, R. W., Munter, S. K., Krage, C. P., and DeJong, J. T. (2019). Liquefaction evaluation of interbedded soil deposit: Cark Canal in 1999 M7. 5 Kocaeli earthquake. *J. Geotech. Geoenviron. Eng.* 145 (9), 05019007. doi:10.1061/(asce)gt.1943-5606.0002089

Burow, K. R., Jurgens, B. C., Kauffman, L. J., Phillips, S. P., Dalgish, B. A., and Shelton, J. L. (2008). Simulations of ground-water flow and particle pathline analysis in the zone of contribution of a public-supply well in Modesto, eastern San Joaquin Valley, California. *Scientific Investigations Report 2008–5035*. Reston, VA: US Geological Survey, 41.

Burow, K. R., Weissmann, G. S., Miller, R. D., and Placzek, G. (1997). Hydrogeologic facies characterization of an alluvial fan near Fresno, California, using geophysical techniques. *Open-file report*. 97-46. Sacramento. CA: US Geological Survey.

Buscheck, T. A., Mansoor, K., Yang, X., Wainwright, H. M., and Carroll, S. A. (2019). Downhole pressure and chemical monitoring for CO_2 and brine leak detection in aquifers above a CO_2 storage reservoir. *Int. J. Greenh. Gas Con.* 91, 102812. doi:10.1016/j.ijggc.2019.102812

Caers, J., (2001). Geostatistical reservoir modelling using statistical pattern recognition. *J. Petrol. Sci. Eng.* 29 (3-4), 177–188. doi:10.1016/S0920-4105(01)00088-2

California Department of Water Resources (1981). Evaluation of groundwater resources, South San Francisco Bay, volume IV, south Santa clara county area. Sacramento, CA: California Department of Water Resources Bulletin, 143.

Carle, S. F. (1996). *A transition probability-based approach to geostatistical characterization of hydrostratigraphic architecture*. PhD dissertation. Davis (CA): University of California.

Carle, S. F. (1997). Implementation schemes for avoiding artifact discontinuities in simulated annealing. *Math. Geol.* 29 (2), 231–244. doi:10.1007/bf02769630

Carle, S. F. (1999). T-PROGS: transition probability geostatistical software version 2.1. Davis: University fo Caifornia. Available at: http://gmsdocs.aquaveo.com/T-PROGS.pdf

Carle, S. F. (2000). UCRL-JC-141551. *Use of a transition probability/Markov approach to improve geostatistical simulation of facies architecture*. Livermore, CA: Lawerence Livermore National Laboratory.

Carle, S. F. (2003). UCRL-JC-153653. *Integration of soft data into categorical geostatistical simulation*. Livermore, CA: Lawrence Livermore National Laboratory.

Carle, S. F. (2007). UCRL-SM-232880. *T-PROGS: transition probability geostatistical software version 2.1*. Livermore, CA: Lawrence Livermore National Laboratory.

Carle, S. F., Esser, B. K., and Moran, J. E. (2006). High-resolution simulation of basin-scale nitrate transport considering aquifer system heterogeneity. *Geosphere.* 2 (4), 195–209. doi:10.1130/ges00032.1

Carle, S. F., and Fogg, G. E. (1996). Transition probability-based indicator geostatistics. *Math. Geol.* 28(4), 453–476. doi:10.1007/bf02083656

Carle, S. F., and Fogg, G. E. (1997). Modeling spatial variability with one and multidimensional continuous-lag Markov chains. *Math. Geol.* 29 (7), 891–918. doi:10.1023/a:1022303706942

Carle, S. F., Labolle, E. M., Weissmann, G. S., Van Brocklin, D., and Fogg, G. E. (1998). "Geostatistical simulation of hydrofacies architecture, a transition probability/Markov approach," in *Hydrogeologic models of sedimentary aquifers, concepts in hydrogeology and environmental geology No. 1*. Editors G. S. Fraser and J. M. Davis (Tulsa, OK: Society for Sedimentary Geology Special Publication), 147–170.

Carle, S. F., and Ramirez, A. (1999). UCRL-JC-136739. *Integrated subsurface characterization using facies models, geostatistics, and electrical resistance tomography*. Livermore, CA: Lawrence Livermore National Laboratory.

Carle, S. F., Ramiriez, A., Daily, W., Newmark, R., and Tompson, A. (1999). UCRL-JC-132943. *High-perforamance computational and geostatistical experiments for testing the capabilities of 3-D electrical resistance tomography*. Livermore, CA: Lawrence Livermore National Laboratory.

Carle, S. F., Tompson, A. F. B., McNab, W. W., Esser, B. K., Hudson, G. B., Moran, J. E., et al. (2004). "Simulation of nitrate biogeochemistry and reactive transport in a California groundwater basin," in *Developments in water science*. New York, NY: Elsevier, Vol. 55, 903–914.

Carle, S. F., Zavarin, M., and Pawloski, G. A. (2002). UCRL-ID-150200. *Geostatistical analysis of spatial variability of mineral abundance and Kd in Frenchman Flat, NTS, Alluvium*. Livermore, CA: Lawrence Livermore National Laboratory.(Accessed November 1, 2002).

Carroll, S. A., Keating, E., Mansoor, K., Dai, Z., Sun, Y., Trainor-Guitton, W., et al. (2014). Key factors for determining groundwater impacts due to leakage from geologic carbon sequestration reservoirs. *Int. J. Greenh. Gas Con.* 29, 153–168. doi:10.1016/j.ijggc.2014.07.007

Chen, G., Sun, Y., Liu, J., Lu, S., Feng, L., and Chen, X. (2018). The effects of aquifer heterogeneity on the 3D numerical simulation of soil and groundwater contamination at a chlor-alkali site in China. *Environ. Earth Sci.* 77 (24), 797. doi:10.1007/s12665-018-7979-0

Cooper, C. A., Chapman, J. B., Zhang, Y., Hodges, R., and Ye, M. (2010). *Update of tritium transport calculations for the Rulison site: report of activities and results during 2009-2010*. Grand Junction, CO: Desert Research Institute Letter Report prepared for Stoller Corporation and United States Department of Energy, Office of Legacy Management.

Dai, Z., Wolfsberg, A., Lu, Z., and Ritzi, R., Jr. (2007). Representing aquifer architecture in macrodispersivity models with an analytical solution of the transition probability matrix. *Geophys. Res. Lett.* 34 (20), 220-230. doi:10.1029/2007gl031608

Damico, J. R., Ritzi, R. W., Gershenzon, N. I., and Okwen, R. T. (2018). Challenging geostatistical methods to represent heterogeneity in CO_2 reservoirs under residual trapping. *Environ. Eng. Geosci.* 24 (4), 357–373. doi:10.2113/EEG-2116

dell'Arciprete, D., Bersezio, R., Felletti, F., Giudici, M., Comunian, A., and Renard, P. (2012). Comparison of three geostatistical methods for hydrofacies simulation: a test on alluvial sediments. *Hydrogeol. J.* 20 (2), 299–311.

Deng, H., Dai, Z., Wolfsberg, A., Lu, Z., Ye, M., and Reimus, P. (2010). Upscaling of reactive mass transport in fractured rocks with multimodal reactive mineral facies. *Water Resour. Res.* 46 (6), 220-230. doi:10.1029/2009wr008363

Deng, H., Stauffer, P. H., Dai, Z., Jiao, Z., and Surdam, R. C. (2012). Simulation of industrial-scale CO2 storage: multi-scale heterogeneity and its impacts on storage capacity, injectivity and leakage. *Int. J. Greenh. Gas Con.* 10, 397–418. doi:10.1016/j.ijggc.2012.07.003

Deutsch, C. V., and Cockerham, P. W. (1994). Practical considerations in the application of simulated annealing in stochastic simulation. *Math. Geol.* 26 (1), 67–82. doi:10.1007/bf02065876

Deutsch, C. V., and Journel, A. G. (1992). *Geostatistical software library and user's guide*. 1st Edn. New York, NY: Oxford University Press.

Deutsch, C. V., and Journel, A. G. (1998). *Geostatistical software library and user's guide*. 2nd Edn. New York, NY: Oxford University Press.

Deutsch, C. V., and Wen, X. H. (2000). Integrating large-scale soft data by simulated annealing and probability constraints. *Math. Geol.* 32 (1), 49–67. doi:10.1023/a:1007502817679

Doherty, J., and Christensen, S. (2011). Use of paired simple and complex models to reduce predictive bias and quantify uncertainty. *Water Resour. Res.* 47 (12), W12534. doi:10.1029/2011wr010763

Doughty, C., and Pruess, K. (2004). Modeling supercritical carbon dioxide injection in heterogeneous porous media. *Vadose Zone J.* 3 (3), 837–847. doi:10.2113/3.3.837

Doveton, J. H. (1994). "Theory and applications of vertical variability measures from Markov chain analysis." in *Stochastic modeling and geostatisitics computer applications in geology, No. 3*. Editors J. H. Yarus and R. L. Chambers (Tulsa, OK: Am. Assoc. Petroleum Geologists), 379.

Dumedah, G., and Schuurman, N. (2008). Minimizing the effects of inaccurate sediment description in borehole data using rough sets and transition probability. *J. Geogr. Syst.* 10 (3), 291–315. doi:10.1007/s10109-008-0066-4

Elshall, A. S., Tsai, F. T. C., and Hanor, J. S. (2013). Indicator geostatistics for reconstructing Baton Rouge aquifer-fault hydrostratigraphy, Louisiana, USA. *Hydrogeol. J.* 21 (8), 1731–1747. doi:10.1007/s10040-013-1037-5

Engdahl, N. B., Ginn, T. R., and Fogg, G. E. (2012). Non-Fickian dispersion of groundwater age. *Water Resour. Res.* 48 (7), W07508. doi:10.1029/2012wr012251

Engdahl, N. B., Vogler, E. T., and Weissmann, G. S. (2010a). Evaluation of aquifer heterogeneity effects on river flow loss using a transition probability framework. *Water Resour. Res.* 46 (1), W01506. doi:10.1029/2009wr007903

Engdahl, N. B., Weissmann, G. S., and Bonal, N. D. (2010b). An integrated approach to shallow aquifer characterization: combining geophysics and geostatistics. *Comput. Geosci.* 14 (2), 217–229. doi:10.1007/s10596-009-9145-y

Erdal, D., Baroni, G., Sánchez-León, E., and Cirpka, O. A. (2019). The value of simplified models for spin up of complex models with an application to subsurface hydrology. *Comput. Geosci.* 126, 62–72. doi:10.1016/j.cageo.2019.01.014

Eslinger, O., Hines, A., Howington, S., Ballard, J., Peters, J., White, B., et al. (2007). "Omicron: rapid mesh generation on hpc platforms for the study of near surface phenomena with remote sensing." in *DoD high performance computing modernization program users group conference*. New York, NY: IEEE, 463–468.

Espinet, A., Shoemaker, C., and Doughty, C. (2013). Estimation of plume distribution for carbon sequestration using parameter estimation with limited monitoring data. *Water Resour. Res.* 49 (7), 4442–4464. doi:10.1002/wrcr.20326

Espinet, A. J., and Shoemaker, C. A. (2013). Comparison of optimization algorithms for parameter estimation of multi-phase flow models with application to geological carbon sequestration. *Adv. Water Resour.* 54, 133–148. doi:10.1016/j.advwatres.2013.01.003

Ezzedine, S., Rubin, Y., and Chen, J. (1999). Bayesian method for hydrogeological site characterization using borehole and geophysical survey data: theory and application to the Lawrence Livermore National Laboratory Superfund site. *Water Resour. Res.* 35 (9), 2671–2683. doi:10.1029/1999wr900131

Falivene, O., Cabrera, L., Muñoz, J. A., Arbués, P., Fernández, O., and Sáez, A. (2007). Statistical grid-based facies reconstruction and modelling for sedimentary bodies. Alluvial-palustrine and turbiditic examples. *Geol. Acta.* 5 (3), 199–230.

Falls, W. F., Baum, J. S., Harrelson, L. G., Brown, L. H., and Jerden, J. L., Jr. (1997). Report 97-4245. *Geology and hydrogeology of cretaceous and tertiary strata, and confinement in the vicinity of the U.S. Department of Energy Savannah River site, South Carolina and Georgia. Water Resources Investigations Report 97-4245*. Columbia, SC: US Geological Survey.

Faulkner, B. R., Brooks, J. R., Forshay, K. J., and Cline, S. P. (2012). Hyporheic flow patterns in relation to large river floodplain attributes. *J. Hydrol.* 448, 161–173. doi:10.1016/j.jhydrol.2012.04.039

Felletti, F., and Beretta, G. P. (2009). Expectation of boulder frequency when tunneling in glacial till: a statistical approach based on transition probability. *Eng. Geol.* 108, 43–53. doi:10.1016/j.enggeo.2009.06.006

Feyen, L., and Caers, J. (2006). Quantifying geological uncertainty for flow and transport modeling in multi-modal heterogeneous formation. *Adv. Water Resour.* 29, 912–929. doi:10.1016/j.advwatres.2005.08.002

Fisher, T. R., Dagdelen, K., and Turner, K. A. (2005). "Modeling 3D grade distributions on the Tarkwa paleoplacer gold deposit, Ghana, Africa." in *Geostatistics banff*. Editors O. Leuangthong and C. V. Deutsch (Dordrecht, NL: Springer), 439–448.

Fleckenstein, J. H., and Fogg, G. E. (2008). Efficient upscaling of hydraulic conductivity in heterogeneous alluvial aquifers. *Hydrogeol. J.* 16 (7), 1239. doi:10.1007/s10040-008-0312-3

Fleckenstein, J. H., Niswonger, R. G., and Fogg, G. E. (2006). River-aquifer interactions, geologic heterogeneity, and low-flow management. *Ground Water.* 44 (6), 837–852. doi:10.1111/j.1745-6584.2006.00190.x

Fogg, G. E. (1986). Groundwater flow and sand-body interconnectedness in a thick, multiple-aquifer system. *Water Resour. Res.* 22 (5), 679–694. doi:10.1029/wr022i005p00679

Fogg, G. E., Carle, S. F., and Green, C. (2000). "Connected network paradigm for the alluvial aquifer system," in *Theory, modelling and field investigation in*

hydrogeology: a special volume in honor of shlomo P. Neuman's 60th birthday, GSA special paper 348. Editors D. Zhang and C. L. Winter (Boulder, CO: Geological Society of America).

Fogg, G. E., Noyes, C. D., and Carle, S. F. (1998). Geologically based model of heterogeneous hydraulic conductivity in an alluvial setting. *Hydrogeol. J.* 6 (1), 131–143. doi:10.1007/s100400050139

Frei, S., Fleckenstein, J. H., Kollet, S. J., and Maxwell, R. M. (2009). Patterns and dynamics of river–aquifer exchange with variably-saturated flow using a fully-coupled model. *J. Hydrol.* 375 (3–4), 383–393. doi:10.1016/j.jhydrol.2009.06.038

Frei, S., Lischeid, G., and Fleckenstein, J. H. (2010). Effects of micro-topography on surface–subsurface exchange and runoff generation in a virtual riparian wetland–A modeling study. *Adv. Water Resour.* 33 (11), 1388–1401. doi:10.1016/j.advwatres.2010.07.006

Galloway, W. E., and Hobday, D. K. (1996). "Facies characterization of reservoirs and aquifers," in *Terrigenous clastic depositional systems*. Berlin, Heidelberg: Springer, 426–444.

Ganot, Y., Holtzman, R., Weisbrod, N., Bernstein, A., Siebner, H., Katz, Y., et al. (2018). Managed aquifer recharge with reverse-osmosis desalinated seawater: modeling the spreading in groundwater using stable water isotopes. *Hydrol. Earth Syst. Sci.* 22 (12), 6323-6333. doi:10.5194/hess-22-6323-2018

Ginn, T. R. (2004). On the application of stochastic approaches in hydrogeology. *Stoch. Environ. Res. Risk Assess.* 18 (4), 282–284. doi:10.1007/s00477-004-0199-z

Giraldo, A. G., Álvarez-Villa, O. D., Monsalve, G., Vélez, J. I., and Blessent, D. (2017). Simulación del transporte de contaminantes en un medio subterráneo heterogéneo mediante el rastreo aleatorio de partículas. *Aqua-LAC.* 9 (2), 15–30.

Glaser, R. E., Johannesson, G., Sengupta, S., Kosovic, B., Carle, S., Franz, G. A., et al. (2004). UCRL-TR-202878. *Stochastic engine final report: applying Markov chain Monte Carlo methods with importance sampling to large-scale data-driven simulation*. Livermore, CA: Lawrence Livermore National Laboratory.

Glinskii, M. L., Pozdniakov, S. P., Chertkov, L. G., Zubkov, A. A., Danilov, V. V., Bakshevskaia, V. A., et al. (2014). Regional flow and transport simulation of liquid radioactive waste disposal at the Siberian chemical combine for long-and super-long-term postinjection periods. *Radiochemistry.* 56 (6), 649–656. doi:10.1134/s1066362214060113

Gomez-Hernandez, J. J., and Srivastava, R. M. (1990). ISIM3D: an ANSI-C three-dimensional multiple indicator conditional simulation program. *Comput. Geosci.* 16 (4), 395–440. doi:10.1016/0098-3004(90)90010-Q

Goovaerts, P. (1996). Stochastic simulation of categorical variables using a classification algorithm and simulated annealing. *Math. Geol.* 28 (7), 909–921. doi:10.1007/bf02066008

Goovaerts, P. (1997). "Geostatistics for natural resources evaluation," in *Applied geostatistics series*. New York, NY: Oxford University Press.

Grasmick, J. G., Mooney, M. A., Trainor-Guitton, W. J., and Walton, G. (2020). Global versus local simulation of geotechnical parameters for tunneling projects. *J. Geotech. Geoenviron. Eng.* 146 (7), 04020048. doi:10.1061/(asce)gt.1943-5606.0002262

Green, C. T., Böhlke, J. K., Bekins, B. A., and Phillips, S. P. (2010). Mixing effects on apparent reaction rates and isotope fractionation during denitrification in a heterogeneous aquifer. *Water Resour. Res.* 46 (8), W08525. doi:10.1029/2009wr008903

Green, C. T., Zhang, Y., Jurgens, B. C., Starn, J. J., and Landon, M. K. (2014). Accuracy of travel time distribution (TTD) models as affected by TTD complexity, observation errors, and model and tracer selection. *Water Resour. Res.* 50 (7), 6191–6213. doi:10.1002/2014wr015625

Guastaldi, E., Carloni, A., Pappalardo, G., and Nevini, J. (2014). Geostatistical methods for lithological aquifer characterization and groundwater flow modeling of the Catania plain quaternary aquifer (Italy). *J. Water Resour. Protect.* 6 (4), 272–296. doi:10.4236/jwarp.2014.64032

Guo, Z., Brusseau, M. L., and Fogg, G. E. (2019a). Determining the long-term operational performance of pump and treat and the possibility of closure for a large TCE plume. *J. Hazard Mater.* 365, 796–803. doi:10.1016/j.jhazmat.2018.11.057

Guo, Z., Fogg, G. E., Brusseau, M. L., LaBolle, E. M., and Lopez, J. (2019b). Modeling groundwater contaminant transport in the presence of large heterogeneity: a case study comparing MT3D and RWhet. *Hydrogeol. J.* 27 (4), 1363–1371. doi:10.1007/s10040-019-01938-9

Guo, Z., Fogg, G. E., and Henri, C. V. (2019c). Upscaling of regional scale transport under transient conditions: evaluation of the multirate mass transfer model. *Water Resour. Res.* 55 (7), 5301–5320. doi:10.1029/2019wr024953

Guo, Z., Henri, C. V., Fogg, G. E., Zhang, Y., and Zheng, C. (2020). Adaptive multirate mass transfer (aMMT) model: a new approach to upscale regional-scale transport under transient flow conditions. *Water Resour. Res.* 56 (2), 1-18. doi:10.1029/2019wr026000

Hansen, A. L., Gunderman, D., He, X., and Refsgaard, J. C. (2014). Uncertainty assessment of spatially distributed nitrate reduction potential in groundwater using multiple geological realizations. *J. Hydrol.* 519, 225–237. doi:10.1016/j.jhydrol.2014.07.013

Harbaugh, J. W., and Bonham-Carter, G. F. (1970). *Computer simulation in geology*. New York: Wiley-Interscience.

Harp, D. R., Dai, Z., Wolfsberg, A. V., Vrugt, J. A., Robinson, B. A., and Vesselinov, V. V. (2008). Aquifer structure identification using stochastic inversion. *Geophys. Res. Lett.* 35 (8), L08404. doi:10.1029/2008gl033585

Harp, D. R., and Vesselinov, V. V. (2010). Stochastic inverse method for estimation of geostatistical representation of hydrogeologic stratigraphy using borehole logs and pressure observations. *Stoch. Environ. Res. Risk Assess.* 24 (7), 1023–1042. doi:10.1007/s00477-010-0403-2

Harter, T. (2005). Finite-size scaling analysis of percolation in three-dimensional correlated binary Markov chain random fields. *Phys. Rev.* 72 (2), 026120. doi:10.1103/physreve.72.026120

Haugen, B. D., Broadfoot, S. W., Wakeley, L. D., Roig-Silva, C., Bourne, S. G., and Talbot, C. A. (2011). Defining 3-D geologic architecture and soil variability for sensor simulations. ERDC/GSL GeoTACS TN-11-1. US Army Corps of Engineers.

He, X., Højberg, A. L., Jørgensen, F., and Refsgaard, J. C. (2015). Assessing hydrological model predictive uncertainty using stochastically generated geological models. *Hydrol. Process.* 29 (19), 4293–4311. doi:10.1002/hyp.10488

He, X., Sonnenborg, T. O., Jørgensen, F., and Jensen, K. H. (2017). Modelling a real-world buried valley system with vertical non-stationarity using multiple-point statistics. *Hydrogeol. J.* 25 (2), 359–370. doi:10.1007/s10040-016-1486-8

Hermans, T., and Irving, J. (2017). Facies discrimination with electrical resistivity tomography using a probabilistic methodology: effect of sensitivity and regularisation. *Near Surf. Geophys.* 15 (1), 13–25. doi:10.3997/1873-0604.2016047

Heywood, C. E. (2013). Simulations of groundwater flow, transport, and age in Albuquerque, New Mexico, for a study of transport of anthropogenic and natural contaminants (TANC) to public-supply wells. *Scientific investigations report 2012–5242*. Reston VA: US Geological Survey, 51.

Hovorka, S. D., Doughty, C. A., Knox, P. R., Green, C. T., Pruess, K., and Benson, S. M. (2001). "Evaluation of brine-bearing sands of the Frio formation, upper Texas Gulf coast for geologic sequestration of CO_2," in First National conference on carbon sequestration, Washington, D.C., May 14–17, 2001 (GCCC Digital Publication Series), 13.

Hoyer, A. S., Jorgensen, F., Foged, N., He, X., and Christiansen, A. V. (2015). Three-dimensional geological modelling of AEM resistivity data – a comparison of three methods. *J. Appl. Geophys.* 115, 65–78. doi:10.1016/j.jappgeo.2015.02.005

Hu, B. X., Wu, J. C., Panorska, A. K., Zhang, D. X., and He, C. M. (2003). Stochastic study on groundwater flow and solute transport in a porous medium with multi-scale heterogeneity. *Adv. Water Resour.* 26 (5), 541–560. doi:10.1016/s0309-1708(03)00003-4

Huang, L., Ritzi, R. W., Jr., and Ramanathan, R. (2012). Conservative models: parametric entropy vs. temporal entropy in outcomes. *Groundwater.* 50 (2), 199–206. doi:10.1111/j.1745-6584.2011.00832.x

Isaaks, E. H., and Srivastava, R. M. (1989). *Applied geostatistics*. New York: Oxford University Press.

Izbicki, J. A. (2002). Geologic and hydrologic controls on the movement of water through a thick, heterogeneous unsaturated zone underlying an intermittent stream in the western Mojave Desert, Southern California. *Water Resour. Res.* 38 (3), 2–14. doi:10.1029/2000wr000197

James, S. C. (2004). SAND2004-3440. *An example uncertainty and sensitivity analysis at the Horonobe site for performance assessment calculations*. United States: Sandia National Laboratories.

Janza, M. (2009). Modeliranje heterogenosti vodonosnika Ljubljanskega polja z uporabo Markovih verig in geostatistike. *Geologija*. 3, 233–240.

Jones, N. L., Green, J. I., and Walker, J. R. (2003). "Stochastic inverse modeling for capture zone analysis," in Probabilistic approaches to groundwater modeling Symposium at World Environmental and Water Resources Congress 2003, Philadelphia, PA, June 23–25, 2003.

Jones, N. L., Walker, J. R., and Carle, S. F. (2002). Using transition probability geostatistics with MODFLOW. In: Korvar, K. and Hrkal, Z., eds. ModelCARE 2002, Proceedings of the 4th international conference on calibration and reliability in groundwater modeling, Prague, Czech Republic; 2002 June 17-20: Acta Universitatis Carolinae - Geologica; 2002, Vol. 46, no. 2/3, 359-364.

Kessler, T. C., Comunian, A., Oriani, F., Renard, P., Nilsonn, B., Klint, K. E., et al. (2013). Modeling fine-scale geological heterogeneity-examples of sand lenses in tills. *Groundwater*. 51, 692–705. doi:10.1111/j.1745-6584.2012.01015.x

Kitanidis, P. (2016). *An advanced joint inversion system for CO_2 storage modeling with large date sets for characterization and real-time monitoring-enhancing storage performance and reducing failure risks under uncertainties*. United States: Stanford University.

Knudby, C., Carrera, J., Bumgardner, J. D., and Fogg, G. E. (2006). Binary upscaling–the role of connectivity and a new formula. *Adv. Water Resour.* 29 (4), 590–604. doi:10.1016/j.advwatres.2005.07.002

Koch, J., He, X., Jensen, K. H., and Refsgaard, J. C. (2014). Challenges in conditioning a stochastic geological model of a heterogeneous glacial aquifer to a comprehensive soft data set. *Hydrol. Earth Syst. Sci.* 18 (8), 2907. doi:10.5194/hess-18-2907-2014

Krage, C. P., DeJoun, J. T., and Boulanger, R. W. (2016). "Identification of geological depositional variations using CPT-based conditional probability mapping," in *Geotechnical and geophysical site characterisation 5-lehane*. Editors H. E. Acosta-Martinez and R. Kelley (Sydney, Australia: Australian Geomechanics Society), 1447–1452.

Krishnan, S., and Journel, A. G. (2003). Spatial connectivity: from variograms to multiple-point measures. *Math. Geol.* 35 (8), 915–925. doi:10.1023/b:matg.0000011585.73414.35

Krumbein, W. C., and Dacey, M. F. (1969). Markov chains and embedded Markov chains in geology. *Math. Geol.* 1 (1), 79–96. doi:10.1007/bf02047072

Kwon, M., Jeong, J., Lee, H., Park, J. B., and Park, E. (2017). A comparative study on the measures determining optimal SAGD locations based on geostatistical and multiphysics simulations. *Econ. Environ. Geol.* 50 (3), 225–238. doi:10.9719/EEG.2017.50.3.225

Labolle, E. M., and Fogg, G. E. (2001). Role of molecular diffusion in contaminant migration and recovery in an alluvial aquifer system. *Transport Porous Media.* 42 (1–2), 155–179. doi:10.1023/a:1006772716244

Labolle, E. M., Fogg, G. E., and Eweis, J. B. (2006). Diffusive fractionation of 3H and 3He in groundwater and its impact on groundwater age estimates. *Water Resour. Res.* 42 (7), W02601. doi:10.1029/2005wr004756

Labolle, E. M., Fogg, G. E., Eweis, J. B., Gravner, J., and Leaist, D. G. (2008). Isotopic fractionation by diffusion in groundwater. *Water Resour. Res.* 44 (7), W07405. doi:10.1029/2006wr005264

Langousis, A., Kaleris, V., Kokosi, A., and Mamounakis, G. (2018). Markov based transition probability geostatistics in groundwater applications: assumptions and limitations. *Stoch. Environ. Res. Risk Assess.* 32 (7), 2129–2146. doi:10.1007/s00477-017-1504-y

Larned, S. T. (2012). Phreatic groundwater ecosystems: research frontiers for freshwater ecology. *Freshw. Biol.* 57 (5), 885–906. doi:10.1111/j.1365-2427.2012.02769.x

Lee, J., Kokkinaki, A., and Kitanidis, P. K. (2018). Fast large-scale joint inversion for deep aquifer characterization using pressure and heat tracer measurements. *Transport Porous Media.* 123 (3), 533–543. doi:10.1007/s11242-017-0924-y

Lee, S. Y. (2004). *Heterogeneity and transport: geostatistical modeling, non-Fickian transport, and efficiency of remediation methods*. dissertation. Davis (CA): University of California.

Lee, S. Y., Carle, S. F., and Fogg, G. E. (2007). Geologic heterogeneity and a comparison of two geostatistical models: sequential Gaussian and transition probability-based geostatistical simulation. *Adv. Water Resour.* 30 (9), 1914–1932. doi:10.1016/j.advwatres.2007.03.005

Leeder, M. R. (1982). *Sedimentology*. London, UK: George Allen and Unwin Ltd, 344.

Li, L., Srinivasan, S., Zhou, H., and Gómez-Hernández, J. J. (2014a). Simultaneous estimation of geologic and reservoir state variables within an ensemble-based multiple-point statistic framework. *Math. Geosci.* 46 (5), 597–623. doi:10.1007/s11004-013-9504-z

Li, L., Srinivasan, S., Zhou, H., and Gomez-Hernandez, J. J. (2015). Two-point or multiple-point statistics? A comparison between the ensemble Kalman filtering and the ensemble pattern matching inverse methods. *Adv. Water Resour.* 86, 297–310. doi:10.1016/j.advwatres.2015.05.014

Li, X., Liu, J., Zhang, J., Wang, W., and Xin, W. (2014b). Soil texture distribution simulation and risk assessment using transition probability-based geostatistics. *Int. Agrophys.* 28 (4), 447-457. doi:10.2478/intag-2014-0035

Liao, H. S., Curtis, Z. K., Sampath, P. V., and Li, S. G. (2020). Simulation of flow in a complex aquifer system subjected to long-term well network Growth. *Groundwater*. 58 (2), 301–322. doi:10.1111/gwat.12918

Linde, N., Renard, P., Mukerji, T., and Caers, J. (2015). Geological realisim in hydrogeological and geophysical inverse modeling: a review. *Adv. Water Resour.* 86, 86–101. doi:10.1016/j.advwatres.2015.09.019

Liu, L., Yi, L., and Cheng, X. (2014). Stochastic simulation of shallow aquifer heterogeneity and it's using in contaminant transport modeling in Tianjin plains. *J. Water Resour. Ocean Sci.* 3 (6), 80–88. doi:10.11648/j.wros.20140306.13

Liu, Y., Wallace, C. D., Zhou, Y., Ershadnia, R., Behzadi, F., Dwivedi, D., et al. (2020). Influence of streambed heterogeneity on hyporheic flow and Sorptive solute transport. *Water*. 12 (6), 1547. doi:10.3390/w12061547

Lu, D., Ye, M., and Curtis, G. P. (2015). Maximum likelihood Bayesian model averaging and its predictive analysis for groundwater reactive transport models. *J. Hydrol.* 529, 1859–1873. doi:10.1016/j.jhydrol.2015.07.029

Lu, S., Molz, F. J., Fogg, G. E., and Castle, J. W. (2002). Combining stochastic facies and fractal models for representing natural heterogeneity. *Hydrogeol. J.* 10 (4), 475–482. doi:10.1007/s10040-002-0212-x

Lu, Z., Zhang, D., and Keating, E. (2001). LA-UR-01-4727. *Applicability of unimodal stochastic approaches in simulating flow in Bimodal heterogeneous formations*. Los Alamos, NM: Los Alamos National Laboratory.

Lu, Z. M., and Zhang, D. X. (2002). On stochastic modeling of flow in multimodal heterogeneous formations. *Water Resour. Res.* 38 (10), 8. doi:10.1029/2001wr001026

Maghrebi, M., Jankovic, I., Weissmann, G. S., Matott, L. S., Allen-King, R. M., and Rabideau, A. J. (2015). Contaminant tailing in highly heterogeneous porous formations: sensitivity on model selection and material properties. *J. Hydrol.* 531, 149–160. doi:10.1016/j.jhydrol.2015.07.015

Maji, R., and Sudicky, E. A. (2008). Influence of mass transfer characteristics for DNAPL source depletion and contaminant flux in a highly characterized glaciofluvial aquifer. *J. Contam. Hydrol.* 102 (1–2), 105–119. doi:10.1016/j.jconhyd.2008.08.005

Maji, R., Sudicky, E. A., Panday, S., and Teutsch, G. (2006). Transition probability/Markov chain analyses of DNAPL source zones and plumes. *Groundwater*. 44 (6), 853–863. doi:10.1111/j.1745-6584.2005.00194.x

Mansoor, K., Carroll, S. A., and Sun, Y. (2014). The role of wellbore remediation on the evolution of groundwater quality from CO_2 and brine leakage. *Energy Procedia*. 63, 4799–4806. doi:10.1016/j.egypro.2014.11.510

Maples, S. R., Fogg, G. E., and Maxwell, R. M. (2019). Modeling managed aquifer recharge processes in a highly heterogeneous, semi-confined aquifer system. *Hydrogeol. J.* 27 (8), 2869–2888. doi:10.1007/s10040-019-02033-9

Maples, S. R., Foglia, L., Fogg, G. E., and Maxwell, R. M. (2020). Sensitivity of hydrologic and geologic parameters on recharge processes in a highly-heterogeneous, semi-confined aquifer system. *Hydrol. Earth Syst. Sci. Discuss.* 27, 1–36. doi:10.5194/hess-24-2437-2020

Mariethoz, G., and Caers, J. (2015). *Multiple-point geostatistics: stochastic modeling with training images*. New Jersey, US: Wiley Blackwell, 364.

Maxwell, R. M., Carle, S. F., and Tompson, A. F. (2008). Contamination, risk, and heterogeneity: on the effectiveness of aquifer remediation. *Environ. Geol.* 54 (8), 1771–1786. doi:10.1007/s00254-007-0955-8

Maxwell, R. M., Carle, S. F., and Tompson, A. F. B. (2000). UCRL-JC-138445. *Risk-Based management of contaminated groundwater: the role of geologic*

heterogeneity, exposure and cancer risk in determining the performance of aquifer remediation. Livermore, CA: Lawrence Livermore National Laboratory.

Maxwell, R. M., Welty, C., and Tompson, A. F. B. (2003). Streamline-based simulation of virus transport resulting from long term artificial recharge in a heterogeneous aquifer. *Adv. Water Resour.* 26 (10), 1075–1096. doi:10.1016/s0309-1708(03)00074-5

McDonald, S., Groncki, J., and Biteman, S. (2005). "Geostatistical evaluation of high frequency hydraulic property and groundwater quality data collected by the waterloo profiler," in *Bringing groundwater quality research to the watershed scale*. Iahs Publication, Vol. 297, 169–175.

McKenna, S. A., and Poeter, E. P. (1995). Field example of data fusion in site characterization. *Water Resour. Res.* 31 (12), 3229–3240. doi:10.1029/95wr02573

Meirovitz, C., M., Fogg, G. E., Weissmann, G. W., Sager, J., Roll, L., and Labolle, E. M. (2017). Non-stationary hydrostratigraphic model of cross-cutting alluvial fans. *Int. J. Hydrol.* 1 (1), 1–10. doi:10.15406/ijh.2017.01.00001

Mi, D., Hao, F., Juo, X., Zheng, F., Shi, Z., and Wu, J. (2016). Comparison of two-dimensional and three-dimensional simulations of DNAPL migration in saturated porous media. *Geol. J. China Univ.* 22 (4), 733–740. doi:10.16108/j.issn1006-7493.2016061

Miall, A. D. (1973). Markov chain analysis applied to an ancient alluvial plain succession. *Sedimentology.* 20 (3), 347–365. doi:10.1111/j.1365-3091.1973.tb01615.x

Miall, A. D. (2013). *Principles of sedimentary basin analysis*. Berlin, Heidelberg: Springer Science and Business Media.

Miller, R. B., Castle, J. W., and Temples, T. J. (2000). Deterministic and stochastic modeling of aquifer stratigraphy, South Carolina. *Ground Water.* 38 (2), 284–295. doi:10.1111/j.1745-6584.2000.tb00339.x

Misut, P. E. (2014). Scientific Investigations Report 2014–5036. *Simulation of zones of contribution to wells at site GM–38, naval weapons industrial reserve plant, Bethpage, New York*. Reston VA: US Geological Survey, 58.

Mukhopadhyay, S., Doughty, C., Bacon, D., Li, J., Wei, L., Yamamoto, H., et al. (2015). The Sim-SEQ project: comparison of selected flow models for the S-3 site. *Transp. Porous Med.* 108 (1), 207–231. doi:10.1007/s11242-014-0361-0

Munter, S. K., Boulanger, R. W., Krage, C. P., and DeJong, J. T. (2017). "Evaluation of liquefaction-induced lateral spreading procedures for interbedded deposits: Cark canal in the 1999 M7. 5 Kocaeli earthquake," in Geotechnical frontiers 2017, Orlando, Florida, March 12–15, 2017, 254–266.

Munter, S. K., Krage, C. P., Boulanger, R. W., DeJong, J. T., and Montgomery, J. (2016). "Potential for liquefaction-induced lateral spreading in interbedded deposits considering spatial variability," in Geotechnical and Structural Engineering Congress 2016, Phoenix, Arizona, February 14–17, 2016, 1484–1494.

Muskus, N., and Falta, R. W. (2018). Semi-analytical method for matrix diffusion in heterogeneous and fractured systems with parent-daughter reactions. *J. Contam. Hydrol.* 218, 94–109. doi:10.1016/j.jconhyd.2018.10.002

Oatfield, W. J., and Czarnecki, J. B. (1989). *Hydrogeologic infernences from drillers' logs and from gravity and resistivity surveys in the Amargosa Desert, southern Nevada*. Open-file report 89-234. Denver, CO: US Geological Survey.

Papapetrou, M., and Theodossiou, N. (2012). "A methodology for the stochastic approach of the geological structure of groundwater aquifers. Application to an aquifer in northern Greece," in International conference on protection and restoration of the environment XI, Thessaloniki, Greece, 316–325.

Paradis, D., Lefebvre, R., Gloaguen, E., and Rivera, A. (2015). Predicting hydrofacies and hydraulic conductivity from direct-push data using a datadriven relevance vector machine approach: motivations, algorithms, and application. *Water Resour. Res.* 51, 481–505. doi:10.1002/2014wr015452

Park, Y. J., Sudicky, E. A., McLaren, R. G., and Sykes, J. F. (2004). Analysis of hydraulic and tracer response tests within moderately fractured rock based on a transition probability geostatistical approach. *Water Resour. Res.* 40 (12), W12404. doi:10.1029/2004wr003188

Pawloski, G. A., A. F. B. Tompson, and S. F. Carle (Editors) (2001). *Evaluation of the hydrologic source term from underground nuclear tests on Pahute Mesa at the Nevada Test Site: the Cheshire test*. Livermore, CA: Lawrence Livermore National Laboratory.

Pedretti, D., and Bianchi, M. (2018). Reproducing tailing in breakthrough curves: are statistical models equally representative and predictive?. *Adv. Water Resour.* 113, 236–248. doi:10.1016/j.advwatres.2018.01.023

Pescimoro, E., Boano, F., Sawyer, A. H., and Soltanian, M. R. (2019). Modeling influence of sediment heterogeneity on nutrient cycling in streambeds. *Water Resour. Res.* 55 (5), 4082–4095. doi:10.1029/2018wr024221

Phillips, S. P., Green, C. T., Burow, K. R., Shelton, J. L., and Rewis, D. L. (2007). *Simulation of multiscale groundwater flow in part of the Northeastern San Joaquin Valley, California*. Scientific investigations report 2007-5009. Reston, VA: US Geological Survey.

Poeter, E. P., and McKenna, S. A. (1995). Reducing uncertainty associated with ground-water flow and transport predictions. *Ground Water.* 33 (6), 899–904. doi:10.1111/j.1745-6584.1995.tb00034.x

Poeter, E. P., and Townsend, P. (1994). Assessment of critical flow path for improved remediation management. *Ground Water.* 32 (3), 439–447. doi:10.1111/j.1745-6584.1994.tb00661.x

Pozdniakov, S. P., Bakshevskaya, V. A., Krohicheva, I. V., Danilov, V. V., and Zubkov, A. A. (2012). The influence of conceptual model of sedimentary formation hydraulic heterogeneity on contaminant transport simulation. *Moscow Univ. Geol. Bull.* 67 (1), 43–51. doi:10.3103/s0145875212010097

Pozdniakov, S. P., Bakshevskaya, V. A., Zubkov, A. A., Danilov, V. V., Rybalchenko, A. I., and Tsang, C. F. (2005). Modeling of waste injection in heterogeneous sandy clay formations. *Dev. Water Sci.* 52, 203–219. doi:10.1016/s0167-5648(05)52017-3

Proce, C. J., Ritzi, R. W., Dominic, D. F., and Dai, Z. (2004). Modeling multiscale heterogeneity and aquifer interconnectivity. *Ground Water.* 42 (5), 658–670. doi:10.1111/j.1745-6584.2004.tb02720.x

Pryshlak, T. T., Sawyer, A. H., Stonedahl, S. H., and Soltanian, M. R. (2015). Multiscale hyporheic exchange through strongly heterogeneous sediments. *Water Resour. Res.* 51 (11), 9127–9140. doi:10.1002/2015wr017293

Purkis, S., Vlaswinkel, B., and Gracias, N. (2012). Vertical-to-lateral transitions among cretaceous carbonate facies -a means to 3-D framework construction via Markov analysis. *J. Sediment. Res.* 82, 232–243. doi:10.2110/jsr.2012.23

Ramgraber, M., Camporese, M., Renard, P., Salandin, P., and Schirmer, M. (2020). Quasi-online groundwater model optimization under constraints of geological consistency based on iterative importance sampling. *Water Resour. Res.* 56 (6), e2019WR026777. doi:10.1029/2019wr026777

Ramirez, A., Friedman, J. F., Dyer, K., and Aines, R. (2006). "Site characterization using joint reconstructions of disparate data types," in CO2 Site Characterization Symposium, Berkeley, CA, March 20–22, 2006, 53–55.

Ramirez, A., Mcnab, W., Carle, S., Hao, Y., White, D., and Johnson, J. (2010). LLNL-TR-464614. Progress report, December 2010: improved site characterization and storage prediction through stochastic inversion of time-lapse geophysical and geochemical data. Livermore, CA: Lawrence Livermore National Laboratory.

Ranjineh Khojasteh, E. (2013). Geostatistical three-dimensional modeling of the subsurface unconsolidated materials in the Göttingen area. Main dissertation. Gottingen (Sweden): Georg-August-University.

Reed, P. M., Ellsworth, T. R., and Minsker, B. S. (2004). Spatial interpolation methods for nonstationary plume data. *Groundwater.* 42 (2), 190–202. doi:10.1111/j.1745-6584.2004.tb02667.x

Refsgaard, J. C., Auken, E., Bamberg, C. A., Christensen, B. S., Clausen, T., Dalgaard, E., et al. (2014). Nitrate reduction in geologically heterogeneous catchments–A framework for assessing the scale of predictive capability of hydrological models. *Sci. Total Environ.* 468, 1278–1288. doi:10.1016/j.scitotenv.2013.07.042

Ritzi, R. W., Jr. (2000). Behavior of indicator variograms and transition probabilities in relation to the variance in lengths of hydrofacies. *Water Resour. Res.* 36 (11), 3375–3381. doi:10.1029/2000wr900139

Roig-Silva, C., Haugen, B. D., and Wakeley, L. D. (2012). ERDC/GSL TR-129. *Defining soil materials for 3-D models of the near surface: preliminary Findings*. Vicksburg, MS: US Army Corps of Engineers.

Ronayne, M. J., Gorelick, S. M., and Caers, J. (2008). Identifying discrete geologic structures that produce anomalous hydraulic response: an inverse modeling approach. *Water Resour. Res.* 44 (8), W08426. doi:10.1029/2007wr006635

Sakaki, T., Frippiat, C. C., Komatsu, M., and Illangasekare, T. H. (2009). On the value of lithofacies data for improving groundwater flow model accuracy in a three-dimensional laboratory-scale synthetic aquifer. *Water Resour. Res.* 45 (11), W11404. doi:10.1029/2008wr007229

Sampath, P. V., Liao, H. S., Curtis, Z. K., Doran, P. J., Herbert, M. E., May, C. A., et al. (2015). Understanding the groundwater hydrology of a geographically-

isolated prairie fen: implications for conservation. *PloS One.* 10 (10), e0140430. doi:10.1371/journal.pone.0140430

Sampath, P. V., Liao, H. S., Curtis, Z. K., Herbert, M. E., Doran, P. J., May, C. A., et al. (2016). Understanding fen hydrology across multiple scales. *Hydrol. Process.* 30 (19), 3390–3407. doi:10.1002/hyp.10865

Sartore, L. (2013). spMC: modelling spatial random fields with continuous lag Markov chains. *The R Journal.* 5 (2), 16-28. doi:10.32614/rj-2013-022

Sartore, L., Fabbri, P., and Gaetan, C. (2016). spMC: an R-package for 3D lithological reconstruction based on spatial Markov chains. *Comput. Geosci.* 94, 40–47. doi:10.1016/j.cageo.2016.06.001

Sawyer, A. H. (2015). Enhanced removal of groundwater-borne nitrate in heterogeneous aquatic sediments. *Geophys. Res. Lett.* 42 (2), 403–410. doi:10.1002/2014gl062234

Schornberg, C., Schmidt, C., Kalbus, E., and Fleckenstein, J. H. (2010). Simulating the effects of geologic heterogeneity and transient boundary conditions on streambed temperatures–Implications for temperature-based water flux calculations. *Adv. Water Resour.* 33 (11), 1309–1319. doi:10.1016/j.advwatres.2010.04.007

Serrano, R. P., Guadagnini, L., Riva, M., Giudici, M., and Guadagnini, A. (2014). Impact of two geostatistical hydro-facies simulation strategies on head statistics under non-uniform groundwater flow. *J. Hydrol.* 508, 343–355. doi:10.1016/j.jhydrol.2013.11.009

Siirila-Woodburn, E. R., and Maxwell, R. M. (2015). A heterogeneity model comparison of highly resolved statistically anisotropic aquifers. *Adv. Water Resour.* 75, 53–66. doi:10.1016/j.advwatres.2014.10.011

Siirila-Woodburn, E. R., Sanchez-Vila, X., and Fernàndez-Garcia, D. (2015). On the formation of multiple local peaks in breakthrough curves. *Water Resour. Res.* 51 (4), 2128–2152. doi:10.1002/2014wr016394

Singh, H. V., Faulkner, B. R., Keeley, A. A., Freudenthal, J., and Forshay, K. J. (2018). Floodplain restoration increases hyporheic flow in the Yakima River Watershed, Washington. *Ecol. Eng.* 116, 110–120. doi:10.1016/j.ecoleng.2018.02.001

Sivakumar, B., Harter, T., and Zhang, H. (2005b). A fractal investigation of solute travel time in a heterogeneous aquifer: transition probability/Markov chain representation. *Ecol. Model.* 182 (3–4), 355–370. doi:10.1016/j.ecolmodel.2004.04.010

Sivakumar, B., Harter, T., and Zhang, H. (2005a). Solute transport in a heterogeneous aquifer: a search for nonlinear deterministic dynamics. *Nonlin. Processes Geophys.* 12 (2), 211–218. doi:10.5194/npg-12-211-2005

Smith, L. N. (2002). "Subsurface geologic mapping from descriptive and petrophysical borehole logs," in *Geoenvironmental mapping–method, theory and practice*. Rotterdam: AA Balkema, 121–145

Soltanian, M. R., Amooie, M. A., Dai, Z., Cole, D., and Moortgat, J. (2016). Critical dynamics of gravito-convective mixing in geological carbon sequestration. *Sci. Rep.* 6 (1), 1–13. doi:10.1038/srep35921

Soltanian, M. R., Amooie, M. A., Gershenzon, N., Dai, Z., Ritzi, R., Xiong, F., et al. (2017b). Dissolution trapping of carbon dioxide in heterogeneous aquifers. *Environ. Sci. Technol.* 51 (13), 7732–7741. doi:10.1021/acs.est.7b01540

Soltanian, M. R., Sun, A., and Dai, Z. (2017a). Reactive transport in the complex heterogeneous alluvial aquifer of Fortymile Wash, Nevada. *Chemosphere.* 179, 379–386. doi:10.1016/j.chemosphere.2017.03.136

Song, X., Chen, X., Ye, M., Dai, Z., Hammond, G., and Zachara, J. M. (2019). Delineating facies spatial distribution by integrating ensemble data assimilation and indicator geostatistics with level-set transformation. *Water Resour. Res.* 55 (4), 2652–2671. doi:10.1029/2018wr023262

Song, X., Chen, X., Ye, M., Dai, Z., and Hammond, G. E. (2015). SAND-2015-6217J. *Delineating hydrofacies spatial distribution by integrating ensemble data assimilation and indicator geostatistics*. Albuquerque, NM: Sandia National Laboratory.

Stevick, E., Pohll, G., and Huntington, J. (2005). Locating new production wells using a probabilistic-based groundwater model. *J. Hydrol.* 303 (1–4), 231–246. doi:10.1016/j.jhydrol.2004.07.016

Strebelle, S. (2002). Conditional simulation of complex geological structures using multiple-point statistics. *Math. Geol.* 34 (1), 1–21. doi:10.1023/a:1014009426274

Sun, A. Y., Ritzi, R. W., and Sims, D. W. (2008). Characterization and modeling of spatial variability in a complex alluvial aquifer: implications on solute transport. *Water Resour. Res.* 44 (4), W04402. doi:10.1029/2007wr006119

Sun, Q., Shao, J., Wang, Y., and Ma, T. (2019). Research on appropriate borehole density for establishing reliable geological model based on quantitative uncertainty analysis. *Arab. J. Geosci.* 12 (13), 410. doi:10.1007/s12517-019-4533-7

Sun, Y., Tong, C., Trainor-Guitton, W. J., Lu, C., Mansoor, K., and Carroll, S. A. (2013). Global sampling for integrating physics-specific subsystems and quantifying uncertainties of CO_2 geological sequestration. *Int. J. Greenh. Gas Con.* 12, 108–123. doi:10.1016/j.ijggc.2012.10.004

te Stroet, C. B. M., and Snepvangers, J. J. J. (2005). Mapping curvilinear structures with local anisotropy kriging. *Math. Geol.* 37 (6), 635–649. doi:10.1007/s11004-005-7310-y

Teramoto, E. H., Chang, H. K., and Caetano-Chang, M. R. (2017). Transporte de solutos em diferentes cenários geológicos gerados por modelos estocásticos de cadeias de Markov. *R. Águas Subter.* 31 (4), 316–326. doi:10.14295/ras.v31i4.28860

Theodossiou, N., and Fotopoulou, E. (2015). Delineating well-head protection areas under conditions of hydrogeological uncertainty. A case-study application in Northern Greece. *Environ. Proc.* 2 (1), 113–122. doi:10.1007/s40710-015-0087-1

Tompson, A. F. B., Bruton, C. J., Pawloski, G. A., Smith, D. K., Bourcier, W. L., Shumaker, D. E., et al. (2002). On the evaluation of groundwater contamination from underground nuclear tests. *Environ. Geol.* 42 (2–3), 235–247. doi:10.1007/s00254-001-0493-8

Tompson, A. F. B., Carle, S. F., Rosenberg, N. D., and Maxwell, R. M. (1999). Analysis of groundwater migration from artificial recharge in a large urban aquifer, a simulation perspective. *Water Resour. Res.* 35 (10), 2981–2998. doi:10.1029/1999wr900175

Trainor-Guitton, W., Mansoor, K., Sun, Y., and Carroll, S. (2016). Merits of pressure and geochemical data as indicators of CO_2/brine leakage into a heterogeneous, sedimentary aquifer. *Intern. J. Greenh. Gas Con.* 52, 237–249. doi:10.1016/j.ijggc.2016.07.002

Traum, J. A., Phillips, S. P., Bennett, G. L., Zamora, C., and Metzger, L. F. (2014). Scientific investigations report 2014-5148. *Documentation of a groundwater flow model (SJRRPGW) for the San Joaquin River RestorationProgram study area, California.* Reston, VA: US Geological Survey.

Troldborg, L. (2004). Report 2004/10. *The influence of conceptual geological models onthe simulation of flow and transport in Quaternary aquifer systems.* PhD dissertation. Kongens Lyngby (Denmark): DTU Environment.

Tsai, F. T. C., and Elshall, A. S. (2013). Hierarchical Bayesian model averaging for hydrostratigraphic modeling: uncertainty segregation and comparative evaluation. *Water Resour. Res.* 49, 5520–5536. doi:10.1002/wrcr.20428

Tsang, Y. W., Tsang, C. F., Hale, F. V., and Dverstorp, B. (1996). Tracer transport in a stochastic continuum model of fractured media. *Water Resour. Res.* 32 (10), 3077–3092. doi:10.1029/96wr01397

Vincent Henri, C., and Harter, T. (2019). Stochastic assessment of nonpoint source contamination: joint impact of aquifer heterogeneity and well characteristics on management metrics. *Water Resour. Res.* 55 (8), 6773–6794. doi:10.1029/2018wr024230

Vistelius, A. B. (1949). On the question of the mechanism of formation of strata. *Dokl. Akad. Nauk SSSR.* 65 (2), 191–194.

Wainwright, H. M., Chen, J., Sassen, D. S., and Hubbard, S. S. (2014). Bayesian hierarchical approach and geophysical data sets for estimation of reactive facies over plume scales. *Water Resour. Res.* 50 (6), 4564–4584. doi:10.1002/2013wr013842

Wallace, C. D., Sawyer, A. H., Soltanian, M. R., and Barnes, R. T. (2020). Nitrate removal within heterogeneous riparian aquifers under tidal influence. *Geophys. Res. Lett.* 47, e2019GL085699. doi:10.1029/2019GL085699

Wang, X., Jardani, A., and Jourde, H. (2017). A hybrid inverse method for hydraulic tomography in fractured and karstic media. *J. Hydrol.* 551, 29–46. doi:10.1016/j.jhydrol.2017.05.051

Weissmann, G. S., Carle, S. F., and Fogg, G. E. (1999). Three-dimensional hydrofacies modeling based on soil surveys and transition probability geostatistics. *Water Resour. Res.* 35 (6), 1761–1770. doi:10.1029/1999wr900048

Weissmann, G. S., and Fogg, G. E. (1999). Multi-scale alluvial fan heterogeneity modeled with transition probability geostatistics in a

sequence stratigraphic framework. *J. Hydrol.* 226 (1–2), 48–65. doi:10.1016/s0022-1694(99)00160-2

Weissmann, G. S., Pickel, A., McNamara, K. C., Frechette, J. D., Kalinovich, I., Allen-King, R. M., et al. (2015). Characterization and quantification of aquifer heterogeneity using outcrop analogs at the Canadian Forces Base Borden, Ontario, Canada. *GSA Bulletin.* 127 (7–8), 1021–1035. doi:10.1130/B31193.1

Weissmann, G. S., Zhang, Y., Fogg, G. E., and Mount, J. F. (2004). "Influence of incised-valley-fill deposits on hydrogeology of a stream-dominated alluvial fan," in *Aquifer characterization*. SEPM Special Publication, Vol. 80, 15–28.

Weissmann, G. S., Zhang, Y., LaBolle, E. M., and Fogg, G. E. (2002). Dispersion of groundwater age in an alluvial aquifer system. *Water Resour. Res.* 38 (10), 16. doi:10.1029/2001wr000907

Wu, L., Wang, J., Pei, X., Fogg, G. E., Yang, T., Yan, X., et al. (2019). Distribution and origination of zinc contamination in newly reclaimed heterogeneous dredger fills: field investigation and numerical simulation. *Mar. Pollut. Bull.* 149, 110496. doi:10.1016/j.marpolbul.2019.110496

Yager, R. M., and Heywood, C. E. (2014). Simulation of the effects of seasonally varying pumping on intraborehole flow and the vulnerability of public-supply wells to contamination. *Groundwater.* 52 (S1), 40–52. doi:10.1111/gwat.12150

Yang, Y. M., Dilmore, R. M., Mansoor, K., Buscheck, T. A., and Bromhal, G. S. (2019). Integration of wellbore pressure measurement and groundwater quality monitoring to enhance detectability of brine and CO_2 leakage. *Int. J. Greenh. Gas Con.* 85, 143–155. doi:10.1016/j.ijggc.2019.04.004

Yang, Z., Chen, Y. F., and Niemi, A. (2020). Gas migration and residual trapping in bimodal heterogeneous media during geological storage of CO_2. *Adv. Water Resour.* 142, 103608. doi:10.1016/j.advwatres.2020.103608

Yang, Z., Tian, L., Niemi, A., and Fagerlund, F. (2013). Upscaling of the constitutive relationships for CO_2 migration in multimodal heterogeneous formations. *Int. J. Greenh. Gas Con.* 19, 743–755. doi:10.1016/j.ijggc.2012.11.015

Ye, M., Cooper, C., Chapman, J., Gillespie, D., and Zhang, Y. (2009). A geologically based Markov chain model for simulating tritium transport with uncertain conditions in a nuclear-stimulated natural gas reservoir. *SPE Reservoir Eval. Eng.* 12 (06), 974–984. doi:10.2118/114920-pa

Ye, M., and Khaleel, R. (2008). A Markov chain model for characterizing medium heterogeneity and sediment layering structure. *Water Resour. Res.* 44 (9), W09427. doi:10.1029/2008wr006924

Yin, M., Zhang, Y., Ma, R., Tick, G. R., Bianchi, M., Zheng, C., et al. (2020). Super-diffusion affected by hydrofacies mean length and source geometry in alluvial settings. *J. Hydrol.* 582, 124515. doi:10.1016/j.jhydrol.2019.124515

Ying, Z. J. (2000). iksim: a fast algorithm for indicator kriging and simulation in the presence of inequality constraints, hard and soft data. *Comput. Geosci.* 26 (5), 493–507. doi:10.1016/s0098-3004(99)00132-6

Yong, H., Hu, K., Li, B., Chen, D., Suter, H. C., and Huang, Y. (2009). Comparison of sequential indicator simulation and transition probability indicator simulation used to model clay content in microscale surface soil. *Soil Sci.* 174 (7), 395–402. doi:10.1097/SS.0b013e3181aea77c

Zetterlund, M., Norberg, T., Ericsson, L. O., and Rosén, L. (2011). Framework for value of information analysis in rock mass characterization for grouting purposes. *J. Construct. Eng. Manag.* 137 (7), 486–497. doi:10.1061/(asce)co.1943-7862.0000265

Zhang, H., Harter, T., and Sivakumar, B. (2006). Nonpoint source solute transport normal to aquifer bedding in heterogeneous, Markov chain random fields. *Water Resour. Res.* 42 (6). doi:10.1029/2004wr003808

Zhang, Y., Benson, D. A., and Baeumer, B. (2007). Predicting the tails of breakthrough curves in regional-scale alluvial systems. *Ground Water.* 45 (4) 473–484. doi:10.1111/j.1745-6584.2007.00320.x

Zhang, Y., and Fogg, G. E. (2003). Simulation of multi-scale heterogeneity of porous media and parameter sensitivity analysis. *Sci. China E.* 46 (5), 459–474. doi:10.1360/02ye0098

Zhang, Y., Green, C. T., and Baeumer, B. (2014). Linking aquifer spatial properties and non-Fickian transport in mobile–immobile like alluvial settings. *J. Hydrol.* 512, 315–331. doi:10.1016/j.jhydrol.2014.02.064

Zhang, Y., Green, C. T., and Fogg, G. E. (2013). The impact of medium architecture of alluvial settings on non-Fickian transport. *Adv. Water Resour.* 54, 78–99. doi:10.1016/j.advwatres.2013.01.004

Zhang, Y., and Meerschaert, M. M. (2011). Gaussian setting time for solute transport in fluvial systems. *Water Resour. Res.* 47 (8), W08601. doi:10.1029/2010wr010102

Zhang, Y., Weissmann, G. S., Fogg, G. E., Lu, B., Sun, H., and Zheng, C. (2018). Assessment of groundwater susceptibility to non-point source contaminants using three-dimensional transient indexes. *Int. J. Environ. Res. Publ. Health.* 15 (6), 1177. doi:10.3390/ijerph15061177

Zheng, C., Bianchi, M., and Gorelick, S. M. (2011). Lessons learned from 25 years of research at the MADE site. *Groundwater.*, 49(5), 649–662. doi:10.1111/j.1745-6584.2010.00753.x

Zhou, Y., Ritzi, R. W., Jr., Soltanian, M. R., and Dominic, D. F. (2014). The influence of streambed heterogeneity on hyporheic flow in gravelly rivers. *Groundwater.* 52 (2), 206–216. doi:10.1111/gwat.12048

Zhu, H. (1991). *Modeling mixture of spatial distributions with integration of soft data*. PhD dissertation. Stanford (CA): Stanford University.

Zhu, L., Dai, Z., Gong, H., Gable, C., and Teatini, P. (2016a). Statistic inversion of multi-zone transition probability models for aquifer characterization in alluvial fans. *Stoch. Environ. Res. Risk Assess.* 30 (3), 1005–1016. doi:10.1007/s00477-015-1089-2

Zhu, L., Franceschini, A., Gong, H., Ferronato, M., Dai, Z., Ke, Y., et al. (2020). The 3-D facies and geomechanical modeling of land subsidence in the Chaobai Plain, Beijing. *Water Resour. Res.* 56 (3), p.e2019WR027026. doi:10.1029/2019wr027026

Zhu, L., Gong, H., Chen, Y., Li, X., Chang, X., and Cui, Y. (2016b). Improved estimation of hydraulic conductivity by combining stochastically simulated hydrofacies with geophysical data. *Sci. Rep.* 6, 22224. doi:10.1038/srep22224

Zovi, F., Camporese, M., Franssen, H. J. H., Huisman, J. A. and Salandin, P. (2017). Identification of high-permeability subsurface structures with multiple point geostatistics and normal score ensemble Kalman filter. *J. Hydrol.* 548, 208–224.

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.

APPENDIX

Expanding **Eq. 14** yields

$$\tilde{t}_{jk}(\mathbf{h}) = \frac{E\left\{\begin{array}{c}\alpha(\mathbf{x})\alpha(\mathbf{x}+\mathbf{h})I_j(\mathbf{x})I_k(\mathbf{x}+\mathbf{h}) + \alpha(\mathbf{x})\beta(\mathbf{x}+\mathbf{h})p_k I_j(\mathbf{x}) \\ +\beta(\mathbf{x})\alpha(\mathbf{x}+\mathbf{h})p_j I_k(\mathbf{x}+\mathbf{h}) + \beta(\mathbf{x})\beta(\mathbf{x}+\mathbf{h})p_j p_k\end{array}\right\}}{E\left\{\left[\alpha(\mathbf{x})I_j(\mathbf{x}) + \beta(\mathbf{x})p_j\right]\right\}}. \tag{22}$$

Assuming stationarity and applying **Eq. 10 and 22** reduceto

$$\tilde{t}_{jk}(\mathbf{h}) = \frac{E\left\{\begin{array}{c}\alpha(\mathbf{x})\alpha(\mathbf{x}+\mathbf{h})I_j(\mathbf{x})I_k(\mathbf{x}+\mathbf{h}) + \alpha(\mathbf{x})\beta(\mathbf{x}+\mathbf{h})p_j p_k \\ +\beta(\mathbf{x})\alpha(\mathbf{x}+\mathbf{h})p_j p_k + \beta(\mathbf{x})\beta(\mathbf{x}+\mathbf{h})p_j p_k\end{array}\right\}}{E\left\{\left[\alpha(\mathbf{x})p_j + \beta(\mathbf{x})p_j\right]\right\}}. \tag{23}$$

Applying **Eqs. 11 and 10** and combining terms, **Eq. 23** reduces to **Eq. 15**:

$$\tilde{t}_{jk}(\mathbf{h}) = \frac{E\left\{\alpha(\mathbf{x})\alpha(\mathbf{x}+\mathbf{h})I_j(\mathbf{x})I_k(\mathbf{x}+\mathbf{h}) + \left[\beta(\mathbf{x}+\mathbf{h}) + \beta(\mathbf{x})\alpha(\mathbf{x}+\mathbf{h})\right]p_j p_k\right\}}{p_j}. \tag{15}$$

frontiers in Earth Science

ORIGINAL RESEARCH
published: 10 November 2020
doi: 10.3389/feart.2020.563122

Stochastic Simulation of the Spatial Heterogeneity of Deltaic Hydrofacies Accounting for the Uncertainty of Facies Proportions

S. Jorreto-Zaguirre[1], *P.A. Dowd*[2]*, *E. Pardo-Igúzquiza*[3], *A. Pulido-Bosch*[1] and *F. Sánchez-Martos*[1]

[1] Water Resources and Environmental Geology Research Group, Department of Biology and Geology, University of Almería, Almería, Spain, [2] Faculty of Engineering, Computer and Mathematical Sciences, University of Adelaide, Adelaide, Australia, [3] Instituto Geológico y Minero de España (IGME), Madrid, Spain

The spatial geological heterogeneity of an aquifer significantly affects groundwater storage, flow and the transport of solutes. In the particular case of coastal aquifers, spatial geological heterogeneity is also a major determining factor of the spatio-temporal patterns of water quality (salinity) due to seawater intrusion. While the hydraulics of coastal hydrogeology can be modeled effectively by various density flow equations, the aquifer geology is highly uncertain. A stochastic solution to the problem is to generate numerical realisations of the geology using sequential stratigraphy, geophysical models or geostatistical approaches. The geostatistical methods (two-point geostatistics, Markov chain models and multiple-point geostatistics) have the advantage of minimal data requirements, e.g., when the only data available are from cores from a few sparsely located boreholes. We provide an extension of sequential indicator simulation by including the uncertainty of the hydrofacies proportions in the simulation approach. We also deal with the problem of variogram estimation from sparse boreholes and we discuss the implicit transition probabilities and the connectivity of simulated realisations of a number of categorical variables. The variogram model used in the simulation of hydrofacies significantly influences the degree of connectivity of the hydrofacies in the simulated model. The choice of model is critical as connectivity determines the amount and extent of seawater intrusion and hence the environmental risk. The methodology is illustrated with a case study of the Andarax river delta, a coastal aquifer in south-eastern Spain. This is a semi-arid Mediterranean region in which the increasing use of, and demand for, groundwater is exacerbated by a transient tourist population that reaches its peak in the summer when the demand for the permanent population is at its highest. The work reported here provides a sound basis for designing flow simulation models for the optimal management of groundwater resources. This paper is an extended version of a presentation given at the 2012 GeoENV Conference held in Valencia, Spain.

Keywords: stochastic simulation, spatial hetereogeneity, hydrofacies, uncertainty, connectivity, sequential indicator simulation

OPEN ACCESS

Edited by:
J. Jaime Gómez-Hernández, Universitat Politècnica de València, Spain

Reviewed by:
Mauro Giudici, University of Milan, Italy
Marijke Huysmans, Vrije University Brussel, Belgium
Alessandro Comunian, University of Milan, Italy

***Correspondence:**
P.A. Dowd
peter.dowd@adelaide.edu.au

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

Received: *18 May 2020*
Accepted: *05 October 2020*
Published: *10 November 2020*

Citation:
Jorreto-Zaguirre S, Dowd PA, Pardo-Igúzquiza E, Pulido-Bosch A, Sánchez-Martos F (2020) Stochastic Simulation of the Spatial Heterogeneity of Deltaic Hydrofacies Accounting for the Uncertainty of Facies Proportions. Front. Earth Sci. 8:563122. doi: 10.3389/feart.2020.563122

INTRODUCTION

Half of the world's population lives in coastal areas and the transient and permanent populations of these areas continue to increase. This generates increasing, and often competing, demands for water. To meet demand, or allocate limited supply, and satisfy environmental and sustainability constraints in these coastal zones requires optimal management of water resources in general and groundwater resources in particular. The problem is exacerbated in the Mediterranean region because of the combined effects of semi-arid climate (high evapotranspiration and low rainfall) and seasonal tourism that increases the demand for water during the summer, which is the period of lowest aquifer recharge. The result is increased depletion of groundwater with the risk of over-extraction.

In coastal areas, over-extraction not only depletes the aquifer but also causes seawater intrusion leading to deterioration of water quality that may ultimately render the aquifer water unfit for human consumption and other uses such as agriculture. In general, mathematical models of aquifers comprise two parts: the medium (the geological materials that comprise the aquifer) and the passage of water (hydraulics) through the geological medium. The hydraulics are modeled effectively by various density flow equations but the geology, particularly the spatial distribution of the hydrofacies, introduces a major source of variability and uncertainty in any model or assessment of an aquifer. In general, the geology is heterogeneous and is unknown apart from limited direct data from sparse boreholes or the indirect geophysical information. Apart from borehole cores, the aquifer is not observable on any meaningful scale and the only realistic approach in such cases is via a stochastic model informed by sparse data and/or surface analogues (outcrops). The prediction of seawater intrusion is thus a stochastic problem.

The spatio-temporal patterns of groundwater quality in coastal aquifers are determined by the spatial heterogeneity and spatial distribution of hydrofacies (Eaton, 2006) as well as by different hydrologic settings and forcings. A general way to address the stochastic seawater intrusion problem is to accommodate the stochastic character of the geology by generating stochastic simulations of aquifer heterogeneity.

Although the techniques discussed here are generally applicable to different sedimentary environments, the focus in this paper is on deltaic environments. The three most common means of generating three-dimensional geological models (in this case, 3D models of hydrofacies) of deltas are sequential stratigraphy (Cabello et al., 2007), geophysical methods (Tercier et al., 2000; Deidda et al., 2006; Barakat, 2010) and geostatistical techniques (dell'Arciprete et al., 2012; Perulero et al., 1997). A good overview of statistical grid-based sedimentary facies reconstruction and modeling methods can be found in Falivene et al. (2007).

In an ideal situation, all of these techniques could be used to integrate all available information to generate a model that is as realistic as possible. However, the amount, type and quality of the available data constrain the choice of method. Of the three methodologies, geostatistical methods are the least demanding in terms of data requirements because they are based on the estimated underlying structural model and its uncertainty (Pardo-Igúzquiza et al., 2009) and they can be applied even when only a few sparsely located boreholes are available.

The geostatistical methods can be classified either as object-based methods (Gouw, 2007) or models based on pixels and voxels (Dubrule and Damsleth, 2001). The latter are the more flexible and, in turn, they can be classified as two-point geostatistical models (Deutsch and Journel, 1998), Markov chain models (Carle and Fogg, 1996) or multiple-point geostatistical models (Strebelle, 2002; Comunian et al., 2011). Multiple-point geostatistical methods require detailed three-dimensional training images, which are not usually available for groundwater applications although other sources such as surface analogues (e.g., Comunian et al., 2011), outcrops, geophysical images, outputs of numerical models can be used. Markov chain models tend to give simulations that are less realistic than the other methods; in particular, the hydrofacies are often too disconnected. For these reasons, we have chosen to use basic, second-order stationary geostatistical models and, in the work presented here, we have used sequential indicator simulation to generate realisations of categorical variables (Goovaerts, 1997). We also considered using truncated plurigaussian simulation (Le Loc'h et al., 1994; Le Loc'h and Galli, 1996; Dowd et al., 2003; Mariethoz et al., 2009) but decided against it because of the inability to infer facies contact relationships with an acceptable level of accuracy from cores from sparse boreholes. This method is more suitable for cases in which there are reliable surface analogues from which detailed contacts can be observed. A comparison of geostatistical methods for hydrofacies simulation is given in dell'Arciprete et al. (2012). The work presented here differs from the latter, which simulates alluvial sediments using vertical facies maps of five almost orthogonal quarry faces and no borehole data whereas we simulate deltaic sediments in a flat area for which there are no outcrops and the only data available are from a few sparsely located boreholes. Dell'Arciprete et al. (2012) compare sequential indicator simulation, transition probability geostatistical simulation and multiple point simulation. We have not used multiple point simulation because of the requirement for 3D training images or at least orthogonal 2D training images. We have not used transition probability simulation because, as can be seen in dell'Arciprete et al. (2012), it can generate unrealistic images of spatial heterogeneity, although its successful use has been reported by other authors (Lee et al., 2007; Bianchi et al., 2011). We have extended sequential indicator simulation to include the uncertainty of the proportions of the hydrofacies. The uncertainty in proportions is propagated as uncertainty in the variograms and thus uncertainty in the connectivity of the hydrofacies. This extended methodology is presented in the following section.

METHODOLOGY

The focus here is on the spatial distribution of hydrofacies in coastal aquifers, which is the major determining factor in seawater intrusion. The physical heterogeneity of the

geological materials that comprise the aquifer is evident from the distribution of hydrofacies that can be observed in outcrops and in borehole cores. In many cases there are no outcrops and, in any case, the distribution of hydrofacies between boreholes remains unknown. Geostatistical simulation is used to generate possible images (as many as desired) of such unknown realities. These simulated images reproduce the experimentally observed (from core samples) spatial variability of the hydrofacies, which is modeled by direct and cross-variograms. These simulations are also conditioned to the experimental data (Chilès and Delfiner, 1999).

A spatial category (or hydrofacies for the applications discussed here) F_i is defined as:

$$F_i = \{\mathrm{u} | f(\mathrm{u}) = i\} \tag{1}$$

where $f(\mathrm{u})$ is the function that assigns to each spatial location $\mathrm{u} = \{x, y, z\}$ a unique hydrofacies from the set $\{i = 1, \ldots, N\}$.

Suppose there are N hydrofacies that are mutually exclusive (that is, at each location u, there is a unique hydrofacies) and defined exhaustively in the three-dimensional space:

$$F_i \cap F_j = \varnothing \quad \forall i, j = 1, \ldots., N; i \neq j \tag{2}$$

$$\bigcup_{i=1}^{N} F_i = \mathbb{R}^3 \tag{3}$$

In practice, only a part of the three-dimensional space is of interest $\chi \subset \mathbb{R}^3$. For each category define an indicator variable:

$$I_i(\mathrm{u}) = \begin{cases} 1 & \text{if } \mathrm{u} \in F_i \\ 0 & \text{if } \mathrm{u} \notin F_i \end{cases} \quad i = 1, \ldots., N \tag{4}$$

From which it follows that:

$$\sum_{i=1}^{N} I_i(\mathrm{u}) = 1 \quad \forall \mathrm{u} \in \chi \tag{5}$$

It is assumed that each indicator variable is a second-order stationary random function (Myers, 1989), that is, the spatial mathematical expectation is constant, and the spatial covariance is a function solely of the distance vector h. The mathematical expectation of each indicator function is equal to the global proportion of the hydrofacies:

$$\mathrm{E}\{I_i(\mathrm{u})\} = p_i \tag{6}$$

with $0 < p_i < 1$, $i = 1, \ldots., N$. The variogram is defined as:

$$\gamma_i(\mathrm{h}) = \frac{1}{2}\mathrm{E}\{[I_i(\mathrm{u}) - I_i(\mathrm{u}+\mathrm{h})]^2\} \tag{7}$$

which is related to the non-centred covariance (Journel and Alabert, 1989) by:

$$C_i(\mathrm{h}) = p_i - \gamma_i(\mathrm{h}) \tag{8}$$

where

$$C_i(\mathrm{h}) = E\{I_i(\mathrm{u})\, I_i(\mathrm{u}+\mathrm{h}) = P\{u \in F_i \text{ and } \mathrm{u}+\mathrm{h} \in F_i\}\} \tag{9}$$

In a similar way the cross-variograms, $\gamma_{ij}(\mathrm{h})$, and cross-covariances can be defined as in Dowd et al. (2003). The condition in **Eq. 5** imposes certain relationships among direct, and cross, variograms (Dowd et al., 2003):

$$\gamma_i(\mathrm{h}) = -\sum_{\substack{j=1 \\ j \neq i}}^{N} \gamma_{ij}(\mathrm{h}) \tag{10}$$

The sill of the variogram is given by the variance of the indicator variable:

$$\gamma_i(\mathrm{h}) \underset{h \to \infty}{\longrightarrow} p_i(1 - p_i) \tag{11}$$

and the sill of the cross-variogram is given by:

$$\gamma_{ij}(\mathrm{h}) \underset{h \to \infty}{\longrightarrow} p_i\, p_j \tag{12}$$

From **Eq. 12** it is evident that all cross-variograms must be negative as the probabilities or proportions on the right-hand side are positive. Furthermore, the non-centered cross-covariance between the indicators of hydrofacies F_i and F_j gives the probability that these two hydrofacies occur at a separation distance h:

$$C_{ij}(\mathrm{h}) = \mathrm{E}\{I_i(\mathrm{u})\, I_j(\mathrm{u}+\mathrm{h})\} = P\{u \in F_i \text{ and } \mathrm{u}+h \in F_j\} \tag{13}$$

Among the many geostatistical simulation methods, we have chosen to use sequential indicator simulation largely because of its simplicity. The method consists of estimating, at each location of interest u (for example, the nodes of a three-dimensional grid), the conditional probability of occurrence of each hydrofacies, subject to the condition:

$$\sum_{i=1}^{N} p_i^*(\mathrm{u}) = 1 \tag{14}$$

The conditional probability $p_i^*(\mathrm{u})$ is estimated by kriging (simple indicator kriging, indicator cokriging, or any other form) using the conditioning data, which initially comprise only the experimental data. A value is simulated by sampling the estimated conditional probability distribution. The simulated value is added to the conditioning data and the process is repeated until values have been simulated at all locations. The algorithm is explained in detail in Deutsch and Journel (1998) and Goovaerts (1997).

As can be seen from **Eqs 10–12** the proportion of each hydrofacies significantly affects the forms of the direct, and cross, variograms. Because the sill of the direct, and cross, variogram depends on the hydrofacies proportions, the ranges fitted to the experimental variograms will also depend on the proportions because the parameters are not fitted independently but simultaneously with a larger sill implying a larger range. In practical applications the real proportions are unknown and must be estimated from sparse data, thus introducing more uncertainty. In order to reproduce the total variability, the uncertainty of the estimated proportions of the hydrofacies must be assessed and taken into account in the simulations. We propose a resampling method to estimate the uncertainty of the proportions. The methodology is summarized in **Figure 1** and comprises the following steps:

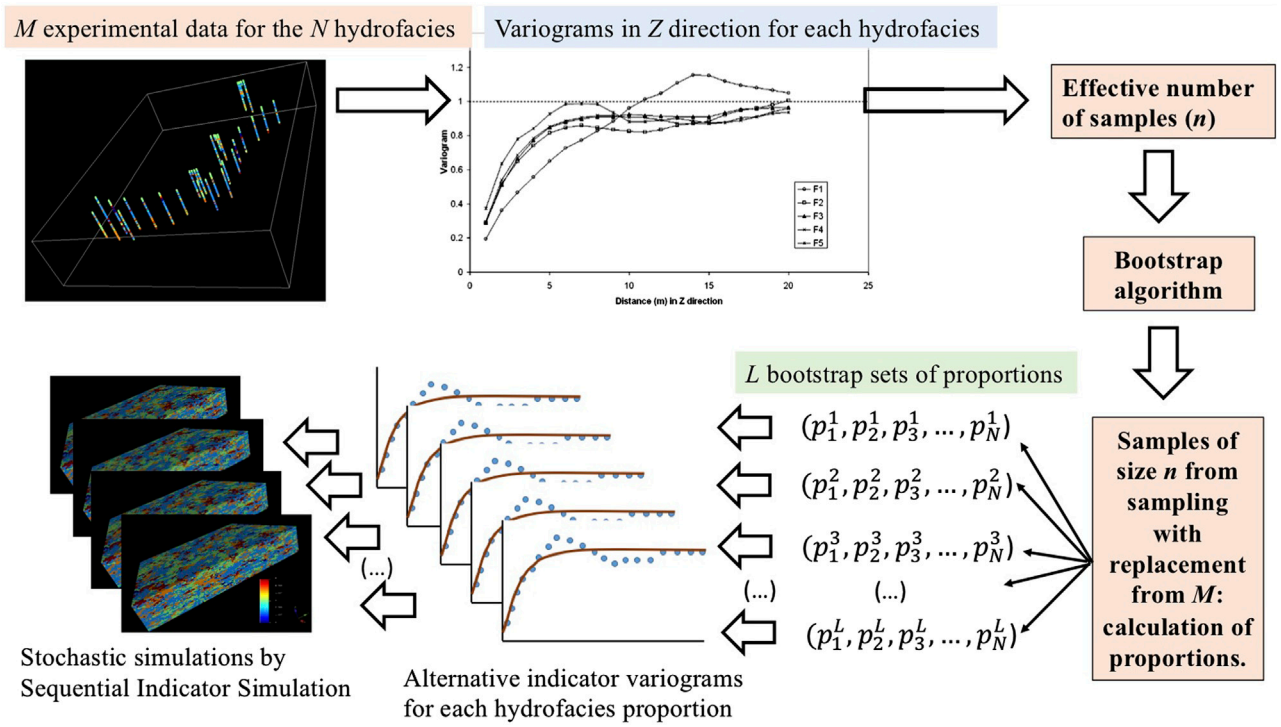


FIGURE 1 | Flowchart for the proposed methodology.

(1) There are M experimental data of N hydrofacies measured along boreholes and piezometers.
(2) Vertical variograms are calculated and used to estimate the effective number of samples n.
(3) A bootstrap algorithm is used to generate L samples of size n by sampling at random and with replacement from the set of M experimental data.
(4) For each bootstrap sample of size n the proportions of the N hydrofacies are calculated to provide a bootstrap sample of proportions.
(5) A variogram model is obtained from the bootstrap sample of proportions.
(6) A realization is generated using sequential indicator simulation and the variogram models.
(7) Go to step 5) using a different bootstrap sample of proportions. Repeat until the desired number of simulations is achieved.

A fundamental issue is the determination of the effective number of samples from the correlated experimental data. We use aligned, contiguous sequences of data, such as strings of borehole cores, to determine the effective number of data to be used for resampling. The effective range of the variogram in the vertical direction is a key parameter because the variability along the boreholes is much greater than the horizontal variability between boreholes. An example of this determination is given in the case study section.

Once the effective number of data is determined, a large number of samples (several thousand) of that size is obtained by sampling with replacement from the full data set. For each sample the hydrofacies proportions are calculated and a histogram is generated for each hydrofacies. Note that, as the proportions of hydrofacies must be used jointly (for example, to satisfy **Eq. 14**), it is the proportions of the resamples that are retained, and each vector of proportions is used in the geostatistical simulation rather than using the estimated proportions from the data.

The sequential simulation algorithm uses the models fitted to the experimental variograms. Emery (2004) provides a critical appraisal of the limitations of sequential indicator simulation in general; those that relate to categorical variables are addressed in the approach described in this paper. There are several possibilities for estimating the probabilities in **Eq. 14**. These range from simple indicator kriging, in which each indicator variable has an anisotropic variogram specific to each indicator variable, to full indicator cokriging. Although the latter is the best and most complete approach, there are serious issues around model inference and some simplification is required. For example, a common model could be used for all cross-variograms with a scaling factor applied to each of them so that the correct sill, as given in **Eq. 12**, is assigned. However, this approach is constrained by the need to satisfy the model validity requirement for a positive definite coregionalization model. For categorical variables a valid model has a very clear physical meaning. Once an arbitrary facies, $F_i = \{u | f(u) = i\}$ is fixed at an arbitrary location u then the same hydrofacies, F_i, occurs at any other arbitrary location $u + h$ with conditional probability

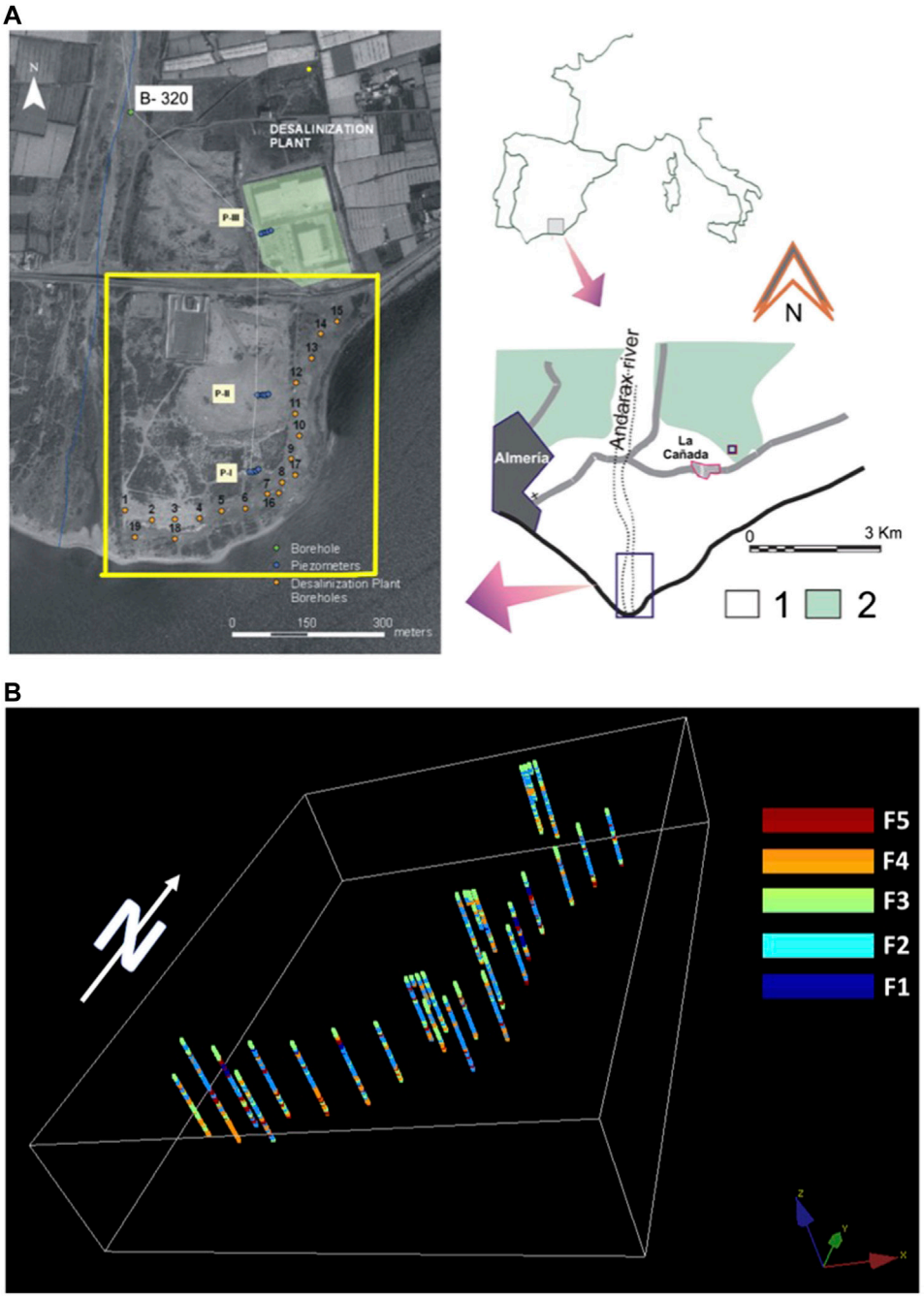


FIGURE 2 | (A) Location of the study area in the Andaraxriver delta. The yellow square is the area of the case study in which there are 19 boreholes and three clusters of four piezometers. The numeral 1 denotes detrital material from the Quaternary and 2 from the Pliocene. **(B)** Data for the five hydrofacies (F1 to F5) within boreholes and piezometers. The color legend ranges from very permeable (dark blue = F1) to very impermeable (dark red = F5).

$\frac{C_{ij}(\mathrm{h})}{p_j} = \frac{(p_i - \gamma_i(\mathrm{h}))}{p_i}$ and hydrofacies F_j (for $j \neq i$) occurs with conditional probability $\frac{C_{ij}(\mathrm{h})}{p_j} = \frac{\gamma_{ij}(\mathrm{h})}{p_j}$. In addition, writing $C_i(\mathrm{h}) = C_{ii}(\mathrm{h})$, the total probability must be:

$$\sum_{j=1}^{N} \frac{C_{ij}(\mathrm{h})}{p_j} = 1 \tag{15}$$

Equation (15), written in terms of the non-centred covariance, can also be written in terms of variograms or in terms of centered covariances, $\sigma(\mathrm{h})$:

$$\sum_{j=1}^{N} \frac{\sigma_{ij}(\mathrm{h}) + p_i p_j}{p_j} = 1 \tag{16}$$

$$-\sum_{j=1}^{N} \gamma_{ij}(\mathrm{h}) = 0 \quad (17)$$

Note that, by writing $\gamma_i(\mathrm{h}) = \gamma_{ii}(\mathrm{h})$, **Eqs 17** and **10** are equivalent. That is, a valid model of covariance and cross-covariances implies the occurrence probabilities are coherent.

From a practical perspective, there is also a need to demonstrate that any improvement in results from the complete cokriging method is sufficient to justify its use over that of simpler approaches. Thus, simple indicator kriging, in which each indicator variable has a specific anisotropic variogram, was chosen for the work presented in this paper.

The technique is illustrated in detail by the case study in the following section.

CASE STUDY

The study area (**Figure 2A**) is the detrital aquifer of the Andarax river delta in the province of Almeria in southern Spain. The aquifer comprises deltaic deposits from the Pleistocene overlain by fluvial and deltaic deposits from the Quaternary (Sánchez-Martos et al., 1999). The Andarax river is ephemeral, with flow usually resulting from big storms, and is an example of rivers in the semi-arid coastal regions of the Mediterranean. Within the study area there are 19 boreholes and three clusters of four piezometers each (Jorreto-Zaguirre et al., 2005) the locations of which are shown in plan view in **Figure 2A**. **Figure 2B** shows the spatial locations of the boreholes and piezometers together with colour-coded plots of the recorded hydrofacies. The vertical resolution of the boreholes and piezomenter cores is one m. The borehole cores have been classified into five types of hydrofacies according to their permeability: very permeable (F1), permeable (F2), low permeability (F3), impermeable (F4), and very impermeable (F5).

The proportions of each hydrofacies measured along the boreholes are: 2.62, 43.72, 27.21, 18.57, and 7.87%, respectively. The problem could be simplified by grouping the hydrofacies into three broader types (aquifer, aquitard and aquiclude) or even into only two types (permeable and impermeable). However, there is value in retaining the five hydrofacies because the very permeable hydrofacies could be associated with high permeability channels while the very impermeable facies could be associated with hydraulic barriers. **Figure 2B** clearly shows the sparsity of the data, with very few borehole cores to represent the total study volume, from which it is reasonable to conclude that the real hydrofacies proportions may differ significantly from the estimated values.

The data comprise hydrofacies observations made on 2,477 one-metre borehole cores. Classical (non-spatial) statistics cannot be used to evaluate the uncertainty of the proportions calculated from these data because spatial correlation implies that neighboring samples are not independent and, therefore, some of the information in the values of these samples is redundant. Variograms can be used to quantify the range of spatial correlation from which it is possible to infer the effective number of (spatially uncorrelated) samples. Variograms calculated along the boreholes (**Figure 3**) indicate an effective range of spatial correlation in this direction of around 8 m, i.e., samples separated by distances greater than, or equal to, 8 m are uncorrelated. Assuming that the average distance between pairs of boreholes is greater than the ranges of spatial correlation in all other directions, the effective number of data can be inferred as approximately 300. The uncertainty of the proportions can now be estimated by a bootstrap procedure, that is, resampling with replacement with a sample size of 300 from the full set of 2,477 data. A total of 5,000 bootstrap samples of size 300 were generated and the bootstrap distribution of the proportion of each of the five categories was calculated. **Figure 4A** shows the bootstrap distribution for the very permeable (F1) hydrofacies. Confidence intervals can be calculated from the bootstrap distributions. For example, although the estimated proportion of the very permeable hydrofacies (F1) is 2.62%, the lower and upper limits of the 95% confidence interval are, respectively, 1% and 4%. **Figure 4B** shows the bootstrap distribution for the very impermeable hydrofacies (F5). The complete results are shown in **Table 1** from which the similarity between mean and median indicates that the distributions are symmetrical.

Variograms were calculated for each of the indicators. As the sills of the direct and cross-variograms depend on the hydrofacies proportions it is useful to standardize the experimental variograms using the estimated proportions so that they can be displayed on the same graph. **Figure 3A** shows the standardized variograms calculated along the boreholes. The variogram for hydrofacies F1 is noticeably different to those of the other hydrofacies; the variogram for F5 also differs from the others but by a smaller amount. None of the variograms in **Figure 3A** reach the theoretical sill of 1.0, which reflects the fact that the estimated proportions differ from the true underlying proportions. The implication of this observation is that models fitted to the sample variograms will be different for each bootstrap sample of proportions. **Table 2** lists the experimental proportions together with the proportions calculated from the first ten bootstrap samples and **Figure 3B** shows the experimental variograms for hydrofacies F1. An indication of the effects of the uncertainty of the proportions is given by comparing the variograms of the experimental proportion of hydrofacies F1 (2.62%) with, for example, the fifth bootstrap value of 3.00%. These two proportions would generate respective variogram sills of 0.0255 and 0.0291 and weighted least squares estimated ranges of 4.37 and 5.62 m, respectively, as shown in **Figure 4B**. The weighted least squares estimated ranges of exponential models for the experimental proportions of the other hydrofacies are 3.00, 2.87, 2.73, and 2.05 m for F2, F3, F4, and F5, respectively, and these values quantify the differences that can be seen graphically in **Figure 3A**. Note that for exponential model variograms the sill is reached asymptotically and it is useful to define an effective, or practical, range at which, for all practical purposes, the sill is reached; the effective range is three times the range parameter in the variogram model.

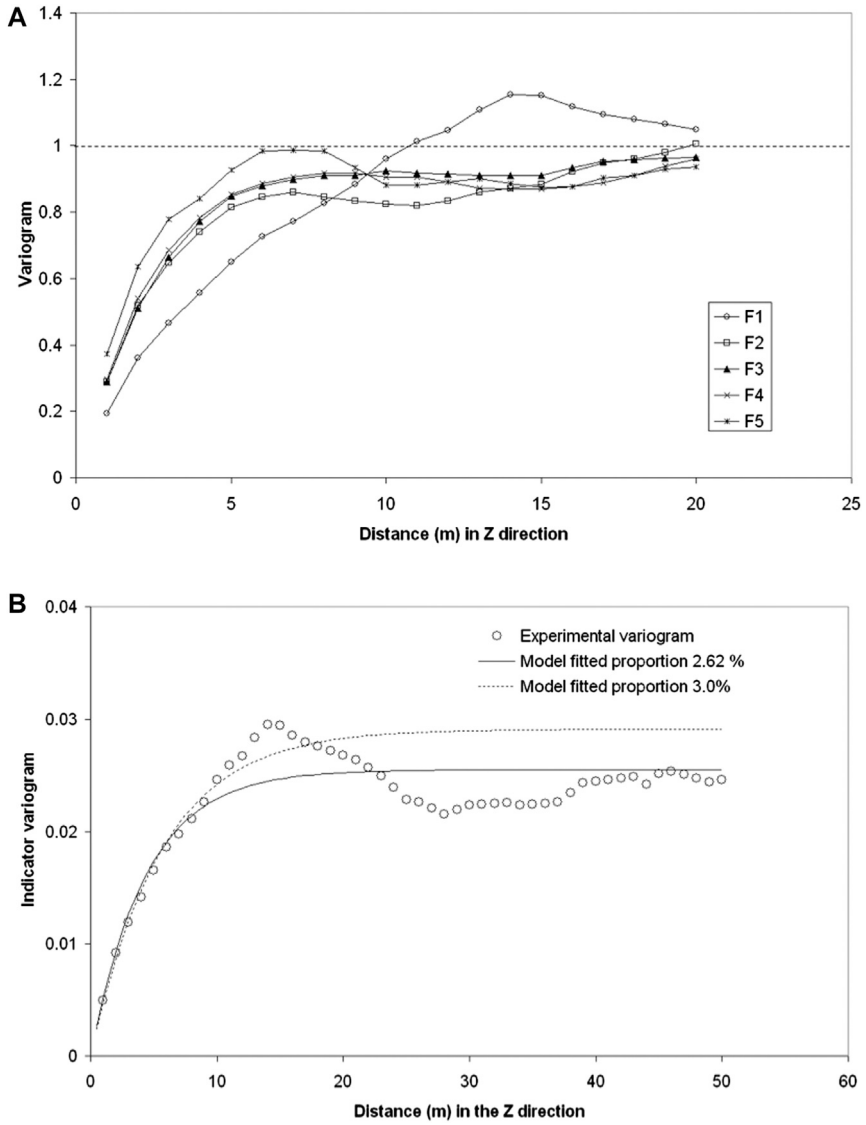


FIGURE 3 | (A) Standardized indicator variograms calculated along the boreholes for the five hydrofacies and standardized by their variances. **(B)** Indicator variogram of facies F1 along the borehole and the two best models fitted which were those for hydro facies F1 proportions of 2.62 and 3.00% which implies variances of 0.0255 and 0.0291, respectively, and fitted ranges of 4.37 and 5.62 m, respectively, for an exponential variogram model.

Although the locations of the boreholes are not ideal for detecting horizontal anisotropies, there is no evidence of anisotropy in the horizontal plane as demonstrated by the example of the east-west and north-south indicator variograms for hydrofacies F5 in **Figure 5**. However, because of the physical structure of deltaic deposits, anisotropies are expected between vertical and horizontal directions and these should be evident in the variograms, i.e., spatial correlation in the horizontal plane should be greater than that in the vertical direction. Given the lack of evidence for anisotropy in the horizontal plane, omni-directional horizontal variograms were calculated and modeled. The exponential models fitted by weighted least squares to the omni-directional variograms in the horizontal plane for the different hydrofacies have ranges of 6.8, 27.5, 35.8, 25.5, and 25.4 m for hydrofacies F1 to F5, respectively. The very high permeability hydrofacies (F1) has an effective range, or spatial correlation length, of 21 m and the rest of the hydrofacies have effective ranges of 75–105 m.

A plausible explanation of these variograms is that the effective ranges quantify the average horizontal extents of permeability channels. If, for example, permeability is due to paleochannels formed from meandering streams then the greater the horizontal extent of the hydrofacies the greater is the likelihood for hydraulic barriers to have formed. Thus, the highest permeability facies F1 has the smallest horizontal extent (21 m) and the lower permeability facies have horizontal extents greater than 75 m.

For the co-regionalization model of the five hydrofacies, **Figure 6** shows the experimental indicator cross-variograms

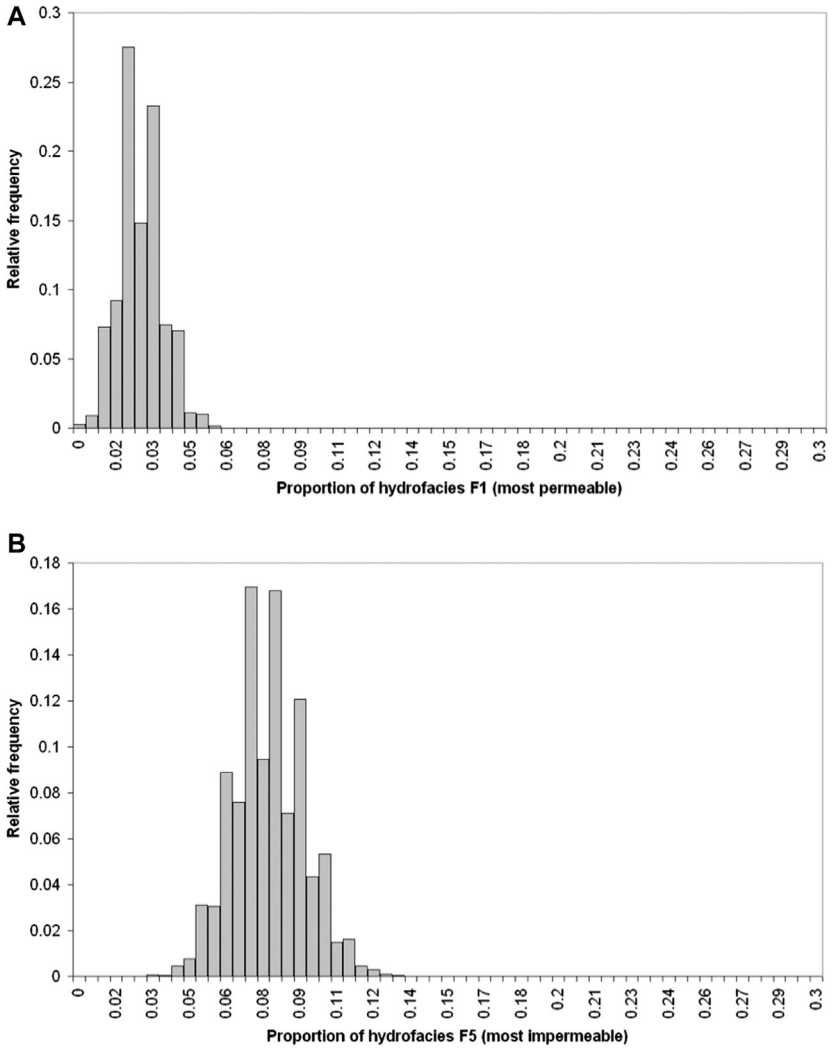


FIGURE 4 | Histogram of the bootstrap distribution for **(A)** F1 hydrofacies and **(B)** F5 hydrofacies.

TABLE 1 | Statistics of the bootstrap distributions for the different hydrofacies.

Hydrofacies		Mean (%)	Median (%)	95% CI lower limit (%)	95% CI upper limit (%)
F1	Very permeable	2.65	2.67	1	4
F2	Permeable	43.70	43.60	38	49
F3	Low permeability	27.10	27.00	22	32
F4	Impermeable	18.50	18.60	14	23
F5	Very impermeable	7.80	7.70	5	11

TABLE 2 | Experimental proportions and first ten bootstrap realisations of proportions.

	Experimental	Bootstrap samples								
F1	0.0262	0.0267	0.0233	0.0367	0.0333	0.0300	0.0233	0.0300	0.0300	0.0267	0.0333
F2	0.4373	0.4467	0.4200	0.4267	0.3967	0.4033	0.4367	0.4333	0.4633	0.3767	0.4000
F3	0.2721	0.2933	0.2867	0.2500	0.2700	0.2700	0.2500	0.2867	0.2533	0.3067	0.2867
F4	0.1857	0.1467	0.1867	0.2067	0.2100	0.2300	0.2033	0.1667	0.1833	0.2233	0.1933
F5	0.0787	0.0867	0.0833	0.0800	0.0900	0.0667	0.0867	0.0833	0.7000	0.0667	0.0867

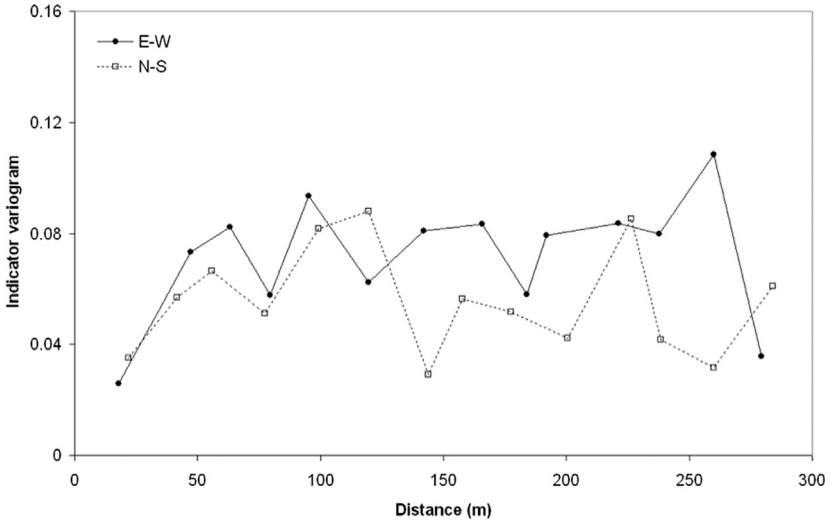


FIGURE 5 | Indicator variograms for hydrofacies F5 for the two perpendicular directions most likely to exhibit anisotropic behavior.

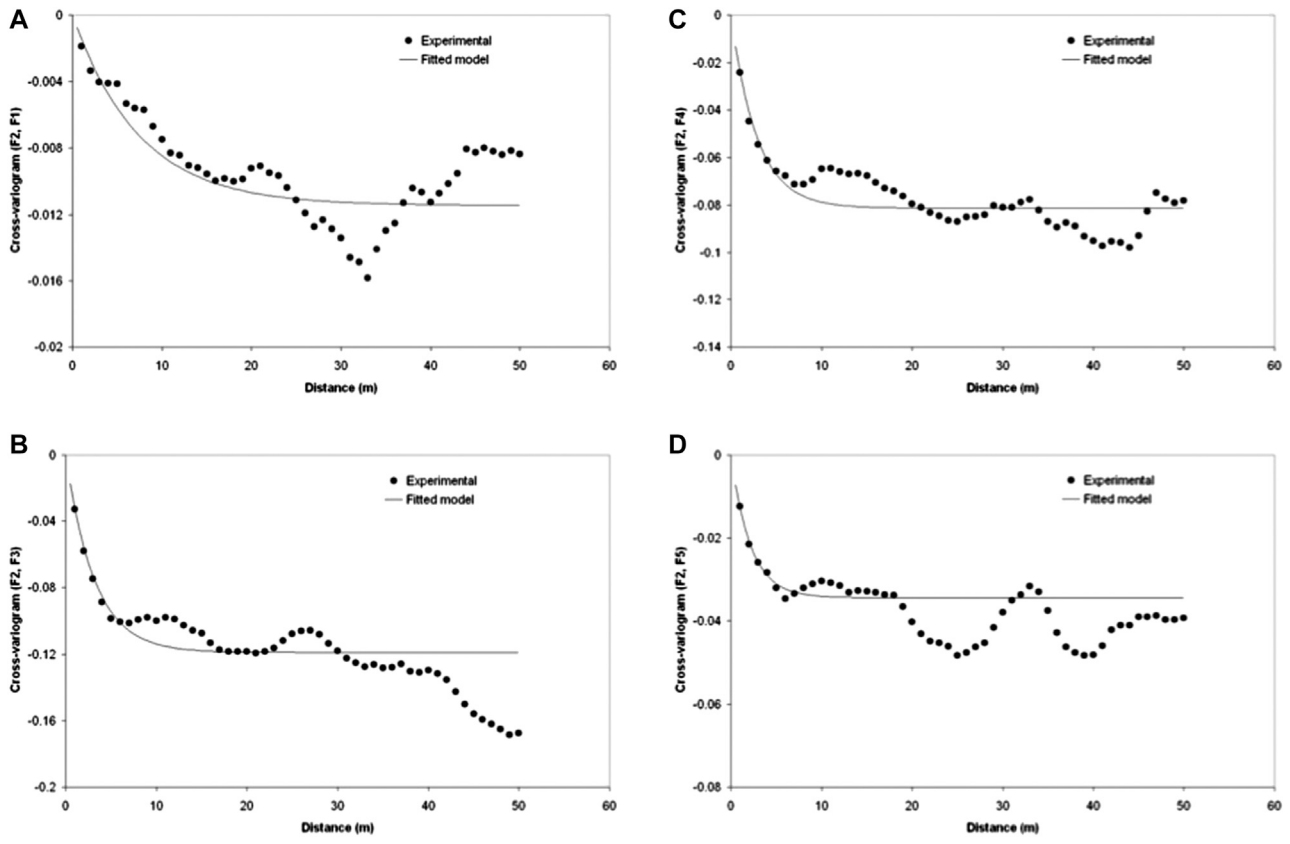


FIGURE 6 | Indicator cross-variograms between hydrofacies F2 and the other hydrofacies and models fitted to them: **(A)** F2-F1; **(B)** F2-F3; **(C)** F2-F4 and **(D)** F2-F5.

along the boreholes between hydrofacies F2 and the other four hydrofacies and the models fitted to them. **Figure 7** shows the experimental indicator variogram along the boreholes of hydrofacies F2 and the model fitted to it. The models were fitted by taking account of the experimental proportions and using weighted least squares to fit an exponential model with no

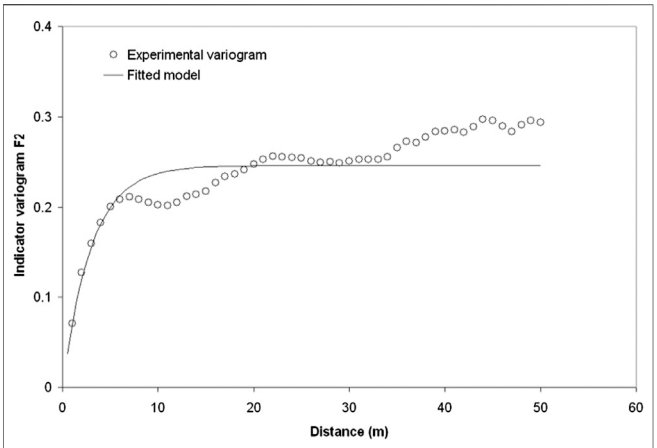


FIGURE 7 | Indicator variograms for hydrofacies F2 and fitted model.

nugget. The model fitted to the indicator cross-variogram between hydrofacies F2 and F1 has a range of 7.45 m while the indicator direct-variograms of those hydrofacies have ranges of 3.0 and 4.37 m, respectively. However, these ranges are not compatible with the linear model of coregionalization in which the range of the cross-variogram cannot be greater than the ranges of the direct variograms. The consequence of failing to meet this requirement can be demonstrated in terms of the conditional probabilities of occurrence. If, for example, hydrofacies F2 is observed at an arbitrary location u, what is the probability that each of the hydrofacies occurs at, say, 5 m further down the borehole?

These probabilities can be calculated from the models fitted to the direct and cross variograms as follows:

Probability of F2 occurring:

$$\frac{C_2(5)}{p_2} = \frac{(p_2 - \gamma_2(5))}{p_2} = \frac{(0.4372 - \gamma_2(5))}{0.4372} = 0.543$$

Probability of F1 occurring:

$$\frac{C_{21}(5)}{p_{21}} = \frac{\gamma_{21}(5)}{p_1} = 0.213$$

Probability of F3 occurring:

$$\frac{C_{21}(5)}{p_{23}} = \frac{-\gamma_{23}(5)}{p_3} = 0.347$$

Probability of F4 occurring:

$$\frac{C_{24}(5)}{p_{24}} = \frac{-\gamma_{24}(5)}{p_4} = 0.361$$

Probability of F5 occurring:

$$\frac{C_{25}(5)}{p_{25}} = \frac{-\gamma_{25}(5)}{p_5} = 0.394$$

The sum of these probabilities is 1.86 rather than 1.0 as it should be and thus the models fitted in **Figures 6, 7** are not valid when taken jointly. The ranges could be modified to ensure that the total probability is 1.0 by increasing the ranges of the indicator cross-variograms. However, this is not a general solution because the example given is very specific: it is only for one hydrofacies, for the Z-direction variogram, using only one conditioning point and for a specific distance of 5 m. Fitting models that would satisfy all possible constraints would be very cumbersome.

If only the direct variograms are used, the probability of F2 occurring remains the same:

$$\frac{C_2(5)}{p_2} = \frac{(p_2 - \gamma_2(5))}{p_2} = \frac{(0.4372 - \gamma_2(5))}{0.4372} = 0.543$$

and the complementary probability 1–0.543 = 0.457 is distributed among F1, F3, F4, and F5 in proportion to their prior probabilities of 0.0262, 0.2721, 0.1857, and 0.0787, respectively. Using these values in the previous equation would reproduce the correct variograms.

The following procedure was adopted for sequential simulation:

(a) Retain each bootstrap realization of the hydrofacies proportions that is inside the 95% confidence interval $\{\widehat{p}_i > 0,\ i = 1, \ldots, 5\}$. This automatically satisfies the requirement: $\sum_{i=1}^{5} \widehat{p}_i = 1$.
(b) Using $\widehat{p}_i(1 - \widehat{p}_i)$ for the sill of each direct variogram and $-\widehat{p}_i\widehat{p}_j$ for the sill of each cross-variogram, determine, by least squares, the ranges that give the best fit of exponential model variograms to the experimental variograms.
(c) Apply the sequential simulation algorithm.
(d) Go to (a) and repeat to generate a specified number of simulations.

For the experimental proportions, the least squares fit to the direct and cross-variograms of an exponential model without a nugget gives the ranges shown in **Table 3**. The effective ranges, or distances beyond which correlations are effectively zero, are three times these values.

There are three general possibilities for modeling the conditional probabilities of categorical variables (Goovaerts, 1994) for the purpose of simulating them by sequential indicator simulation: median indicator kriging, multiple indicator kriging and indicator co-kriging. Median indicator kriging uses the same indicator variogram model

TABLE 3 | Ranges of an exponential variogram model fitted to the experimental variogram by weighted least squares taking into account the experimental proportions.

Hydrofacies	Range in direction Z (m)	Range in X-Y plane (m)	Sill
F1 (p_1 = 0.0262)	4.37	6.78	$p_1(1 - p_1)$ = 0.0255
F2 (p_2 = 0.4373)	3.00	27.50	$p_2(1 - p_2)$ = 0.2461
F3 (p_3 = 0.2721)	2.87	35.78	$p_3(1 - p_3)$ = 0.1981
F4 (p_4 = 0.1857)	2.73	25.48	$p_4(1 - p_4)$ = 0.1512
F5 (p_5 = 0.0787)	2.05	25.41	$p_5(1 - p_5)$ = 0.0725

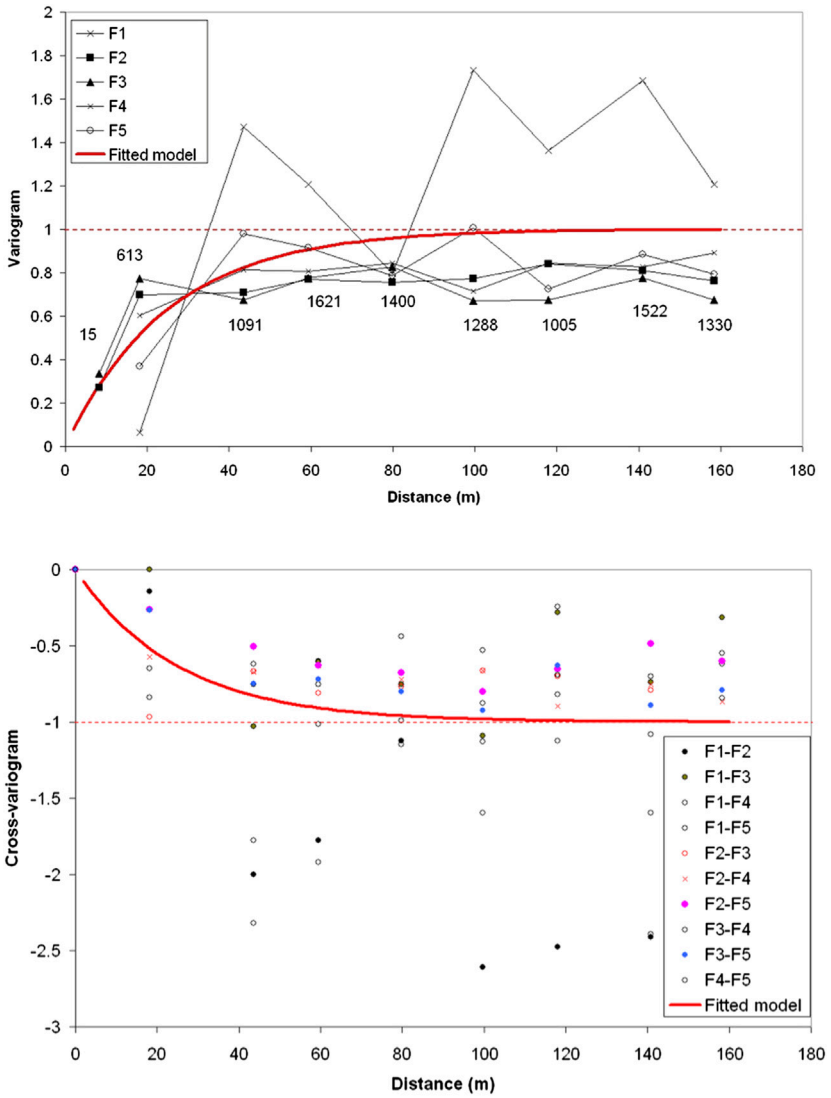


FIGURE 8 | (A) Standardized indicator direct variogram for the five hydrofacies and fitted model. **(B)** Standardized indicator cross-variograms for the five hydrofacies: ten cross-variograms and fitted model. The fitted model is an intrinsic model of coregionalization. The model is exponential with range 25 m (effective range of 75 m).

for all categories; this common model is a type of mean, or median, model. The second possibility is multiple indicator kriging using variogram models fitted to each hydrofacies indicator.

The third possibility is indicator co-kriging in which the main difficulty is to fit a valid coregionalization model. The simplest coregionalization model is the linear model in which all direct, and cross, variograms are linear combinations of a set of basic direct variograms. The intrinsic coregionalization model is a special case of the linear model in which the basic direct variograms for each variable are identical. Allowing for the noisy nature of the empirical variograms and cross-variograms in **Figures 8A,B** for the horizontal plane (shown as dots in the figure), it is reasonable to fit the same model for all directions as shown by the red lines. A different model is fitted for the variograms in the vertical direction, but the model is the same for the direct and cross-variograms. As this model is an intrinsic model of coregionalization and, because all variables (the five indicators of the five hydrofacies) have been measured at all locations, it has the property of auto-krigeability (Wackernagel, 1994), which means that cokriging a variable from the set of coregionalized variables yields the same value as kriging.

The intrinsic correlation model can be written as:

$$\Gamma(h) = C\,\gamma(h) \tag{18}$$

where $\gamma(h)$ is an exponential model variogram with range 25 m as shown in **Figure 8**, and C is a positive definite matrix of coefficients:

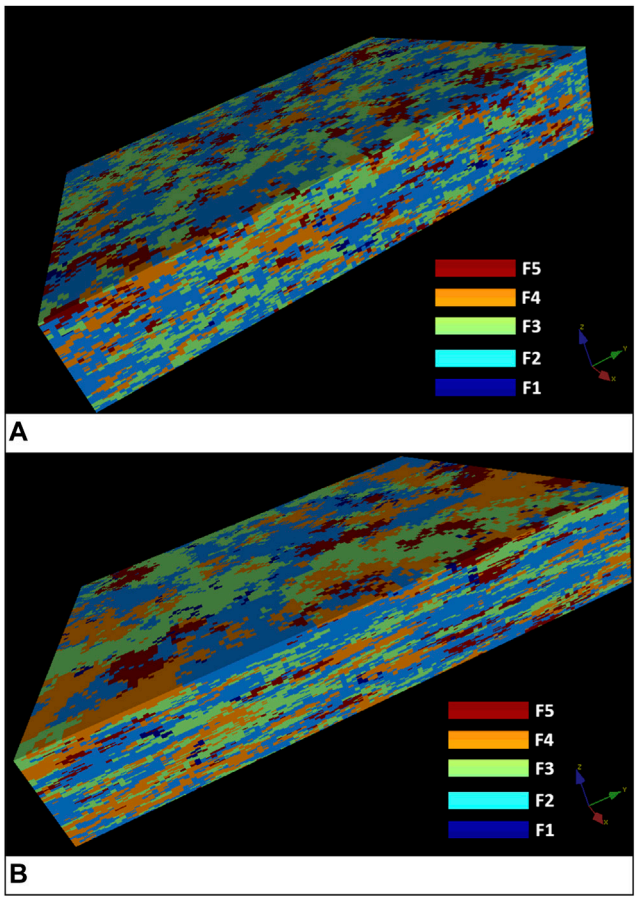


FIGURE 9 | 3D view of a realization generated by using the experimental proportions and **(A)** median indicator kriging and **(B)** multiple indicator kriging.

$$C = \begin{bmatrix} p_1(1-p_1) & -p_1p_2 & -p_1p_3 & -p_1p_4 & -p_1p_5 \\ -p_2p_1 & p_2(1-p_2) & -p_2p_3 & -p_2p_4 & -p_2p_5 \\ -p_3p_1 & -p_3p_2 & p_3(1-p_3) & -p_3p_4 & -p_3p_5 \\ -p_4p_1 & -p_4p_2 & -p_4p_3 & p_4(1-p_4) & -p_4p_5 \\ -p_5p_1 & -p_5p_2 & -p_5p_3 & -p_5p_4 & p_5(1-p_5) \end{bmatrix} \quad (19)$$

Furthermore, from **Eq. 10** these coefficients must be such that:

$$\begin{aligned} \gamma_1(h) &= -\gamma_{12}(h) - \gamma_{13}(h) - \gamma_{14}(h) - \gamma_{15}(h) \\ \gamma_2(h) &= -\gamma_{21}(h) - \gamma_{23}(h) - \gamma_{24}(h) - \gamma_{25}(h) \\ \gamma_3(h) &= -\gamma_{31}(h) - \gamma_{32}(h) - \gamma_{34}(h) - \gamma_{35}(h) \\ \gamma_4(h) &= -\gamma_{41}(h) - \gamma_{42}(h) - \gamma_{43}(h) - \gamma_{45}(h) \\ \gamma_5(h) &= -\gamma_{51}(h) - \gamma_{52}(h) - \gamma_{53}(h) - \gamma_{54}(h) \end{aligned} \quad (20)$$

Using median indicator kriging gives unrealistic outputs, in particular hydrofacies models with connectivities that are too low to generate observed flows. In this work, multiple indicator kriging was used together with the auto-krigeability model in which the cross-variogram of each variable with all other variables is proportional to the variogram of that variable.

This approach allows for different spatial variability for each hydrofacies and it generates realisations with significantly higher connectivity than those generated by median indicator kriging as can be seen by comparing **Figures 9A,B**. **Figure 10** shows three-dimensional views of four simulations using the experimental proportions; these views provide a better appreciation of the relationships between the different hydrofacies and of the real heterogeneity of the aquifer. This heterogeneity may influence the spatial patterns of seawater intrusion and thus condition the spatial distribution of water quality.

Some examples of the direct- and cross-variograms reproduced by the method are shown in **Figures 11, 12** for the Z-direction and the horizontal plane, respectively. The ergodic fluctuations of the variograms of the simulation may not seem large enough to include the experimental variogram, but it should be remembered that only the experimental proportions were used and not the bootstrap proportions. When the bootstrap proportions are used, the sills of the variogram models used in the simulation change, as do the ranges, and this increases the ergodic fluctuation to reflect the unknown real variability of the hydrofacies. This would be apparent in an application in which thousands of simulations are used in order to include the spatial uncertainty of the hydrofacies in the simulation.

Finally, **Figures 13A,B** show the connectivity function (Pardo-Igúzquiza and Dowd, 2003) for the five hydrofacies in the vertical direction and on the horizontal plane, respectively. For hydrofacies facies F1, for example, there are connectivities of up to 40 m in the vertical direction and 100 m in the horizontal direction, which, because these are very high permeability channels, may have important consequences for water intrusion and/or for rapid propagation of contaminants.

DISCUSSION

The conditional geostatistical simulation of hydrofacies provides numerical models of aquifers that reproduce the observed spatial variability of the geological structures. These patterns of spatial variability significantly influence flow and transport modeling in coastal aquifers and, as a consequence, influence the assessment of seawater intrusion. This is because the spatial variability of the hydrofacies influences the mechanical macro-dispersion of the mixed zone in seawater intrusion.

This work shows that the uncertainty in hydrofacies proportions estimated from sparse data can be included in sequential simulation. These proportions condition the sills of the variogram models, the variability of which provides a measure of the impact of the uncertainty in fitting theoretical models.

Various approaches have been used to simulate categorical variables using indicators. Each of these uses different types of indicator kriging in the sequential simulation algorithm. The work presented in this paper uses indicator kriging, in which each hydrofacies has its own model, as a compromise between simpler options and the most complete option of full cokriging. The

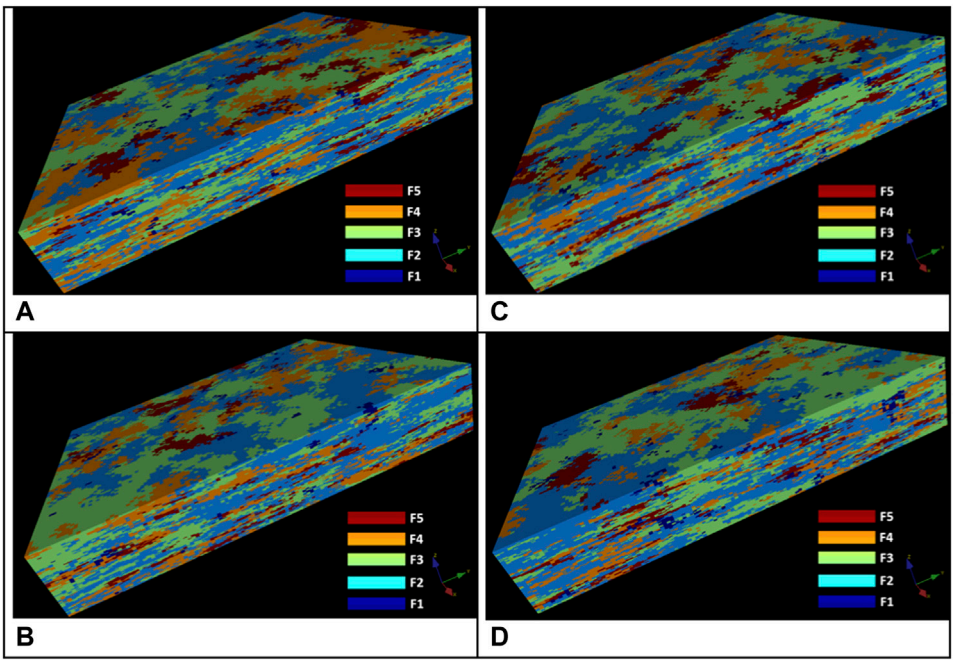


FIGURE 10 | 3D views of four simulations using the experimental proportions **(A)** simulation 1; **(B)** simulation 2; **(C)** simulation 3; **(D)** simulation 4.

connectivity function given in Pardo-Igúzquiza and Dowd (2003) has been used as a measure of connectivity although other connectivity indices could have been used, for example, Vassena et al. (2010) and Renard and Allard (2013).

Although the work presented here is limited to the simulation of categorical variables, it can easily be extended to include the generation of quantitative variables (e.g., hydraulic conductivity, porosity) by assigning a probability density function of the variable to each categorical hydrofacies. Borehole logging (natural gamma and resistivity are available) could be used to construct the probability functions. Another extension would be to use pumping test data for screening the simulated realisations. All these lines of research are left open for future work.

A fundamental question that arises from this study is whether the inclusion of the uncertainty of the hydrofacies proportions in the simulation (a large number of realisations considering the bootstrap proportions) gives results that are significantly different to those that would be obtained from a standard simulation (i.e. a large number of realisations with the same global proportion, but ignoring the uncertainty of the hydrofacies proportions). To answer this question, we used a simplified experiment to compare the outputs from the proposed extended sequential indicator simulation with those that would be obtained from standard sequential indicator simulation. The experiment comprised:

- A two-dimensional problem.
- Only two facies: the most permeable phase (0) and all others (1).
- Non-conditional simulations.
- 1,000 independent simulations.
- Identical variograms for the proposed extended sequential indicator simulation and the standard sequential indicator simulation.
- The same experimental proportions were used for all of the 1,000 standard sequential indicator simulations.
- 1,000 bootstrapped proportions were used for the 1,000 simulations using the proposed extended sequential indicator simulations.

An assessment of the connectivity of the two approaches showed that the means of the connectivity function statistics and the means of the connectivity function itself are very similar. However, the minimum and maximum numbers of connected components in a single simulation are significantly different with a greater range of variation in the simulations generated by the proposed extension. The minimum and maximum number of connected components for a standard sequential indicator simulation were 19 and 65, respectively whereas for the extended sequential indicator simulation they were 4 and 95, respectively. Thus, the variability is significatively larger when the uncertainty of the facies proportions is included. In addition, the same constant variograms (type of variogram and range) were used in both sets of 1,000 realisations. However, the ranges of the variograms will change in the proposed approach because the proportions are related to the sill and thus the ranges fitted to the experimental variograms may be different. This would further increase the variability of the extended sequential indicator simulations.

In this test, the propagation of uncertainty from the facies proportions into the set of simulations is significant and we conclude that the same must be so in the three-dimensional

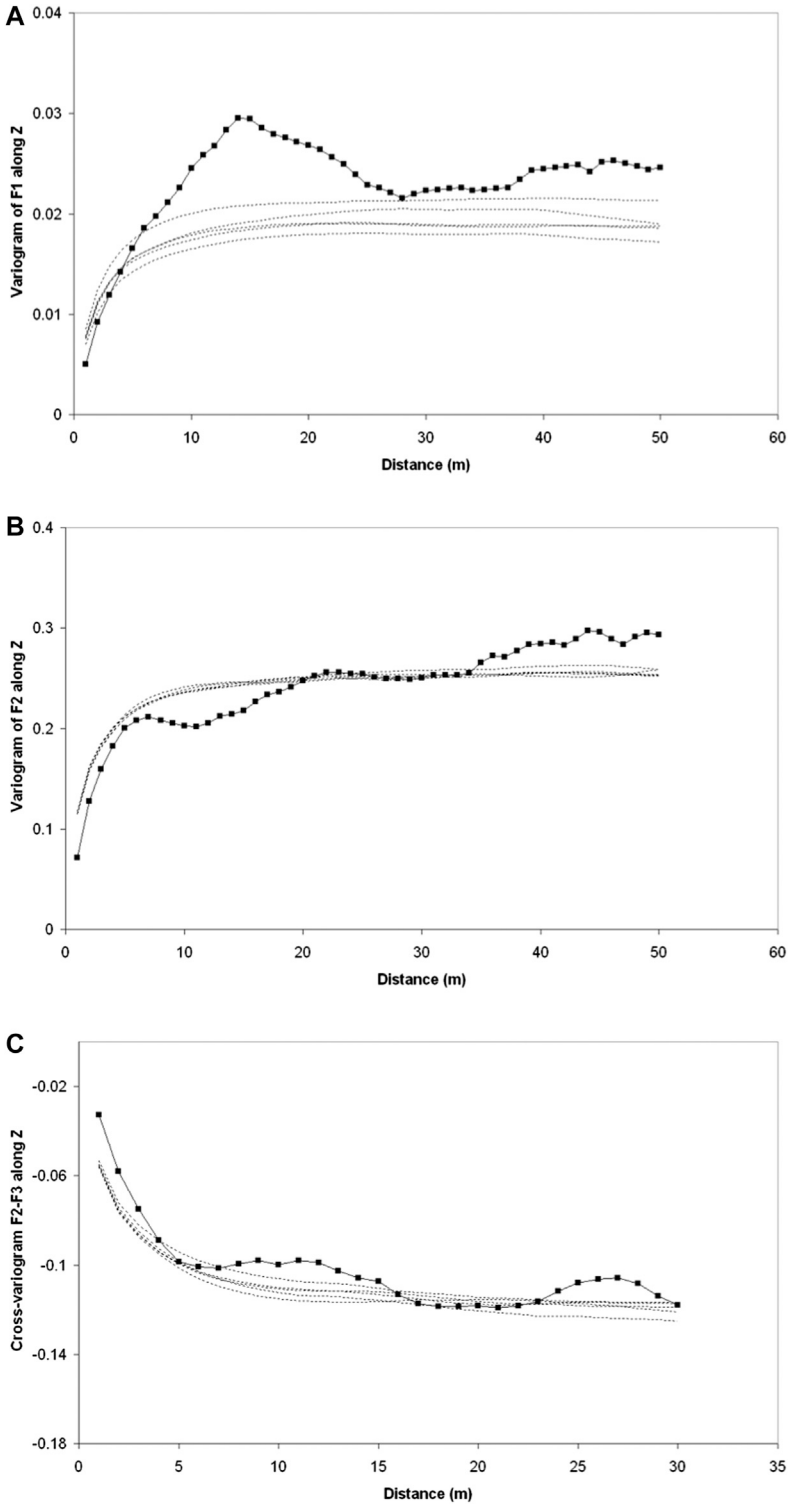


FIGURE 11 | Variograms of the borehole data (solid line) and the simulations (dashed lines) calculated for the vertical direction. **(A)** Variogram of hydrofacies F1. **(B)** Variogram of hydrofacies F2 and **(C)** Cross-variogram of hydrofacies F2 and F3.

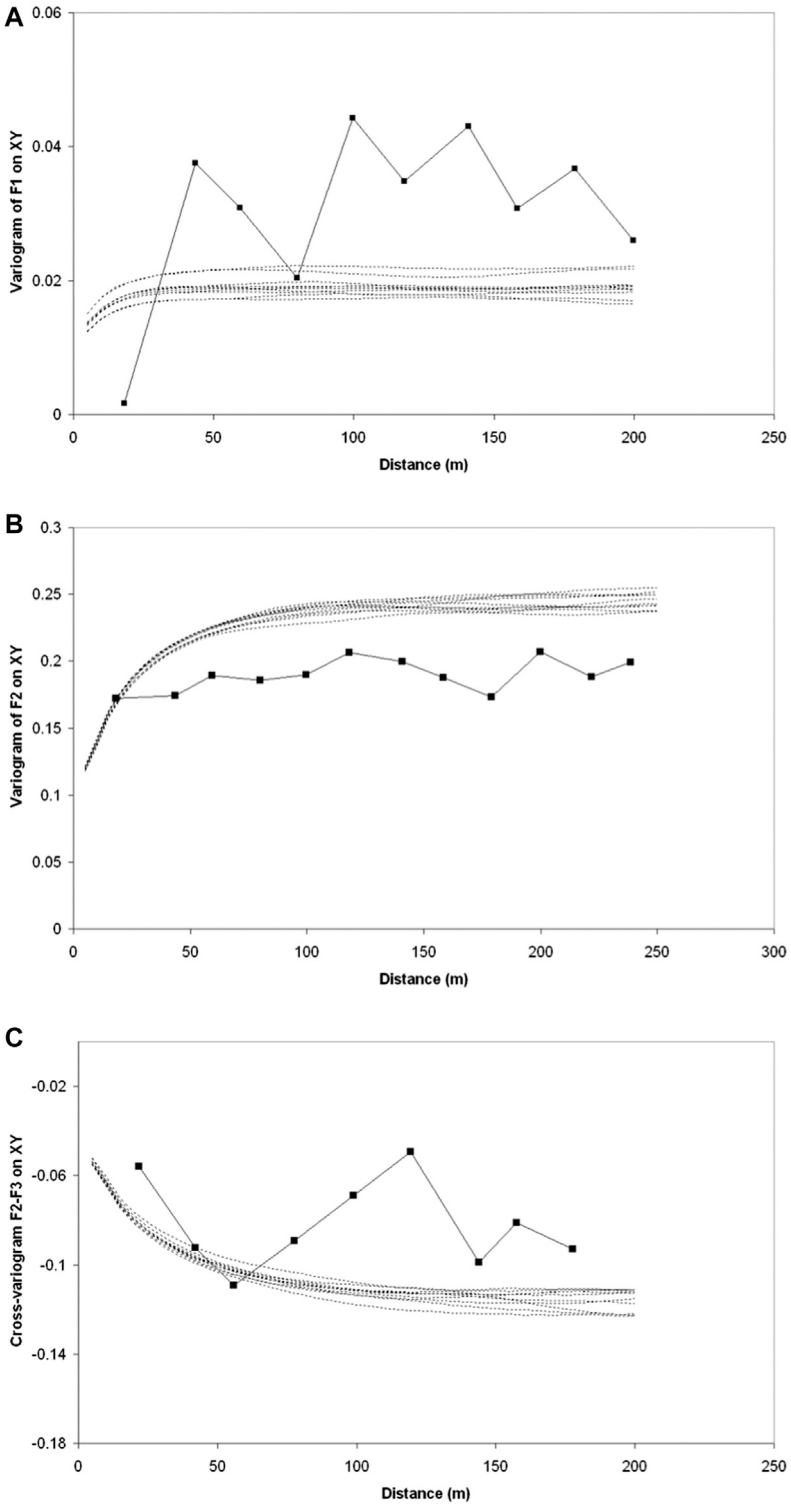


FIGURE 12 | Variogram of the borehole (solid line) data and from the simulations (dashed lane) calculated along the horizontal direction. **(A)** Variogram of hydrofacies F1. **(B)** Variogram of hydrofacies F2 and **(C)** cross-variogram of hydrofacies F2 and F3.

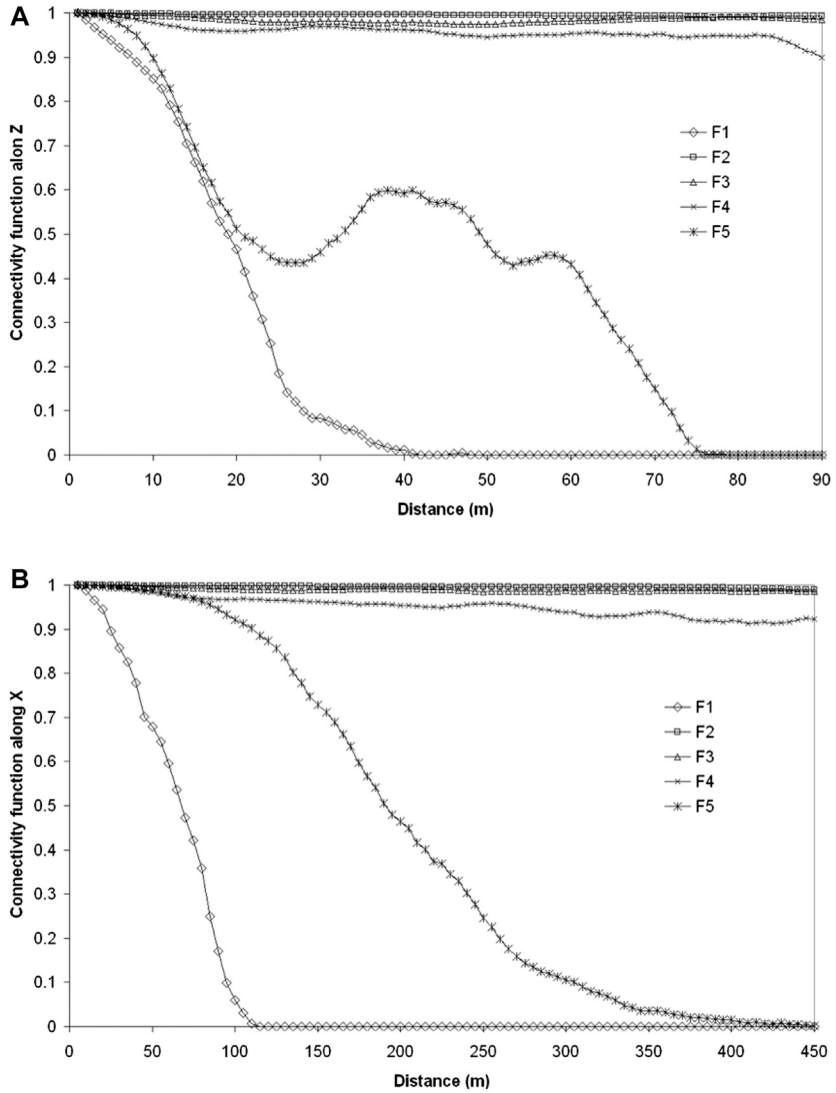


FIGURE 13 | Connectivity function for the different hydrofacies. **(A)** Vertical direction. **(B)** *X* direction.

case with significantly more hydrofacies than were used in this demonstration.

CONCLUSIONS

An extension of sequential indicator simulation for simulating realisations of the three-dimensional distribution of deltaic hydrofacies has been proposed in this paper. The extension is the inclusion of the uncertainty of the proportions of each hydrofacies by using a bootstrap algorithm that generates feasible realisations of the proportions using an effective number of samples to evaluate that uncertainty. The actual proportions of the hydrofacies influence the direct and cross-variograms of the hydrofacies (using indicators) and thus affect the connectivity between them. This extension will allow a more realistic evaluation of the uncertainty of the underground geological medium which, in turn, will affect the simulation of flow and transport in coastal aquifers.

DATA AVAILABILITY STATEMENT

All datasets presented in this study are included in the article/ **Supplementary Material**.

AUTHOR CONTRIBUTIONS

SJ collected the data and studied the hydrofacies as part of her PhD. She participated in the design and structure of the paper. PD participated in the geostatistical analysis of the data and the writing of the paper. EP conducted the geostatistical simulations and participated in the writing of the paper. AP participated in the

design of the data collection, their interpretation and the drawing of the figures. FS participated in the geological interpretation of the study area and the data as well as in the interpretation of the results.

FUNDING

This work was supported by Research Projects CGL 2010–15498 and CGL 2007–63450 del Ministerio de Economía y Competitividad, Spain and by the Australian Research Council Discovery Grant DP110104766.

ACKNOWLEDGMENTS

We thank the four reviewers for their constructive criticism that has helped significantly to improve the final version of this paper.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feart.2020.563122/full#supplementary-material

REFERENCES

Barakat, M. K. A. (2010). Modern geophysical techniques for constructing a 3D geological model on the Nile Delta, Egypt. MS Dissertation. Berlin (Germany): Technische Universität Berlin, 158.

Bianchi, M., Zheng, C., Wilson, C., Tick, G. R., Liu, G., and Gorelick, S. M. (2011). Spatial connectivity in a highly heterogeneous aquifer: from cores to preferential flow paths. *Water Resour. Res.* 47, W05524. doi:10.1029/2009wr008966

Cabello, P., Cuevas, J. L., and Ramos, E. (2007). 3D modelling of grain size distribution in quaternary deltaic deposits (Llobregat Delta, NE Spain). *Geol. Acta.* 5, 231–244. doi:10.1344/105.000000296

Carle, S. F., and Fogg, G. E. (1996). Transition probability-based indicator geostatistics. *Math. Geol.* 28, 453–476. doi:10.1007/bf02083656

Chilès, J.-P., and Delfiner, A. (1999). *Geostatistics: modelling spatial uncertainty*. New York, NY: Wiley-Interscience, 695.

Comunian, A., Renard, P., Straubhaar, J., and Bayer, P. (2011). Three-dimensional high resolution fluvio-glacial aquifer analog - Part 2: geostatistical modeling. *J. Hydrol.* 405, 10–23. doi:10.1016/j.jhydrol.2011.03.037

Deidda, G. P., Ranieri, G., Uras, G., Consentino, P., and Martorana, R. (2006). Geophysical investigations in the flumendosa river delta, Sardinia (Italy) — seismic reflection imaging. *Geophysics* 71, 1JA–Z75. doi:10.1190/1.2213247

dell'Arciprete, D., Bersezio, R., Felleti, F., Giudici, M., Comunian, A., and Renard, P. (2012). Comparison of three geostatistical methods for hydrofacies simulation: a test on alluvial sediments. *Hydrogeol. J.* 20, 299–311. doi:10.1007/s10040-011-0808-0

Deutsch, C. V., and Journel, A. G. (1998). *GSLIB: geostatistical software library and user´s guide*. 2nd Edn. New York, NY: Oxford University Press, 340.

Dowd, P. A., Pardo-Igúzquiza, E., and Xu, C. (2003). Plurigau: a computer program for simulating spatial facies using the truncated plurigaussian method. *Comput. Geosci.* 29, 123–141. doi:10.1016/s0098-3004(02)00070-5

Dubrule, O., and Damsleth, E. (2001). Achievements and challenges in petroleum geostatistics. *Petrol. Geosci.* 7, S1–S7. doi:10.1144/petgeo.7.s.s1CrossRef Full Text |

Eaton, T. T. (2006). On the importance of geological heterogeneity for flow simulation. *Sediment. Geol.* 184, 187–201. doi:10.1016/j.sedgeo.2005.11.002

Emery, X. (2004). Properties and limitations of sequential indicator simulation. *Stoch. Environ. Res. Risk Assess.* 18, 414–424. doi:10.1007/s00477-004-0213-5

Falivene, O., Cabrera, L., Muñoz, J. A., Arbués, P., Frenández, O., and Sáez, A. (2007). Statistical grid-based facies reconstruction and modelling for sedimentary bodies. Alluvial-palustrine and turbiditic examples. *Geol. Acta.* 5, 199–230. doi:10.1344/105.000000295

Goovaerts, P. (1994). "Comparison of COIK, IK and MIK performances for modelling conditional probabilities of categorical variables," in *Geostatistics for the next century*. Editor R. Dimitrakopoulos, The Netherlands: Kluwer Academic Publishers, 18–29.

Goovaerts, P. (1997). *Geostatistics for natural resources evaluation*. Oxford, UK: Oxford University Press, 502.

Gouw, M. J. P. (2007). Alluvial architecture of fluvio-deltaic successions: a review with special reference to Holocene settings. *The Netherlands J. Geosci.* 86, 211–227. doi:10.1017/s0016774600077817

Jorreto-Zaguirre, S., Pulido-Bosch, A., Gisbert-Gallego, J., and Sánchez-Martos, F. (2005). Las diagrafías y la caracterización de la influencia de los bombeos de agua de mar sobre el acuífero del delta del Andarax (Almería). *Ind. Miner.* 362, 15–21.

Journel, A., and Alabert, F. (1989). Non-gaussian data expansion in the earth sciences. *Terra. Nova.* 1, 123–134, doi:10.1111/j.1365-3121.1989.tb00344.x.

Lee, S.-Y., Carle, S. F., and Fogg, G. E. (2007). Geologic heterogeneity and a comparison of two geostatistical models: sequential Gaussian and transition probability-based geostatistical simulation. *Adv. Water Resour.* 30, 1914–1932. doi:10.1016/j.advwatres.2007.03.005

Le Loc'h, G., Beucher, H., Galli, A., and Doligez, B. (1994). Improvement in the truncated gaussian method: combining several gaussian functions. Conference proceedings ECMOR IV, Fourth european conference on the mathematics of oil recovery, Røros, Norway, June, 7–10 1994 (European Association of Geoscientists & Engineers).

Le Loc'h, G., and Galli, A. (1996). "Truncated Pluri-Gaussian method: theoretical and practical points of view," in *Geostatistics Wollongong'96*. Editors Baafi, et al., (Dordrecht, Netherlands: Kluwer), Vol. 1, 211–222.

Mariethoz, G., Renard, P., Cornaton, F., and Jaquet, O. (2009). Truncated plurigaussian simulations to characterize aquifer heterogeneity. *Ground Water.* 47, 13–24. doi:10.1111/j.1745-6584.2008.00489.x

Myers, D. E. (1989). To be or not to be. . . stationary? That is the question. . .Stationary? That is the question. *Math. Geol.* 21, 347–362. doi:10.1007/bf00893695

Pardo-Igúzquiza, E., and Dowd, P. A. (2003). CONNEC3D: a computer program for connectivity analysis of 3D random set models. *Comput. Geosci.* 29, 775–785. doi:10.1016/s0098-3004(03)00028-1

Pardo-Iguzquiza, E., Chica-Olmo, M., Luque-Espinar, J. A., and Garcia-Soldado, M. J. (2009). Using semi-variogram parameter uncertainty in hydrogeological applications. *Groundwater* 41 (1), 25–34. doi:10.1111/j.1745-6584.2008.00494.x

Perulero, R., Guadagnini, L., Riva, M., Giudici, M., and Guadagnini, A. (2014). Impact of two geostatistical hydro-facies simulation strategies on head statistics under non-uniform groundwater flow. *J. Hydrol.* 508, 343–355. doi:10.1016/j.jhydrol.2013.11.009

Renard, P., and Allard, D. (2013). Connectivity metrics for subsurface flow and transport. *Adv. Water Resour.* 51, 168–196. doi:10.1016/j.advwatres.2011.12.001

Sanchez Martos, F., Bosch, A. P., and Calaforra, J. M. (1999). Hydrogeochemical processes in an arid region of Europe (Almeria, SE Spain). *Appl. Geochem.* 14, 735–745. doi:10.1016/s0883-2927(98)00094-8

Strebelle, S. (2002). Conditional simulation of complex geological structures using multiple-point statistics. *Math. Geol.* 34, 1–21. doi:10.1023/a:1014009426274

Tercier, P., Knight, R., and Jol, H. (2000). A comparison of the correlation structure in GPR images of deltaic and barrier-spit depositional environments. *Geophysics* 64, 1028–1340. doi:10.1190/1.1444807

Vassena, C., Cattaneo, L., and Giudici, M., (2010). Assessment of the role of facies heterogeneity at the fine scale by numerical transport experiments and connectivity indicators. *Hydrogeol. J.* 18, 651–668. doi:10.1007/s10040-009-0523-2

Wackernagel, H. (1994). Cokriging versus kriging in regionalized multivariate data analysis. *Geoderma* 62, 83–92. doi:10.1016/0016-7061(94)90029-9

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.

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

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 $\alpha = \mathsf{L}^{-1}$), 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:

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

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

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

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

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

in which h_{fix} is a known hydraulic head (dim $h_{\mathrm{fix}} = \mathsf{L}$), $\mathbf{n}$ is the dimensionless unit normal vector, q_{fix} is a known normal flux (dim $q_{\mathrm{fix}} = \mathsf{L}\,\mathsf{T}^{-1}$), h_{sim} is the simulated head if the boundary was considered a no-flow boundary ($h_{\mathrm{sim}} = \mathsf{L}$), z_{surf} is the surface elevation (dim $z_{\mathrm{surf}} = \mathsf{L}$), C_{L} is a river conductance (dim $C_{\mathrm{L}} = \mathsf{L}^2\,\mathsf{T}^{-1}$), V is the volume related to the source/sink term (dim $V = \mathsf{L}^3$), h_{riv} is a known river head (dim $h_{\mathrm{riv}} = \mathsf{L}$), C_{D} is a drainage conductance (dim $C_{\mathrm{D}} = \mathsf{L}^2\,\mathsf{T}^{-1}$) and Δz is a pressure difference threshold (dim$\Delta z = \mathsf{L}$). 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_{riv} = dim $w_{\mathrm{riv}} = \mathsf{L}$), L_{sed} is the thickness of the sediment bed (dim $L_{\mathrm{sed}} = \mathsf{L}$), and K_{sed} is the sediment's hydraulic conductivity (dim $K_{\mathrm{sed}} = \mathsf{L}\,\mathsf{T}^{-1}$).

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

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 ($\dim \mathbf{v} = \mathsf{L}\mathsf{T}^{-1}$), $\mathbf{x}_i(t)$ is the position vector ($\dim \mathbf{x}_i(t) = \mathsf{L}$) 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

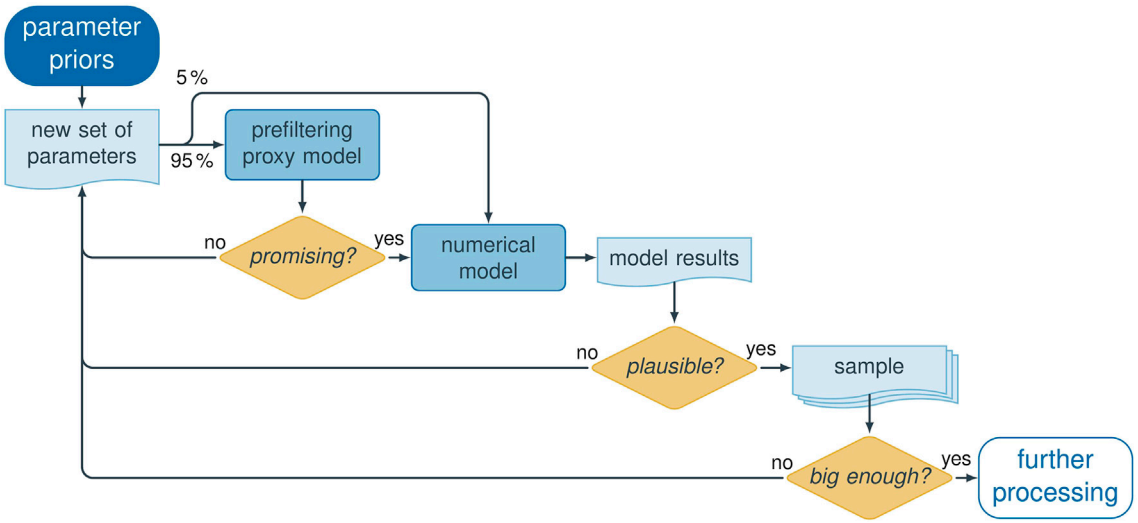


FIGURE 1 | Procedure used to generate a sample of physically plausible model realizations for the optimal experimental design analysis.

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 $\mathbf{z}$ 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 $\mathbf{z}$ 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, $\mathbf{z}$ 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:

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

in which $A_{2\text{D}}$ 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):

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

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 $1 \times n_{\text{par}}$

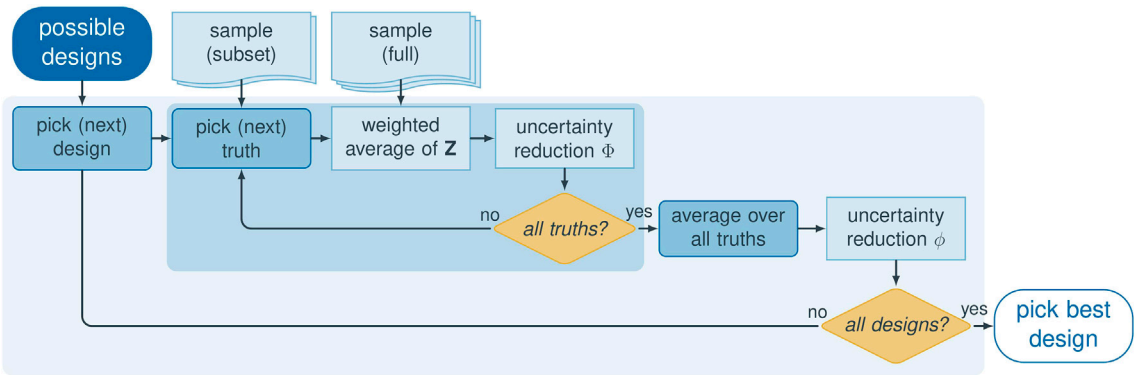


FIGURE 2 | Schematic illustration of the general preposterior data impact assessor procedure. Inner loop in dark blue, outer loop in light blue.

vector of prediction variables **z** (binary particle-fate map) as discussed above. The prediction solely depends on the input parameter vector **S** and is independent of the measurement design **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 **z** implies that the sample average of it equals the vector of probabilities that the individual elements of **z** are one.

After acquiring n_{sample} stage-2-accepted sample members of the parameter vector **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 **z** (which can be summarized in a $n_{\text{sample}} \times n_{\text{par}}$ matrix **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 **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 **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:

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

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

in which n_y is the number of virtual measurements according to the current design **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 **Z**, conditioned on the observations $\mathbf{y}_j(\mathbf{d})$ of the temporary true parameter set $\mathbf{S}_j$ according to the current design **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 **z** are one, conditioned on the vector of observations $\mathbf{y}_j(\mathbf{d})$ of realization j using the design **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 **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:

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

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 $\mathbf{d}_{\text{opt}}$ 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 $3 \cdot 10^{12}$ 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

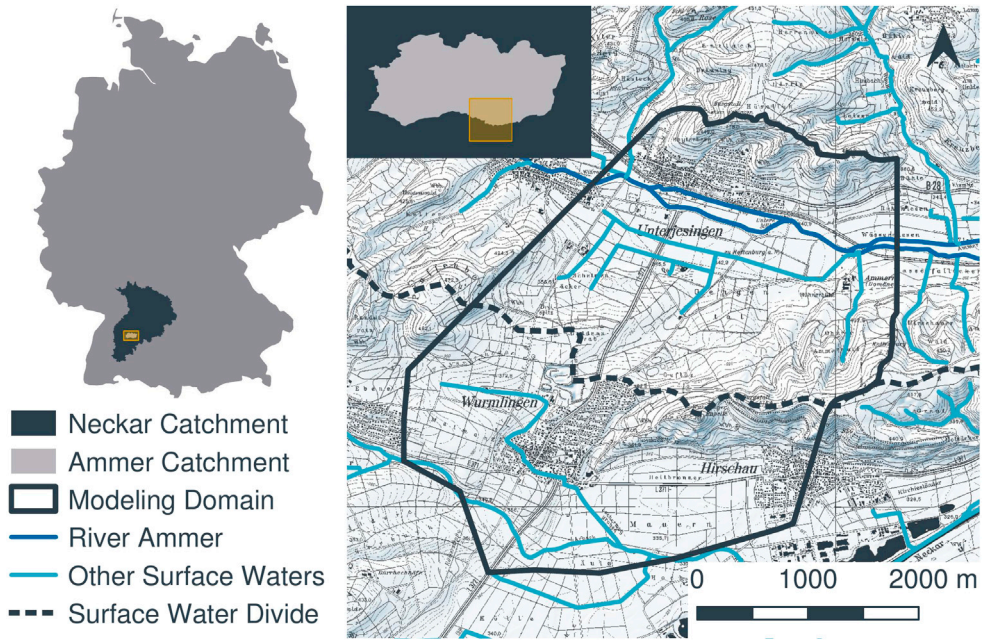


FIGURE 3 | Two-dimensional overview of the model domain and its location in Germany.

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.

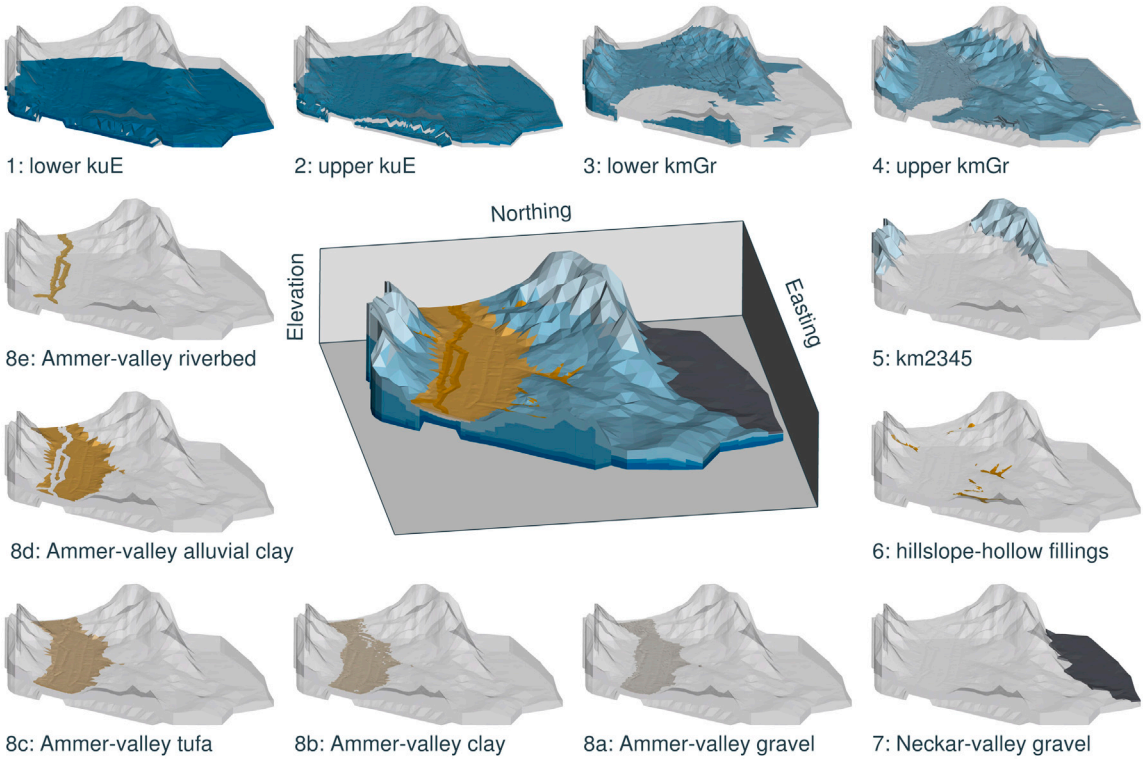


FIGURE 4 | Three-dimensional overview of the subsurface-flow model. The vertical dimension is exaggerated by a factor of five to enhance distinctness.

(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.

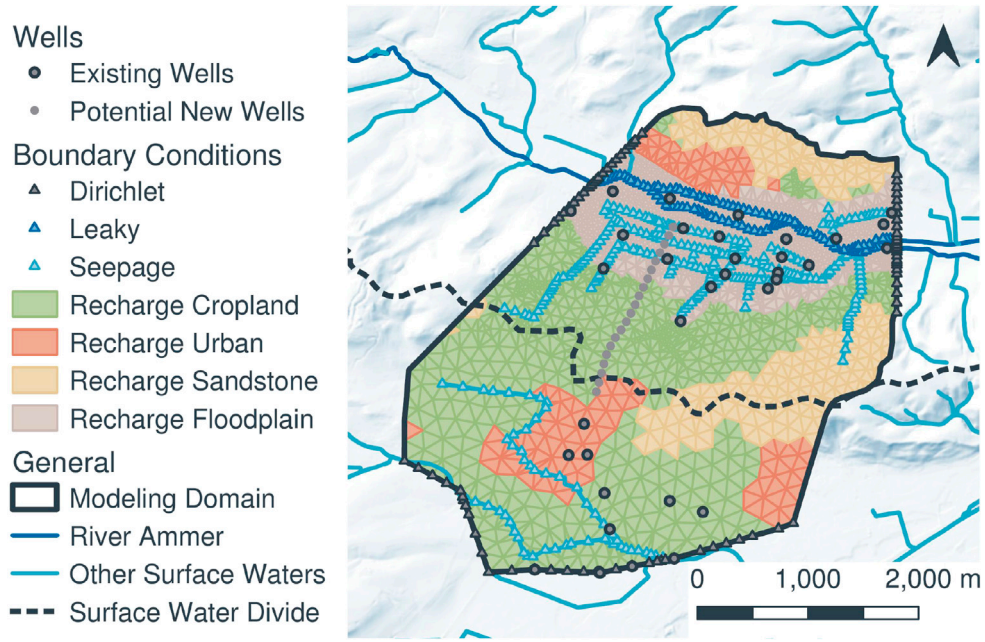


FIGURE 5 | Boundary conditions and two-dimensional discretization of the underlying subsurface flow model.

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 $S_{wr} = \Theta_r/\Theta_s$. 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_v/K_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: $S_{wr,1-8} = 5\,\%$, $S_{wr,9} = 17\,\%$, $S_{wr,10} = 18\,\%$, $S_{wr,11-12} = 25\,\%$.

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 [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, $h_{Ammer,in}$ and $h_{Ammer,out}$, 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 $K_{8d,h}$. For the drainage boundaries outside of the floodplain, we assume a soil layer of 0.20 m thickness and a hydraulic conductivity of $1 \cdot 10^{-6}$ m s^{-1}. The river boundary condition (see **Eq. 12**) uses K_{8e} for its conductivity and the geometry parameters $L_{riv} = 40$ m, $w_{riv} = 3$ m, and $L_{sed} = 0.5$ 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	$\log_{10}K_{1,h}$	−8.0	−6.0	$m\,s^{-1}$	—
#P2	$K_{2,h}$	$1/250 \cdot K_{1,h}$	$1/2 \cdot K_{1,h}$	$m\,s^{-1}$	—
#P3	$\log_{10}K_{3,h}$	−9.0	−6.3	$m\,s^{-1}$	—
#P4	$K_{4,h}$	$K_{3,h}$	$10^3 \cdot K_{3,h}$	$m\,s^{-1}$	—
#P5	$\log_{10}K_{5,h}$	−8.3	−7.0	$m\,s^{-1}$	—
#P6	$\log_{10}K_{6,h}$	−9.0	−3.0	$m\,s^{-1}$	—
#P7	$\log_{10}K_{7,h}$	−5.3	−3.0	$m\,s^{-1}$	—
#P8	$\log_{10}K_{8a,h}$	−5.3	−3.0	$m\,s^{-1}$	—
#P9	$\log_{10}K_{8b,h}$	−10.0	−7.0	$m\,s^{-1}$	—
#P10	$\log_{10}K_{8c,h}$	−5.3	−3.0	$m\,s^{-1}$	—
#P11	$\log_{10}K_{8d,h}$	−9.0	−5.3	$m\,s^{-1}$	—
#P12	$\log_{10}K_{8e}$	−8.0	−3.0	$m\,s^{-1}$	—
#P13	$K_{1,v}/K_{1,h}$	1/15	1	—	—
—	$K_{2,v}/K_{2,h}$	1/15	1	—	Coupled to #P13
#P14	$K_{3,v}/K_{3,h}$	1/15	1	—	—
#P15	$K_{4,v}/K_{4,h}$	1/15	1	—	—
#P16	$K_{5,v}/K_{5,h}$	1/15	1	—	—
#P17	$K_{6,v}/K_{6,h}$	1/5	1	—	—
#P18	$K_{7,v}/K_{7,h}$	1/5	1	—	—
#P19	$K_{8a,v}/K_{8a,h}$	1/5	1	—	—
#P20	$K_{8b,v}/K_{8b,h}$	1/15	1	—	—
#P21	$K_{8c,v}/K_{8c,h}$	1/15	1	—	—
#P22	$K_{8d,v}/K_{8d,h}$	1/15	1	—	—
#P23	α_{1-6}	0.50	5.00	m^{-1}	—
#P24	α_{8b}	0.01	0.10	m^{-1}	—
#P25	α_{8c}	8.00	12.00	m^{-1}	—
#P26	α_{8d}	0.50	0.70	m^{-1}	—
#P27	N_{1-6}	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	—	—

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

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

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 $\frac{1}{3} \leq \frac{|Q_{in}|}{|Q_{out}|} \leq 3$.

(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.

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

ID	Name	Minimum	Maximum	Unit	Comment
#B1	$R_{cropland}$	$1.5 \cdot 10^{-9}$	$8.0 \cdot 10^{-9}$	$m\,s^{-1}$	—
#B2	$R_{floodplain}/R_{cropland}$	0	1	—	Coupled to #B1
#B3	$R_{mud/sandstone}/R_{cropland}$	0	1	—	Coupled to #B1
#B4	$R_{urban}/R_{cropland}$	0.25	1	—	Coupled to #B1
#B5	Δh_{Neckar}	−0.50	0.50	m	—
#B6	Δh_{river}	−0.25	0.25	m	—
#B7	$h_{Ammer,in}$	346.0	347.0	m	—
#B8	$h_{Ammer,out} - h_{Ammer,in}$	−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

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

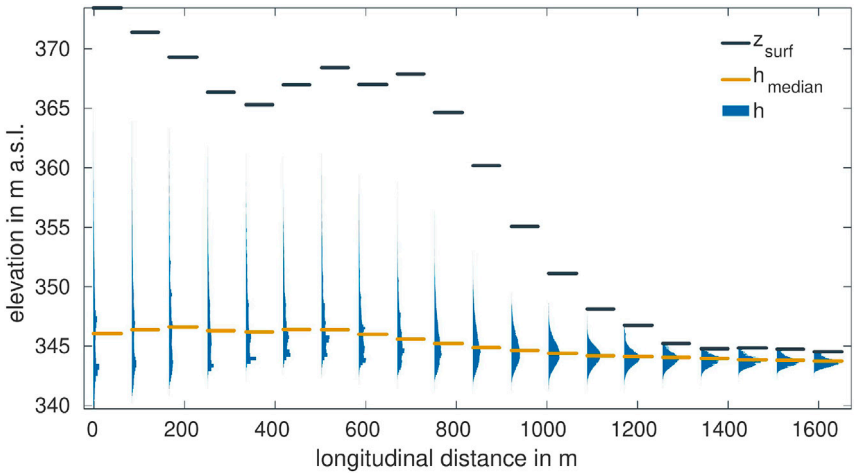


FIGURE 6 | Distributions of virtual hydraulic-head observations using the sample of stage-2-accepted realizations at all twenty potential locations along the transect.

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:

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

in which n_{pts} is the number of potential observation points. With $n_{\text{pts}} = 20$, **Eq. 28** results in a total of $n_{\text{des}} = 20 + 190 + 1140 = 1350$ 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 $n_{\text{sample}} = 50\,000$ 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 $n_{\text{sub}} = 10\,000$ 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

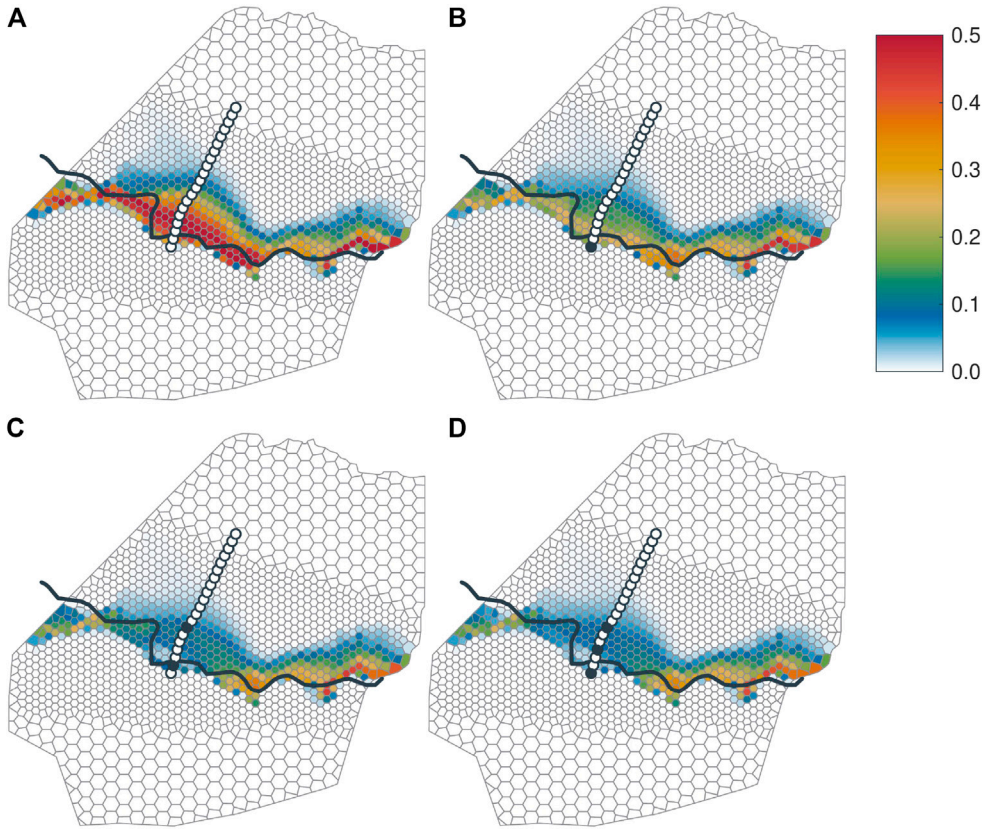


FIGURE 7 | Probability maps of misclassifying the attribution to the Ammer and Neckar subsurface-catchments $P_{mc} = 2P \cdot (1 - P)$. **(A)** Prior P_{mc}; **(B)**: P_{mc} for the best design with one additional piezometer; **(C)**: with two additional piezometers; **(D)**: with three additional piezometers.

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

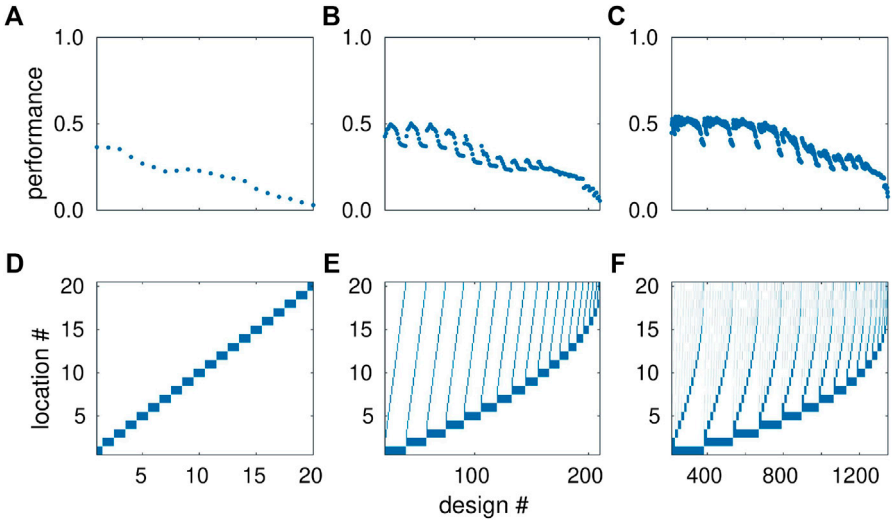


FIGURE 8 | Performance of all 1,350 investigated monitoring designs. Top row **(A–C)**: normalized utility function $\phi(\mathbf{d})$ of the given design according to **Eq. 26**; bottom row (**D–F**): piezometer combination of the given design. **(A,D)** Designs with one additional piezometer; **(B,D)** 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 $\approx$ 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 ($AESS_1 = 859.7$, $AESS_2 = 179.7$ and $AESS_3 = 68.1$), which means the sample of $n_{sample} = 50\,000$ 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,

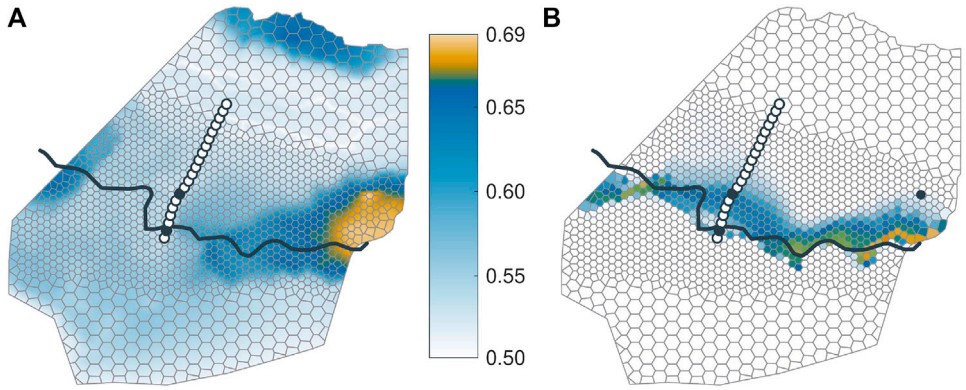


FIGURE 9 | Assessment of measurement designs with the third piezometer being placed outside the transect. **(A)** Performance ϕ of the design as a function of where the third observation well is placed; **(B)** map of the misclassification probability P_{mc} for the optimal three-piezometer design with 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.

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

frontiers in Earth Science

ORIGINAL RESEARCH
published: 27 November 2020
doi: 10.3389/feart.2020.565613

Early Uncertainty Quantification for an Improved Decision Support Modeling Workflow: A Streamflow Reliability and Water Quality Example

Brioch Hemmings[1]*, *Matthew J. Knowling*[1,2] and *Catherine R. Moore*[1]

[1]GNS Science, Wellington, New Zealand, [2]School of Civil, Environmental and Mining Engineering, University of Adelaide, Adelaide, South Australia, Australia

Effective decision making for resource management is often supported by combining predictive models with uncertainty analyses. This combination allows quantitative assessment of management strategy effectiveness and risk. Typically, history matching is undertaken to increase the reliability of model forecasts. However, the question of whether the potential benefit of history matching will be realized, or outweigh its cost, is seldom asked. History matching adds complexity to the modeling effort, as information from historical system observations must be appropriately blended with the prior characterization of the system. Consequently, the cost of history matching is often significant. When it is not implemented appropriately, history matching can corrupt model forecasts. Additionally, the available data may offer little decision-relevant information, particularly where data and forecasts are of different types, or represent very different stress regimes. In this paper, we present a decision support modeling workflow where early quantification of model uncertainty guides ongoing model design and deployment decisions. This includes providing justification for undertaking (or forgoing) history matching, so that unnecessary modeling costs can be avoided and model performance can be improved. The workflow is demonstrated using a regional-scale modeling case study in the Wairarapa Valley (New Zealand), where assessments of stream depletion and nitrate-nitrogen contamination risks are used to support water-use and land-use management decisions. The probability of management success/failure is assessed by comparing the proximity of model forecast probability distributions to ecologically motivated decision thresholds. This study highlights several important insights that can be gained by undertaking early uncertainty quantification, including: i) validation of the prior numerical characterization of the system, in terms of its consistency with historical observations; ii) validation of model design or indication of areas of model shortcomings; iii) evaluation of the relative proximity of management decision thresholds to forecast probability distributions, providing a justifiable basis for stopping modeling.

Keywords: uncertainty quantification, decision support, data assimilation, groundwater modeling, stream depletion, contaminant transport, resource management

OPEN ACCESS

Edited by:
Andrés Alcolea,
Independent Researcher, Switzerland

Reviewed by:
Alexander Maria Rogier Bakker,
Directorate-General for Public Works and Water Management, Netherlands
Tony Zhao,
Sun Yat-sen University, China

***Correspondence:**
Brioch Hemmings
b.hemmings@gns.cri.nz

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

Received: *25 May 2020*
Accepted: *05 October 2020*
Published: *27 November 2020*

Citation:
Hemmings B, Knowling MJ and Moore CR (2020) Early Uncertainty Quantification for an Improved Decision Support Modeling Workflow: A Streamflow Reliability and Water Quality Example. Front. Earth Sci. 8:565613. doi: 10.3389/feart.2020.565613

1. INTRODUCTION

Numerical models are routinely used to inform environmental management decision making by exploring possible system responses to proposed management strategies. Probabilistic assessment of these system responses are a further requirement of model-based decision support (e.g., Freeze et al., 1990; Doherty and Simmons, 2013). This allows the likelihood of any undesired impacts to be assessed alongside relevant management decision thresholds which are defined on the basis of ecological, economic and/or cultural objectives.

It is widely considered that history matching (also known as "model calibration" or "data assimilation") is a prerequisite for such decision support model deployment (e.g., Barnett et al., 2012). This follows the philosophy "How can a model be robust if it isn't calibrated?" This philosophy has its basis in the expectation that history matching, which can be considered an implementation of Bayes equation, will result in a reduction of parameter and predictive uncertainty (often expressed in terms of predictive variance; e.g., Moore and Doherty, 2005; Dausman et al., 2010). However, the ability of the history matching process to improve the reliability of parameter estimations, and to appropriately reduce decision-relevant forecast uncertainty, may be limited by a number of important factors.

First, the information content of observation datasets used for history matching may be deficient for inferring values of physically-based model parameters that represent spatially distributed subsurface properties (e.g., Vasco et al., 1997; Vasco et al., 1998; Clemo et al., 2003; Moore and Doherty, 2006), with subsequent implications for reducing forecast uncertainty. This is particularly the case where there is significant hydraulic property heterogeneity, which is typically the major cause of predictive uncertainty in groundwater models. Even where there is significant data available, a lack of forecast relevant data may still inhibit efforts to reduce uncertainty through history matching.

Second, a number of studies have shown that there is potential for history matching to corrupt the quantification of predictive uncertainty, when considering the inevitable presence of model defects (e.g., Doherty and Christensen, 2011; Brynjarsdóttir and O'Hagan, 2014; White et al., 2014; Oliver and Alfonzo, 2018). For example, Knowling et al. (2019) show that there is potential to induce bias and corruption in estimates of predictive variance, as a consequence of inappropriate parameter compensation when history matching "under-parameterized" models. Ultimately, this can compromise the reliability of decisions made on the basis of model forecasts. Additionally, Knowling et al. (2020) showed that there is significant potential to induce predictive bias due to the inability of an imperfect model to appropriately assimilate information-rich data, such as environmental tracer observations. Therefore, when faced with using a model with prediction relevant imperfections, a modeler may wish to critically consider whether history matching is appropriate.

History matching efforts can also place unexpected requirements on labor and computational resources. This is because balancing the expression of prior knowledge with information contained within observation data, during history matching, is not straightforward and is often undertaken in a highly iterative process. This adds significant complexity to the modeling workflow, and risk to project budgets and time-lines. This is particularly the case for large-scale models of real-world systems which are prone to numerical stability issues and long simulation durations.

An additional consideration when examining the usefulness of a history matching effort in a decision support context, is the proximity of forecast probability distributions to predefined management decision thresholds (e.g., Knowling et al., 2019; White et al., 2020b). Evaluation of the relative positions of forecast distributions and decision thresholds can be used to provide a measure of apparent "decision difficulty" (e.g., White et al., 2020b). Where, on the basis of the prior information alone, a forecast probability distribution lies far from the critical threshold (i.e decision difficulty trends to zero), a modeler may elect not to progress with history matching, as it is unlikely to alter the evaluation of management strategy effectiveness.

Finally, uncertainty quantification that underpins decision support relies upon robust definition of the prior parameter probability distributions that express system expert knowledge (i.e., before history matching). Verification of robust prior probability distributions is difficult, especially for complex real-world numerical models, often requiring prohibitively expensive paired model analyses (e.g., Doherty and Christensen, 2011; Gosses and Wöhling, 2019). However, an indication of an inappropriate expression of prior parameter uncertainty, with the potential to undermine model-based decision support, can be provided through comparison of historical system observations with simulated outputs generated on the basis of the prior parameter distributions, in a "prior-data conflict assessment" (e.g., Nott et al., 2016). Formally, prior-data conflict can be identified when there is no overlap between an ensemble of observation data, accounting for potential observation error, and the prior ensemble of simulated outputs. This analysis can be undertaken early in a modeling project, so that indications of an inappropriate prior can be identified and addressed ahead of a full comprehensive history matching process. We note that this criterion alone is not sufficient to demonstrate that the initial prior ensemble is valid, as model error may also be the cause of such prior-data conflict, requiring additional statistical tests (e.g., Brynjarsdóttir and O'Hagan, 2014; Alfonzo and Oliver, 2019).

In this paper we explore the benefits of recasting the typical modeling workflow, which starts with a conceptual system model and ends with a calibrated numerical model (e.g., Barnett et al., 2012), such that uncertainty quantification is undertaken at an early stage in the project, before attempting comprehensive history matching. The recast workflow involves exploration of the prior decision-relevant forecast uncertainty relative to decision thresholds, as well as prior distributions of model outputs relative to system observations. The workflow then involves undertaking an abridged history matching and a preliminary posterior uncertainty assessment, to help identify the extent to which available data informs estimated parameter

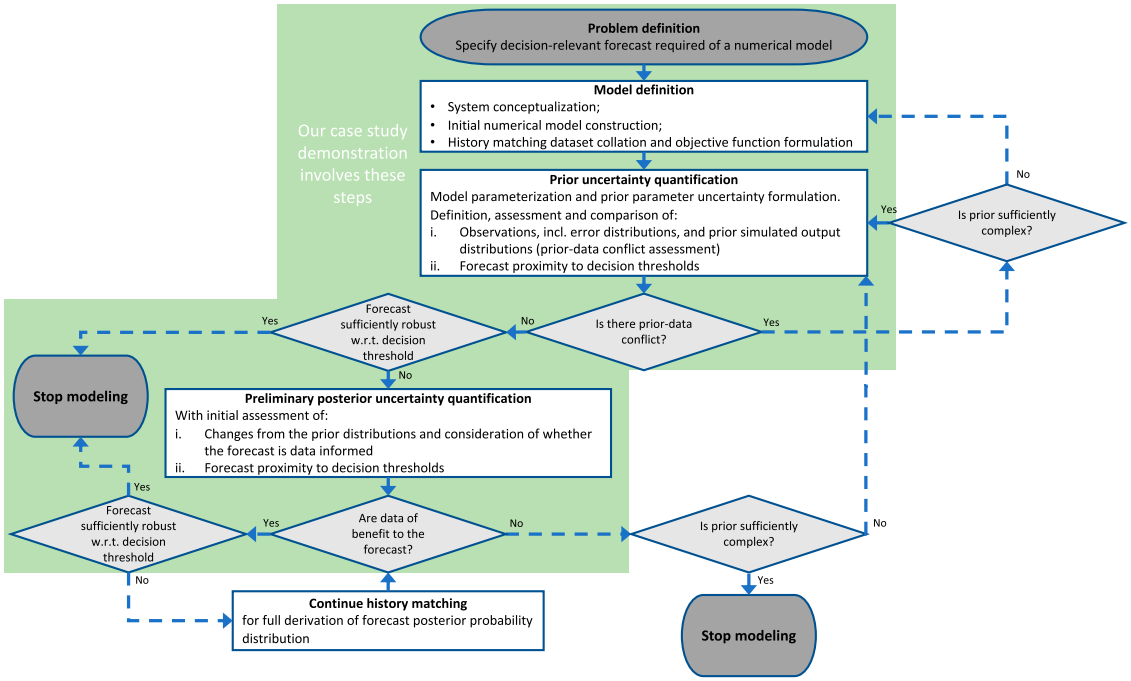


FIGURE 1 | Graphical presentation of the proposed decision support modeling workflow, which incorporates uncertainty quantification at an early stage before undertaking more comprehensive history matching. The steps involved in the case study presented herein are highlighted with the green background.

values and model forecasts, and provide insights into the benefits, or potential hazards, of continuing history matching.

This study is motivated by, and follows, the forecast focused workflow promoted in Doherty (2015) and White (2017). It is also consistent with the recommendations in Doherty and Moore (2019), which suggest undertaking uncertainty quantification early in the decision support modeling process, to identify the extent that available historical observations can inform decision-salient parameter values. Doherty and Moore (2019) outline how this information can then be used to guide the modeling process, from model conceptualization to deployment for decision support. This study adopts this fundamental recommendation, and explores the implications of doing so, for a real-world case study.

The case study contributes to the small body of real-world decision support worked examples that are currently available in the international literature (e.g., Kunstmann et al., 2002; Enzenhoefer et al., 2014; Sepúlveda and Doherty, 2015; Brouwers et al., 2018; Sundell et al., 2019; White et al., 2020a).

2. METHODOLOGY

This section describes the proposed workflow for the early assessment of the uncertainty surrounding model predictions that are of particular management interest ("decision-relevant"), which are herein referred to as "forecasts". **Figure 1** provides a flowchart outlining the major steps and decision points in the workflow.

2.1. Problem Definition

"Problem definition" defines the decision context for numerical modeling. It involves the definition of the management problem and the undesirable outcome that the management strategy is designed to prevent, e.g., land-use consent restriction or regulation to prevent ecological degradation of a stream system due to nutrient contamination. This step also includes the specification of forecasts that can be used to evaluate the efficacy of management decisions, and the definition of decision thresholds against which the success (or failure) of a management strategy can be evaluated.

2.2. Model Definition

This step involves defining the pertinent processes and components of the hydrogeological system that the forecasts are likely to be sensitive to. On this basis, an initial numerical model is constructed, ensuring sufficient complexity to represent these aspects of the system (e.g., Hunt et al., 2007). This step also includes formulation of the model objective function, which involves collation and processing of historical system observations and definition of a weighting scheme. The weighting scheme ideally aims to

maximize the flow of information during the history matching process, from observation data to model parameters, and subsequently to model forecasts, while accounting for both measurement and model error (Doherty and Welter, 2010; Doherty, 2015).

2.3. Prior Uncertainty Quantification

"Prior uncertainty quantification" is underpinned by the definition of uncertain model parameters and a formulation of prior parameter uncertainty. Simulation of the model with parameter realizations that reflect this prior parameter uncertainty, allows an initial quantification of simulated output and forecast uncertainty. As the relationship between model parameters and simulated outputs can rarely be fully known, a priori, it is advantageous to employ high parameter dimensionality (e.g., Hunt et al., 2007; Fienen et al., 2010; Doherty and Simmons, 2013; Knowling et al., 2019), in concert with a conservative expression of parameter uncertainty, informed by "expert knowledge" of the conceptualized system.

Comparison of simulated output distributions with historically observed system behavior supports an early assessment of the appropriateness of the prior parameter probability distributions. The presence of prior-data conflict may indicate underestimation of measurement error, underestimation of parameter variances and/or insufficient complexity in the numerical model and prior parameterization. If the model and prior parameterization is deemed to be sufficiently complex with respect to the processes that the forecasts are sensitive to, prior-data conflict may be resolved by inflation of parameter variances or observation error, as appropriate. If however, the prior is regarded to represent insufficient forecast relevant complexity, the prior parameterization itself may also need to be revisited. For example, model boundaries may need to be parameterized in a different way, if the forecast is sensitive to these boundary conditions. Or if the connectedness of high hydraulic conductivity facies is relevant to the forecasts being made, a parameter representation of these connected facies may need to be adopted in place of, for example, a multi-variate Gaussian parameter representation.

This part of the workflow also accommodates the situation where there is found to be no forecast relevant data available. In this situation the prior parameterization becomes more important, as it is now the only source of information in the model. In this situation a modeler is freed from all history matching burdens, and may choose to adopt a very complex model parameterization, to ensure a robust probability distribution (Doherty and Moore, 2019).

If prior-data conflict is not present, the assessment of the proximity of the prior forecast probability distributions to the decision threshold may be sufficient for addressing the management decision and the workflow can move to the "Stop modeling" option (left-hand side of **Figure 1**). However, if reduction in forecast uncertainty is desired to support the management decision, the workflow moves toward a preliminary approximation of the posterior uncertainty, as discussed in **Section 2.4** below.

2.4. Preliminary Posterior Uncertainty Quantification

The "Preliminary posterior uncertainty quantification" is undertaken with a view to reducing forecast uncertainty through a preliminary formal assimilation of observation data. This step provides an opportunity to assess the ability of the observation data to inform model parameters and forecasts; a number of modeling workflow decisions may then be made on the basis of this preliminary posterior uncertainty quantification.

If forecast uncertainty reduction through the preliminary conditioning of influential parameters is sufficient for addressing the management decision, the workflow can move to "Stop modeling." If the assimilation of data is deemed beneficial for reducing forecast uncertainty, but further uncertainty reduction is desired, then this preliminary posterior uncertainty quantification provides justification for continuing to assimilate observation data and deriving more advanced posterior forecast distributions. If, however, the data do not inform forecasts, after verifying that the model, its parameterization, and its prior probability representation, is sufficiently complex to adequately represent the forecast uncertainty, the modeling workflow should move to "Stop modeling." If the level of forecast uncertainty is not satisfactory for decision support purposes, then the workflow can be re-initiated with options for recasting forecasts, model design, parameterization, and collection or reprocessing of system observations to improve the flow of information to model parameters and forecasts. Alternatively, the decision maker can choose to consider an alternative management scenario, on the basis of this initial iteration through the workflow.

The workflow presented here is essentially agnostic with respect to the methodologies adopted for the uncertainty quantification, with the following proviso: that the uncertainty quantification method must be computationally efficient, as its purpose is to provide guidance on whether or not the investment in a more completed and rigorous data assimilation effort is necessary.

The model simulation and uncertainty quantification methods adopted for this study are described in detail in the following case study section (**Section 3**). The case study addresses all major steps (boxes with green background) in **Figure 1**, before shifting the "Stop modeling' option (left-hand side of **Figure 1**), based on the results. The alternative options would be relevant for other case studies with differing forecast-data relationships; examples of when these options may be favorable are discussed.

3. CASE STUDY

A numerical modeling case study is used to demonstrate the workflow outlined above. The case study, set within the Ruamāhanga catchment in the Wairarapa Valley, North Island, New Zealand, involves simulation-based forecasting of groundwater abstraction impacts on stream low flows and land-use change impacts on groundwater nutrient concentrations.

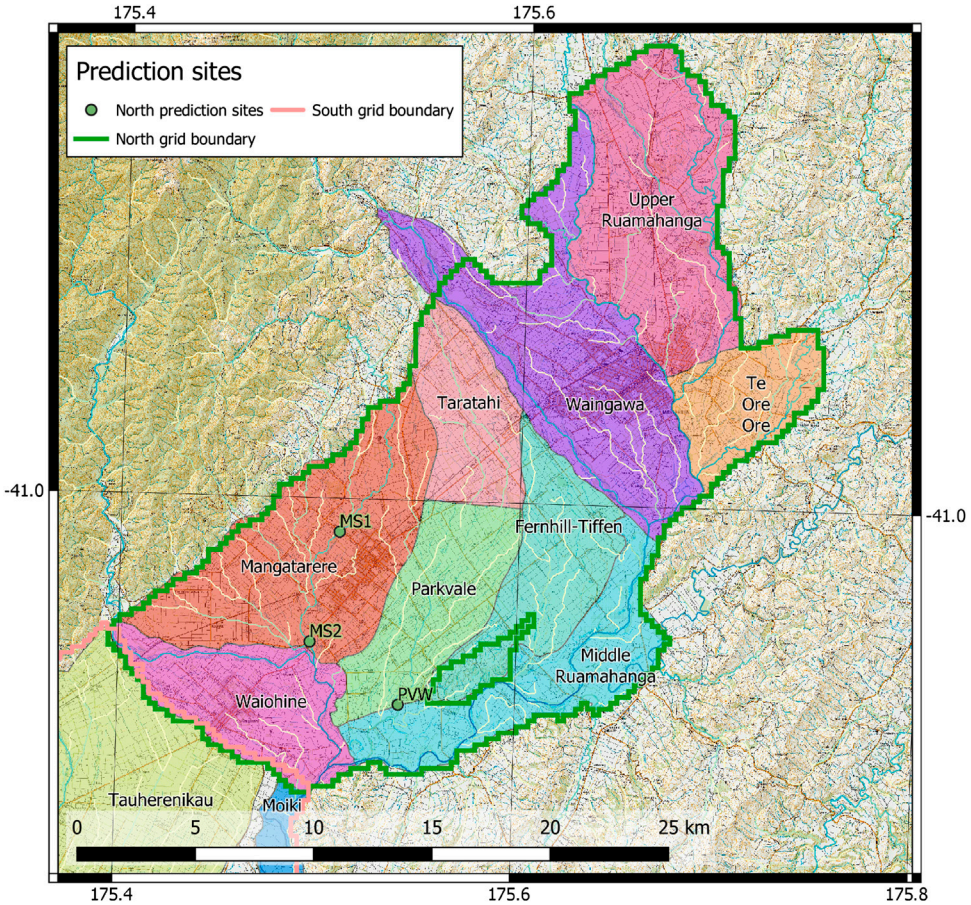


FIGURE 2 | Decision-relevant forecast locations. Streamflow reliability forecasts are indicated by green dots. Groundwater management zones (GMZs) for mean zonal nitrate concentration forecasts are delimited by colored polygons.

3.1. Problem Definition

3.1.1. Decision-Relevant Forecasts

The forecasts considered relate to water availability and water quality issues which are of particular significance within the Wairarapa Valley. Increased groundwater and surface water abstraction in Wairarapa Valley are believed to have impacted stream flows; ensuring stream flows are sufficient to meet environmental and ecological criteria is therefore an important management constraint, when setting groundwater allocation limits. For the security of fish habitats critical criteria are the frequency and persistence of days in which low streamflow (low-flow) conditions exist. The specific streamflow reliability forecasts explored here are the number of low-flow days within a 7-year time period (low-flow frequency) and the maximum number of consecutive low-flow days (low-flow persistence), at three sites (PVW, MS1, and MS2; locations provided in **Figure 2**). A low-flow day is defined as a day with simulated streamflow that is exceeded for 95% of the simulated period (Q95), under natural conditions (i.e., when no abstraction from either groundwater or surface water occurs).

Land-use practices can present freshwater contamination risks and management strategies often aim to limit this contamination potential. Of particular interest in the Wairarapa Valley is nitrate-nitrogen (hereafter referred to as nitrate) concentrations in groundwater. The specific water quality forecast simulated in this case study is mean nitrate concentration within defined groundwater management zones (GMZs; as presented in **Figure 2**).

The case study considers the forecasts described above as absolute quantities and also in a relative sense, as differences, relative to a baseline scenario; e.g., percentage-changes in the simulated output values in response to water-use- and land-use-change management scenarios. The full list of the case study forecasts is presented in **Table 1**.

3.1.2. Management Scenarios

The scenarios explored constitute simplified examples of water- and land-use management strategies employed in practice, e.g., to satisfy water-supply needs or to reduce water contamination risk. The streamflow reliability scenario reflects the estimated spatial and temporal variation in water-use in the case study area which reflects a mean (in-time) groundwater abstraction rate of 82,000 m^3d^{-1}. Changes in the frequency and persistence of low-flow days are compared with a "naturalized" baseline scenario with no abstraction.

TABLE 1 | Decision-relevant forecasts. PVW, MS1 and MS2 relate to the streamflow reliability forecast sites in **Figure 1**. GMZs refers to groundwater management zones.

Forecast	Code	Detail	Decision threshold	
			PVW	MS1, MS2
Streamflow reliability				
Low-flow frequency	*nday<Q95*	Number of days below Q95	155	155
	%Δnday<Q95	Percent change in number of days below Q95	10%	10%
Low-flow persistence	*nconsecday<Q95*	Maximum number of consecutive days below Q95	85	20
	%Δnconsecday<Q95	Percent change in maximum number of consecutive days below Q95	10%	10%
Water quality				
mean zonal concentration	*mean.conc*	Mean concentration in GMZs	5.65 mg L^{-1}	
	%Δmean.conc	Percent change in mean zonal concentration in GMZs	10%	

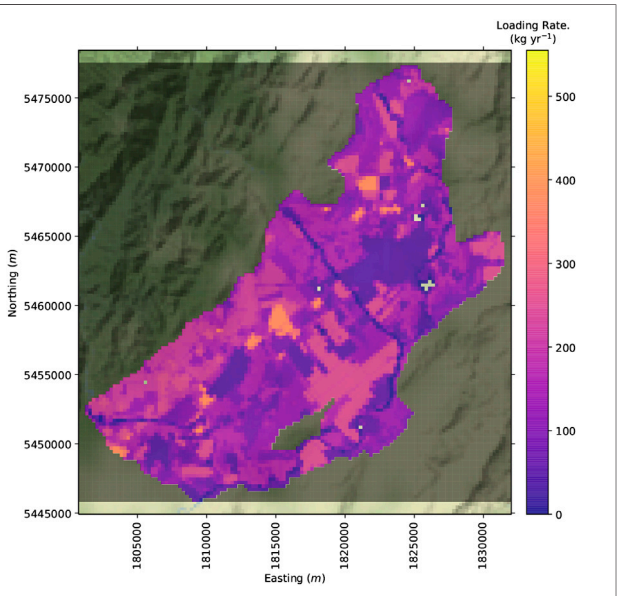


FIGURE 3 | Land-use management scenario nitrate loading rate.

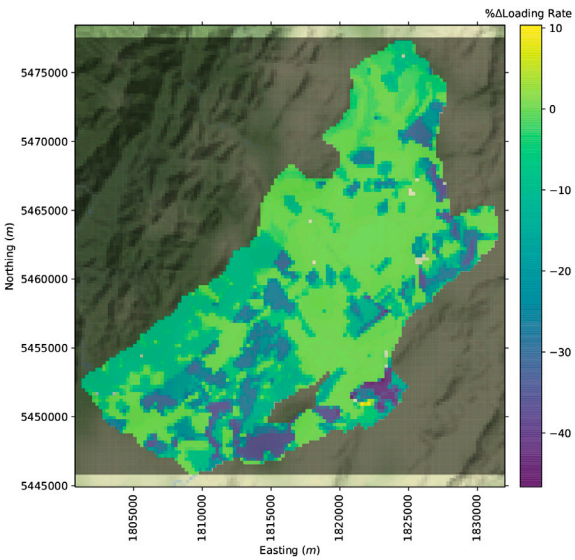


FIGURE 4 | Percentage change in nitrate loading rate from baseline.

The land-use scenario is combined with the abstraction scenario and is represented with a spatially distributed nitrate loading rate (**Figure 3**) reflecting a real-world nitrate loading scenario. This nitrate loading rate is a reduction from the current land-use baseline (**Figure 4**).

3.1.3. Decision Thresholds

The decision thresholds considered are listed in **Table 1**. For streamflow reliability forecasts, the simulated groundwater abstraction scenario should not increase the frequency or persistence of low-flow days by greater than 10%. For water quality forecasts, mean zonal nitrate concentrations should not exceed 50% of the maximum acceptable value (MAV) for nitrate-nitrogen in drinking water (5.65 mg L^{-1}; Ministry of Health 2018) and the management goal is that the land-use scenario should result in a 10% decrease in the zonal mean concentration.

3.2. Model Definition

3.2.1. System Conceptualization

The hydrogeological system in the Wairarapa Valley is characterized by successions of unconsolidated, late Quaternary and Holocene, alluvial sediments (Gyopari and McAlister, 2010a; Gyopari and McAlister, 2010b). The general hydraulic gradient results in groundwater flow from the north-east to south-west. Interactions between aquifer units are complicated by variable degrees of compaction, reworking and faulting. Active fault systems that splay from the major Wairarapa Fault, which bounds the western side of the valley, are anticipated to compartmentalize the groundwater system and potentially modify the interaction between groundwater and surface water (Gyopari and McAlister, 2010a; Gyopari and McAlister, 2010b); springs occur along the Masterton and Carterton Fault features. Rainfall recharge within the valleys is supplemented by significant river inflows, especially from the Tararua Range to the west. Nitrate loading is mostly derived from

land-use within the plains, with low concentrations anticipated in the inflowing rivers. The Wairarapa Valley is divided into two distinct groundwater domains. Groundwater is discharged to surface water before leaving the northern portion of the valley (425 km^2; **Figure 2**) *via* the Ruamāhanga River.

3.2.2. Numerical Model Construction

To address the management challenges outlined above, a regional-scale, integrated groundwater/surface-water flow and nitrate transport model was developed. The flow and transport processes were simulated using MODFLOW-NWT (Niswonger et al., 2011) and MT3D-USGS (Bedekar et al., 2016), respectively. The flow and transport simulations were undertaken on a regular finite-difference grid, consisting of five layers, 127 rows and 125 columns at 250 m spacing. Lateral model boundaries were defined as no-flow. Flow and contaminant mass could enter the model *via* recharge, streamflow and injection wells, and leave the model through streamflow, well abstraction and first-order reactions implemented using the MT3D-USGS reaction package (RCT).

Two simulation periods were considered, a "history matching period" from 1992 to 2007, and a "prediction period" from 2007 to 2014. The history matching period was discretized into 771 *weekly* stress periods (with a constant time-step of 1.75 days), with varying hydrological stresses (e.g., recharge, well pumping and stream inflows). The prediction period was divided into 2,821 *daily* stress periods (with daily time-steps), again with varied hydrological stresses.

Both periods consisted of a transient flow model simulation, followed by a 20-years transport simulation with temporally constant (though spatially variable) nitrate mass inputs. The flow solution for the transport simulations were provided by steady-state flow simulations using temporally averaged hydrological stresses for each of the considered time periods. The steady-state flow simulations were also used to provide initial conditions for the respective transient flow simulations.

Additional relevant aspects of the model include:

- surface-water flow and contaminant routing, simulated using the Streamflow-Routing (SFR) (Niswonger and Prudic, 2005) and Stream-Flow Transport (SFT) packages of MODFLOW-NWT and MT3D-USGS, respectively;
- potential fault-bound compartmentalization of the groundwater system, simulated using the Horizontal Flow Barrier package of MODFLOW-NWT (HFB6); and
- use of the total-variation-diminishing (TVD) solver scheme MT3D-USGS to maximize model stability and to reduce numerical dispersion. Note, the use of TVD restricts the transport time step size to honor a Courant number of 1.

3.2.3. History Matching Dataset and Objective Function Formulation

The history matching observations included nitrate concentration measurements from groundwater and surface water, groundwater elevation, in-stream flows and estimates of groundwater to surface water exchanges for the history matching time period. The data included transient time series records, of varied lengths and sampling frequency, and more occasional survey observations. The available observation dataset was composed as follows (observation locations are provided in **Supplementary Figure SI 1.1**):

- 35 groundwater level sites, including time series (of varied lengths); providing a total of 20,702 observations. The 35 observation sites provided the definition of 35 observation groups. Observations for these sites are provided in **Supplementary Section SI 4.1.1**.
- 88 surface water gauging sites. Including, continuous, repeat and spot gauging observations; providing a total of 4,385 streamflow observations. Surface water flow observations were partitioned into 12 observation groups, 11 of which relate to sites with 50 or more observations, and the remaining group composed of sites with a low observation count. Observations for these sites are also provided in **Supplementary Section SI 4.1.2**.
- 23 simultaneous streamflow gaugings, providing a total of 26 surface water to groundwater exchange estimate observations. These observations were assigned to a single observation group.
- 203 groundwater quality monitoring sites, providing a total of 203 mean nitrate concentration observations, assigned to a single observation group.
- 14 surface water quality monitoring sites, providing a total of 14 mean nitrate concentration observations, assigned to a single observation group.

The model objective function was formulated as the sum of weighted, squared residuals between simulated outputs and the historical system observations. Initial observation weights were defined to reflect the uncertainty in system measurements, as the inverse of the estimated measurement standard deviation. For water level observation the standard deviation was defined at 0.5 m; for streamflow observations the standard deviation was 20% of the observed value; for groundwater to surface water exchange estimates, the standard deviation was 10% of the observed value. For concentration observations, where values for each location were averages of different lengths of measurement record, the standard deviations where scaled to account for the number of observation in the record (count). For groundwater concentration the standard deviation used was defined as the standard deviation of the values in the record, multiplied by 100/count. For surface water concentrations the standard deviation used was defined as 100/count multiplied by the mean of the values in the record. The numerator scaling attempts to account for the use of temporally sparse, point concentration measurements, as average system observations. The measurement uncertainties defined on the basis of these standard deviation are illustrated in **Supplementary Sections SI 4.1.1, 4.1.2**, where red shading denotes three standard deviations from observed values. For the objective function formulation, the initial observations weights were re-balanced such that each

group of observations contributed equally to the initial total objective function (e.g., Doherty and Hunt, 2010).

3.3. Prior Uncertainty Quantification

3.3.1. Model Parameterization

A "highly parameterized" parameterization scheme was adopted, as advocated by Hunt et al. (2007); Doherty and Hunt (2010). Model input uncertainty was represented using probability distributions for a total of 2,129 adjustable parameters, all of which are expressed as multipliers on parameter values (e.g., McKenna et al., 2020). The uncertain model parameters considered are as follows:

- horizontal and vertical hydraulic conductivity—520 pilot-point and 82 zonal multipliers
- specific storage—260 pilot-point and 16 zonal multipliers
- specific yield—65 pilot-point and four zonal multipliers
- porosity—260 pilot-point and five layer-constant multipliers
- denitrification rate—260 pilot-point and 15 zonal multipliers
- groundwater recharge—one global multiplier; 65 pilot-point multipliers; and 12 temporal, monthly multipliers
- nitrate loading rate—1 global multiplier and 65 pilot-point multipliers
- stream-bed conductivity—115 spatial multipliers
- surface-water inflow—117 temporally constant multipliers
- surface-water inflow concentrations—230 spatial multipliers
- groundwater abstraction—one global multiplier; 11 multipliers, by usage type; and 12 temporal, monthly multipliers
- surface-water abstraction—two multipliers, by usage type

Initial model parameter values (to which multipliers were applied) were informed by "expert knowledge" based on hydrogeological assessments incorporating information from, for example, bore logs, aquifer pumping test data and literature studies (e.g., Gyopari and McAlister, 2010a; Gyopari and McAlister, 2010b).

More detail on the spatial and temporal distributions of parameters and the initial native model parameter values is provided in the **Supplementary Section SI 3**.

3.3.2. Parameter Uncertainty

Prior parameter uncertainty was specified through a block-diagonal prior covariance matrix with prior covariances informed by expert knowledge. Diagonal elements of the prior covariance matrix represent expected individual parameter variances; off-diagonal elements define the correlations between parameters. For the spatially variable parameter types (e.g., hydraulic conductivity, porosity, recharge) the correlation between these parameters were derived from geostatistical analysis of hydrogeolgical field data (e.g., pumping test data; after Moore et al., 2017). Geospatial correlation was defined by exponential variogram with a sill proportional to the expected prior variance (with a proportionality constant of 0.45), and range of 3,500 m. This geospatial correlation estimation also supported the pilot-point parameter interpolation (Doherty, 2003; Doherty et al., 2011). Temporal and non-spatially distributed parameters were defined as uncorrelated.

Two hundred prior realizations were drawn from the prior parameter covariance matrix using a Monte-Carlo multivariate Gaussian sampling approach (e.g., Tarantola, 2005) to produce the prior parameter ensemble. The statistics for the prior parameter ensemble (in terms of parameters groups) are provided in **Supplementary Table SI 3.1** Grouped prior parameter PDFs are illustrated in **Supplementary Figure SI 3.8**.

3.3.3. Propagation of Prior Uncertainty

Parameter uncertainty was propagated to simulated outputs and forecasts through forward model runs using the 200 prior parameter realizations in the prior parameter ensemble. An assessment of the resultant prior simulated output uncertainty distribution, relative to system observations, was undertaken for the history matching period to identify potential model deficiencies. Predictive period simulations using the same parameter realizations provided forecast distributions which were evaluated against management decision thresholds to provide a prior probabilistic expression of management scenario success or failure.

3.4. Preliminary Posterior Uncertainty Quantification

The preliminary approximation of the posterior forecast uncertainty was derived after assimilation of state observations using an abridged single-iteration history matching and a preliminary approximation of the posterior parameter uncertainty.

3.4.1. Abridged History Matching

History matching was undertaken with a single iteration of the Gauss-Levenberg-Marquardt (GLM) algorithm, using PESTPP-GLM (Welter et al., 2015). The single GLM parameter iteration upgraded the initial parameter vector, effectively constituting and update of the mean of the prior parameter distributions (**Supplementary Figure SI 3.8**).

The use of a single iteration reduces the computational resource burden compared to undertaking multiple iterations but it is sufficient to provide a Jacobian matrix, which is required to derive the preliminary posterior (see **Section 3.4.2**), and to support an analysis of the relevance of the available data to the forecasts. The abridged history matching required a total of 2,195 model forward simulations; 2,130 to populate the first-order sensitivity (i.e., Jacobian) matrix, and 65 simulations to test parameter upgrades (aimed at minimizing an objective function). The Jacobian matrix was populated using 1% two-point derivative increments on all parameters, except for surface-water inflow parameters, which were necessarily offset by a value of 100 and used a 0.01% derivative increment (e.g., Doherty, 2016).

3.4.2. Preliminary Posterior Uncertainty Using Linearized Bayes Equation

Using a linear form of Bayes equation, the posterior parameter covariance matrix was approximated *via* the Schur complement (e.g., Tarantola 2005; Doherty 2015). This approach required: the prior parameter covariance matrix, as introduced above; the Jacobian matrix; and the epistemic noise covariance matrix, which was defined as diagonal with elements specified on the basis of observation "weights." To account for the effect of model error in an approximate manner, these observations weights were adjusted on the basis of the model-to-measurement residuals after the abridged history matching; weights were defined such that observations within each group contributed equally to the objective function and the total contribution from each group was the same (e.g., Doherty, 2015). Note, this differs from the weighting scheme employed for history matching. More detail on the linearized Bayes equation and the Schur complement method is provided in the **Supplementary Section SI 2**.

3.4.3. Propagation of Preliminary Posterior Parameter Uncertainty

As with the propagation of prior uncertainty, propagation of the preliminary posterior parameter uncertainty to preliminary posterior forecast uncertainty was achieved through predictive period model forward simulations for each management scenario using 200 parameter realizations, sampled from the preliminary posterior parameter covariance matrix. Sampling was undertaken using the same multi-variate Gaussian Monte-Carlo method (e.g., Tarantola, 2005), outlined above (**Section 3.3.2**).

The effectiveness of the abridged history matching was evaluated by comparing prior and preliminary posterior parameter and simulated output distributions. The resulting preliminary posterior forecast distributions were evaluated against the defined management decision thresholds. On this basis, the decision points in the workflow were negotiated, leading to a "Stop modeling" outcome.

4. RESULTS

Ongoing or iterative model design, history matching and uncertainty quantification decisions are made on the basis of the "Prior uncertainty quantification" and "Preliminary posterior uncertainty quantification" steps of the workflow depicted in **Figure 1**. These are decisions that can be informed by an early uncertainty quantification, and therefore, the case study results focus on these parts of the workflow, i.e., after the modeling problem has been defined and a model has been built. To simplify the presentation of the case study results, and to facilitate comparisons, the prior and preliminary posterior uncertainty quantification results figures have been combined.

Note that the case study is used to illustrate the proposed workflow, and the process of negotiating the decision points, within the workflow. It is specifically used to evaluate and demonstrate the insights gained through the steps that constitute early uncertainty quantification ("Prior uncertainty quantification" and "Preliminary posterior uncertainty quantification"). The case study does not iterate through all possible decision pathways of the workflow, but instead demonstrates the workflow in general.

4.1. Prior Uncertainty Quantification

Recall, that purpose of this analysis is twofold. Firstly to check that model outputs, simulated on the basis of the prior parameter probability distributions, encompass the measured observations. Secondly, to assess the proximity of model forecasts to decision thresholds.

4.1.1. Comparison Between Simulated Outputs and Historical Observations

Comparison plots of simulated outputs (both prior and preliminary posterior) and system observations for the history matching period are provided in the **Supplementary Sections SI 4.1.1, 4.1.2**. The prior simulated output distributions generally encompass their associated observation values—i.e., gray bars in **Supplementary Section SI 4.1.1** span the diagonal one-to-one line in the left-hand plot and the zero residual line in right-hand plots; and prior realization outputs in **Supplementary Section SI 4.1.2** (gray lines) generally overlap with observations, accounting for observation uncertainty (red shaded areas).

This consistency between the observations and the system conceptualization, as expressed through the numerical model and its prior parameters, was particularly robust for those observations that were of a similar nature and location to the model forecasts. However, this fortunate outcome was not absolute, as some prior-data conflict is evident in the results, e.g., water-level observations for two closely located wells, "s26_0298" and "s26_0308" (highlighted in **Supplementary Figure SI 1.1A**, see associated plots in **Supplementary Sections SI 4.1.1, 4.1.2**). Because these observations relate to water levels at relatively distant locations from the streamflow reliability forecast locations, this prior-data conflict is not considered critical to the model forecasts.

Some streamflow observations also demonstrated a degree of prior-data conflict, particularly when simulating short duration, extreme events. For example, the simulated outputs for "fo_s056," which is close to the streamflow reliability forecast location MS1, fail to reproduce the magnitude of the extreme flow events, generally over predicting high flows (**Supplementary Sections SI 4.1.1, 4.1.2**). However, the simulated outputs do reproduce the lower flow conditions. This may relate to simplified representation of quick-flow and runoff processes in the numerical model, coupled with the seven-day stress period duration in the history matching period simulation and the relatively sparse temporal parametrization. As the forecasts relate to stream low-flows, this prior-data conflict also not considered critical.

With respect to the workflow schematic in **Figure 1**, the prior uncertainty analysis indicates that there is no prior-data conflict that is significant for the forecast, and therefore the

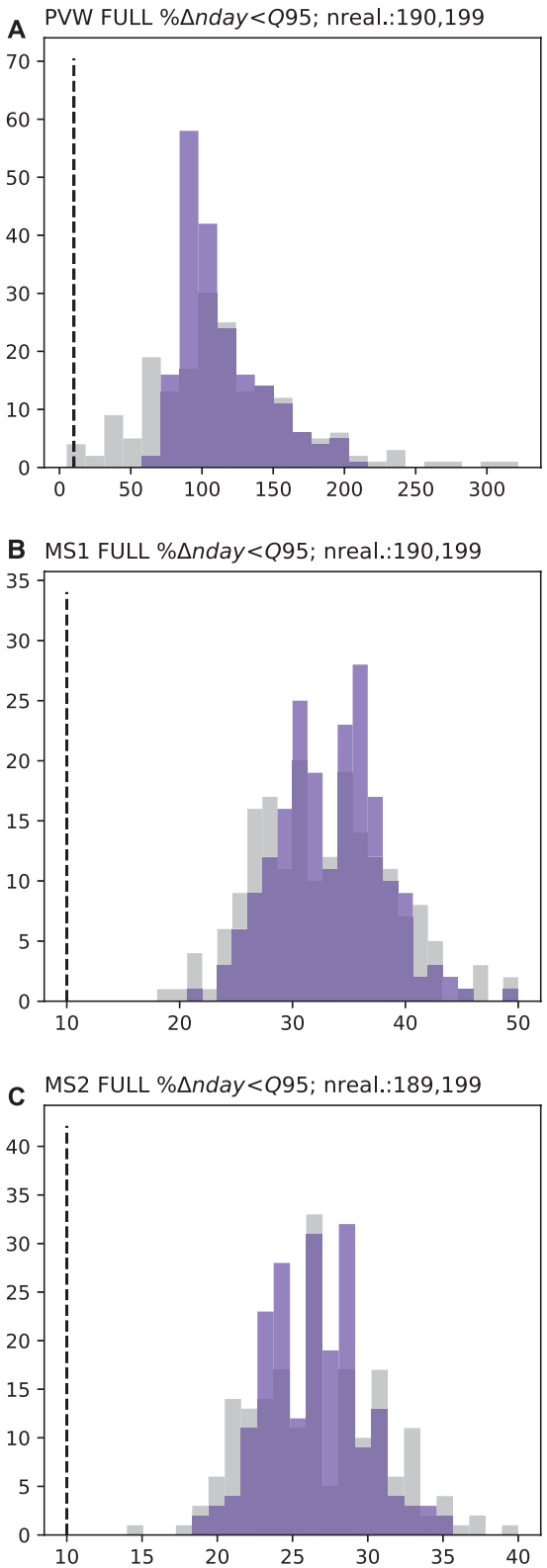


FIGURE 5 | Percent change low-flow frequency (*%nday<Q95*) prior (grey) and posterior (blue) forecast PDFs at PWV **(A)**, MS1 **(B)** and MS2 **(C)**. Dashed black lines represents the 10% change decision threshold. "nreal." details the number of successful realizations (therefore the number of data points represented in the PDF) for prior and posterior ensembles, respectively.

decision was made to progress to assessing prior forecast uncertainty and (if necessary) approximation of preliminary posterior uncertainty quantification.

4.1.2. Prior Forecast Uncertainty

The prior forecast probability density functions (PDFs) (presented in gray in **Figures 5–9**) provide the initial indication of the magnitude of the uncertainty surrounding model forecasts and the proximity of these forecast distributions to the decision thresholds.

4.1.2.1. Streamflow reliability forecasts

The streamflow reliability forecasts relate to the frequency and persistence of low-flow conditions at the three sites of interest

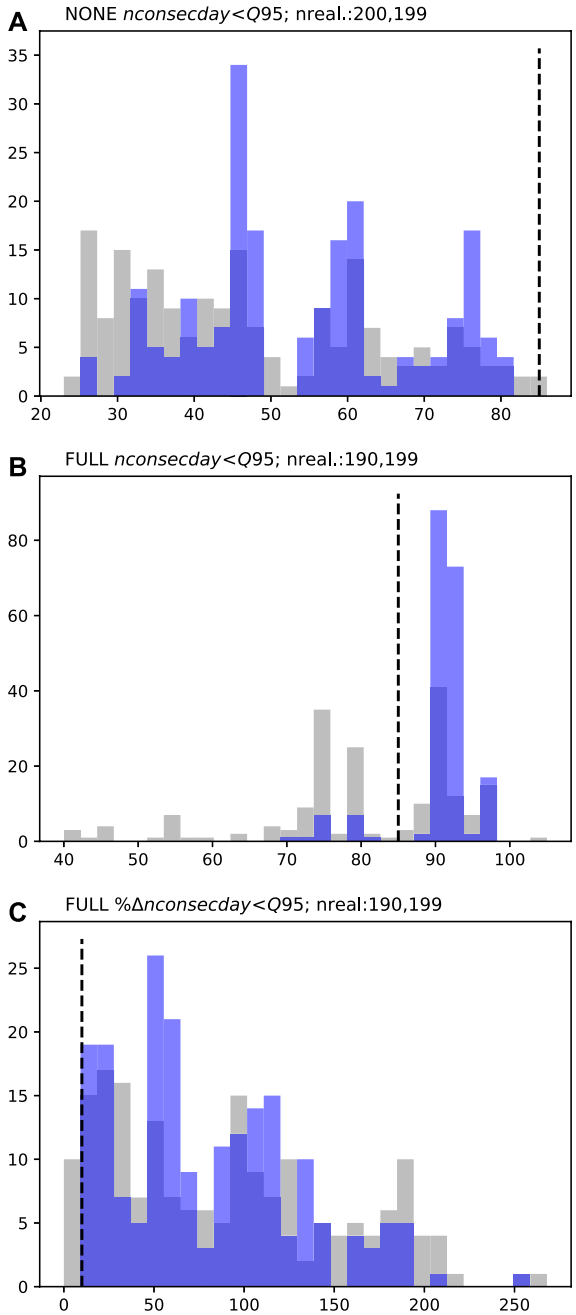


FIGURE 6 | PVW low-flow persistence (*nconsecday<Q95*; **A** and **B**) and percent change low-flow persistence (*%nconsecday<Q95*; **C**) prior (grey) and posterior (blue) forecast PDFs. Dashed black lines represent the respective decision thresholds (85 days in **A** and **B** and 10% change in **C**). "NONE" represents the "naturalized" baseline and "FULL" the abstraction scenario. "nreal." details the number of successful realizations (therefore the number of data points represented in the PDF) for prior and posterior ensembles, respectively.

(PVW, MS1, and MS2; **Figure 2**), as well as the percentage changes in these forecasts, associated with abstraction; see **Table 1**).

4.1.2.1.1. Low-Flow frequency forecasts. For simplicity, the following presentation of the low-flow frequency forecast results will focus on the percentage change forecast (**Figure 5**) at each location. The full results for the low-flow frequency forecasts are presented in the **Supplementary Section SI 6**.

For this forecast the prior PDFs vary between sites. The forecast uncertainty is highest at site PVW (**Figure 5A**). At this site the prior forecast PDF extends (just) below the decision threshold, indicating that there is potential (albeit with low probability) for the abstraction to comply with the management limit at this site. At MS1 and MS2 sites, the forecast distributions are entirely above their respective decision thresholds (**Figures 5B,C**), supporting the assumption of certainty that abstraction will cause exceedance of the management limit, i.e., abstraction is predicted to increase the number of low-flow days by more than 10%.

On this basis, the results of the prior uncertainty quantification are sufficient for supporting the management decision relating to low-flow frequency objectives at sites MS1 and MS2. Consequently, when considering these forecasts, the prior uncertainty quantification provides justification for moving to "Stop modeling" in the workflow. However, at site PVW the results indicate that benefit may be gained by reducing forecast uncertainty and the prior uncertainty quantification provides justification for moving to "Preliminary posterior uncertainty quantification."

4.1.2.1.2. Low-Flow persistence forecasts. For the low-flow persistence forecasts, the PDFs are multi-modal (**Figures 6–8**). This is a consequence of their discontinuous nature, where, for example, for any given realization, two consecutive periods low-flow periods may, or may not, be bridged, and when bridging occurs this results in one, significantly longer, low-flow period. This multi-modal nature is most evident for the PVW site (**Figure 6**) where the under naturalized conditions, the prior forecast PDF spans 23–86 days, beyond the decision threshold, with no single dominant mode apparent (**Figure 6A**). Under the abstraction scenario, the forecast at this site also indicates high uncertainty and spans the decision threshold (**Figure 6B**). The prior forecast uncertainty is also high for the percentage change forecast at PVW, extending below the 10% decision threshold, to 0% change, thus indicating the potential for abstraction to have little effect on the persistence of low-flow conditions at this site. These results suggest that the forecast is not sufficiently robust to allow the workflow to progress to "Stop modeling," indicating the "Preliminary posterior uncertainty quantification" should be undertaken.

For the MS1 and MS2 sites, the prior PDFs for the low-flow persistence forecasts are similar (**Figures 7, 8**). Both sites display a single dominant mode in the forecast PDFs, with and without abstraction. For the baseline scenario, the prior PDFs at both sites span the decision threshold (**Figures 7A, 8A**), this indicates that even under "natural" conditions the number of consecutive low flow days may exceed 20 days. With abstraction, the prior PDFs indicate certainty that this decision threshold will be exceeded, i.e., the prior distributions are wholly above the decision threshold (**Figures 7B, 8B**). On the basis of these absolute forecast results, there would be justification for moving to "Stop modeling" in the workflow. However, for the percentage change forecast, the minimum of the prior PDF at MS1 is at the decision threshold (**Figure 7C**), while for MS2 it extends below the decision threshold (**Figure 8C**); this suggests that for these normalized difference forecasts, "Preliminary posterior uncertainty quantification" should be undertaken.

4.1.2.2. Water Quality Forecasts

For the mean zonal concentration forecasts, prior forecast PDFs demonstrate high variance and indicate a potential for mean concentrations to exceed the 5.65 mg L^{-1} decision threshold in all zones, under both the land-use management scenario and baseline conditions (**Figures 9A,B**). The prior forecast distributions indicate significant spatial variability in the predicted success of the management strategy, with respect to the decision threshold of a desired 10% reduction in mean zonal concentration (**Figure 9C**). In some zones (e.g., Te Ore Ore, Parkvale, Middle Ruamāhanga, Waiohine) the percent change forecasts indicate that there is a possibility for management scenario success (forecast PDFs at least span the decision threshold). However, for the other zones these results indicate failure of the land-use management scenario. Generally, for the water quality forecasts, the prior uncertainty quantification indicates that the workflow should move to "Preliminary posterior uncertainty quantification."

4.2. Preliminary Posterior Uncertainty Quantification

The preliminary posterior uncertainty quantification was undertaken after an abridged history matching and provides approximate posterior parameter distributions which are propagated to approximate posterior forecast distributions.

4.2.1. Parameter Uncertainty

To assist in the interpretation of the uncertainty surrounding the 2,129 model parameters they are grouped by type (**Figure 10**). Prior parameter distributions (gray PDFs in **Figure 10**) reflect sampled parameter covariances that are based on expert knowledge alone. Posterior parameter distributions (blue PDFs in **Figure 10**) reflect the approximated parameter uncertainty after the abridged history matching. Together, the distributions presented in **Figure 10** provide an indication of the level of parameter conditioning achieved through the abridged history matching. The prior and posterior distribution statistics for each parameter group are provided in **Supplementary Table SI 3.1**.

While the grouping of parameters in **Figure 10** and **Supplementary Table SI 3.1** can obscure visual identification

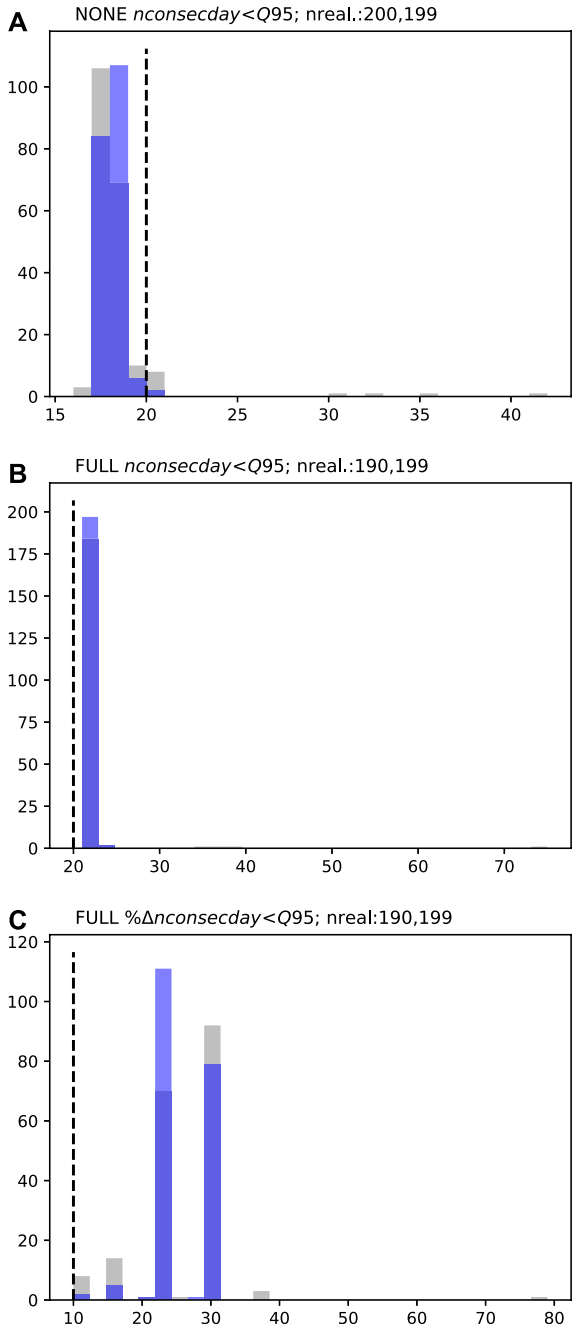


FIGURE 7 | MS1 low-flow persistence (*nconsecday<Q95*; **A** and **B**) and percent change low-flow persistence (*%nconsecday<Q95*; **C**) prior (grey) and posterior (blue) forecast PDFs. Dashed black lines represent the respective decision thresholds (20 days in **A** and **B** and 10% change in **C**). "NONE" represents the "naturalized" baseline and "FULL" the abstraction scenario. "nreal." details the number of successful realizations (therefore the number of data points represented in the PDF) for prior and posterior ensembles, respectively.

of the extent of the conditioning of individual parameters within a group, it is clear that the abridged history matching has achieved only minor reductions in the uncertainty of most parameters. Generally, reductions in uncertainty are more notable for the lumped, global or layer-wide, parameters, than for the spatially distributed parameters. Illustrations of the spatial distribution of the variance reduction for both zone and pilot point parameters are shown in **Supplementary Section SI 5**. These plots also highlight that the abridged history matching achieved only minor reduction in parameter uncertainty over large areas of the numerical model spatial domain, coinciding with areas lacking in observation data (**Supplementary Figures SI 5.1-5.18**). The streamflow reliability forecast locations are coincident with areas where the reduction in parameter variance is minimal (e.g., see **Supplementary Figures SI 5.5-5.14**).

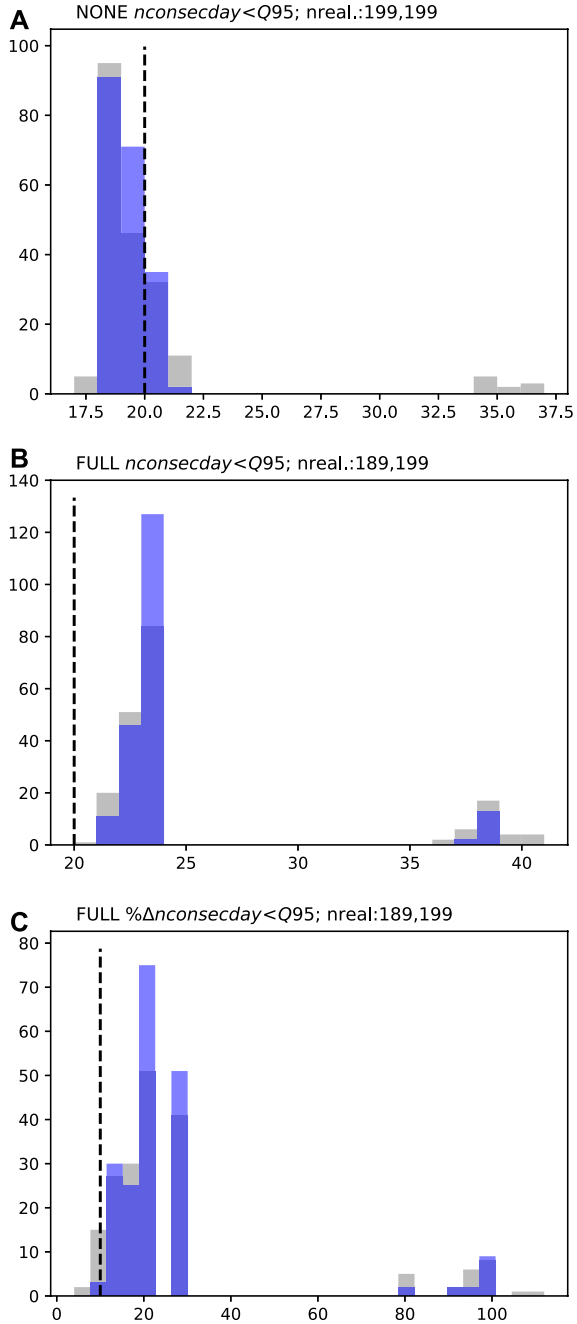


FIGURE 8 | MS2 low-flow persistence (*nconsecday<Q95;* **A** and **B**) and percent change low-flow persistence (*%nconsecday<Q95;* **C**) prior (grey) and posterior (blue) forecast PDFs. Dashed black lines represent the respective decision thresholds (20 days in **A** and **B** and 10% change in **C**). "NONE" represents the "naturalized" baseline and "FULL" the abstraction scenario. "nreal." details the number of successful realizations (therefore the number of data points represented in the PDF) for prior and posterior ensembles, respectively.

Collectively these results indicate that the conditioning of model parameters reflects significant parameter non-uniqueness, which is as a combined result of data scarcity and the distributed parameterization required to represent forecast relevant system detail (e.g., Sanford, 2011; Erdal and Cirpka, 2016; Knowling and Werner, 2016).

Despite the fact that variance reductions in the parameter PDFs are very subtle, a number of parameter groups do show uncertainty reduction. The most notable parameter variance reduction occurs for the global nitrate surface-loading rate parameter (**Figure 10H**). The apparent posterior uncertainty for this parameter is significantly constrained, with the approximated posterior probability distribution reducing from a range of 0.13–10 to a range of 0.45–1.57, with a variance reduction of 91%. This reflects the highly aggregated nature of this parameter; it captures information from all available nitrate concentration observations.

Only a few parameter groups show any evidence of first moment changes after history matching, i.e., the mean of the prior and posterior distributions are generally similar. Differences between mean values of prior and posterior parameter distributions indicate that either the mean in the prior distribution is inadequately defined, or alternatively, that there is a potential for bias due to inadequate model or parameter complexity. The only parameter group that shows a significant first moment response to data assimilation is the layer-constant porosity group (**Figure 10J**) which displays a first moment shift to higher values; the distribution mean for this porosity multiplier group shifts from 1.01 to 1.17. Higher porosity values effectively reduce simulated flow velocities, which in turn cause increased contaminant reaction (denitrification, in this instance). This inference of higher porosity results from the flow of information in the nitrate concentration observations to groundwater and surface water concentration simulated outputs which are the only simulated outputs that are sensitive to groundwater flow velocities. Over 90% of the 203 groundwater concentration observations are below 10 mg L^{-1}; prior simulated outputs corresponding to many of these observations, however, spanned significantly higher values (up to 80 mg L^{-1}; see **Supplementary Section SI 4.1.1**). So, while the prior distribution of simulated nitrate concentrations span the observed concentrations (i.e., there is no prior-data conflict), this indicates that the prior distributions for nitrate-loading and layer-based porosity parameters require some refinement.

4.2.2. Comparison of Simulated Outputs With Historical Observations

We note that the prior and posterior simulated output distributions in **Supplementary Section SI 4.1.1** indicate that, for these outputs at least, the prior distribution is conservative relative to the posterior distribution, i.e., the prior distributions (gray in **Supplementary Sections SI 4.1.1, 4.1.2**) fully encapsulate the posterior distributions (blue in **Supplementary Sections SI 4.1.1, 4.1.2**).

However, after the preliminary posterior uncertainty quantification some simulated output distributions no longer overlap with observations (blue bars and lines in **Supplementary Section SI 4.1.1, 4.1.2**). An example of this behavior can be seen with the preliminary posterior simulated outputs that relate to observations from well "s26_0656" (see **Supplementary Section SI 4.1.2**). Discrepancy between posterior simulated output distributions and observed system behavior may be a manifestation of observation error, causing tension in the history

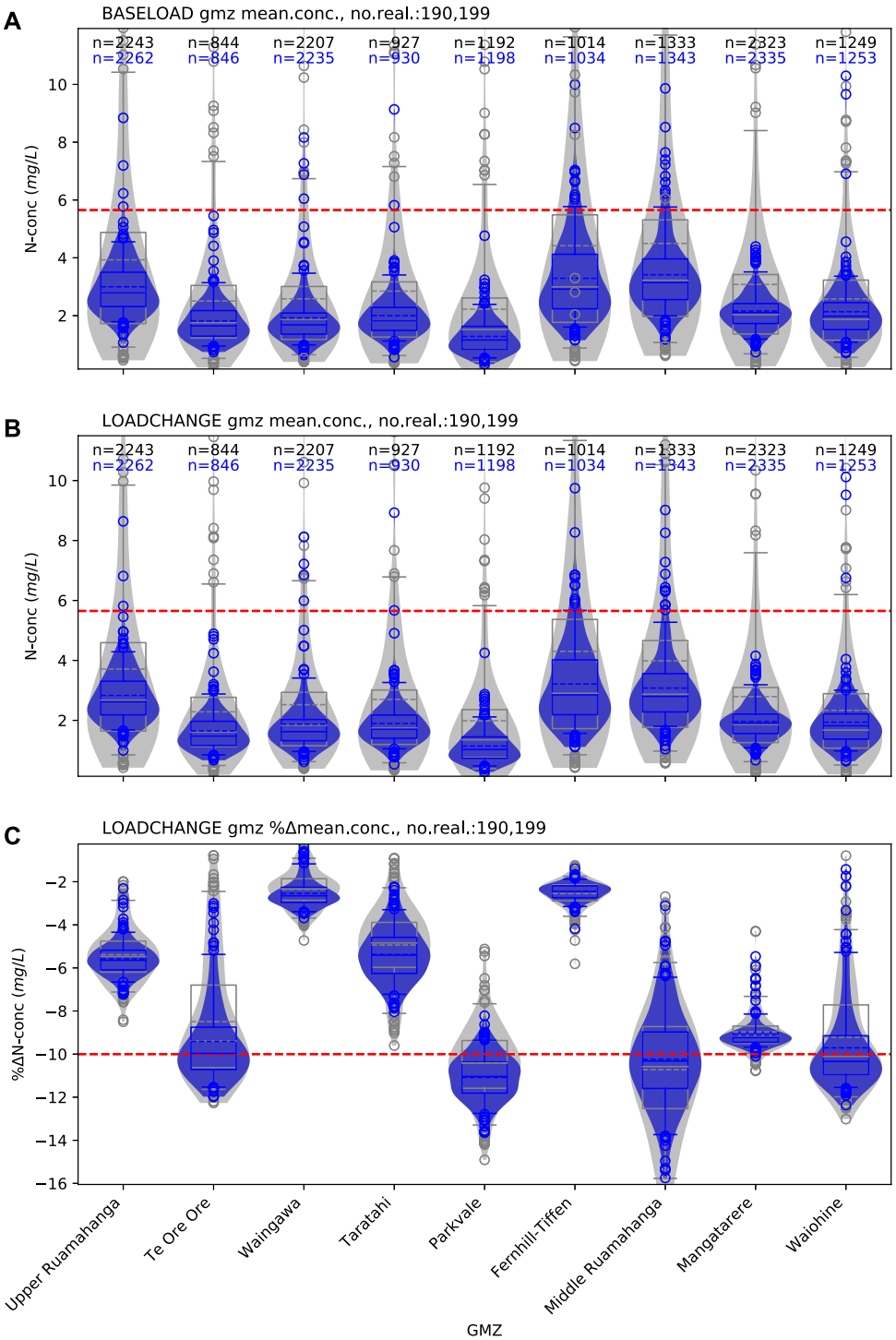


FIGURE 9 | Water quality prior (grey) and posterior (blue) forecasts distributions for mean zonal concentration in GMZs (see **Figure 2**). The width of the distribution indicates the distribution density. Box plots are also displayed to indicate the inter-quartile range (IQR, Q3-Q1, box), median (line in box), and 5th and 95th percentile (whiskers) of the distributions. Red dashed lines represent decision threshold values; for absolute mean zonal concentration forecasts (*mean.conc*, **A** and **B**) this threshold is 5.65mg l–1; for percentage-change forecasts (*%mean.conc*, **C**), the threshold is10% reduction in concentration."no.real." details the number of realization completed—therefore the number represented in the PDF—for prior and posterior ensembles, respectively. In **A** and **B**, "n" represents the mean number of active model cells in each zone, across all prior (black text) and posterior (blue text) realizations. "LOADCHANGE" represents the land-use management scenario explored relative to the baseline (denoted as "BASELOAD").

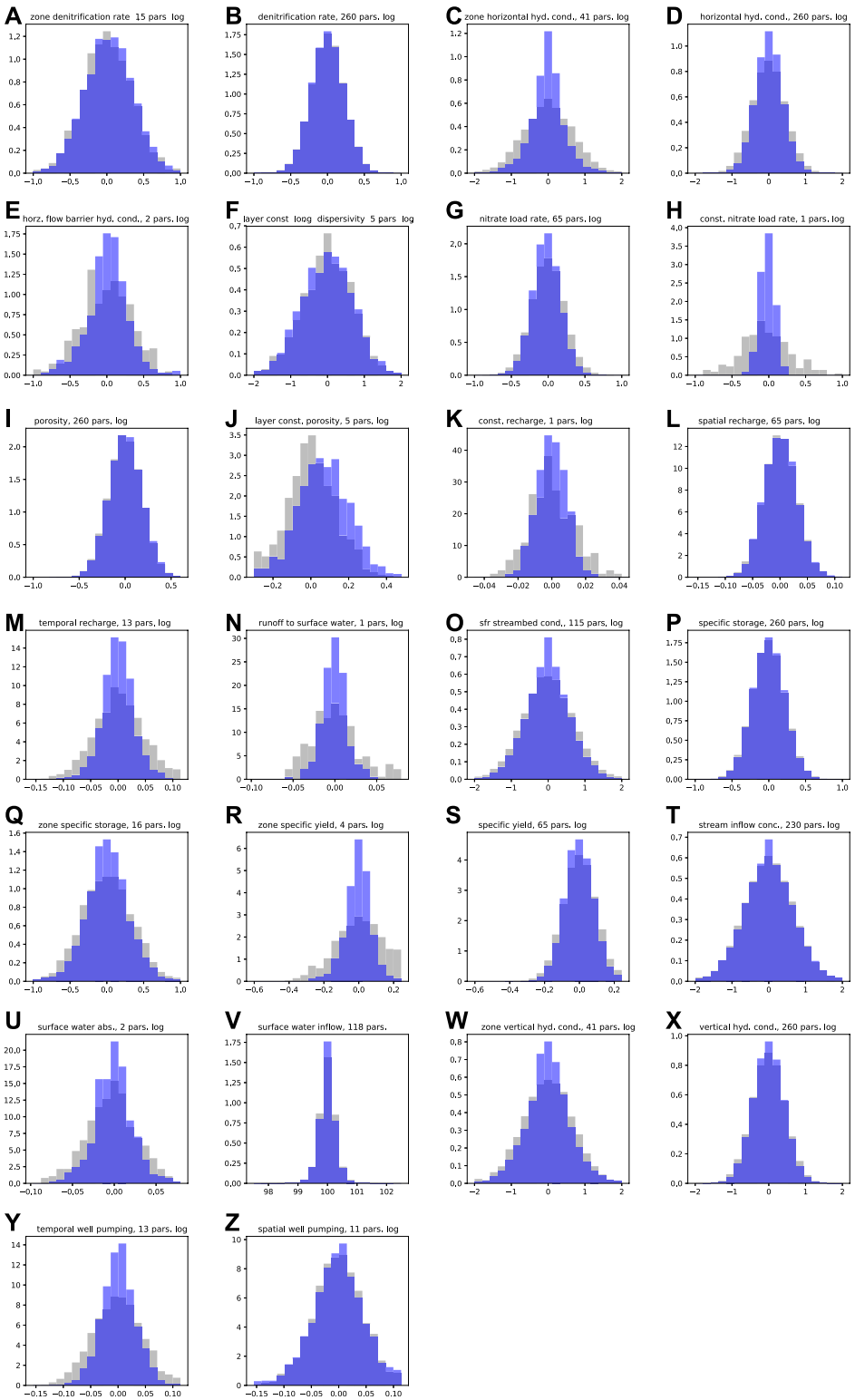


FIGURE 10 | Posterior parameter distributions (blue) plotted over prior parameter distribution (grey). The total number of parameters in the grouped distributions are detailed in the individual plot titles. If distributions refer to the log10 transformed parameters values, this is also detailed in the individual plots. Plot of grouped prior parameter distributions alone are presented in **Supplementary Figure SI 3.8**.

matching, or it may be an indication of insufficient complexity in the model parameter conceptualization, which inhibits simulation of system behaviors that are captured by the data. Evaluation of the source and implication of such results is important for assessing the robustness of posterior model forecast distributions.

For the specific example of well "s26_0656" observations, the preliminary posterior simulated outputs for the nearby "s26_0155" well effectively capture the observed water levels. Well "s26_0155" is in model layer 1, while "s26_0656" is in model layer 4, and as such it is anticipated to be more important for informing for the streamflow reliability forecasts. This supports continuing to the assessment of preliminary posterior forecast uncertainty. However, as the presence of such conflict could be an indication of insufficient parameter complexity, this assessment could provide justification for returning to the "Prior uncertainty quantification" step in the workflow and potentially reassessing discretization as well as parameterization, especially before any further history matching is considered.

4.2.3. Preliminary Posterior Forecast Uncertainty

4.2.3.1. Streamflow Reliability Forecasts

4.2.3.1.1. Low-Flow Frequency Forecasts. The abridged history matching, undertaken as part of the preliminary posterior uncertainty quantification, has reduced the uncertainty of the low-flow frequency percentage change forecasts, as demonstrated by narrower posterior PDFs (blue) compared to prior PDFs (gray) in **Figure 5**. However, the magnitude of the forecast uncertainty reduction varies between sites. The reduction in uncertainty is greatest at site PVW (**Figure 5A**) and results in the posterior forecast PDF moving above the decision threshold, indicating a certainty that the abstraction scenario will exceed the decision threshold; such certainty was not achievable based on the prior forecast PDF. At MS1 and MS2, the reduction in forecast uncertainty is relatively low (**Figures 5B,C**). However, as the prior forecast PDF is already wholly above the decision threshold, this limited reduction in uncertainty is of little consequence in the specific decision support context considered here. On the basis of these results, the workflow could move to "Stop modeling" after this preliminary posterior uncertainty quantification.

4.2.3.1.2. Low-Flow Persistence Forecasts. For the low-flow persistence forecasts, some apparent uncertainty reduction is achieved by the abridged history matching. At PVW, the reduction in forecast uncertainty moves the baseline scenario forecast away from the decision threshold (**Figure 6A**). However, for all other forecasts (at all sites), the abridged history matching has had limited impact on the position of forecast distributions, relative to their respective decision thresholds, and therefore provides little additional benefit for decision support. For these forecasts the observation dataset may be considered to be of little benefit for reducing forecast uncertainty. In this case, continued history matching would not be expected to sufficiently reduce forecast uncertainty and the modeling process should move to "Stop modeling."

4.2.3.2. Water Quality Forecasts

The abridged history matching reduces the uncertainty for the absolute mean zonal concentration forecasts, for all zones (blue PDFs in **Figures 9A,B**). For the land-use management scenario the variance reduction is such that only forecast outliers (beyond the 95th percentile) extend beyond the decision threshold (**Figure 9B**). For some zones the variance reduction is such that the posterior PDFs move entirely below the decision threshold (e.g., Te Ore Ore, Parkvale and Mangatarere). While the preliminary posterior forecast PDFs indicate low probability that the nitrate loading management scenario will result in mean zonal concentrations above the decision threshold of 5.65 mg L^{-1}, they are not sufficiently constrained through the abridged history matching to provide apparent certainty. However, the reduction in forecast uncertainty indicates that the observation dataset contains information that is beneficial for this forecast and the preliminary posterior uncertainty quantification provides justification to "Continue history matching."

For the percentage change forecast, there is relatively little reduction in uncertainty as a result of the abridged history matching (**Figure 9C**). Consequently, the variability of the predicted outcome of the management strategy, between zones, still persists. This result suggests that although the absolute forecast is informed by the observation data, the conditioning of parameters appears to have little influence on the normalized change forecast. This indicates that if uncertainty reduction for the percentage change in mean zonal concentration forecast is desired, there would be little benefit in continuing history matching with this combination of numerical model and observation data.

5. DISCUSSION

The role of uncertainty quantification in modeling for decision support is widely recognized in the literature (e.g., Freeze et al., 1990; Gupta et al., 2006; Moore and Doherty, 2006; Vrugt, 2016; Ferré, 2017). A requirement for the proposed decision support workflow, is that the forecasts of interest, and the decision thresholds associated with them, are defined at the outset, as outlined by the "Problem definition" step in **Figure 1**. This follows from a growing recognition that for decision support applications, model design should be based on the pre-definition of decision-relevant forecasts and the hypothesis to be tested (e.g., Guthke, 2017; White, 2017; Doherty and Moore, 2019).

We extend this 'start from the problem and work backwards' approach, to one that explicitly considers the environmental threshold that a management decision is seeking to avoid. In this context we demonstrate the important insights provided by a workflow which incorporates a preliminary assessment of forecast uncertainty. This can then support a number of ongoing modeling decisions, e.g.:

- the potential to assess the ability of a model and its associated prior parameter uncertainty to robustly simulate historical system observations, particularly those that are aligned with forecasts (i.e., to evaluate prior-data conflict);
- an opportunity to revisit model design and parameterization to allow simulated outputs to better reproduce system observations that are pertinent to forecasts;
- an opportunity to assess the extent to which forecasts are informed by observation data;
- an opportunity to process system observation data to provide the most appropriate information to inform the parameters that forecasts are sensitive to;
- provision of a defensible basis for undertaking (and continuing or stopping) history matching; and
- potentially, an opportunity to forgo history matching altogether, where it is not necessary to further support the decision making process.

As a result of the benefits listed above, the proposed workflow has the potential to reduce the time and effort required for decision support model deployment. For example, the prior uncertainty quantification and preliminary posterior uncertainty quantification (if required) can be achieved at low computational cost, relative to a more complete derivation of posterior uncertainty after a more exhaustive history matching effort. In the case study example, further history matching iterations would require 2,195 model runs per iteration, this equates to a computational cost of around 300 CPU days per iteration.

It should be noted however, that the history matching method used here follows a traditional approach, employing finite-difference gradients. Alternative approaches exist that support efficient data assimilation and uncertainty quantification with relativity few model forward simulations; for example, using ensemble based approximations (e.g., Chen and Oliver, 2013; White, 2018). Nevertheless, such approaches still require additional model forward simulations to approximate and propagate posterior parameter uncertainty. As the general workflow presented in **Figure 1** is agnostic of the specific uncertainty quantification methods employed, it still provides the potential for time and effort savings potential as well as protection (and potential for mitigation) against model failure induced by inappropriate history matching.

5.1. Assumption of a Conservative Prior

The proposed workflow is underpinned by the assumption that prior parameter distributions, and therefore prior forecast distributions, are "conservative." In this context, a conservative prior means that a forecast distribution will tend to inflate the inherent uncertainty in the modeled system behavior. This follows the need to avoid uncertainty variance under-estimation in modeling for decision support (e.g., Doherty and Simmons, 2013). Unfortunately, verification of whether the prior is sufficiently conservative is challenging, requiring, for example, paired model analysis, whereby prior forecast distributions are compared for pairs of models of differing complexity in order to highlight and expose predictive bias or variance corruption induced through model simplification (e.g., Doherty and Christensen, 2011; Gosses and Wöhling, 2019). In real-world decision support applications, undertaking such an analysis is rarely considered due to time and resource limitations, and the uncommon availability of multiple models of varying complexity.

Notwithstanding the challenges associated with formulating and verifying a conservative prior, a number of strategies to circumvent the effects of a non-conservative prior have been postulated in previous studies. These include: adopting high parameter dimensionality (e.g., Hunt et al., 2007; Knowling et al., 2019), with parameterization expressing system uncertainty at different spatial and temporal scales (e.g., White et al., 2020a; McKenna et al., 2020), and processing or transforming simulated outputs to minimize uncertainty, and thereby also the effects of an inadequate prior (e.g., Sepúlveda and Doherty, 2015; Knowling et al., 2019). Deploying such strategies is an important component of "Model definition" and prior formulation in "Prior uncertainty quantification" in the proposed workflow. Accordingly, in the case study presented herein, a highly parameterized scheme was adopted (e.g., Hunt et al., 2007), incorporating a combination of pilot points and zone- and layer-based parameters, with expert knowledge-based parameter variance and correlation; this aligns with an "intermediate" parameterization scheme in Knowling et al. (2019) which was found to be relatively robust for making predictions related to the depletion of low streamflows in response to groundwater pumping. We also considered "differenced" forecasts (e.g., percentage-change forecasts) in an effort to reduce possible ill-effects due the potential presence of large-scale boundary condition errors (e.g., Doherty and Welter, 2010; Sepúlveda and Doherty, 2015).

5.2. Insights and Advantages of Prior Uncertainty Quantification

The "Prior uncertainty quantification" step in the proposed workflow provides insights into the appropriateness of the conservative prior assumption, through comparison of the prior uncertainty surrounding simulated outputs with observations of system behavior (e.g., assessing the presence of prior-data conflict). Through this process the prior uncertainty quantification can also reveal errors in observation data at an early point in the workflow, before effort is wasted trying to match model outputs to errant data. Under the assumption that the observation errors are minimal, and are appropriately represented by the observation weights and the epistemic noise covariance matrix (see **Section 3.4.2**), the presence of prior-data conflict between simulated outputs and system observations may be an indication of potential inadequacy in the specified prior parameter uncertainty and/or in the way that the forecast relevant aspects of the system are parameterized (e.g., Nott et al., 2016). Where conflicts exist for data that are closely

related, in type and/or location, to the decision-relevant forecast, this inadequacy has the potential to propagate to the expression of forecast uncertainty—with the associated risk of incurring "model failure" (Doherty and Simmons, 2013).

In the presented case study example, the "Prior uncertainty quantification" step does indicate instances of prior-data conflict, for example, for some water level and high streamflow observations, as highlighted in **Section 4.1**. For the specific forecasts in the case study, the conflicts were not considered critical; i.e., the location and/or nature of the conflict between simulated outputs and system observations was not expected to impact on the reliability of the forecasts.

However, where model inadequacy, highlighted through the assessment of prior-data conflict, is deemed to risk the reliability of forecasts, additional work can be undertaken to improve the model parameter representation of forecast relevant aspects of the system and/or formulation of the prior parameter probabilities. The appropriate remedy depends on an assessment of the sufficiency of the model, its parameters, and of the expression of the prior parameter uncertainty, to adequately represent the system processes that are pertinent to the forecasts. If the parameter representation of the modeled system is deemed to be sufficiently complex to support full expression of the processes and process uncertainty that forecasts are sensitive to, the prior-data conflict indicates underestimated prior parameter variances. This can be remedied by revisiting the "Prior uncertainty quantification" step and inflating parameter variances. If however, the prior-data conflict is deemed to be due to an inadequate parameterization of specific processes, essentially creating a type of model structural defect, then the workflow should also return to the "Prior uncertainty quantification" step to revisit the parameterization scheme itself, as well as its probability distribution. In some cases, it will be the model definition itself that needs to be revisited, if the parameterization of a simulated process is not the issue, but rather the complete omission of the process. Iteration through these steps may be required. Importantly, undertaking this prior uncertainty quantification provides the necessary insights for identifying and rectifying model and parameter inadequacies at early stage in the workflow, before significant damage is done to the decision making process.

Once the model and prior parameter distribution is considered to adequately represent forecast relevant system processes, the prior uncertainty quantification also provides an opportunity to forgo a formal history matching effort. In the case study presented here, an example of this justification for forging history matching, is provided by the low-flow frequency forecast at MS1 and MS2 (**Figures 5B,C**). For this example, the prior forecast PDFs are wholly above the 10% decision threshold indicating that breaching of the management limit is almost certain to occur. In this case, and under the important assumption that the prior forecast uncertainty is conservative, the management decision could be made from a prior stance, with very low apparent risk of an incorrect assessment; efforts to reduce the uncertainty of the prediction through data assimilation would be unnecessary, as this would not change the evaluation of decision threshold exceedance, and the workflow can proceed to "Stop modeling."

Unfortunately, however, the capability of prior uncertainty quantification to provide a sufficient basis for a particular decision support context, is expected to be highly case specific. It is likely to depend on, for example, the inherent subjective definition of the decision threshold, the system parameterization and how the uncertainty of these parameters is defined (especially in regional-scale numerical models where parameters are abstractions of true natural system properties; Watson et al., 2013), as well as the risk tolerance of the decision-maker. Where the prior quantification of forecast uncertainty fails to provide sufficient certainty to support decision making (and prior-data conflict issues are minimal) the modeler may choose to undertake history matching with the aim of reducing model forecast uncertainty (e.g., the low-flow frequency forecast at PVW and the mean zonal concentration forecasts in the case study). In this instance the results of prior uncertainty quantification provides justification for moving to the "Preliminary posterior uncertainty quantification" step of the workflow.

5.3. Undertaking Approximate Preliminary Posterior Uncertainty Quantification

If history matching is considered necessary, approximation of the preliminary posterior uncertainty surrounding simulated outputs and forecasts at an early stage of the process provides another opportunity to assess the relationship between the model, its parameters, observation data, and ultimately, forecasts. Specifically, this early approximation of the posterior can provide insights into the ability of the observation data to sufficiently and appropriately inform model parameters that forecasts are sensitive to. Additionally, the resulting forecast uncertainty may be such that the preliminary posterior uncertainty quantification is sufficient for decision support model deployment.

For example, from the presented case study, the prior distribution of the low-flow frequency forecast at PVW extends just below the 10% decision threshold (**Figure 6A**). Consequently, for a decision maker with low risk tolerance (i.e., they desire low probability that the critical condition will not be violated), the prior uncertainty quantification does not provide sufficient support for the specific management decision. However, for this forecast, conditioning of parameters through the abridged history matching in the "Preliminary posterior uncertainty quantification" step was effective at achieving a sufficient reduction in forecast variance to support an assessment of management scenario outcome, with relative certainty; i.e., after the abridged history matching, the posterior forecast distribution was entirely above the decision threshold (**Figure 6A**). For this example, further history matching is not required, as no further uncertainty reduction is necessary to support management decision making. The modeling workflow could move to "Stop modeling."

As with the "Prior uncertainty quantification" step of the workflow, the potential to move to "Stop modeling" after

"Preliminary posterior uncertainty quantification" is also likely to be highly case specific. For example, for the absolute mean zonal concentration forecasts in this case study (**Figure 9**), the uncertainty remaining after abridged history matching, is still too high to provide certainty that the tested management scenarios will not result in conditions that exceed the decision threshold. However, in this instance the prior uncertainty quantification and subsequent preliminary approximation of the posterior uncertainty was still beneficial. It provided; i) an early indication of the ability of the model and prior parameterization to reproduce system observations that are pertinent to the forecasts; ii) an indication of the ability for system observations to inform model parameters to promote a sufficient reduction in forecast uncertainty; and, iii) indication of the likely implications of continuing history matching in an attempt to further reduce forecast uncertainty.

For the water quality forecasts in the case study, the data that were most directly relevant for the forecasts were measurements of groundwater nitrate concentration. While these observations are generally encapsulated by the prior simulated output distributions, and although some protection against the ill-effects associated with history matching may have been be gained by employing relatively high parameter dimensionality at multiple-scales, together with the spatially aggregated (mean) nature of the forecasts (e.g., Doherty and Christensen, 2011; White et al., 2014; McKenna et al., 2020), the simulated output distributions associated with the approximated posterior no longer reproduce higher concentration observations (see **Supplementary Section 4.1.1**). This indicates that the model parameterization and its prior probability distribution may not be sufficiently complex, and potentially may even require a refinement of the model discretization, which for the purposes of the workflow, can be considered part of a parameterization scheme. While further history matching iterations may result in the necessary variance reduction to provide the apparent certainty desired for the water quality forecasts, given the current model parameterization, the efforts to assimilate system data may induce bias in simulated outputs; therefore, further history matching (i.e., moving to the "Continue history matching" step) risks model failure (e.g., Doherty and Simmons, 2013).

6. CONCLUDING REMARKS

This study demonstrates the critical role of uncertainty quantification in model-based decision support. The quantification of uncertainty of simulated decision-relevant forecasts underpins the assessment of risks associated with management scenarios. Our case study demonstrates that the suggested workflow, which includes undertaking uncertainty quantification before comprehensive history matching efforts, brings significant advantages.

Quantification of model forecast uncertainty on the basis of prior parameter uncertainty is generally straightforward and relatively computationally efficient. It may be sufficient for informing an evaluation of the efficacy of a management strategy without the need for history matching. Depending on the proximity and considered robustness of the prior forecast probability distribution with respect to the decision threshold, it may be possible to conclude a modeling project this stage.

Prior uncertainty quantification also provides an opportunity to assess the validity of prior parameter distributions, through comparison of historical observations with prior simulated output distributions. Where prior-data conflict exists for forecast relevant observations, a modeler has the opportunity to revisit the model conceptualization and specification, including reformulation of prior parameter uncertainty, before embarking on history matching with a potentially flawed conceptualization.

Using a computationally efficient, albeit approximate, method for quantifying preliminary posterior uncertainty provides a defensible basis for undertaking further history matching. Importantly, this insight can be provided at an early stage of the modeling workflow.

Additionally, undertaking uncertainty quantification early in the modeling workflow can provide important insights into how the information in observation data flows to model parameters and forecasts. These insights can guide further model design and parameterization efforts, objective function formulation and observation processing, if deemed necessary, for example, on the basis of prior-data conflict.

We recommend adjusting the traditional modeling workflow, so that decision support modeling projects can benefit from the insights and potentially significant cost savings afforded by this early uncertainty quantification approach.

In summary, the approach provides:

- an early indication of the capacity of the numerical model and the chosen parameterization to numerically represent the decision-relevant forecasts, and therefore support decision making;
- an indication of the ability of observations to inform model parameters that the decision-relevant forecasts are sensitive to, and the opportunity to rectify incompatibilities between the model (parameters or structural design), observations and forecasts;
- the potential that the forecast uncertainty is sufficiently constrained to support a management decision or assess management strategy effectiveness without requiring history matching; and
- a quantitative and defensible basis for undertaking (and stopping) history matching for the purpose of reducing forecast uncertainty, with respect to management decision thresholds.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

The modeling work was led by BH with significant contribution to methods and results interpretation by all authors. All authors contributed to the writing of all components of the manuscript text, after an initial draft by BH.

FUNDING

This work was funded, in part, through the "Smart Models for Aquifer Management" (SAM) research program Grant Number: C05X1508 funded by the New Zealand Ministry of Business, Innovation and Employment, and was also supported by GNS Science Groundwater Strategic Science Investment Fund (SSIF).

ACKNOWLEDGMENTS

The case study example development was supported by Greater Wellington Regional Council (GWRC). The authors also thank Jeremy White for assistance throughout and a constructive review of preliminary versions of this manuscript. We also thank the helpful and constructive review of two reviewers whose efforts have helped to improve this manuscript considerably.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feart.2020.565613/full#supplementary-material

REFERENCES

Alfonzo, M., and Oliver, D. S. (2019). Evaluating prior predictions of production and seismic data. *Comput. Geosci.* 23, 1331. doi:10.1007/s10596-019-09889-6

Barnett, B., Townley, L., Post, V., Evans, R., Hunt, R., Peeters, L., et al. (2012). Waterlines Report. Australian groundwater modelling guidelines. Canberra: National Water Commission.

Bedekar, V., Morway, E. D., Langevin, C. D., and Tonkin, M. (2016). Report, U.S. Department of the Interior, U.S. Geological Survey. MT3D-USGS version 1: a U.S. Geological survey release of MT3DMS updated with new and expanded transport capabilities for use with MODFLOW.

Brouwers, M., Martin, P. J., Abbey, D. G., and White, J. (2018). Groundwater modeling with nonlinear uncertainty analyses to enhance remediation design confidence. *Groundwater* 56, 562–570. doi:10.1111/gwat.12669

Brynjarsdóttir, J., and O'Hagan, A. (2014). Learning about physical parameters: the importance of model discrepancy. *Inverse Probl.* 30, 114007. doi:10.1088/0266-5611/30/11/114007

Chen, Y., and Oliver, D. S. (2013). Levenberg-Marquardt forms of the iterative ensemble smoother for efficient history matching and uncertainty quantification. *Comput. Geosci.* 17, 689–703. doi:10.1007/s10596-013-9351-5

Clemo, T., Michaels, P., and Lehman, R. M. (2003). Transmissivity resolution obtained from the inversion of transient and pseudo-steady drawdown measurements. Tech. Rep. Boise, ID: Boise State University.

Dausman, A. M., Doherty, J., Langevin, C. D., and Sukop, M. C. (2010). Quantifying data worth toward reducing predictive uncertainty. *Groundwater* 48, 729–740. doi:10.1111/j.1745-6584.2010.00679.x

Doherty, J. (2015). *Calibration and uncertainty analysis for complex environmental models, PEST: complete theory and what it means for modelling the real world*. Brisbane, Australia: Watermark Numerical Computing.

Doherty, J., and Christensen, S. (2011). Use of paired simple and complex models to reduce predictive bias and quantify uncertainty. *Water Resour. Res.* 47. doi:10.1029/2011WR010763

Doherty, J. E., Fienen, M. N., and Hunt, R. J. (2011). Report 2010-5168. Approaches to highly parameterized inversion: pilot-point theory, guidelines, and research directions.

Doherty, J. (2003). Ground water model calibration using pilot points and regularization. *Ground Water* 41, 170–177. doi:10.1111/j.1745-6584.2003.tb02580.x

Doherty, J., and Hunt, R. (2010). Report 2010-5169. Approaches to highly parameterized inversion—a guide to using PEST for groundwater-model calibration.

Doherty, J., and Moore, C. (2019). Decision support modeling: data assimilation, uncertainty quantification, and strategic abstraction. *Groundwater* 58, 327. doi:10.1111/gwat.12969

Doherty, J. (2016). *PEST user manual part I and II*. 6th Edn. Brisbane, Australia: Watermark Numerical Computing.

Doherty, J., and Simmons, C. T. (2013). Groundwater modelling in decision support: reflections on a unified conceptual framework. *Hydrogeol. J.* 21, 1531–1537. doi:10.1007/s10040-013-1027-7

Doherty, J., and Welter, D. (2010). A short exploration of structural noise. *Water Resour. Res.* 46. doi:10.1029/2009WR008377

Enzenhoefer, R., Bunk, T., and Nowak, W. (2014). Nine steps to risk-informed wellhead protection and management: a case study. *Groundwater* 52, 161–174. doi:10.1111/gwat.12161

Erdal, D., and Cirpka, O. A. (2016). Joint inference of groundwater-recharge and hydraulic-conductivity fields from head data using the ensemble kalman filter. *Hydrol. Earth Syst. Sci.* 20, 555–569. doi:10.5194/hess-20-555-2016

Ferré, T. P. A. (2017). Revisiting the relationship between data, models, and decision-making. *Groundwater* 55, 604–614. doi:10.1111/gwat.12574

Fienen, M. N., Doherty, J. E., Hunt, R. J., and Reeves, H. W. (2010). Scientific Investigations Report 2010-5159. Using prediction uncertainty analysis to design hydrologic monitoring networks: example applications from the Great Lakes water availability pilot project. U.S. Geological Survey.

Freeze, R. A., Massmann, J., Smith, L., Sperling, T., and James, B. (1990). Hydrogeological decision analysis: 1. a framework. *Groundwater* 28, 738–766. doi:10.1111/j.1745-6584.1990.tb01989.x

Gosses, M., and Wöhling, T. (2019). Simplification error analysis for groundwater predictions with reduced order models. *Adv. Water Resour.* 125, 41–56. doi:10.1016/j.advwatres.2019.01.006

Gupta, H. V., Beven, K. J., and Wagener, T. (2006). *Model calibration and uncertainty estimation*. Atlanta, GA: American Cancer Society, Chap. 131. doi:10.1002/0470848944.hsa138

Guthke, A. (2017). Defensible model complexity: a call for data-based and goal-oriented model choice. *Groundwater* 55, 646–650. doi:10.1111/gwat.12554

Gyopari, M., and McAlister, D. (2010a). *Wairarapa Valley groundwater resource investigation: Middle Valley catchment hydrogeology and modelling*. Technical Publication GW/EMI-T-10/73. Greater Wellington Regional Council.

Gyopari, M., and McAlister, D. (2010b). *Wairarapa Valley groundwater resource investigation: upper Valley catchment hydrogeology and modelling*. Technical Publication GW/EMI-T-10/74. Greater Wellington Regional Council.

Hunt, R. J., Doherty, J., and Tonkin, M. J. (2007). Are models too simple? Arguments for increased parameterization. *Ground Water* 45, 254–262. doi:10.1111/j.1745-6584.2007.00316.x

Knowling, M. J., and Werner, A. D. (2016). Estimability of recharge through groundwater model calibration: insights from a field-scale steady-state example. *J. Hydrol.* 540, 973–987. doi:10.1016/j.jhydrol.2016.07.003

Knowling, M. J., White, J. T., Moore, C. R., Rakowski, P., and Hayley, K. (2020). On the assimilation of environmental tracer observations for model-based decision support. *Hydrol. Earth Syst. Sci.* 24, 1677–1689. doi:10.5194/hess-24-1677-2020

Knowling, M. J., White, J. T., and Moore, C. R. (2019). Role of model parameterization in risk-based decision support: an empirical exploration. *Adv. Water Resour.* 128, 59–73. doi:10.1016/j.advwatres.2019.04.010

Kunstmann, H., Kinzelbach, W., and Siegfried, T. (2002). Conditional first-order second-moment method and its application to the quantification of uncertainty in groundwater modeling. *Water Resour. Res.* 38, 6–1–6–14. doi:10.1029/2000WR000022

McKenna, S. A., Akhriev, A., Echeverría Ciaurri, D., and Zhuk, S. (2020). Efficient uncertainty quantification of reservoir properties for parameter estimation and production forecasting. *Math. Geosci.* 52, 233–251. doi:10.1007/s11004-019-09810-y

Ministry of Health (2018). Tech. Rep. Drinking-water standards for New Zealand 2005 (revised 2018). New Zealand Ministry of Health.

Moore, C., and Doherty, J. (2005). Role of the calibration process in reducing model predictive error. *Water Resour. Res.* 41. doi:10.1029/2004WR003501

Moore, C., and Doherty, J. (2006). The cost of uniqueness in groundwater model calibration. *Adv. Water Resour.* 29, 605–623. doi:10.1016/j.advwatres.2005.07.003

Moore, C., Gyopari, M., Toews, M., and Mzila, D. (2017). Consultancy Report 2016/162. Ruamāhanga catchment groundwater modelling. GNS Science.

Niswonger, R. G., Panday, S., and Ibaraki, M. (2011). "MODFLOW-NWT, a Newton formulation for MODFLOW-2005," in *Techniques and Methods*. (Reston, VA: U.S. Geological Survey), Vol. 6, book section 37.

Niswonger, R., and Prudic, D. (2005). Report, U.S. Geological Survey Techniques and Methods. Documentation of the Streamflow-Routing (SFR2) Package to include unsaturated flow beneath streams—a modification to SFR1.

Nott, D. J., Wang, X., Evans, M., and Englert, B.-G. (2016). Tech. Rep. Checking for prior-data conflict using prior to posterior divergences.

Oliver, D. S., and Alfonzo, M. (2018). Calibration of imperfect models to biased observations. *Comput. Geosci.* 22, 145–161. doi:10.1007/s10596-017-9678-4

Sanford, W. (2011). Calibration of models using groundwater age. *Hydrogeol. J.* 19, 13–16. doi:10.1007/s10040-010-0637-6

Sepúlveda, N., and Doherty, J. (2015). Uncertainty analysis of a groundwater flow model in east-central Florida. *Groundwater* 53, 464–474. doi:10.1111/gwat.12232

Sundell, J., Haaf, E., Tornborg, J., and Rosén, L. (2019). Comprehensive risk assessment of groundwater drawdown induced subsidence. *Stoch. Environ. Res. Risk Assess.* 33, 427–449. doi:10.1007/s00477-018-01647-x

Tarantola, A. (2005). "Inverse problem theory and methods for model parameter estimation." in *Other Titles in Applied Mathematics*. (Philadelphia, PA: Society for Industrial and Applied Mathematics). doi:10.1137/1.9780898717921

Vasco, D. W., Datta-Gupta, A., and Long, J. C. S. (1997). Resolution and uncertainty in hydrologic characterization. *Water Resour. Res.* 33, 379–397. doi:10.1029/96WR03301

Vasco, D. W., Peterson, J. E., Majer, E. L., and Majer, E. L. (1998). Resolving seismic anisotropy: sparse matrix methods for geophysical inverse problems. *Geophysics* 63, 970–983. doi:10.1190/1.1444408

Vrugt, J. A. (2016). Markov chain Monte Carlo simulation using the dream software package: theory, concepts, and matlab implementation. *Environ. Model. Software* 75, 273–316. doi:10.1016/j.envsoft.2015.08.013

Watson, T. A., Doherty, J. E., and Christensen, S. (2013). Parameter and predictive outcomes of model simplification. *Water Resour. Res.* 49, 3952–3977. doi:10.1002/wrcr.20145

Welter, D. E., White, J. T., Hunt, R. J., and Doherty, J. E. (2015). Report 7-C12. Approaches in highly parameterized inversion-PEST++ version 3, a parameter ESTimation and uncertainty analysis software suite optimized for large environmental models.

White, J. T. (2018). A model-independent iterative ensemble smoother for efficient history-matching and uncertainty quantification in very high dimensions. *Environ. Model. Software* 109, 191–201. doi:10.1016/j.envsoft.2018.06.009

White, J. T., Doherty, J. E., and Hughes, J. D. (2014). Quantifying the predictive consequences of model error with linear subspace analysis. *Water Resour. Res.* 50, 1152–1173. doi:10.1002/2013WR014767

White, J. T. (2017). Forecast first: an argument for groundwater modeling in reverse. *Groundwater* 55, 660–664. doi:10.1111/gwat.12558

White, J. T., Foster, L. K., Fienen, M. N., Knowling, M. J., Hemmings, B., and Winterle, J. R. (2020a). Toward reproducible environmental modeling for decision support: a worked example. *Front. Earth Sci.* 8, 50. doi:10.3389/feart.2020.00050

White, J. T., Knowling, M. J., and Moore, C. R. (2020b). Consequences of groundwater-model vertical discretization in risk-based decision-making. *Groundwater* 58, 695–709. doi:10.1111/gwat.12957

Conflict of Interest: This research was initiated while all authors were employed by the company GNS Science, a Crown Research Institute owned by the New Zealand Government.

frontiers in Earth Science

ORIGINAL RESEARCH
published: 02 February 2021
doi: 10.3389/feart.2020.586247

Recovering the Effects of Subgrid Heterogeneity in Simulations of Radionuclide Transport Through Fractured Media

Thomas Williams[1]*, *Jordi Sanglas*[2], *Paolo Trinchero*[2], *Guanqun Gai*[1], *Scott L. Painter*[3] and *Jan-Olof Selroos*[4]

[1]Jacobs, Harwell Campus, Didcot, United Kingdom, [2]AMPHOS 21 Consulting S.L., Barcelona, Spain, [3]Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, United States, [4]Svensk Kärnbränslehantering AB (SKB), Solna, Sweden

Groundwater flow and contaminant transport through fractured media can be simulated using Discrete Fracture Network (DFN) models which provide a natural description of structural heterogeneity. However, this approach is computationally expensive, with the large number of intersecting fractures necessitated by many real-world applications requiring modeling simplifications to be made for calculations to be tractable. Upscaling methods commonly used for this purpose can result in some loss of local-scale variability in the groundwater flow velocity field, resulting in underestimation of particle travel times, transport resistance and retention in transport calculations. In this paper, a transport downscaling algorithm to recover the transport effects of heterogeneity is tested on a synthetic Brittle Fault Zone model, motivated by the problem of large safety assessment calculations for geological repositories of spent nuclear fuel. We show that the variability in the local-scale velocity field which is lost by upscaling can be recovered by sampling from a library of DFN transport paths, accurately reproducing DFN transport statistic distributions and radionuclide breakthrough curves in an upscaled model.

Keywords: groundwater modeling, radionuclide transport, discrete fracture network, upscaling, downscaling

OPEN ACCESS

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

Reviewed by:
Jeffrey Hyman, Los Alamos National Laboratory, United States
Philip Stauffer, Los Alamos National Laboratory (DOE), United States

***Correspondence:**
Thomas Williams
tom.williams@jacobs.com

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

Received: *22 July 2020*
Accepted: *06 November 2020*
Published: *02 February 2021*

Citation:
Williams T, Sanglas J, Trinchero P, Gai G, Painter SL and Selroos J-O (2021) Recovering the Effects of Subgrid Heterogeneity in Simulations of Radionuclide Transport Through Fractured Media. Front. Earth Sci. 8:586247. doi: 10.3389/feart.2020.586247

1. INTRODUCTION

Understanding the fate and migration of dissolved contaminants is important for practical applications such as the optimal design of aquifer remediation strategies (Bolster et al., 2009), the delineation of protection areas around wells used for drinking water production (Trinchero et al., 2008) or the safety assessment of a repository for spent nuclear fuel (SKB, 2010).

Contaminant transport depends on the groundwater flow velocity field and this is in turn affected by the underlying heterogeneous distribution of hydraulic conductivity. Fully characterizing this distribution is impossible and thus upscaled macrodispersion models are typically used (Gelhar and Axness, 1983). However, these models are only valid for mild to low degrees of heterogeneity, whereas natural sedimentary aquifers often exhibit variations of up to 12 orders of magnitude in hydraulic conductivity (Bear, 1972; Sanchez-Vila et al., 2006). This requires the use of alternative upscaling methods (Hakoun et al., 2019).

In fractured crystalline rocks, transport patterns are particularly complex as they depend on both the topological configuration of the sparse network of fractures (network-scale heterogeneity) and the variable distribution of fracture openings (fracture-scale heterogeneity). Moreover, a large volume of these fractured

systems that is not affected by fluid flow is still accessible by solutes via a mechanism called matrix diffusion (Neretnieks, 1980). The coupling between solute advection along the open conductive fractures and diffusion into the rock matrix is a function of the underlying groundwater flow velocity field. It turns out that a proper characterization of groundwater velocity patterns is crucial to properly simulate contaminant transport (along conductive fractures) and retention (in the intact rock matrix).

An appealing approach for simulating groundwater flow and advective transport in fractured media is by means of Discrete Fracture Network (DFN) models (Cacas et al., 1990). DFN models account for the structural heterogeneity of the site, by means of field-derived statistical properties, and could also include fracture internal heterogeneity (Makedonska et al., 2016). In these models, contaminant retention in the rock matrix can be accounted for by using time domain particle tracking calculations (Painter et al., 2008).

Despite being methodologically attractive, DFN models are computationally challenging and sometimes simplifications are required, particularly when multiple stochastic realisations must be assessed. These simplifications are often carried out at the expense of a careful characterization of the local groundwater flow velocity field. Hence, methods to reconstruct the underlying variability of groundwater velocity are required. In particular it is important to recast the significant spatial persistence of high/low velocity regions along connected fractures, which was observed in previous works (Benke and Painter, 2003; Painter and Cvetkovic, 2005; Comolli et al., 2019; Hakoun et al., 2019). Different approaches have been proposed to represent that persistence (e.g., Benke and Painter, 2003; Painter and Cvetkovic, 2005; Comolli et al., 2019; Hakoun et al., 2019). Although these approaches differ in operational details, they are all based on some type of spatial Markov approximation for the Lagrangian velocity. This means that spatial persistence in velocity is simulated by assuming that the underlying Lagrangian velocity is a spatial Markov process, which implies that velocity in one fracture segment depends only on the velocity of the preceding segment.

A few approaches have been proposed to extrapolate or reconstruct velocity fluctuations, and related travel times, in large-scale DFN models. For instance, recently Hyman et al., 2019 have shown that heavy-tailed first passage time distributions can be reproduced using time domain random walk simulations based on transition times between fracture intersections. In Hyman et al., 2019 velocity variations within fractures are neglected and statistical ergodicity is invoked to sample velocity from a parametric distribution. The approach is shown to reproduce well the tail of a first passage time distribution but it is inadequate to describe the abrupt rising limb caused by early arrivals, which are mostly controlled by few large and transmissive fractures and are thus clearly not ergodic. A similar approach was presented earlier by Painter and Cvetkovic, 2005 to extrapolate velocity distributions from small-scale to much larger-scale DFN models. The main difference is that in the method by Painter and Cvetkovic, 2005 empirical distribution functions of the velocity field are derived from DFN models statistically representative of the large-scale fracture intensity, orientation and distribution. Therefore, the method is suited to reproduce the full travel time distribution and not only the tail.

Here, the algorithm of Painter and Cvetkovic, 2005 is used to reconstruct the lost variability in velocity by means of stochastic simulations. The proposed approach is used in combination with a computationally challenging DFN model of a Brittle Fault Zone (BFZ). In the DFN models used to assess the safety of proposed sites for geological repositories of spent nuclear fuel in fractured crystalline rock, the representation of BFZ is a topic of particular attention, due to the importance of such zones in determining site-scale flow and transport pathways (Hartley et al., 2018).

Radionuclide transport calculations are here carried out using particle tracking in the time domain (Painter et al., 2008). That and closely related approaches have been used in combination with discrete fracture networks in previous studies related to safety of geological disposal options (e.g., Selroos and Painter, 2012; Poteri et al., 2014; Trinchero et al., 2016, 2020), radionuclide transport at contaminated sites (Kwicklis et al., 2019), and production of natural gas from unconventional reservoirs (Karra et al., 2015). The focus here is on stochastic approaches to recover the transport effects of the subgrid velocity variability that is lost when using upscaled representations of fracture swarms associated with BFZs in large-scale flow models. In particular, we evaluate and refine the transport downscaling algorithm of Painter and Cvetkovic, 2005. All the radionuclide transport calculations are carried out using the computer code MARFA (Painter et al., 2008; Painter and Mancillas, 2013).

2. METHODS

2.1. Representation of Brittle Fault Zones as Fracture Swarms

Structurally, BFZ typically consist of a damage zone with finite thickness, containing many small fractures, about a central fault core (Aaltonen et al., 2016). Many BFZ are hydraulically active, and extend for multiple kilometres, providing hydraulic connectivity over long ranges with significant implications for the safety of geological repositories for spent nuclear fuel. In the groundwater flow and transport modeling calculations performed for repository safety assessments, the most commonly used approach has historically been to represent each BFZ as either a large planar fracture (in a DFN model) or a homogeneous volume (in a continuum model). This is the approach adopted in the previous generation of site descriptive modeling studies by both the Swedish (Joyce et al., 2010) and Finnish (Posiva, 2013) radioactive waste management organisations. In the DFN case, both stochastic variability and depth-dependency of flow and transport properties has sometimes been assigned to the plane using a relatively simple scaling-law formulation, relating transmissivity to fracture size, together with a normally-distributed random component, as in Joyce et al. (2010).

In the most recent iteration of the Olkiluoto DFN Version 3 (ODFN3) model, describing the site of the repository currently

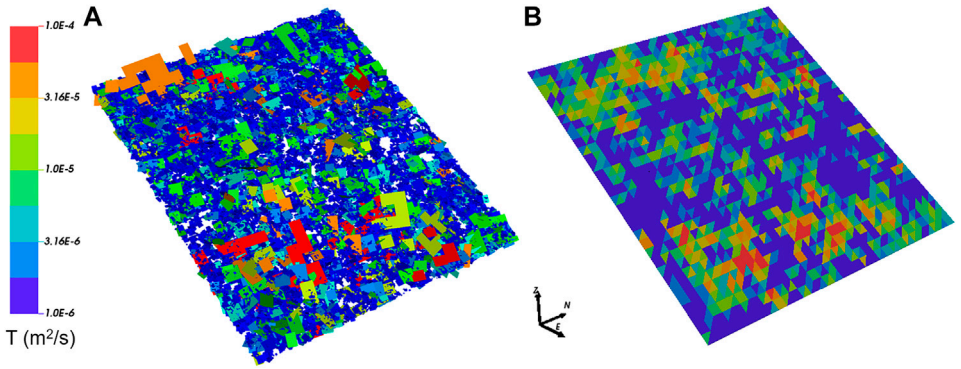


FIGURE 1 | Comparison of BFZ representations; **(A)** is a single realization of the underlying DFN model; **(B)** is the plane-projected representation of the same realization. Fractures colored by transmissivity.

under construction in Finland, a new representation of BFZ was developed (Hartley et al., 2018). In this formulation, BFZ damage zones are explicitly modeled as finite-thickness swarms of fractures stochastically generated around each fault core. Three-dimensional heterogeneity of the flow field within a zone can therefore be simulated, with effects such as choking of flow included. While this is a major advance, the added realism of this description comes at a computational cost. Methods for simplifying this BFZ model, while retaining heterogeneity in flow and transport properties, are desirable for the large number of particle tracking simulations required for a safety assessment calculation. This study therefore investigates several such methods by applying them to a simple test DFN model and comparing the results of flow and transport simulations.

2.2. Discrete Fracture Network Model of a Single Brittle Fault Zone

A simplified DFN model, consisting of a single BFZ fracture swarm overlaying a randomly oriented background fracture network is used as the basis of these simulations. Parameters are chosen to match a typical BFZ in the Olkiluoto DFN Version 3 (ODFN3) model (Hartley et al., 2018). One stochastic realization of the BFZ fracture swarm component of this model is shown in subfigure A of **Figure 1**.

The BFZ fracture swarm is stochastically generated about a midplane of dip angle 30°, dipping in the *x*-direction, with side length 577.4 m. This is located within a modeling domain of dimensions 500 × 500 × 500 m, intersecting the outer faces of the domain which lie in the *xz* and *yz* planes.

For the BFZ fracture set, fracture intensity is normally distributed in the direction orthogonal to the midplane, with a peak value at the midplane; in the ODFN3 model, the width of this distribution is adjusted to correspond to the observed damage zone width. Here, a peak volumetric intensity of 4.28 $m^2m^{\{-3\}}$ and standard deviation σ of 2.69 m are selected, corresponding to a typical ODFN3 BFZ. Fracture orientations are sampled from a Fisher distribution about the parallel to the midplane, with a Fisher dispersion constant *K* of 20; the resulting swarm fractures are therefore subparallel to the midplane.

For the background fracture set, a homogeneous volumetric intensity of 0.2 $m^2m^{\{-3\}}$ is used throughout the modeling domain, with a uniform orientation distribution over all possible angles.

In both BFZ and background sets, fractures are square in shape, with side length sampled from a truncated power-law probability distribution with exponent $k_r = 2.67$, minimum side length $r_{min} = 1.77$ m and maximum side length $r_{max} = 200$ m. Hydraulic aperture e_h follows a "semi-correlated" relationship to fracture length, consisting of a power-law distribution with a log-normally distributed random component:

$$\log_{10}(e_h) = \log_{10}(a) + b\log_{10}(L) + \alpha N(0,1) \quad (1)$$

where $a = 3.345 \cdot 10^{-5}$ m, $b = 0.5$ and $\alpha = 0.3$. A simple linear scaling of transport aperture e_t with hydraulic aperture is used ($e_t = 10e_h$). When solving for flow, the cubic law is used to determine fracture transmissivity: $T = e_h^3/12\mu$ where μ is fluid viscosity.

Following the ODFN3 methodology, both BFZ and background fractures of side length greater than 20 m are tessellated and a randomly selected portion of the resulting tessellates deleted, with the retained tessellates representing that portion of fracture surface area which is hydraulically open to flow. The retained portion of open fracture surface area $\omega(r,z)$ is determined by the following prescription:

$$\omega(r,z) = \begin{cases} (0.6 + 0.002z) + 0.05\left(\log_{10}(r) + \dfrac{\log_{10}(\pi)}{2}\right) & \text{if } z > 300\text{ m} \\ 0.2 + 0.05\left(\log_{10}(r) + \dfrac{\log_{10}(\pi)}{2}\right) & \text{if } z \le 300\text{m} \end{cases} \quad (2)$$

hence decreasing linearly with depth in the upper 200 m of the modeling domain, and constant below this point for a fracture of given length.

Flow and transport simulations are carried out for 10 stochastic realisations of the resulting DFN, using the ConnectFlow

groundwater modeling software suite (Jacobs, 2018). A pressure gradient across the modeling domain is created by imposing fixed pressure boundary conditions of 15,000 Pa at the inlet ($+x$) boundary and 0 Pa at the outlet ($-x$) boundary of the domain, and a steady-state flow solution obtained for each of the 10 realisations. Pathline tracking simulations using this flow field are performed for an ensemble of particles injected uniformly over the inlet boundary, at the intersections between the BFZ swarm and the inlet boundary.

For the MARFA downscaling method described in 2.4 to be applied, samples must be taken from a library of pathline legs calculated within a smaller "sampling cube" model. The DFN parameters within this 100 × 100 × 100 m model are homogeneous and equal to those for fractures at the BFZ midplane (i.e., peak intensity). For this set of parameters, 10 realisations of the sampling cube model are generated, and flow and transport calculations carried out in each of the three axial directions, with an applied hydraulic gradient equal to that applied in the full DFN model.

2.3. Plane-Projected Model

For comparison to the DFN model described in section 2.2, a plane-projected representation is produced using the upscaling and projection methods developed in Baxter et al., 2019. The first stage of this process requires upscaling of the DFN to an equivalent continuous porous medium (ECPM) model in ConnectFlow. The DFN model is subdivided into submodels (5 × 5 × 5 m cubes); in each of these, a linear hydraulic gradient is applied in each of the three axial directions, and the flux through the portion of the fracture network within the submodel calculated. A hydraulic permeability tensor **k** is then evaluated to fit the calculated fluxes for each submodel. Where cross-flows (non-parallel to the head gradient) arise, indicating anisotropy due to the orientation, connectivity and transmissivity of the underlying DFN, this is approximated by introducing nonzero off-diagonal elements to the tensor.

In addition to **k**, which is used to calculate flux in an ECPM model, transport simulations require two additional quantities to be evaluated; equivalent kinematic porosity ϕ and equivalent flow-wetted fracture surface area a_r. These are calculated using the flow-based transport upscaling method described in Baxter et al., 2019, which was shown to provide a more accurate estimate of these quantities when compared to the simple geometric methods previously used, as only those portions of the fracture network through which flow and transport actually occurs are included. Once **k** has been determined, particle tracking is performed in each of the three axial directions within each submodel and the distributions of travel time and flow-related transport resistance are evaluated over the ensemble of particles. The median of each distribution is then used to calculate the directional components of each of these quantities:

$$\phi_i = \frac{Q_i}{A_i}\left\langle \frac{\tau_{ij}}{L_{ij}} \right\rangle_m \qquad a_{r,i} = \frac{Q_i}{A_i}\left\langle \frac{\beta_{ij}}{L_{ij}} \right\rangle_m \tag{3}$$

where Q_i is the flux out of the downstream face and A_i is the transverse cross-sectional area in direction i, and τ_{ij}, L_{ij} and β_{ij} are respectively the travel time, path length and flow-related transport resistance (defined in **section 2.4**) for path j when a pressure gradient is applied in direction i. The subscript m indicates that the median value of these distributions is taken. Although **Eq. 3** results in vector quantities for kinematic porosity and flow-wetted surface, in practice, transport simulations (including those described here) typically require them to be scalar and hence the geometric mean of each is used.

An ECPM model can be used directly for flow and transport simulations, but this paper only considers it as an intermediate step between DFN and plane-projected representations. The process of projecting ECPM properties onto a plane results in a further simplification in which spatial heterogeneity in flow and transport properties is lost orthogonal to the BFZ midplane, but retained within the plane, averaged on the scale of triangulation. The resulting planar BFZ representation provides a computationally less expensive framework for flow and transport simulations in a heterogeneous BFZ damage zone, and is particularly attractive for use in safety assessment calculations over many stochastic realisations, as geostatistical methods such as sequential Gaussian simulation can be applied to a single plane-projected realization to generate many more, without the computational expense of DFN generation, upscaling and plane-projection for each realization.

The plane projection method, described in detail in Baxter et al., 2019, can be summarized as follows:

(1) For all ECPM elements, construct line geometries passing through the element centroids in each of the three axial directions (this method requires that elements form a Cartesian grid).
(2) Triangulate the BFZ midplane at the appropriate level of discretization (in this case, right-angled triangles with legs of length 1.8 m).
(3) For each triangle, determine the largest axial component of the triangle normal; this is referred to as the "collapse direction".
(4) Find those line geometries calculated in Step 1 which are parallel to the collapse direction and intersect the triangle. Calculate the points at which these intersections occur; these are referred to as the "collapse points".
(5) For each collapse point, aggregate a list of those elements located along the line geometry which are within one damage zone half-thickness (projected into the collapse direction) from the midplane.
(6) Calculate effective properties at the collapse points for these lists of elements as follows:

- Calculate the rotation matrix **Q** corresponding to a rotation between the triangle normal and unit normal in the collapse direction, and use this to calculate the rotated permeability tensor: $\mathbf{k}' = \mathbf{Q}\,\mathbf{k}\,\mathbf{Q}^{\mathbf{T}}$.
- Calculate the magnitude of the in-plane component of the rotated permeability tensor in each element, sum over the elements and convert to transmissivity: $T = H\frac{\rho g}{\mu}\sum_{i=1}^{n} k'_{\text{plane}}$ where H is the damage zone width, ρ is fluid density, μ is fluid viscosity and g is gravitational acceleration.

- For porosity and flow-wetted surface, the arithmetic mean over elements is taken: $\overline{\phi} = \frac{1}{n}\sum_{i=1}^{n}\phi_i$ and $\overline{a_r} = \frac{1}{n}\sum_{i=1}^{n} a_{r,i}$.

(7) Aggregate the properties for all collapse points within the triangle. For porosity and flow-wetted surface, an arithmetic mean over collapse points is taken. For transmissivity, a geometric mean is taken and multiplied by the proportion of collapsed grid cells which are active for flow (i.e. have a non-zero permeability value).

The resulting heterogeneous planar representation for one realization of the model is shown in subfigure B of **Figure 1**, for comparison with the underlying DFN model shown in subfigure A.

As for the DFN model in section 2.2, flow and transport simulations are carried out for 10 stochastic realisations of the equivalent plane-projected model, using identical boundary conditions and modeling parameters.

2.4. Downscaling Method

In sparsely fractured media, such as crystalline rocks, contaminant transport occurs along discrete transport pathways that connect the source location to a receiver location or compliance boundary. These pathways lie on fracture planes and are made up of consecutive segments, each segment being delimited by fracture-to-fracture intersections. Contaminants are displaced by groundwater flow along the pathway segments (advection) and can access the adjacent intact rock matrix by a mechanism called matrix diffusion. Contaminant retardation in the rock matrix might be further enhanced by sorption onto the available mineral surfaces.

In a Lagrangian sense, advection is described by the groundwater travel time, τ [T], whereas the coupling between advection and matrix diffusion is characterized by the so-called transport resistance, β [T L^{-1}] (also sometimes denoted by *F*), which is defined as:

$$\beta(\tau) = \int_0^{\tau} \frac{d\xi}{b(\xi)} d\xi \tag{4}$$

where *b* [L] is the fracture half-aperture. From **Eq. 4** it is evident that β and τ are highly correlated.

The simplification of DFN models by means of upscaling techniques (e.g. **section 2.3**) provides good approximations of the underlying groundwater fluxes (Jackson et al., 2000) but it can lead to biased distributions of τ and β (Painter and Cvetkovic, 2005) due to lost variability. Hence, Painter and Cvetkovic, 2005 developed a method for restoring the lost variability by stochastic simulation. In the original context, they considered the stochastic simulation as a way of extrapolating from small scale DFN models to much larger models. Here, we use the Painter and Cvetkovic, 2005 approach as a downscaling algorithm to reconstruct this lost variability based on sample DFN models. This algorithm, which is used in the calculations presented in the next sections, is here briefly described.

Each single transport pathline connecting the source to the boundary is divided into a number of sub-segments, with each sub-segment being fully defined by the triplet $|\Delta_i, \tau_i, \beta_i|$, where Δ [L] is the sub-segment length and sub-script *i* identifies the *i*th sub-segment. The total pathline length, L_T [L], is approximated as $L(n) = \sum_{i=1}^{N_T}\Delta_i$ with $N_T = \min\{n : L(n) \geq L_T\}$ and the total cumulative groundwater travel time and transport resistance are given by:

$$\tau(N_T) = \sum_{i=1}^{N_T} \tau_i \qquad \beta(N_T) = \sum_{i=1}^{N_T} \beta_i \tag{5}$$

The triplet $|\Delta, \tau, \beta|$ is a random process subordinated to N_T, which is also a random variable. At this point, no assumption has been made about the distribution of $|\Delta_i, \tau_i, \beta_i|$ along the transport pathway. Evidence has pointed out a significant spatial persistence of high/low velocity regions along connected fractures (Benke and Painter, 2003; Painter and Cvetkovic, 2005; Comolli et al., 2019; Hakoun et al., 2019). Several approaches have been developed to represent that persistence (Benke and Painter, 2003; Painter and Cvetkovic, 2005; Comolli et al., 2019; Hakoun et al., 2019). Although those approaches differ significantly in operational details, they all rely on some type of spatial Markov approximation for the Lagrangian velocity. That is, spatial persistence in velocity is accounted for by assuming that Lagrangian velocity is a spatial Markov process, meaning that velocity in one segment depends only on the velocity of the preceding segment. To implement that spatial dependence, we adopt the algorithm of Painter and Cvetkovic, 2005 for its simplicity and to take advantage of its implementation in the Marfa software (Painter et al., 2008; Painter and Mancillas, 2013). Operationally, the velocity space is discretized into velocity classes, which define an internal state, S_j, for particles transiting through a segment of transport pathway. Evolution of a particle's state is then modeled as a discrete-state Markov chain. In a discrete state Markov chain, changes of states are governed by a transition matrix: P_{jk}. The probability of transitioning from state S_j to all other states is 1, thus $\sum_{k=1}^{N_S} P_{jk} = 1$, where N_S is the total number of states. The sum over all the preceding states gives instead the probability for state *k*: $\sum_{j=1}^{N_S} P_{jk} = P_k$. Here we construct the transition matrix from a set of DFN training simulations and use the transition matrix to simulate the discrete state Markov chain. After determining a state in this algorithm, a triplet (Δ, τ, β) is then sampled from the probability distribution conditional on the internal state. In this Markov model the probability of a triplet (Δ, τ, β) is conditional on the state and thus the unconditional probability for the first sub-segment is given by:

$$f(\Delta_1, \tau_1, \beta_1) = \sum_{j=1}^{N_S} f\left(\Delta_1, \tau_1, \beta_1 | S_j\right) P_j \tag{6}$$

The joint distribution of triplets at downstream locations is:

$$f(\Delta_1, \tau_1, \beta_1, \Delta_2, \tau_2, \beta_2, \dots) = \sum_{j=1}^{N_S} f\left(\Delta_1, \tau_1, \beta_1 | S_j\right) P_j \cdot \sum_{k=1}^{N_S} f\left(\Delta_2, \tau_2, \beta_2 | S_k\right) P_{jk} \cdot \dots \tag{7}$$

This model is referred to as Markov-directed random walk (MDRW). If spatial persistence in velocity is not accounted

for, the framework reduces to a random walk (RW) model and its joint distribution is given by:

$$f(\Delta_1, \tau_1, \beta_1, \Delta_2, \tau_2, \beta_2, \dots) = f(\Delta_1, \tau_1, \beta_1) f(\Delta_2, \tau_2, \beta_2) \cdot \dots \quad (8)$$

The practical implementation of both MDRW and RW is carried out by constructing a number of equiprobable DFN models. These DFN's are generated in a cubical domain and thus are denoted as *sampling cubes*. The fracture recipe of each DFN is based on the statistical properties of the core of the BFZ, which implies that statistical homogeneity is invoked. This hypothesis is in contrast to the assumption used to construct the BFZ, which considers that fracture intensity is continuously decreasing at increasing distances from the core plane of the BFZ (see **section 2.2**). The implication of this simplification is discussed in **section 3.2**.

A synthetic global gradient, ϵ^g [-] is applied along the three principal directions of the sampling cubes and particles are traced from inlets to outlets, thus providing libraries of triplets $|\Delta_i, \tau_i, \beta_i|_{\boldsymbol{u}_j}$, where $\boldsymbol{u}_j$ is the unit vector that indicates the jth principal direction. These libraries are of direct use for the RW downscaling calculations. For MDRW, following the procedure illustrated in Painter and Cvetkovic, 2005, sub-segments are classified according to the local particle velocity. This classification provides the internal state of the considered sub-segment.

The downscaling approach is implemented as follows (notice that a *segment* refers here to the transport pathways computed using the BFZ model whereas a *sub-segment* is related to the libraries generated with the sampling cubes).

(1) The nth particle is launched by randomly sampling a starting pathway and a starting start time t_0 [T].
(2) The particle is placed at the inlet of the first segment, $\boldsymbol{r}_1$, and the library with closest direction is chosen: $\boldsymbol{u} = \min\left\{\left|\boldsymbol{u}_j \cdot \frac{\boldsymbol{r}_1}{\|\boldsymbol{r}_1\|} - 1\right|\right\}$. The length counter is set to zero: $L = 0$.
(3) The triplet $|\Delta_0, \tau_0^*, \beta_0^*|$ is randomly selected from the library. The asterisk indicates that the variables have been scaled by the ratio between the local gradient along the segment and the global gradient applied in the sampling cube, ϵ^1/ϵ^g. The length counter is updated ($L = L + \Delta_0$), the retention time in the matrix (t_D [T]) is sampled conditioned on β_0^* (see Painter et al., 2008 for further details) and the particle clock is updated accordingly ($t_n = t_0 + t_D + \tau_0^*$). The state of the following sub-segment, S_{0+1} is recorded.
(4) The triplet $|\Delta_1, \tau_1^*, \beta_1^*|$ is drawn from all the sub-segments belonging to class S_{0+1}. The length counter and particle clock are updated and the state of the following sub-segment, S_{1+1} is recorded.
(5) Step 4 is repeated until $L > ||\boldsymbol{r}_1||$, which is when the particle is placed at the inlet of the following segment ($\boldsymbol{r}_2$). The library is updated according to the direction of the new segment (step 2) and step 4 is repeated until hitting the outlet of the new segment and so on until reaching the outlet boundary. The exceeding sub-segment length, and the related exceeding τ and β are used for the following segment.

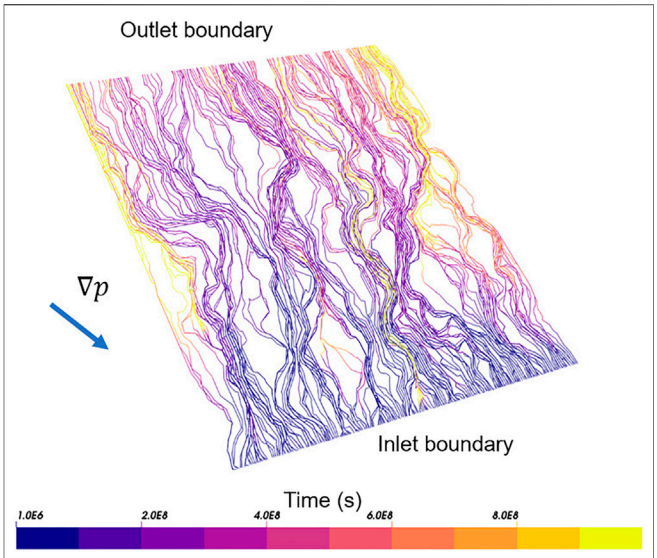


FIGURE 2 | Pathlines from ConnectFlow transport simulation in plane-projected representation, colored by travel time. Direction of pressure gradient with respect to inlet and outlet boundaries is shown.

The algorithm described in steps 1-5 is the same presented in Painter and Cvetkovic, 2005 and is from now on denoted as *standard downscaling algorithm* (SDA). As already discussed above, this Markov chain algorithm is subordinated to the length of the segment and, thus, the total length of the transport pathway. It turns out that, to properly reconstruct the signature of the original DFN model, the SDA must be applied to upscaled transport pathways that somehow preserve the tortuosity of the original set of trajectories. The collapsed plane model is supposed to preserve the trajectory tortuosity by means of the representation of the DFN heterogeneity onto the BFZ plane. However, for the single trajectory of the homogeneous plane, tortuosity is lost and needs to be somehow reconstructed. To this end, we propose here an alternative implementation of the algorithm, denoted as the *projected sub-segment algorithm* (PSSA). In PSSA, steps 3 to 5 are modified by projecting the drawn value of Δ along the selected principal direction (step 2): $\Delta^* = \Delta \cdot \boldsymbol{u}$.

3. RESULTS

3.1. Initial Particle Tracking Results

Once ConnectFlow has computed the flow field within the model, it is used to carry out particle tracking simulations; pathline results for a single realization of the plane-projected model are shown in **Figure 2**. The initial positions of particles are evenly distributed along the inlet boundary. As particles cross the plane toward the outlet boundary, their pathlines are observed to cluster due to the in-plane heterogeneity of properties (shown in **Figure 1**) creating preferential paths for flow and transport. In some regions of the plane, choking of flow appears to occur, with resulting retardation to certain pathlines (those with longer travel times).

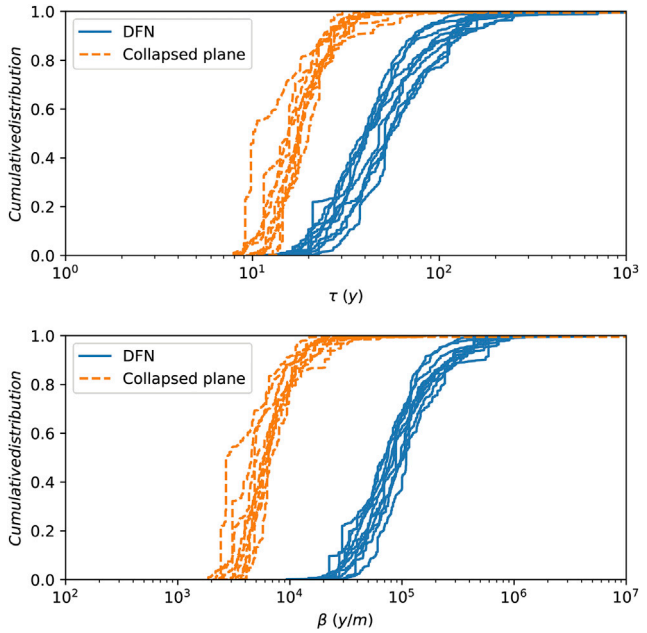


FIGURE 3 | Cumulative distribution of τ and β of the trajectories of all DFN realizations.

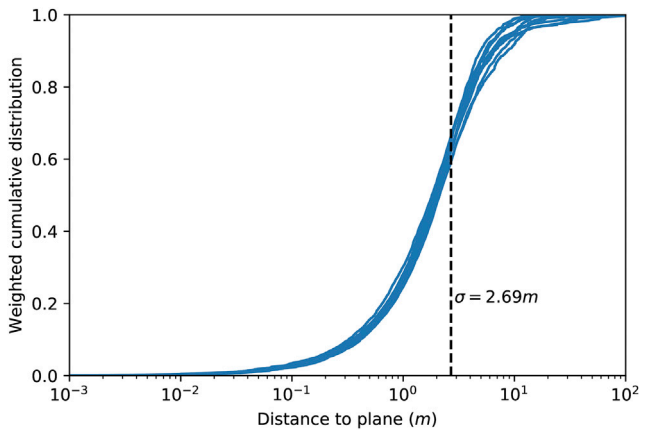


FIGURE 4 | Cumulative distribution of the distance to plane of the DFN realizations' segments, weighted by length. The vertical black line shows the sigma distance.

This results in distributions of τ (travel time) and β transport resistance over all particles, which are plotted in **Figure 3**. However, a comparison of results for the underlying DFN and plane-projected representations shows that a significant mismatch exists for both of these transport statistics in all stochastic realisations; the median value of each distribution is shifted by approximately half an order of magnitude for τ and a full order of magnitude for β, with the plane-projected results displaying shorter travel times and less transport resistance. The distributions are also narrower in the plane-projected case.

TABLE 1 | Parameters of the rock matrix used in the *Conservative case* and the *Decay Chain case*.

Parameter	Value
Rock density	2,600 kg/m^3
Matrix depth	4.5 m
Matrix porosity	0.018
Matrix effective diffusivity	8.5×10^{-7} m^2/y

These results indicate that the method of projection to a midplane described in **section 2.3**, while capturing some of the heterogeneity of the underlying DFN swarm, has not retained a full description of the transport properties of the simulated test BFZ. Particles transit through the BFZ too quickly, and thus corrections must be applied to the pathline results.

3.2. Downscaling Tests

The ability of the MARFA downscaling algorithm to reintroduce lost heterogeneity in the plane-projected BFZ representation and hence resolve the observed mismatches with τ and β in the DFN representation is investigated.

In **figure 4** the cumulative distribution of the distance of all DFN realizations' segments to the main BFZ plane can be seen. The cumulative distribution has been weighted by length, so that is independent of how the trajectories are discretized in segments. The vertical line shows the value of sigma distance σ. It can be seen that most of the length of the trajectories is found between the BFZ plane and a distance *s*. For that reason, the core sampling cubes are the ones used for the downscaling algorithm.

As described in **section 3.1**, transport pathways are traced using ConnectFlow by means of particle tracking calculations with particles being placed uniformly over the inlet boundary. In all the calculations presented hereafter a correction is applied to account for flux-weighted particle injection. Moreover, the results of the ten realisations of the sampling cubes are merged into single sub-segment libraries. Details on the pre-processing procedures used are provided in **Appendix 1: Processing Transport Pathways**.

Two test cases are carried out. In the first test, denoted the *Conservative case*, a short pulse of a non-sorbing non-decaying radionuclide is injected at the BFZ boundary. Radionuclide transport is computed for each of the ten BFZ realizations using the related transport pathways and the rock matrix parameters listed in **Table 1**. The exercise is repeated first using the ten realisations of the collapsed plane without the downscaling algorithm and then by applying the downscaling algorithm on the same set of pathways. Moreover, the downscaling approach is also carried out using the single straight pathway representative of the homogeneous plane. The same procedure is used for the second test, denoted the *Decay Chain case*, where the 4n+2 chain of U238 is simulated. The decay chain is simplified as follows:

$$\text{Cm246} \rightarrow \text{Pu242} \rightarrow \text{U238} \rightarrow \text{U234} \rightarrow \text{Th230} \rightarrow \text{Ra226}$$

The radionuclide-specific parameters are listed in **Table 2**, whereas the parameters of the rock matrix are the same as in

TABLE 2 | Radionuclide-specific parameters for the *Decay Chain case*.

Radionuclide	K_d [m^3/kg]	λ [y^{-1}]
Cm246	$1.43 \cdot 10^{-2}$	$1.47 \cdot 10^{-4}$
Pu242	$1.43 \cdot 10^{-2}$	$1.86 \cdot 10^{-6}$
U238	$5.09 \cdot 10^{-2}$	$1.55 \cdot 10^{-10}$
U234	$5.09 \cdot 10^{-2}$	$2.82 \cdot 10^{-6}$
Th230	$5.09 \cdot 10^{-2}$	$9.19 \cdot 10^{-6}$
Ra226	$1.93 \cdot 10^{-4}$	$4.33 \cdot 10^{-4}$

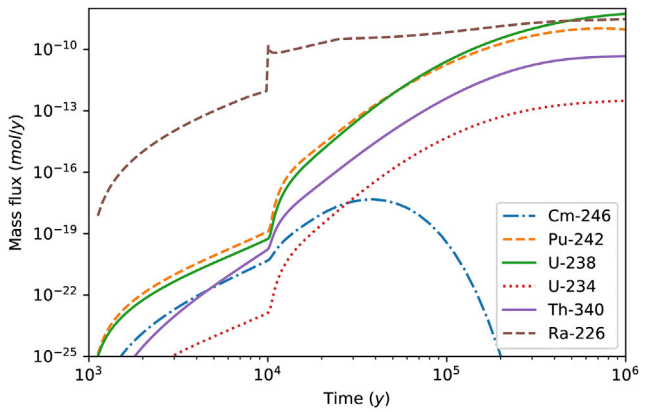


FIGURE 5 | Plot of the source function of all radionuclides used in the *Decay Chain case*.

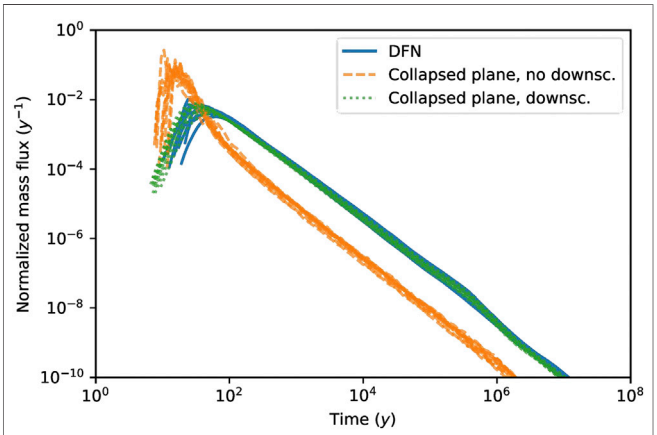


FIGURE 6 | *Conservative case*: breakthrough curves of a non-sorbing non-decaying radionuclide computed using transport pathways obtained using the different realisations of the DFN model (blue continuous curves) compared with related calculations carried out using transport pathways derived from the collapsed plane model and without downscaling (orange dashed lines) and with related calculations where downscaling is applied (green dotted line).

the *Conservative case* (**Table 1**). The source functions (**Figure 5**) are randomly selected from the 6,234 functions used to describe a hypothetical release at canister location in the so-called Forsmark Growing Pinhole Scenario (FGPS) (Selroos and Painter, 2012; Trinchero et al., 2020). FGPS was one of the modeling cases used in SR-Site, the safety assessment study for the planned repository for spent nuclear fuel at Forsmark, Sweden.

TABLE 3 | Results of the Kolmogorov–Smirnov (KS) test for the different variant simulations of the *Conservative case*. The KS test is run for each DFN realization and related collapsed plane or homogeneous model and the results are provided in terms of maximum ($D^{max}_{n,m}$), minimum ($D^{min}_{n,m}$) and average ($\langle D_{n,m} \rangle$) test statistic.

Cases	$D^{max}_{n,m}$	$D^{min}_{n,m}$	$\langle D_{n,m} \rangle$
DFN vs. collapsed plane no downscaling	$8.79 \cdot 10^{-1}$	$6.66 \cdot 10^{-1}$	$7.81 \cdot 10^{-1}$
DFN vs. collapsed plane with downscaling	$2.24 \cdot 10^{-1}$	$5.12 \cdot 10^{-2}$	$9.29 \cdot 10^{-2}$
DFN vs. homogeneous SDA	$4.29 \cdot 10^{-1}$	$3.43 \cdot 10^{-1}$	$4.01 \cdot 10^{-1}$
DFN vs. homogeneous PSSA	$3.24 \cdot 10^{-1}$	$1.53 \cdot 10^{-1}$	$2.23 \cdot 10^{-1}$

In both tests the downscaling simulations are carried out using the standard algorithm (SDA). In the calculations based on the homogeneous plane the projected sub-segment algorithm (PSSA) is also tested. It is worthwhile noting that in this latter set of calculations (i.e. straight pathway representative of a homogeneous BFZ), the local hydraulic gradient along the straight pathway is equal to the global gradient applied in the BFZ model ($\epsilon^i = 3.06 \cdot 10^{-3} \cos 30^\circ$).

3.2.1. Conservative Case

The breakthrough curves computed for each BFZ realization (i.e. BFZ modeled as an explicit DFN) are first compared with the results obtained using transport pathways derived from the collapsed plane model and without downscaling. This comparison (**Figure 6**) shows that the collapsed plane model significantly under-estimates the radionuclide retention capacity of the BFZ system. This evidence is consistent with the statistical analysis (**figure 3**) which points out that transport pathways computed using the collapsed plane model are characterized by lower groundwater travel time and transport resistance. The low groundwater travel times affect the leading edge of the breakthrough curves and partly explain the observed higher peak values. The low values of transport resistance have an impact on mass exchange processes and thus mostly affect the late time part of the breakthrough curves, which explains the observed low level of the tail. When the downscaling option is used, the agreement with the DFN-based solutions is very good (**Figure 6**), which points out that the downscaling algorithm can properly reconstruct the variability lost during the DFN representation onto the plane. Moreover, this good agreement obtained using the standard downscaling algorithm (SDA) indicates that the collapsed plane heterogeneity is properly capturing the original geometrical tortuosity of the transport pathways. These conclusions, which are based on visual inspections of the breakthrough curves, are confirmed by the results of the Kolmogorov–Smirnov (K-S) test (**Table 3**) that gives significantly lower statistical values when the downscaling algorithm is applied.

The comparison exercise is repeated using the single straight trajectory representative of the homogeneous plane. When the downscaling is performed using the SDA, the downscaled solution is characterized by an earlier breakthrough and a

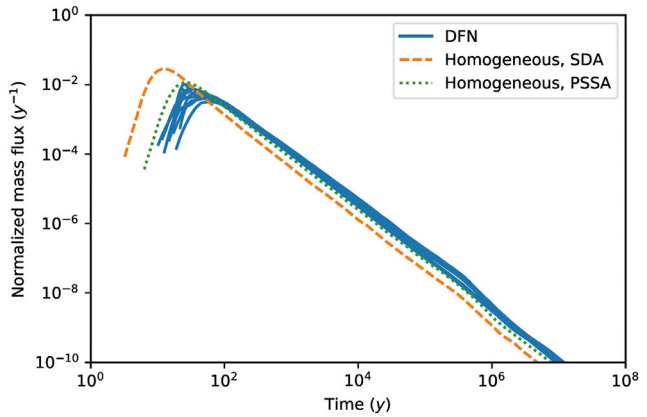


FIGURE 7 | *Conservative case*: breakthrough curves of a non-sorbing non-decaying radionuclide computed using transport pathways obtained using the different realisations of the DFN model (blue continuous curves) compared with the calculation carried out using the single straight transport pathways derived from the homogeneous plane model and with downscaling applied using the standard algorithm (SDA) (orange dashed line) and with downscaling applied using the projected sub-segment algorithm (PSSA) (green dotted line).

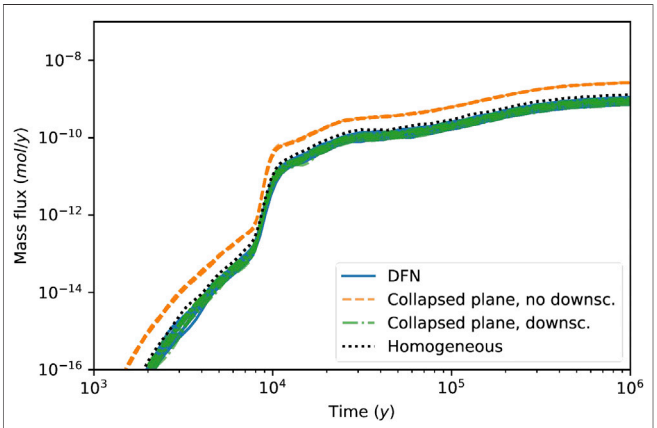


FIGURE 8 | *Decay Chain case*: breakthrough curves of Ra226 computed using transport pathways obtained using the different realisations of the DFN model (blue continuous curves) compared with related calculations carried out using transport pathways derived from the collapsed plane model and without downscaling (orange dashed lines), with related calculations carried out using transport pathways derived from the collapsed plane model and with downscaling applied using the standard algorithm (SDA) (green dashed-dotted lines) and with calculations carried out using the single straight transport pathways derived from the homogeneous plane model and with downscaling applied using the projected sub-segment algorithm (PSSA) (orange dotted line).

higher peak (**Figure 7**). This poor agreement is due to the tortuosity lost in the homogeneous representation. However, when the projected sub-segment algorithm (PSSA) is used, tortuosity is reconstructed by the downscaling algorithm and the agreement is significantly improved (**Figure 7**). The superior performance of the PSSA is also confirmed by the results of the KS test (**Table 3**).

TABLE 4 | Results of the mass test for the different variant simulations of the *Decay Chain case*. The test is run for each DFN realization and related collapsed plane or homogeneous model and the results are provided in terms of maximum ($\varepsilon_{n,m}^{max}$), minimum ($\varepsilon_{n,m}^{min}$) and average ($\langle\varepsilon_{n,m}\rangle$) test statistic.

Cases	$\varepsilon_{n,m}^{max}$	$\varepsilon_{n,m}^{min}$	$\langle\varepsilon_{n,m}\rangle$
DFN vs. collapsed plane no downscaling	2.87	1.31	1.94
DFN vs. collapsed plane with downscaling	$5.03 \cdot 10^{-1}$	$-3.66 \cdot 10^{-1}$	$-5.06 \cdot 10^{-3}$
DFN vs. homogeneous PSSA	$8.36 \cdot 10^{-1}$	$1.6 \cdot 10^{-1}$	$4.36 \cdot 10^{-1}$

3.2.2. Decay Chain Case

The results of the *Decay Chain case* are discussed here in terms of breakthrough curves of the daughter radionuclide Ra226. The results of the parent radionuclides are not shown for the sake of brevity.

When downscaling is not applied (**Figure 8**) and trajectories and related values of τ and β are taken from the collapsed planes, significantly larger mass fluxes are seen at the outlet boundary. These higher mass fluxes are related to the lower retention capacity of the BFZ system as represented with the collapsed plane model (**Figure 3**). Similar findings were already observed in the *Conservative case*.

When downscaling is applied, using the different sets of trajectories computed from the different realisations of the collapsed plane model and SDA, the observed Ra226 breakthrough curves agree well with the corresponding solutions computed using the explicit DFN representation of the BFZ (**Figure 8**). A good agreement is also observed when using the PSSA and the single straight trajectory representative of the homogeneous plane (see same figure).

Due to the presence of decay, a good indicator for the goodness of fit between the different models is the relative error in recovered mass. For a collapsed plane or single straight trajectory model n compared to a DFN model m, the relative error $\varepsilon_{n,m}$ is defined as:

$$\varepsilon_{n,m} = \frac{M_n - M_m}{M_m} \tag{9}$$

where M_n and M_m are the values of recovered mass for models n and m.

The results for the recovered mass test are shown in **Table 4**. Lower relative errors are observed in the collapsed plane model when the downscaling algorithm is applied. The good agreement between the single straight trajectory with PSSA and the DFN model is also confirmed by the test results.

4. DISCUSSION

As expected, the model obtained by projecting the flow and transport properties of a 3D fracture swarm onto a plane does not fully replicate the particle transport behavior of the full DFN. By removing heterogeneity normal to the midplane and restricting pathlines to a 2D surface, retention within the system is underestimated and particle travel times and transport resistance are systematically shorter.

TABLE 5 | DFN model runtime in seconds for the *Conservative case* and *Decay Chain case*, calculated as a mean over 10 realisations ± 1σ and separated into ConnectFlow and MARFA components.

Cases	ConnectFlow	MARFA	Total
Conservative case	1154.8 ± 338.9	36.1 ± 0.9	1190.8 ± 338.7
Decay Chain case	1154.8 ± 338.9	318.8 ± 6.5	1473.6 ± 341.0

TABLE 6 | Collapsed plane runtime in seconds for the *Conservative case* and *Decay Chain case*, calculated as a mean over 10 realisations ± 1σ and separated into ConnectFlow and MARFA components. Approximate speedup over the DFN model runtime in **Table 5** is also provided.

Cases	ConnectFlow	MARFA	Total	Total speedup (x)
Conservative case	6.7 ± 2.0	54.0 ± 0.5	60.7 ± 2.2	20 ± 6
Decay Chain case	6.7 ± 2.0	482.4 ± 8.0	489.1 ± 8.0	3 ± 1

A correction method, based on a previously published downscaling algorithm (Painter and Cvetkovic, 2005), has been tested using two test cases focused on the transport of specific radionuclides. The algorithm operates by statistically sampling libraries of triplets $|\Delta_i, \tau_i, \beta_i|$, which are in turn built using simplified but statistically representative DFN models (referred to as sampling cubes). Correlation in high/low velocity zones is accounted for by using a Markov chain. When the downscaling algorithm is used, the geometric information of the collapsed-plane-based transport pathways is used whereas the particles are advanced according to the sampled triplet values. An additional set of calculations has been carried out where the downscaling algorithm has been applied to a single straight trajectory representative of a homogeneous plane.

When applied to the heterogeneous collapsed plane, the downscaling approach is shown to accurately reconstruct the lost transport heterogeneity and the resulting radionuclide breakthrough curves agree well with the DFN-based solution. However, when the single straight trajectory is used, the resulting breakthrough curves show early arrivals and, in the case of the decaying radionuclide Ra-226, higher mass flux. This mismatch is caused by the lost tortuosity of the homogeneous model. Hence, an alternative implementation of the downscaling algorithm has been proposed, in which trajectory segments are projected along the principle direction. The modified algorithm is shown to provide very good results when applied to transport pathways derived from homogeneous models.

The generality of the Markov model applied here means that it is in principle applicable to a DFN based on any kind of empirical relationship and thus not only limited to the set of parameters and correlations described in **Section 2.2**. As the method is based on libraries which are constructed using an identical recipe to the DFN model, this consistency is ensured. The central assumption of the Markov method here is that the triplet $|\Delta, \tau, \beta|$ for each pathline segment depends only on the preceding segment. This method may therefore be less appropriate for DFN models in which some longer-range persistence along a pathline is present that cannot be captured by the Markov approximation (Painter and Cvetkovic, 2005).

The simple case presented in this article consists of a single synthetic BFZ fracture swarm with a single set of DFN parameter values; only one library of pathline legs is therefore required for the MARFA downscaling method to be applied. More than one library must be generated in order to apply the method to the realistic site-scale case, in which multiple BFZ are present, with DFN parameter values varying between BFZ and in some cases between distinct layers or regions within a single BFZ. The increased computational overhead of multiple library generation can be mitigated by grouping similar BFZ into classes based on their orientation and fracture intensity, with the pathline legs within each class described by a single library. This approach is now being tested in a subsequent phase of this study.

The computational speedup provided by the use of the collapsed plane model in comparison to the full DFN model is illustrated by the runtime provided in **Tables 5**, **6**. For the *Conservative case*, the time taken by ConnectFlow to solve for flow and generate pathlines is reduced by more than two orders of magnitude. Although there is an increase in the MARFA radionuclide tracking runtime due to the application of the downscaling algorithm, this is small in comparison and so a mean speedup of 20x is achieved across the 10 realisations. For the *Conservative case*, the ConnectFlow calculation is identical but the MARFA runtime accounts for a greater proportion of the total runtime due to the additional expense of modeling a decay chain, reducing the mean speedup to 3x. It should be noted that, for this simple synthetic model, the relative cost of constructing a library of pathline segments is high in comparison to the time taken to perform radionuclide tracking calculations in the full DFN model, with a mean ConnectFlow runtime for one realization of the sampling cube DFN model approximately 3 times greater than the runtime for the full DFN model. This is a result of the relative sizes of the models (homogeneous DFN in a 100 × 100 × 100 m domain for the sampling cube, as compared to a narrow fracture swarm in a 500 × 500 × 500 m domain for the full DFN model), and therefore not representative of the relative cost of constructing the library when scaled to the realistic case in which the full DFN model is much larger and more complex relative to the sampling cube. Further to this point, a pathline library must only be constructed once for each class of BFZ in the model (as defined in the previous paragraph), whereas real-world applications such as safety assessments of geological repositories of spent nuclear fuel typically necessitate multiple iterations of particle tracking calculations with varying boundary conditions. In such a case, the cost of library construction is thus expected to be much smaller relative to the total computation time saved by the use of the collapsed plane model.

5. SUMMARY AND CONCLUSION

A method for the simplification of complex DFN models of BFZ damage zones has been tested. The method consists of first upscaling the DFN to an ECPM model, using a flux-based methodology. The resulting ECPM model is then projected

onto the mid-plane of the BFZ. The plane-projected model is much more computationally tractable and provides an accurate description of global groundwater flow patterns. However, during the upscaling, local heterogeneity in flow and transport properties and related local choking effects are lost and thus related transport calculations are biased toward smaller values of groundwater travel time and transport resistance.

We have shown that the lost transport heterogeneity can be reconstructed by combining the plane-projected models with libraries of the two transport controlling parameters: the groundwater travel time and the transport resistance. These libraries are pre-computed using small DFN models (sampling cubes) with fracture statistics equal to those at the core of the BFZ model. The reconstruction of the groundwater velocity field and the transport resistance is carried out by assuming that the Lagrangian velocity follows a Markov process; i.e. the velocity in a fracture segment depends only on the velocity of the preceding segment. This sampling methodology was originally developed to extrapolate velocity distributions from small-to much larger-scale DFN models. To adapt the methodology to the foreseen application, a modification has been introduced, which consists of projecting fracture segments of the sampling cubes along principal directions. This modification is needed to account for the tortuosity of transport pathways of DFN models, which is lost in homogeneous representations of the BFZ. When the BFZ is represented as a heterogeneous collapsed plane this modification is not needed and the original algorithm provides good agreement with the synthetic solution.

The proposed downscaling approach is shown to provide a very accurate description of radionuclide breakthrough curves at compliance or outlet boundaries. The promising results shown in this work are of interest for a broad number of applications related to contaminant transport in fractured rocks. These include safety assessment calculations for spent nuclear fuel repositories, studies for the remediation of polluted sites or, in the context of nuclear nonproliferation, the detection of chemical signatures following underground explosions.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

ConnectFlow simulations were carried out by TW and GG, and MARFA simulations by JS and PT. All authors contributed to the design of the study, interpretation of numerical experiments, and preparation of the manuscript.

FUNDING

This work was funded by Svensk Kärnbränslehantering AB (SKB).

REFERENCES

Aaltonen, I., Kosunen, P., Mattila, J., Engström, J., Paananen, M., Paulamäki, S., et al. (2016). Tech. Rep. POSIVA Report 2016-16. *Geology of Olkiluoto*. Eurajoki, Finland: Posiva Oy.

Baxter, S., Hartley, L., Hoek, J., Myers, S., Tsitsopoulos, V., and Williams, T. (2019). Tech. Rep. SKB Report R-19-01. *Upscaling of brittle deformation zone flow and transport properties*. Stockholm, Sweden: Svensk Kärnbränslehantering AB (SKB).

Bear, J. (1972). *Dynamics of fluids in porous media*. New York, NY: American Elsevier.

Benke, R., and Painter, S. (2003). Modeling conservative tracer transport in fracture networks with a hybrid approach based on the Boltzmann transport equation. *Water Resour. Res.* 39. doi:10.1029/2003WR001966

Bolster, D., Barahona, M., Dentz, M., Fernandez-Garcia, D., Sanchez-Vila, X., Trinchero, P., et al. (2009). Probabilistic risk analysis of groundwater remediation strategies. *Water Resour. Res.* 45. doi:10.1029/2008WR007551

Cacas, M. C., Ledoux, E., de Marsily, G., Tillie, B., Barbreau, A., Durand, E., et al. (1990). Modeling fracture flow with a stochastic discrete fracture network: calibration and validation: 1. the flow model. *Water Resour. Res.* 26, 479–489. doi:10.1029/WR026i003p00479

Comolli, A., Hakoun, V., and Dentz, M. (2019). Mechanisms, upscaling, and prediction of anomalous dispersion in heterogeneous porous media. *Water Resour. Res.* 55, 8197–8222. doi:10.1029/2019WR024919

Gelhar, L. W., and Axness, C. L. (1983). 3-dimensional stochastic-analysis of macrodispersion in aquifers. *Water Resour. Res.* 19, 161–180. doi:10.1029/WR019i001p00161

Hakoun, V., Comolli, A., and Dentz, M. (2019). Upscaling and prediction of Lagrangian velocity dynamics in heterogeneous porous media. *Water Resour. Res.* 55, 3976–3996. doi:10.1029/2018WR023810

Hartley, L., Appleyard, P., Baxter, S., Hoek, J., Joyce, S., Mosley, K., et al. (2018). Tech. Rep. POSIVA working Report 2017-32. *Discrete fracture network modelling (Version 3) in support of Olkiluoto site description 2018*. Eurajoki, Finland: Posiva Oy.

Hyman, J. D., Dentz, M., Hagberg, A., and Kang, P. K. (2019). Emergence of stable laws for first passage times in three-dimensional random fracture networks. *Phys. Rev. Lett.* 123, 248501. doi:10.1103/PhysRevLett.123.248501

Jackson, C., Hoch, A., and Todman, S. (2000). Self-consistency of a heterogeneous continuum porous medium representation of a fractured medium. *Water Resour. Res.* 36, 189–202. doi:10.1029/1999WR900249

Jacobs (2018). Tech. Rep. Jacobs report Jacobs/ENV/CONNECTFLOW/15. ConnectFlow technical summary release 12.0.

Joyce, S., Simpson, T., Hartley, L., Applegate, D., and Hoek, J. (2010). Tech. Rep. R-09-20. *Groundwater flow modelling of periods with temperate climate conditions: Forsmark*. Stockholm, Sweden: Svensk Kärnbränslehantering AB (SKB).

Karra, S., Makedonska, N., Viswanathan, H. S., Painter, S. L., and Hyman, J. D. (2015). Effect of advective flow in fractures and matrix diffusion on natural gas production. *Water Resour. Res.* 51, 8646–8657. doi:10.1002/2014WR016829

Kwicklis, E., Lu, Z., Middleton, R., Miller, T., Bourret, S., and Birdsell, K. (2019). Numerical evaluation of unsaturated-zone flow and transport pathways at Rainier Mesa, Nevada. *Vadose Zone J.* 18, 190005. doi:10.2136/vzj2019.01.0005

Makedonska, N., Hyman, J. D., Karra, S., Painter, S. L., Gable, C. W., and Viswanathan, H. S. (2016). Evaluating the effect of internal aperture variability on transport in kilometer scale discrete fracture networks. *Adv. Water Resour.* 94, 486–497. doi:10.1016/j.advwatres.2016.06.010

Neretnieks, I. (1980). Diffusion in the rock matrix: an important factor in radionuclide retardation. *J. Geophys. Res.* 85, 4379–4397. doi:10.1029/JB085iB08p04379

Painter, S., Cvetkovic, V., Mancillas, J., and Pensado, O. (2008). Time domain particle tracking methods for simulating transport with retention and first-order transformation. *Water Resour. Res.* 44 (1), W01406. doi:10.1029/2007WR005944

Painter, S., and Cvetkovic, V. (2005). Upscaling discrete fracture network simulations: an alternative to continuum transport models. *Water Resour. Res.* 41. doi:10.1029/2004WR003682

Painter, S., and Mancillas, J. (2013). Tech. Rep. POSIVA working Report 2013-01. *MARFA user's manual: migration analysis of radionuclides in the far field*. Eurajoki, Finland: Posiva Oy.

Posiva (2013). Tech. Rep. POSIVA Report 2011-02. *Olkiluoto site description 2011*. Eurajoki, Finland: Posiva Oy.

Poteri, A., Nordman, H., Pulkkanen, V.-M., and Smith, P. (2014). Tech. Rep. Working Report 2017-23. *Radionuclide transport in the repository near-field and far-field*. Eurajoki, Finland: Posiva Oy.

Sanchez-Vila, X., Guadagnini, A., and Carrera, J. (2006). Representative hydraulic conductivities in saturated groundwater flow. *Rev. Geophys.* 44, RG3002. doi:10.1029/2005RG000169

Selroos, J.-O., and Painter, S. L. (2012). Effect of transport-pathway simplifications on projected releases of radionuclides from a nuclear waste repository (Sweden). *Hydrogeol. J.* 20, 1467–1481. doi:10.1007/s10040-012-0888-5

SKB (2010). *Radionuclide transport report for the safety assessment SR-Site*. Tech. Rep. TR-10–50, Stockholm, Sweden: Svensk Kärnbränslehantering AB (SKB).

Trinchero, P., Painter, S., Ebrahimi, H., Koskinen, L., Molinero, J., and Selroos, J.-O. (2016). Modelling radionuclide transport in fractured media with a dynamic update of Kd values. *Comput. Geosci.* 86, 55–63. doi:10.1016/j.cageo.2015.10.005

Trinchero, P., Painter, S. L., Poteri, A., Sanglas, J., Cvetkovic, V., and Selroos, J.-O. (2020). A particle-based conditional sampling scheme for the simulation of transport in fractured rock with diffusion into stagnant water and rock matrix. *Water Resour. Res.* 56, e2019WR026958. doi:10.1029/2019WR026958. E2019WR026958

Trinchero, P., Sánchez-Vila, X., and Fernàndez-Garcia, D. (2008). Point-to-point connectivity, an abstract concept or a key issue for risk assessment studies? *Adv. Water Resour.* 31, 1742–1753. doi:10.1016/j.advwatres.2008.09.001

Conflict of Interest: TW and GG were employed by Jacobs United Kingdom Ltd. JS and PT were employed by AMPHOS 21 Consulting S.L. J-OS was employed by Svensk Kärnbränslehantering AB (SKB).

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

APPENDIX
PROCESSING TRANSPORT PATHWAYS

The set of transport pathways, computed by ConnectFlow assuming uniform particle injection, is re-sampled to mimic a flux-weighted injection. The re-sampling algorithm is as follows (algorithm 1):

(1) Calculate water flux at the inlet of each transport pathway, q_i $[L^3T^{-1}]$.
(2) Sort the set of pathways in order of increasing flux at the inlet and construct the cumulative distribution (CD) function by assigning a probability of q_i/Q_{tot} to the ith pathway, where $Q_T = \sum_{i=1}^{N} q_i$ and N is the total number of transport pathways.
(3) Randomly sample the CD and store the selected transport pathway in memory.
(4) Repeat step 3 M times and write the new set of transport pathway in an ASCII file.

In the calculations of **section 3** $M = 1000$ whereas N depends on the realization and varies between 820 and 1,652 while for the collapsed planes around 300 pathways are used. A sensitivity analysis to M has been carried out to ensure that statistical ergodicity is attained. It is worthwhile noting that the methodology above implies that some of the original transport pathways are sampled multiple times whereas other pathways, with low flux at the inlet, are never sampled.

The libraries used for the downscaling algorithm are obtained from ten equi-probable sampling cubes. The related set of particle trajectories (one set for each realization and each principal direction) are merged into a single library for each principal direction using the following methodology:

(1) Calculate the inlet flux for every transport pathway and for all the 10 sampling cubes.
(2) Construct a CD function for each sampling cube as in step 2 of algorithm 1.
(3) Randomly select a sampling cube assuming a uniform distribution.
(4) Select a transport pathway from the chosen sampling cube as in step 3 of algorithm 1.
(5) Repeat steps 1–3 M times until enough trajectories are sampled and write the library file.
(6) The original set of transport pathways of each sampling cube consists of 1,000 particle trajectories and the library file is built by setting $M = 10000$.

frontiers in Earth Science

ORIGINAL RESEARCH
published: 08 February 2021
doi: 10.3389/feart.2021.613011

Estimating Watershed Subsurface Permeability From Stream Discharge Data Using Deep Neural Networks

*Erol Cromwell[1], Pin Shuai[1], Peishi Jiang[1], Ethan T. Coon[2], Scott L. Painter[2], J. David Moulton[3], Youzuo Lin[3] and Xingyuan Chen[1]**

[1]Pacific Northwest National Laboratory (DOE), Department of Energy National Laboratories, Department of Energy's Office of Science, Richland, WA, United States, [2]Oak Ridge National Laboratory, Oak Ridge, TN, United States, [3]Los Alamos National Laboratory (DOE), Department of Energy National Laboratories, Los Alamos, NM, United States

Subsurface permeability is a key parameter in watershed models that controls the contribution from the subsurface flow to stream flows. Since the permeability is difficult and expensive to measure directly at the spatial extent and resolution required by fully distributed watershed models, estimation through inverse modeling has had a long history in subsurface hydrology. The wide availability of stream surface flow data, compared to groundwater monitoring data, provides a new data source to infer soil and geologic properties using integrated surface and subsurface hydrologic models. As most of the existing methods have shown difficulty in dealing with highly nonlinear inverse problems, we explore the use of deep neural networks for inversion owing to their successes in mapping complex, highly nonlinear relationships. We train various deep neural network (DNN) models with different architectures to predict subsurface permeability from stream discharge hydrograph at the watershed outlet. The training data are obtained from ensemble simulations of hydrographs corresponding to an permeability ensemble using a fully-distributed, integrated surface-subsurface hydrologic model. The trained model is then applied to estimate the permeability of the real watershed using its observed hydrograph at the outlet. Our study demonstrates that the permeabilities of the soil and geologic facies that make significant contributions to the outlet discharge can be more accurately estimated from the discharge data. Their estimations are also more robust with observation errors. Compared to the traditional ensemble smoother method, DNNs show stronger performance in capturing the nonlinear relationship between permeability and stream hydrograph to accurately estimate permeability. Our study sheds new light on the value of the emerging deep learning methods in assisting integrated watershed modeling by improving parameter estimation, which will eventually reduce the uncertainty in predictive watershed models.

Keywords: DNN, machine learning, inverse modeling, subsurface permeability, stream discharge

OPEN ACCESS

Edited by:
Teng Xu,
Hohai University, China

Reviewed by:
Pingping Luo,
Chang'an University, China
Hongbo Su,
Florida Atlantic University, United States

****Correspondence:***
Xingyuan Chen
xingyuan.chen@pnnl.gov

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

Received: *01 October 2020*
Accepted: *05 January 2021*
Published: *08 February 2021*

Citation:
Cromwell E, Shuai P, Jiang P, Coon ET, Painter SL, Moulton JD, Lin Y and Chen X (2021) Estimating Watershed Subsurface Permeability From Stream Discharge Data Using Deep Neural Networks. Front. Earth Sci. 9:613011. doi: 10.3389/feart.2021.613011

1 INTRODUCTION

Subsurface flows formed by infiltration of precipitation and snow melt play a significant role in controlling the magnitude and timing of stream flows, especially in forested headwater watersheds (Scanlon et al., 2000). The permeability of soil and geologic formations determine both the infiltration rate and lateral subsurface flow rates, and ultimately the stream discharges. Integrated watershed models that mechanistically simulate both surface and subsurface flows with spatially distributed parameters and inputs are expected to provide better predictions of stream flow given sufficient data for parameterization and model calibration (Chen et al., 2020). Such distributed models also require a significantly larger number of unknown model parameters to be specified or estimated. Subsurface permeability is one of the key parameters that determine the subsurface flow and transport processes in watershed models. However, this parameter is difficult and expensive to measure directly at the spatial extent and resolution required by fully distributed, physics-based watershed models. The linkages between permeability and stream flow provide a new opportunity to estimate subsurface permeability from stream flow monitoring data that are made available through monitoring networks.

Inverse modeling has been used extensively to infer permeability from indirect subsurface measurements such as groundwater table in wells (Carrera et al., 2005) using optimization based methods. Parameter ESTimation software PEST (Doherty, 2010) has been a popular inverse modeling tool for performing uncertainty analyses and inverse modeling, including applications in integrated surface-subsurface models (Ala-aho et al., 2017). Ensemble-based approaches, including but not limited to Ensemble Kalman filter (EnFK) (Evensen, 1994; Evensen, 2003) and Ensemble Smoother (ES) (van Leeuwen and Evensen, 1996), have been adopted to estimate both model parameters and states (Moradkhani et al., 2005; Wen and Chen, 2006; Clark et al., 2008; Bailey and Baù, 2010; Vogt et al., 2012; Chen et al., 2013; Chen and Oliver, 2013; Emerick and Reynolds, 2013; Song et al., 2019). Despite its ease of implementation and proved efficiency in inverse modeling through various applications, ES is based on linear estimation theory and its performance may suffer in highly nonlinear problems (Evensen, 2018; Zheng et al., 2019). New approaches are needed to assist inverse modeling associated with highly nonlinear processes while maintaining computational efficiency.

Recent advances in machine learning, especially deep learning models including deep neural networks (DNNs), shed new light on inverse modeling by providing new ways to map the nonlinear relationships between model inputs and outputs (Shen, 2018; Mo et al., 2019). Neural networks have been an area of interest in hydrology over the past several decades. Early applications of neural networks to hydrology included a wide range of problems, such as rainfall-runoff modeling, streamflow prediction, groundwater modeling, water quality, water management, precipitation forecasting, hydrological time series, and other hydrologic applications (ASCE Task Committee, 2000). These neural networks were shallow and were typically composed of three layers: an input layer, a hidden layer, and an output layer. Limited by the computational power, the hidden layers were relatively small (5–20 nodes) and the model occasionally used a second hidden layer. Neural networks have been found to outperform other traditional statistical methods in a wide variety of applications in water resources domains (e.g., forecasting daily streamflows) as recently reviewed by Oyebode and Stretch (2018). However, such shallow neural networks may not be sufficient when it comes to estimating parameters that are related to indirect observations in a complex, highly nonlinear manner, for which multiple layers with larger sets of neurons are necessary. Neural networks capture nonlinearity by using nonlinear activation functions in between layers. Increasing the depths of the network, i.e., the number of hidden layers, could improve its ability to represent more complex system behaviors (Raghu et al., 2017; Shen, 2018), especially for mapping highly nonlinear relationships between the model inputs and outputs. Mo et al. (2019) successfully employed a deep autoregressive neural network-based surrogate approach to estimate the heterogeneous aquifer permeability as well as groundwater contaminant sources with high accuracy and computational efficiency. Canchumuni et al. (2019) compared convolutional variational autoencoder and the ensemble smoother with multiple data assimilation (ES-MDA) for the parameterization of facies in a geological reservoir with complex spatial distributions. They found that the DNN-based method outperformed the standard ES-MDA in reconstructing the spatial distribution of geologic facies.

The main objective of this study is to develop DNN-based inverse modeling method to estimate the subsurface permeability of a watershed from stream discharge data and test the accuracy and robustness of the new approach. DNNs are built to map the relationship between stream discharge time series and subsurface permeabilities for several soil and geologic facies. An ensemble of watershed simulations are performed using the Advanced Terrestrial Simulator (ATS), a spatially distributed, fully coupled surface and subsurface hydrologic model, to provide training and validation datasets for the DNNs. The performance of the DNN-based inversion is compared against the ES method in terms of their estimation accuracy and computational efficiency.

2 METHODOLOGY

2.1 Model Architecture

In this subsection, we describe the DNN-based models to estimate five unknown subsurface permeability parameters from the discharge time series data. A general description of neural networks is available in the **Supplementary Material**. We experiment with single-task learning (STL) and multi-task learning (MTL) models. For this work, a task is estimating a permeability parameter. An STL model estimates a single permeability parameter from the discharge data using a DNN, while an MTL model estimates all five parameters using a shared DNN. MTL models may improve their performance by, as summarized by Caruana (1997), "leveraging the domain-

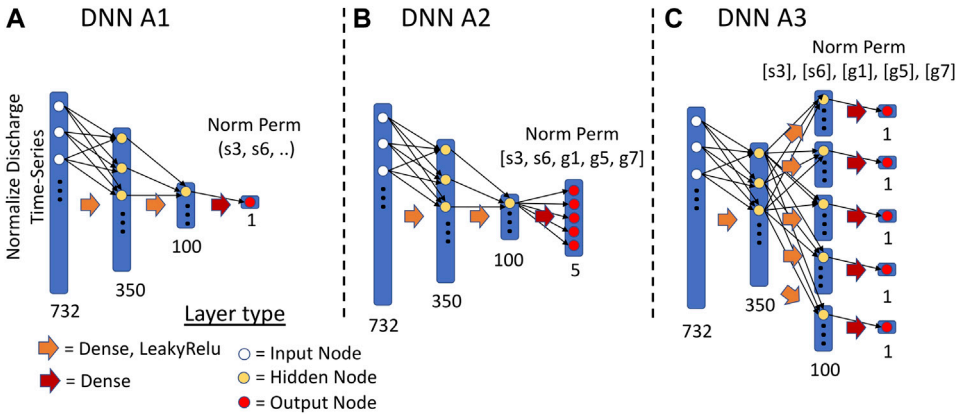


FIGURE 1 | Illustration of three DNN architectures. **(A)** The DNN A1 architecture; **(B)** The DNN A2 architecture; **(C)** The DNN A3 architecture. The input to each of the models is the normalized discharge time-series data (see **Sections 4.1** and **4.2**). The output from DNN A1 is a single normalized permeability parameter. The outputs from DNN A2 and A3 are five normalized permeability parameters.

specific information contained in the training signals of related tasks." In other words, having one model learn multiple tasks allows each task to benefit from the information used to train the other tasks by developing unique features that will emerge from estimating all the parameters together (Caruana, 1997). Furthermore, MTL has been shown to reduce the risk of model over-fitting (Baxter, 1997).

We built three different DNN architectures as shown in **Figure 1**: 1) a multi-layer dense network which estimates a single permeability parameter (DNN A1); 2) a multi-layer dense network which estimates all five parameters using a shared DNN with a single output layer with five nodes (DNN A2), with each node responsible for estimating one permeability parameter; and 3) a multi-layer dense network which estimates all five parameters in a shared DNN with a sub-network for each output node (DNN A3). DNN A1 is an example of STL and DNN A2 and A3 are examples of MTL. Different from DNN A2, DNN A3 branches out into five sub-networks after the first hidden layer, with each sub-network responsible for one of the outputs. The architectural design of DNN A3 not only allows for any shared features developed from estimating multiple parameters to be captured by the hidden layers before branching, it also allows each of the sub-networks to develop features that are more specific to the individual permeability parameters. Compared to DNN A2, DNN A3 is a larger model with more model weights to train. Consequently, it may require more data and more computational resources to train.

3 STUDY SITE AND TRAINING DATA GENERATION

To test the performance of the proposed DNNs in estimating subsurface properties from discharge data, we applied the method to a small catchment, the Rock Creek watershed, in Colorado (CO). Ensemble forward simulations were performed to provide sufficient training data for the DNNs that map the stream discharge hydrograph to the permeabilities of various soil and geologic layers. This section describes the study site, the forward watershed model implemented using the Advanced Terrestrial Simulator (ATS), and the generation of training data from ensemble ATS simulations.

3.1 Study Site: Rock Creek

Rock Creek is a small ($3\,km^2$) primary catchment in the East River watershed near Crested Butte, CO. The watershed is a high alpine, snow dominated catchment, characterized as Dfb, or warm summer, humid continental climate on the Koppen classification system. It is characterized by majority aspen, meadow, and mixed conifer vegetation types, and receives approximately 70 cm of precipitation per year. This site is part of the Watershed Function Science Focus Area of the Department of Energy, and has a significant history of hydrologic studies on mountainous, primary watersheds (Carroll et al., 2018; Hubbard et al., 2018).

3.2 ATS Forward Model

ATS is an integrated, distributed hydrologic code that solves the diffusion wave approximation of the St. Vernant equations for surface flow coupled to Richards equation for flow in variably saturated porous media in the subsurface (Coon et al., 2019). This coupling is achieved through a continuous pressure, continuous flux approach described by Coon et al. (2020) and Painter et al. (2016). This code leverages the Mimetic Finite Difference method to ensure accurate, efficient solution of the equations, in mixed, conservative form [e.g., Celia et al. (1990)] on meshes that allow distorted, arbitrary polyhedra, including the triangular prisms used here. The resulting equations are solved using an implicit, backward Euler time integration scheme which is solved using the Nonlinear Krylov Acceleration approach of Carlson and Miller (1998); Calef et al. (2013) and preconditioned using the Boomer Algebraic Multigrid package in HYPRE (Falgout and Yang, 2002). This code has been benchmarked against a variety of hydrologic codes in Kollet et al. (2017), and is shown to be

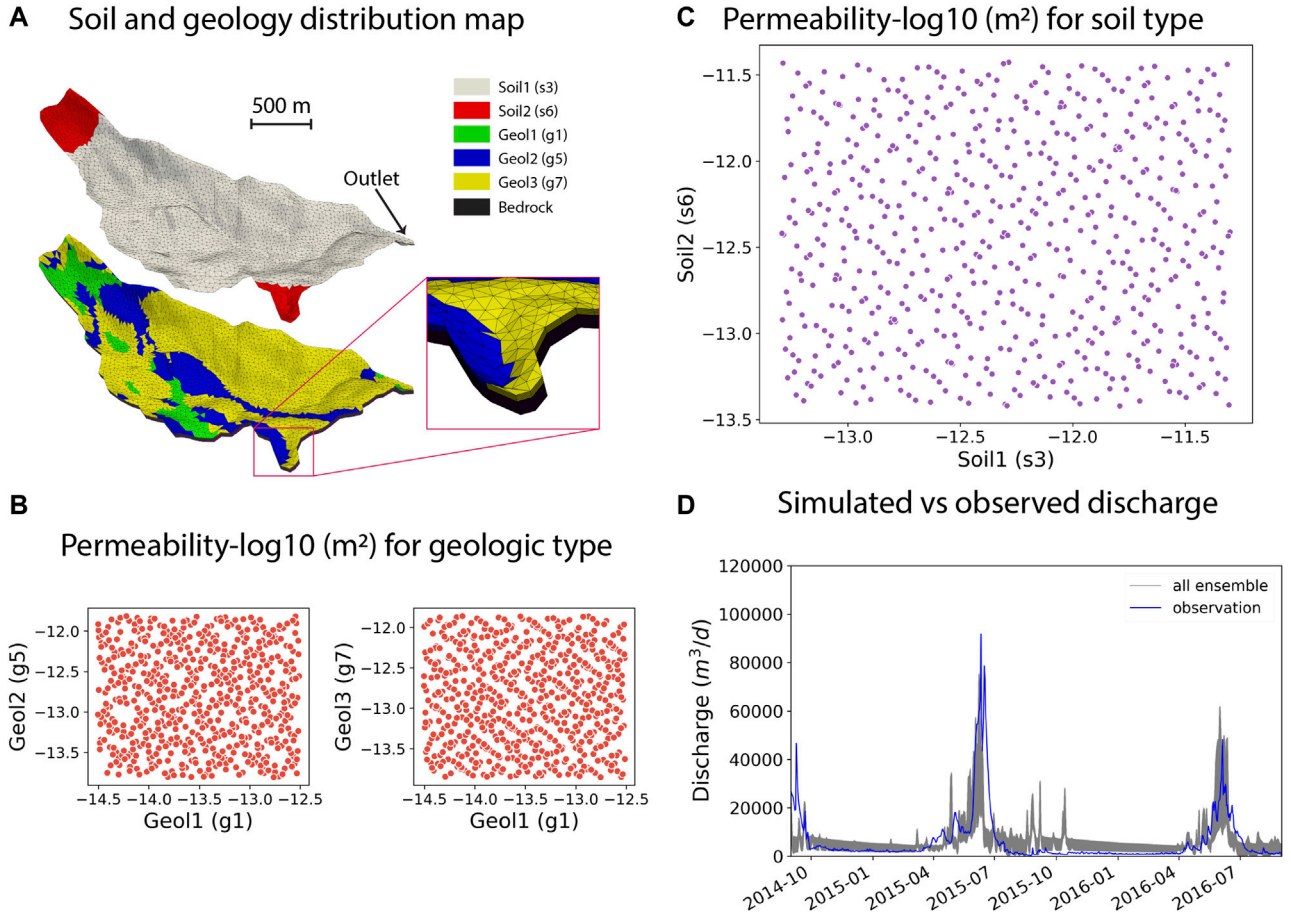


FIGURE 2 | (A) Soil and geology distribution within the Rock Creek watershed. The zoom-in box shows the triangle meshes. **(B)** Permeability [log10 (m^2)] realizations for three geologic types. **(C)** Permeability [log10 (m^2)] realizations for two soil types. **(D)** Ensemble of simulated discharge hydrographs from 597 realizations compared with the observed discharge hydrograph.

appropriate for solving problems of integrated hydrology (e.g., watersheds).

The baseline simulation was developed by using the Watershed Workflow package (Coon, 2020) to bring together a variety of data streams, delineate the catchment, and generate a variable resolution mesh with refined resolution at the stream network. Resolutions ranged from typical cell areas of 5,000 m^2 at the stream to 10,000 m^2 away from the stream network. This triangular surficial mesh was then elevated using Digital Elevation Model (DEM) from the National Elevation Dataset (NED) 9 m resolution dataset. In the work of Pribulick (2015), a base subsurface structure was defined by three stratigraphic layers–a soil layer of 1 m at the surface of the mesh, a near-surface geology layer 9 m thick below the soil layer, and a bedrock layer 20 m thick below the geologic layer. Based on the National Resources Conservation Service (NRCS) soils database, two soil types were identified and mapped within the soil layer. Using a surface geology dataset from the United States Geological Survey geological maps, three geologic material types were identified and mapped within the geologic layer. The spatial distribution of the soil and geological layers is demonstrated in **Figure 2A**. The vertical resolution of the mesh gradually increased from Δz = 5 cm at the surface to 2 m at the 2 m depth, and it remained constant at 2 m until the bottom of the model domain at a depth of 30 m.

The model was first run for 20,000 days with constant precipitation (~ 556 mm/yr) and the permeabilities for all the soil and geological layers adopted from the work of Pribulick (2015) as the spin-up that resulted in steady state model outputs at the final timestep, which was then used as the initial condition for an 8-year transient simulation (2010 to 2017) driven by daily-averaged meteorological data from the DayMet (Thornton et al., 2016) dataset. The DayMet forcing is a 1 km raster that covers the entire North America with only three pixels over our modeling domain. Time series of precipitation, air temperature, incoming shortwave radiation, and relative humidity were mapped onto the mesh, and prescribed throughout the simulation. Additional setup/inputs required for ATS are described in the **Supplementary Text S3**.

To develop the training, validation and testing datasets for the DNNs, we completed an ensemble of 597 ATS transient simulations for the Rock Creek watershed, each simulation

corresponded to a given set of soil and geology permeability parameters randomly generated from their probability distributions. Uniform distributions were assumed for all the parameters, with the lower and upper bounds chosen to be one order of magnitude below and above their values used in the baseline transient simulations. Quasi-Monte Carlo sampling method (Lemieux, 2009) was used to generate the permeability realizations shown in **Figures 2B,C**. The ensemble of simulated discharges in **Figure 2D** show reasonable match with the observed discharges at the outlet with RMSE ranging from 4.3 to 5.2 mm/day), which is a strong evidence for a reasonable conceptual model and sensitivity of the discharge to variations in soil and geologic layer permeabilities. We only used the discharge data from August 31, 2014 to August 31, 2016 to train the DNNs such that the trained model can be used to estimate the permeability of the real watershed from the limited field observation data available in this region during the same time window.

4 DNN TRAINING

4.1 Data Preprocessing and Preparation

Prior to training the DNNs, the discharge data D was log-transformed and normalized via zero-mean and unit variance using the following equation:

$$N(D) = \left[\ln(D) - \overline{\ln(D_t)}\right] \Big/ \sigma[\ln(D_t)], \tag{1}$$

where $\overline{\ln(D_t)}$ and $\sigma[\ln(D_t)]$ are the mean and standard deviation of the log-transformed discharge in the training dataset, respectively. Five of the permeability parameters were also log-transformed and normalized via zero-mean and unit variance using the same equation N, performed independently for each of them.

4.2 Training of DNNs

The DNNs were designed and implemented with the Keras python module (Chollet, 2015). In order to train the models, we divided the ensemble of 597 ATS runs into training, validation, and testing sets using roughly 70-15-15% divide (477, 60, 60 runs), respectively. Each run contains 732 daily measurements from August 31, 2014 to August 31, 2016, which were used as the input for all the DNNs. The DNN models were trained on the training set of 477 runs for 1,500 epochs with a batch size of 10, using mean-squared error (MSE) as the loss function and an Adam optimizer (Kingma and Ba, 2014). For each DNN type, we performed a hyperparameter search by varying the learning rate for training and the size of the layers (see **Supplementary Table S1** for the combinations of learning rate and layer size). After the training was completed, each DNN model configuration was run on the validation set to choose the best model hyperparameters (i.e., model configuration) of a given DNN architecture. Each DNN configuration was trained using four different initialization seeds and results were averaged during validation. For the DNN A1 models, we trained separate models to estimate each of the five permeability parameters. **Supplementary Table S2** shows the best hyperparameter combination for each model type based on the MSE on the validation set. Then, the best model configuration of each DNN architecture (e.g., DNN A1, A2, A3) was run on the testing set, and their performance was compared against each other.

4.3 Ensemble Smoother

We compared the DNNs against the ES approach in estimating permeabilities to assess the importance of dealing with nonlinearity between model inputs and outputs in inverse modeling. We followed a similar training-testing strategy when applying the ES method as used in the DNNs. We took the 537 realizations used in training and validation sets for DNNs as the prior ensembles for the permeabilities and modeled responses, then we perturbed each simulated hydrograph in the testing set (60 realizations in total) by adding random observation errors and used them as the synthetic observations, which were assimilated by the ES to generate posterior ensemble of permeabilities. The testing process yielded 60 sets of posterior permeability ensemble given an observation error, which is a hyperparameter for the ES approach. We considered a range of relative errors for discharge observations: 0.005, 0.01, 0.015, 0.02, 0.03, 0.05, and 0.1. For each observation error, we computed the correlation coefficient and root mean square error (RMSE) between the 60 synthetic true permeability and the corresponding 60 estimations from each of the 537 posterior realizations. We found that the ES trained with relative error of 0.05 performed the best in terms of both correlation coefficient and RMSE. More details about this analysis can be found in the **Supplementary Text S2**. Therefore, we chose the ES trained with a relative error of 0.05 for the comparison against the DNNs and for estimating the permeability from real observations.

5 RESULTS AND DISCUSSION

5.1 Training Results of DNNs

Overall, we were able to successfully train the three DNN types to estimate the permeability parameters from the simulated stream discharge. As can be seen in the overall training loss for the DNN A3 model (**Figure 3A**), the DNN model learns to estimate the permeabilities during the training period. The overall and individual validation losses plateau before or around epoch 600 as shown in **Figure 3**. For the permeability of s6 (**Figure 3C**), its training loss continues to decrease after its validation loss plateaus, indicating an overfitting problem due to insufficient training data or lack of useful information in the discharge data to constrain the permeability. Similarly, the DNN A3 model also overfits on the g1 permeability (**Figure 3D**), but to a lesser extent as the validation loss is much closer to the training loss. No overfitting problems were found for the permeabilities of s3, g5, and g7 as their validation losses closely follows the training losses as seen in **Figures 3B,E,F**. The training and validation losses for the DNN A1 and DNN A2 models have similar performance and the figures are available in the

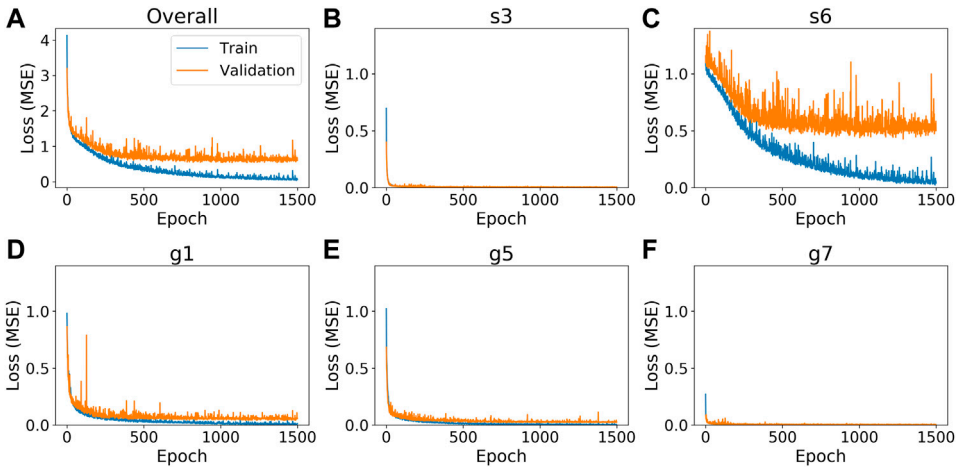


FIGURE 3 | Training loss of the best DNN A3 model. The blue line is the loss from the training set and the orange line is the loss from the validations set. The *x*-axis is the epoch number. The *y*-axis is the model loss (MSE). **(A)** overall training loss of best DNN A3 model. The overall loss is the sum of the loss of the five permeability parameters **(B–F)**; **(B)** training loss for the s3 permeability parameter; **(C)** training loss for the s6 permeability parameter; **(D)** training loss for the g1 permeability parameter; **(E)** training loss for the g5 permeability parameter; **(F)** training loss for the g5 permeability parameter.

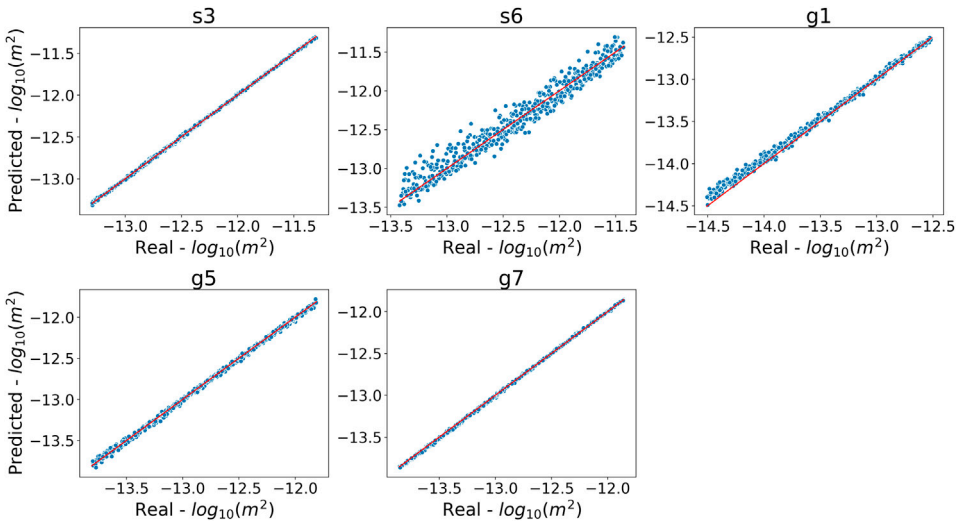


FIGURE 4 | One-to-one plot of the DNN A3 model permeability estimation when compared to the real estimation for the training set. Each plot is the DNN A3 model estimation for the given permeability parameter. Each dot represents a realization from the training set of ensembles. The *x*-axis is the $\log_{10}$ real permeability value, the y-axis is the $\log_{10}$ estimated permeability value from the model. The red line is the one-to-one line.

Supplementary Figures S3, S4. Therefore, the overall overfitting of the DNNs can be mainly attributed to the overfitting on the s6 permeability, and to a lesser extent on the g1 permeability.

On the training set, all the three DNN types perform very well in estimating the permeability parameters. As shown in **Figure 4**, the one-to-one plots between the permeabilities estimated by the DNN A3 model and the real permeabilities are distributed closely along the 1:1 lines. The DNN A1 and A2 models achieve similar levels of performance and their plots are available in the **Supplementary Figures S5, S6**, respectively. All three models yield R^2 values greater than 0.99 for the s3, g1, g5, and g6 permeabilities (see **Supplementary Table S3**). All three DNNs perform relatively worse in estimating the s6 permeability on the training set, as evidenced by more scatters drifted from the 1:1 line in **Figure 4**, resulting in R^2 values of about 0.96 from the DNN A2 and A3 models, and an R^2 of 0.89 from the DNN A1 model.

5.2 Testing Results of DNNs

For the testing set, all the three DNN types perform very well in estimating the s3, g5, and g7 parameters, with R^2 values between the predicted and true permeabilities for all DNN models above

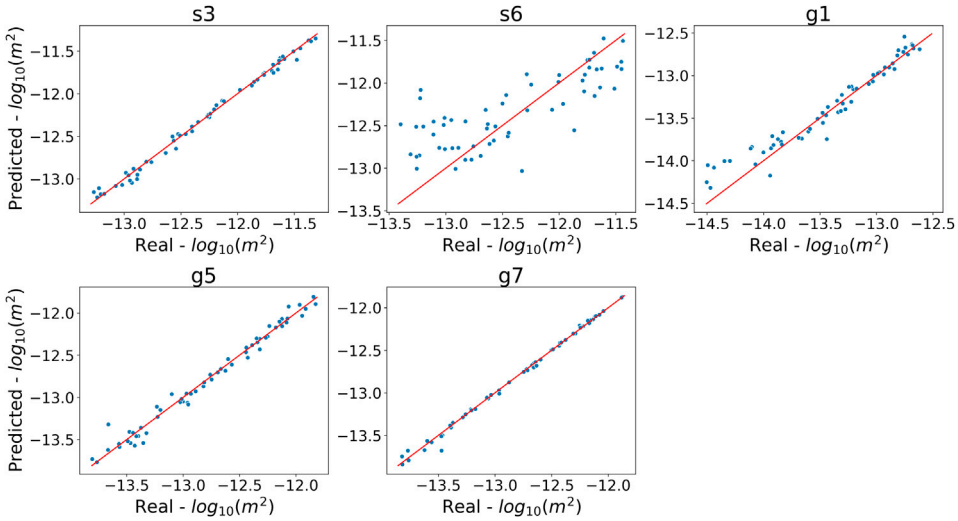


FIGURE 5 | One-to-one plot for permeability estimated from the DNN A3 model against the true permeability in the testing set. Each plot is for a given permeability parameter, and each data point represents a realization from the testing set. The *x*-axis is the $\log_{10}$ real permeability value, the y-axis is the $\log_{10}$ estimated permeability value from the model. The red line is the one-to-one line.

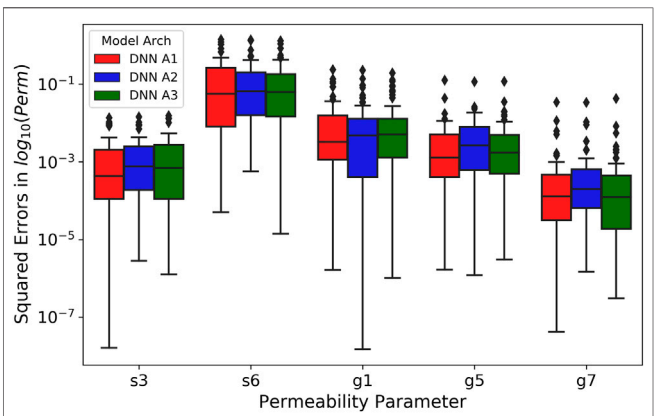


FIGURE 6 | Box plot of the squared error of the DNNs on the testing set. The *x*-axis is the permeability parameter. The *y*-axis is the squared error of the estimated $\log_{10}$ of the permeability parameter compared to their true values. The *y*-axis is plotted in log-scale for showing differences spanning over several orders of magnitude. The results for different DNN architectures are represented by the colors filling the box plots.

0.98. However, the R^2 drops to $\sim$ 0.93 for the g1 permeability and below 0.6 for the s6 permeability (see **Supplementary Table S4**). The one-to-one scatter plots for DNN A3 in **Figure 5** further demonstrates the disparity in estimation accuracy between the group of s3, g5, and g7 and the group of s6 and g1, which is similar to the training results with increased deviation from the exact 1:1 line. The one-to-one plots for the DNN A1 and DNN A2 models follow the same pattern and the plots are available in the **Supplementary Figures S7, S8**, respectively.

The difficulty in estimating the s6 and g1 permeabilities might be explained by the small areas covered by these two soil/geologic layers (**Figure 2A**) and their distances from the outlet where the discharges are measured. Thus, they may not contribute as much to the simulated discharge compared to the other parameters. In other words, the simulated discharges are not as sensitive to the s6 and g1 permeabilities, and consequently the discharges at the outlet are not sufficiently informative for estimating the s6 and g1 permeabilities.

The performance of all three DNN architectures on the testing set was compared using the squared errors between the estimated and true log10-transformed permeabilities. **Figure 6** shows the three quartiles and ranges of the squared errors calculated from the 60 realizations in the testing set using boxplots, while their means and standard deviations are provided in **Supplementary Table S4**. All the DNN architectures yield very similar results in terms of the statistics of the squared errors for all the parameters. Their mean squared errors (MSEs) differ by less than 1.1×10^{-4} for the g7 permeability, around 4.8×10^{-4} for the s3 permeability, and within 1.3×10^{-3} for the g5 permeability. Overall, the DNN A1 models slightly outperform the DNN A2 and DNN A3 models in terms of the means and medians of squared errors provided in **Figure 6** and **Supplementary Table S4**. The MTL models show slight improvement over the STL models in MSE and R^2 for the s6 permeability, suggesting the potential for the less sensitive parameters to benefit from the joint features developed in MTL models at little or no expense to the estimation accuracy of other parameters. The performance difference between the DNN A2 and A3 models appears negligible. Thus, in this case study, it may not be necessary to add additional model complexity using the sub-networks to develop parameter-specific features.

5.3 Estimation Sensitivity to Observation Errors

To assess how sensitive the DNN-based parameter estimation is to observation errors in the data used for inverse modeling, we randomly selected a realization from the testing set and generated

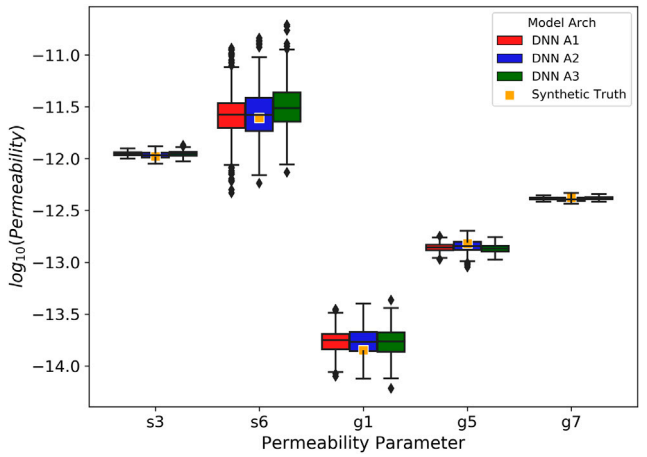


FIGURE 7 | Sensitivity of model types to noise added to the simulated discharge for ensemble 537. The boxplots represent the distribution of the $\log_{10}$ estimated permeability value from the 100 noisy discharge vectors. The yellow square dot are the synthetic permeability values for the simulation (ensemble 537).

100 realizations of noisy observed discharge time series ($\mathbf{d_n}$) by adding random observation errors to the simulated discharges ($\mathbf{d}$), i.e.,

$$\mathbf{d_n} = \mathbf{d} + \epsilon * \mathbf{d} * \mathbf{r}, \tag{2}$$

where $\mathbf{r}$ is a vector of the same size as $\mathbf{d}$ composed of random samples drawn from a standard normal (Gaussian) distribution with a mean of 0 and a standard deviation of 1, and ϵ is the standard deviation of the noise, which is usually taken as 1/3 of the observation error. For this study, we set $\epsilon = 0.0166$ for a 5% observation error. Then, we obtained 100 realizations of estimated permeability set from each of the DNN model types with the best configuration (the same ones used in **Section 5.2**) from the 100 realizations of noisy discharges. The variability in estimated permeabilities are shown in boxplots in **Figure 7**. It is observed across all the DNN model types that the parameters that can be accurately estimated (i.e., s3, g5, and g7) from the discharge data show more robust performance under the presence of the observation errors, whereas those that are estimated less successfully (i.e., s6 and g1) are also more vulnerable to the observation errors.

As discussed in **Section 5.2**, it is highly likely that there is less information available in the discharge data for estimating parameters with smaller spatial coverage, which consequently limits DNNs in their ability to generalize beyond the training data. The observation error further contaminates the useful information (i.e., signal) in the data, thus exacerbating the estimating inaccuracy.

5.4 Comparison With Ensemble Smoother

We compared the performance of DNN-based methods against the ES method (with relative error of 0.05) in estimating the permeability from the same discharge data using the squared errors of the estimated permeability on the testing set, as shown in **Figure 8**. To ensure a fair comparison, for each realization in the testing set, we generated 100 noisy discharge time series assuming a 5% observation error, which led to 100 instances of estimated permeability set that are compared to their corresponding synthetic truth with squared errors calculated. Therefore, each boxplot in **Figure 8** associated with the DNN methods was generated from 6,000 samples of squared errors, whereas that for the ES method was generated from 32,220 samples because each testing realization has an ensemble of 537 posterior estimations. The comparison shows that the DNNs significantly outperform the ES with much smaller squared errors for all the five permeabilities. The MSE across all five permeabilities for the DNNs (0.0568 for DNN A1, 0.051 for DNN A2, and 0.045 for DNN A3) is an order of magnitude smaller compared to the MSE for the ES method (0.634). Therefore, DNNs are promising alternatives for inverse modeling, especially for using indirect data that are nonlinearly related to the parameters of interest.

Additionally, we investigated the amount of training data needed to achieve a similar level of accuracy using the full amount of training data for the DNNs and the ES approach. We trained the three DNNs using 50, 100, 200, and 300 realizations from the training set with the same hyperparameters listed in **Supplementary Table S2**. Then, we performed the ES-based estimations with a relative error of 0.05 with the same four training sets of various sizes. Finally, permeabilities on the testing set were estimated using the DNNs trained on different amount of data and the ES of various ensemble sizes and compared. Both sets of estimations assumed a relative observation error of 5%. As shown in **Figures 9A–C** that the performance of DNNs keeps improving with the increasing amount of training data increases. Nevertheless, the gain in performance diminishes when the training data size is increased from 300 to 477. Thus, we consider 300 realizations as sufficient for achieving good training results for the DNNs in this case study. The ES approach, on the other hand, does not show as much improvement in estimation accuracy when increasing the ensemble size (**Figure 9D**). Interestingly, the DNNs trained with 50 realizations yield a lower MSE for all the s3, g1, g5, and g7 permeability parameters than all of the ES variations, while achieving equivalent performance for the s6 permeability. Therefore, the DNNs may require less training information to achieve equivalent or better performance than the ES. Moreover, the DNNs can more effectively utilize the information in larger training dataset than the ES approach to capture nonlinear relationships.

We also compared the computation times needed for both methods. The computational cost for DNN models is spent on both the training and prediction phases, whereas the ES approach does not have an explicit training phase. We were able to perform the DNN-based and ES-based inversions on laptops without involving the Graphics Processing Unit (GPU) or parallel computing. The DNN A1 and DNN A2 models finished training for 1,500 epochs in under 2 min. The DNN A3 finished training in just over 5 min. For estimating the permeabilities from a single time series of stream discharge, the DNNs took less than 0.4 ms on average. When estimating

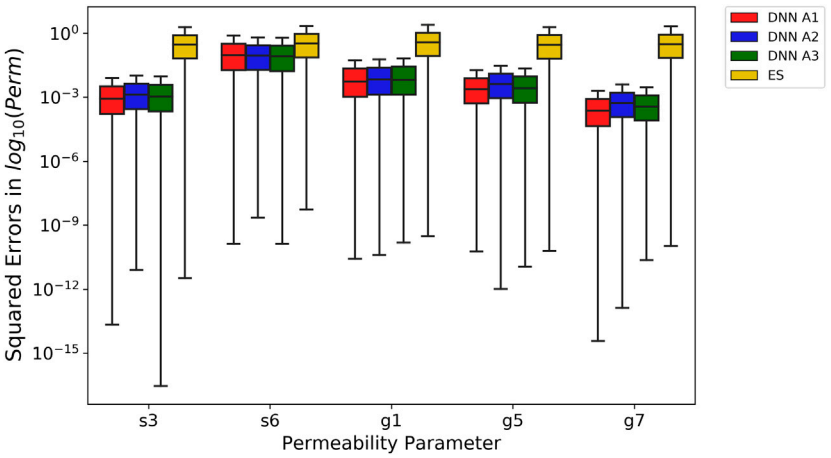


FIGURE 8 | Box plots of squared errors on permeabilities estimated by DNNs and ES for the testing set using relative observation error 0.05. The *x*-axis is the permeability parameter. The *y*-axis is the squared errors of the $\log_{10}$-transformed permeability parameters. The *y*-axis is shown in log-scale. The boxplots are filled with different colors corresponding to different estimation methods.

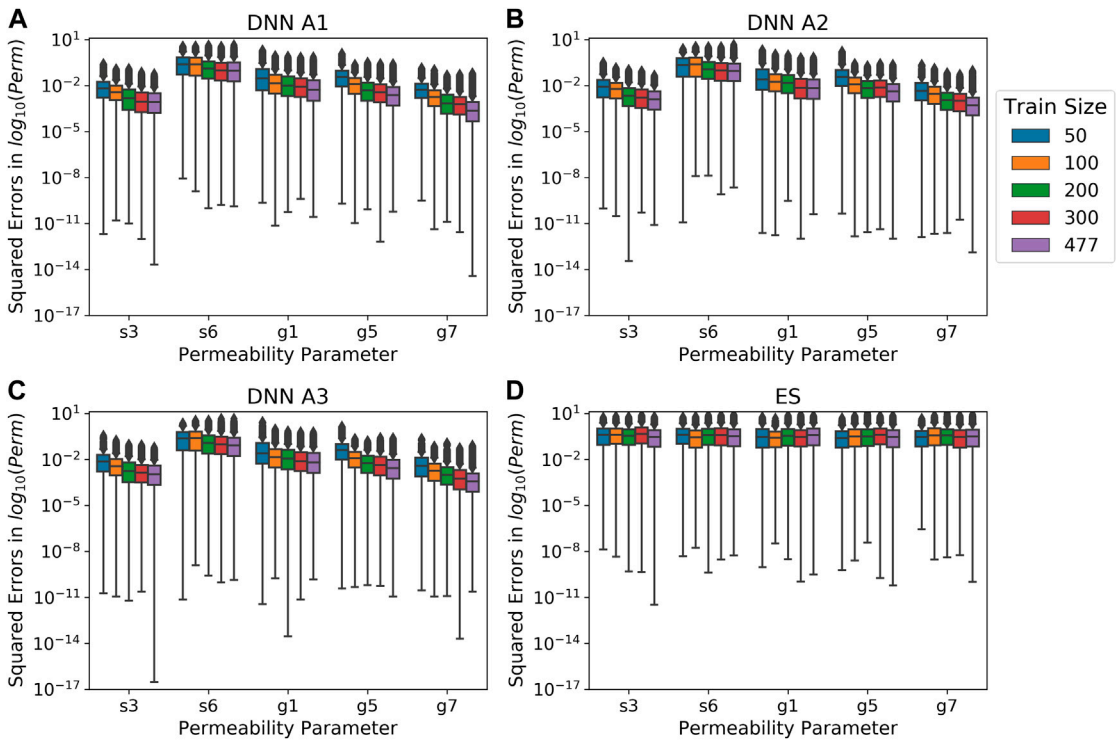


FIGURE 9 | Box plots of squared errors on permeabilities estimated by DNNs and ES for the testing set using relative observation errors of 0.05 with different training data sizes. The *x*-axis is the permeability parameter. The *y*-axis is the squared errors of the $\log_{10}$-transformed permeability parameters. The *y*-axis is shown in log-scale. Each color represents a number of realizations used to train the DNNs or ES (50, 100, 200, 300, 477). **(A)** the squared errors for the DNN A1 model; **(B)** the squared errors for the DNN A2 model; **(C)** the squared errors for the DNN A3 model; **(D)** the squared errors for the ES.

all 60 stream discharges in the testing set, DNNs A1, DNN A2, and DNN A3 took 0.87, 0.89, and 1.17 ms, respectively, on average. For estimating the permeability for a single discharge time series, the ES took approximately 40 ms to update the permeability ensemble by assimilating the corresponding discharge observations through matrix operations. The computational costs spend on the inversion using either of the methods are negligible compared to the computing resources

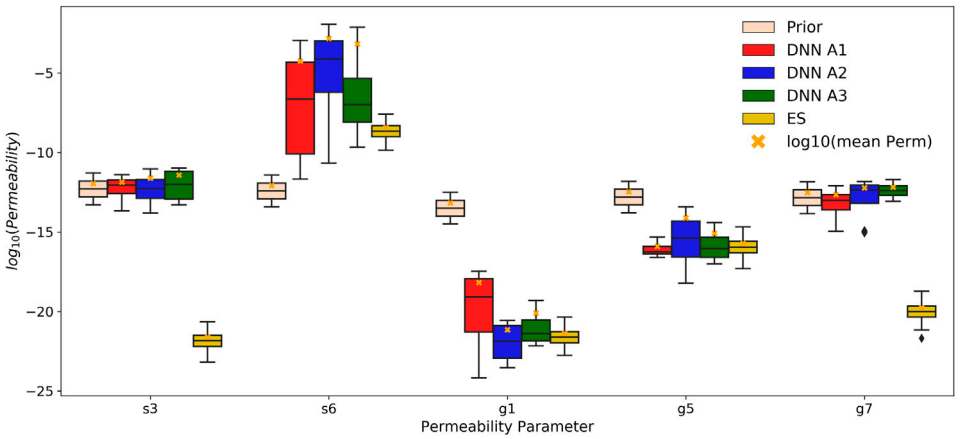


FIGURE 10 | Permeability estimations from the observed discharges from the DNNs and the ES using relative observation error of 0.05. The peach-colored boxplots are the prior distributions for each permeability parameter. The gold-colored boxplots are the distribution of permeability estimations from the ES. Each of the remaining colored boxplots is the corresponding distributions of DNN estimations. The orange crosses are the log_{10}-transformed mean permeability value for the corresponding distribution.

required to generate the training, validation and tesing datasets from ensemble ATS simulations, which were approximately 324,768 cpu hours using the supercomputing resources at the National Energy Research Scientific Computing Center (NERSC).

5.5 Permeability Estimation From Observed Discharge for Real Systems

After achieving the adequate estimation performance on training and testing the DNNs, we moved forward with estimating the soil and geologic permeability parameters of the Rock Creek catchment from the real observed discharge (**Figure 2D**) using the DNNs trained and tested in **Section 5.2**. We generated 100 realizations of noisy discharge time series from the real observed discharge assuming a relative observation error of 5%. The ensembles of the estimated permeability parameters by DNNs were used to generate the boxplots in **Figure 10**, compared against the posterior estimates from the ES approach with a relative observation error of 5%. Note that the prior boxplots were generated from the prior parameter ensemble for the ES approach, which was also used as the DNN training dataset. It can be observed from **Figure 10** that the DNN estimations for the s3 and g7 permeabilities are distributed in the similar ranges as in their training data, while the DNN estimations for the g5, g1 and s6 permeabilities are significantly shifted away from their ranges in the training dataset. The variability in results obtained from different DNN types is substantially greater for the g5, g1 and s6 parameter group than the s3 and g7 group, which is consistent with the difference in their sensitivity to the discharge time series as revealed during the training and testing stages. Thus, the estimated permeabilities for s3 and g7 are likely more accurate (i.e., reasonable) than those for g5, g1 and s6.

Figure 10 also demonstrates that the ES approach yielded vastly different estimations from the DNN models, except for the g5 permeability, with relatively tighter distributions. The considerable differences between the prior and posterior

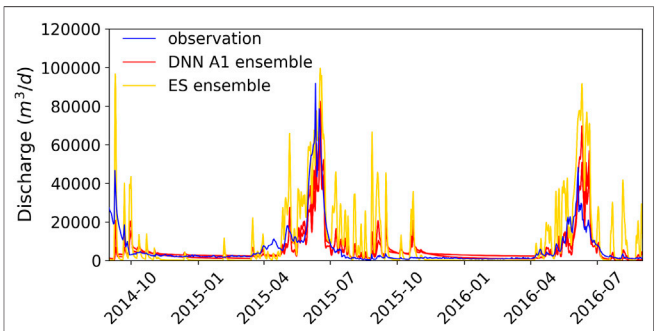


FIGURE 11 | Comparison between the observed discharge and the predicted discharge from ATS using the estimated permeability values from the DNN A1 and ES, respectively.

distributions in parameters suggest that the estimations are extrapolated from the training/prior data for this real application, which is related to the fact that the ensemble of the simulated hydrographs failed to encapsulate the observed hydrograph in major recession periods as shown in **Figure 2D**. DNNs appear to provide more realistic estimations for the s3 and g7 permeability than the ES. The s3 permeability estimated by the ES is unrealistically low for soils, even lower than the geologic layers, which may be caused by extrapolation errors based on linearized relation between model parameters and outputs. DNNs, on the other hand, are better able to generalize in this case by capturing the nonlinear relationships. To evaluate how the estimated permeabilities change the model predictions of the hydrograph at the outlet, we randomly selected 30 realizations from the ensembles of permeabilities estimated by the DNN A1 and the ES to generate updated ensembles of predicted hydrographs, which were then compared to the observations to assess the improvement in model performance. We encountered numerical model convergence issues for some

permeability combinations with substantial contrast between the low-permeability and high-permeabilities layers. Thus, not all simulations were completed within the assigned wall clock time. The ensemble simulation results in **Figure 11** contain nine and five sets of completed simulations for the DNN A1 and the ES, respectively. Although the ensemble sizes of model runs may not be sufficient for representing the full range of uncertainty in the updated model predictions, we expect them to adequately represent the mean model behaviors. Overall, the permeabilities estimated by DNN A1 lead to much improved prediction during year 2015 as compared to the prior predictions in **Figure 2D**, whereas they cause considerable overpredictions in peak discharges during year 2016. In contrast, the models with the permeabilities estimated by the ES consistently overpredict the discharges owing to the unrealistically low permeability estimated for s3. It is worth noting that none of the simulated hydrographs reproduces the same level of inter-annual variability manifested in the real observations, which implies potential deficiencies in the numerical model representation of the real system. Further investigations are needed to identify additional processes and parameters that may contribute to such inter-annual variability.

6 CONCLUSION

In this paper, we developed a DNN-based inversion method to estimate permeabilities of multiple soil and geologic layers within a watershed from the observed stream discharge time series. We successfully trained DNNs to map from the stream discharges to the permeability set using the training/validation/testing data generated through the ensemble watershed simulations. In doing so, we found that the accuracy and robustness of DNN-based estimations are influenced by the relevant information content contained in the observation data with respect to the parameters. In the watershed system we studied, permeability for soil and geologic layers with larger spatial coverage can be estimated more accurately from the observed discharge data, and their estimations were more robust to observation errors.

In comparing the parameters estimated by the DNNs and the traditional ES method from the same observation data, we found that the DNNs consistently outperformed the ES algorithm. On the testing set, the DNNs achieved an overall MSE an order of magnitude lower than the ES method. The DNNs is more effective in utilizing the information provided by larger training dataset than the ES approach. By capturing the nonlinear relationships between the model inputs and outputs through multiple layers of neurons, DNNs yielded more realistic permeability estimations for the real watershed system, leading to improved match between model predicted and observed stream discharges. However, improving the permeability used in the model alone does not enable the models to capture the inter-annual variability in the discharges, future work is needed to identify additional processes and parameters that may contribute to the unresolved inter-annual variability in system responses.

Note that the accuracy of DNN-based estimation of permeability will be impacted by the accuracy of the mapped distributions of soil and geologic layers, which directly impacts how well a numerical model can represent a real system. During training, the DNNs learn to estimate the permeabilities from the simulated stream discharge for that given distribution map of soil and geologic formation types. Therefore, a less accurate distribution map will result in less accurate estimations of the permeabilities for the watershed, leading to biases that cannot be resolved by the inversion method. A facies-based approach [e.g., Song et al. (2019)] can be adopted to estimate the distribution of soil types and geologic layers along with their permeabilities.

Our study has demonstrated that the DNNs can potentially be a powerful tool to estimate parameters from indirect, relevant observations. The success in linking permeability with stream discharges using DNNs presents new opportunities to improve the subsurface characterization of large-scale watersheds, which has been limited by scarce subsurface characterization and observation data. Our work also paves the way for developing more general model calibration strategies that involve multiple parameters and multiple types of observation data for complex systems. Our next step is to expand the study to assist the multi-process modeling for larger watersheds with more complex subsurface structures. New DNN architectures with deeper and bigger networks might be required to deal with the increasing dimensionality in both model inputs and outputs. Substantial computational resources may also be required to generate sufficient training data for the DNN models if high-resolution distributed models are used.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found on ESS-Dive: https://doi.org/10.15485/1756193.

AUTHOR CONTRIBUTIONS

XC and EC conceived and designed the study with input from SLP and ETC. EC lead the writing of the manuscript, developed DNNs architectures and performed DNNs analysis. ETC and SLP developed the reference case forward simulation of Rock Creek using ATS. PS conducted the ensemble of numerical simulations using ATS and provided data inputs for DNNs training. PJ performed the ensemble smoother. EC provided model related details and background information. SLP, JDM, YL, and ETC provided guidance and overall direction and participated in the interpretation of results. All authors provided critical feedback and inputs to the manuscript.

FUNDING

This work was funded by the ExaSheds project, which was supported by the United States Department of Energy, Office of Science, Office of Biological and Environmental Research,

Earth and Environmental Systems Sciences Division, Data Management Program, under Award Number DE-AC02-05CH11231.

ACKNOWLEDGMENTS

The authors are grateful to Ahmad Jan for helping with model setup for the reference case ATS model of Rock Creek and to Zhufeng Fang for helping developed the reference case forward simulation of Rock Creek using ATS. This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the United States Department of Energy under contract DE-AC02-05CH11231. Pacific Northwest National Laboratory is operated for the DOE by Battelle Memorial Institute under contract DE-AC05-76RL01830. Oak Ridge National Laboratory is managed by UT-Battelle, LLC for the U.S. Department of Energy under Contract Number DE-AC05-00OR22725. Los Alamos National Laboratory is operated by Triad National Security, LLC, for the National Nuclear Security Administration of U.S. Department of Energy (Contract No. 89233218CNA000001). This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the United States Department of Energy or the United States Government. This manuscript has been co-authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paidup, irrevocable, world-wide license to publish, or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feart.2021.613011/full#supplementary-material.

REFERENCES

Ala-aho, P., Soulsby, C., Wang, H., and Tetzlaff, D. (2017). Integrated surface-subsurface model to investigate the role of groundwater in headwater catchment runoff generation: a minimalist approach to parameterisation. *J. Hydrol.* 547, 664–677. doi:10.1016/j.jhydrol.2017.02.023

ASCE Task Committee (2000). Artificial neural networks in hydrology. ii: hydrologic applications. *J. Hydrol. Eng.* 5, 124–137. doi:10.1061/(ASCE)1084-0699(2000)5:2(124)

Bailey, R., and Baù, D. (2010). Ensemble smoother assimilation of hydraulic head and return flow data to estimate hydraulic conductivity distribution. *Water Resour. Res.* 46, W12543. doi:10.1029/2010WR009147

Baxter, J. (1997). A Bayesian/information theoretic model of learning to learn via multiple task sampling. *Mach. Learn.* 28, 7–39. doi:10.1023/a:1007327622663

Calef, M. T., Fichtl, E. D., Warsa, J. S., Berndt, M., and Carlson, N. N. (2013). Nonlinear Krylov acceleration applied to a discrete ordinates formulation of the k-eigenvalue problem. *J. Comput. Phys.* 238, 188–209. doi:10.1016/j.jcp.2012.12.024

Canchumuni, S. W. A., Emerick, A. A., and Pacheco, M. A. C. (2019). Towards a robust parameterization for conditioning facies models using deep variational autoencoders and ensemble smoother. *Comput. Geosci.* 128, 87–102. doi:10.1016/j.cageo.2019.04.006

Carlson, N. N., and Miller, K. (1998). Design and application of a gradient-weighted moving finite element code I: in one dimension. *SIAM J. Sci. Comput.* 19, 728–765. doi:10.1137/S106482759426955X

Carrera, J., Alcolea, A., Medina, A., Hidalgo, J., and Slooten, L. J. (2005). Inverse problem in hydrogeology. *Hydrogeol. J.* 13, 206–222. doi:10.1007/s10040-004-0404-7

Carroll, R. W. H., Bearup, L. A., Brown, W., Dong, W., Bill, M., and Willlams, K. H. (2018). Factors controlling seasonal groundwater and solute flux from snow-dominated basins. *Hydrol. Process.* 32, 2187–2202. doi:10.1002/hyp.13151

Caruana, R. (1997). Multitask learning. *Mach. Learn.* 28, 41–75. doi:10.1023/a:1007379606734

Celia, M. A., Bouloutas, E. T., and Zarba, R. L. (1990). A general mass-conservative numerical solution for the unsaturated flow equation. *Water Resour. Res.* 26, 1483–1496. doi:10.1029/WR026i007p01483

Chen, X., Hammond, G. E., Murray, C. J., Rockhold, M. L., Vermeul, V. R., and Zachara, J. M. (2013). Application of ensemble-based data assimilation techniques for aquifer characterization using tracer data at hanford 300 area. *Water Resour. Res.* 49, 7064–7076. doi:10.1002/2012WR013285

Chen, X., Lee, R. M., Dwivedi, D., Son, K., Fang, Y., Zhang, X., et al. (Forthcoming 2020). Integrating field observations and process-based modeling to predict watershed water quality under environmental perturbations. *J. Hydrol.* doi:10.1016/j.jhydrol.2020.125762

Chen, Y., and Oliver, D. S. (2013). Levenberg-Marquardt forms of the iterative ensemble smoother for efficient history matching and uncertainty quantification. *Comput. Geosci.* 17, 689–703. doi:10.1007/s10596-013-9351-5

Chollet, F., et al. (2015). Keras. Available at: https://keras.io.

Clark, M. P., Rupp, D. E., Woods, R. A., Zheng, X., Ibbitt, R. P., Slater, A. G., et al. (2008). Hydrological data assimilation with the ensemble Kalman filter: use of streamflow observations to update states in a distributed hydrological model. *Adv. Water Resour.* 31, 1309–1324. doi:10.1016/j.advwatres.2008.06.005

Coon, E. (2020). Watershed workflow: a suite of tools for generating hyper-resolution hydrology simulations. Oak Ridge National Laboratory, Tech. Rep. Available at: https://github.com/ecoon/watershed-workflow/.

Coon, E., Moulton, J., Kikinzon, E., Berndt, M., Manzini, G., Garimella, R., et al. (2020). Coupling surface and subsurface flow in complex soil structures using mimetic finite differences. *Adv. Water Resour.* 144, 103701. doi:10.1016/j.advwatres.2020.103701

Coon, E., Svyatsky, D., Jan, A., Kikinzon, E., Berndt, M., Atchley, A., et al. (2019). Advanced terrestrial simulator. [Computer Software]. doi:10.11578/dc.20190911.1

Doherty, J. (2010). *Pest user-manual: model-independent parameter estimation*. 5th Edn. Brisbane, QLD, Australia: Watermark Numerical Computing.

Emerick, A. A., and Reynolds, A. C. (2013). Ensemble smoother with multiple data assimilation. *Comput. Geosci.* 55, 3–15. doi:10.1016/j.cageo.2012.03.011

Evensen, G. (2018). Analysis of iterative ensemble smoothers for solving inverse problems. *Comput. Geosci.* 22, 885–908. doi:10.1007/s10596-018-9731-y

Evensen, G. (1994). Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.* 99, 10143–10162. doi:10.1029/94JC00572

Evensen, G. (2003). The Ensemble Kalman Filter: theoretical formulation and practical implementation. *Ocean Dyn.* 53, 343–367. doi:10.1007/s10236-003-0036-9

Falgout, R. D., and Yang, U. M. (2002). "Hypre: a library of high performance preconditioners," in *Computational science—ICCS 2002 lecture notes in computer science*. Editors P. M. A. Sloot, A. G. Hoekstra, C. J. K. Tan, and J. J. Dongarra (Berlin, Heidelberg, Germany: Springer), 632–641.

Hubbard, S. S., Williams, K. H., Agarwal, D., Banfield, J., Beller, H., Bouskill, N., et al. (2018). The East River, Colorado, watershed: a mountainous community testbed

for improving predictive understanding of multiscale hydrological-biogeochemical dynamics. *Vadose Zone J.* 17, 180061. doi:10.2136/vzj2018.03.0061

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kollet, S., Sulis, M., Maxwell, R. M., Paniconi, C., Putti, M., Bertoldi, G., et al. (2017). The integrated hydrologic model intercomparison project, IH-MIP2: a second set of benchmark results to diagnose integrated hydrology and feedbacks. *Water Resour. Res.* 53, 867–890. doi:10.1002/2016WR019191

Lemieux, C. (2009). "Monte Carlo and quasi-Monte Carlo sampling," in *Springer series in statistics*. Editors P. Bühlmann, P. Diggle, U. Gather, and S. Zeger (New York, NY: Springer).

Mo, S., Zabaras, N., Shi, X., and Wu, J. (2019). Deep autoregressive neural networks for high-dimensional inverse problems in groundwater contaminant source identification. *Water Resour. Res.* 55, 3856–3881. doi:10.1029/2018WR024638

Moradkhani, H., Sorooshian, S., Gupta, H. V., and Houser, P. R. (2005). Dual state-parameter estimation of hydrological models using ensemble Kalman filter. *Adv. Water Resour.* 28, 135–147. doi:10.1016/j.advwatres.2004.09.002

Oyebode, O., and Stretch, D. (2018). Neural network modeling of hydrological systems: a review of implementation techniques. *Nat. Resour. Model.* 32 (1), e12189. doi:10.1002/nrm.12189

Painter, S. L., Coon, E. T., Atchley, A. L., Berndt, M., Garimella, R., Moulton, J. D., et al. (2016). Integrated surface/subsurface permafrost thermal hydrology: model formulation and proof-of-concept simulations. *Water Resour. Res.* 52, 6062–6077. doi:10.1002/2015WR018427

Raghu, M., Poole, B., Kleinberg, J., Ganguli, S., and Sohl-Dickstein, J. (2017). "On the expressive power of deep neural networks," in Proceedings of the 34th international conference on machine learning (PMLR), Sydney, NSW, Australia, August 6-11, 2017. Editors D. Precup and Y. W. Teh (Sydney, NSW: International Convention Centre), Vol. 70, 2847–2854.

Scanlon, T. M., Raffensperger, J. P., Hornberger, G. M., and Clapp, R. B. (2000). Shallow subsurface storm flow in a forested headwater catchment: observations and modeling using a modified TOPMODEL. *Water Resour. Res.* 36, 2575–2586. doi:10.1029/2000WR900125

Shen, C. (2018). A transdisciplinary review of deep learning research and its relevance for water resources scientists. *Water Resour. Res.* 54, 8558–8593. doi:10.1029/2018WR022643

Song, X., Chen, X., Ye, M., Dai, Z., Hammond, G., and Zachara, J. M. (2019). Delineating facies spatial distribution by integrating ensemble data assimilationand indicator geostatistics with level-set transformation. *Water Resour. Res.* 55, 2652–2671. doi:10.1029/2018WR023262

Pribulick, C. (2015). *Propagating climate and vegetation change through the hydrologic cycle in a mountain headwaters catchment*. MS thesis (Hydrology). Golden (CO): Colorado School of Mines.

Thornton, P., Thornton, M., Mayer, B., Wei, Y., Devarakonda, R., Vose, R., et al. (2016). Data from: Daymet: daily surface weather data on a 1-km grid for North America Artwork size: 711509.8892839993 MB medium: NetCDF publisher: ORNL distributed active archive. Version 3, 711509.8892839993 MB 10.3334/ORNLDAAC/1328 Center Version Number: 3.4.

van Leeuwen, P. J., and Evensen, G. (1996). Data assimilation and inverse methods in terms of a probabilistic formulation. *Mon. Wea. Rev.* 124, 2898–2913. doi:10.1175/1520-0493(1996)124<2898:DAAIMI>2.0.CO;2

Vogt, C., Marquart, G., Kosack, C., Wolf, A., and Clauser, C. (2012). Estimating the permeability distribution and its uncertainty at the EGS demonstration reservoir Soultz-sous-Forêts using the ensemble Kalman filter. *Water Resour. Res.* 48, W08517. doi:10.1029/2011WR011673

Wen, X.-H., and Chen, W. H. (2006). Real-time reservoir model updating using ensemble Kalman filter with confirming option. *SPE J.* 11, 431–442. doi:10.2118/92991-PA

Zheng, Q., Zhang, J., Xu, W., Wu, L., and Zeng, L. (2019). Adaptive multifidelity data assimilation for nonlinear subsurface flow problems. *Water Resour. Res.* 55, 203–217. doi:10.1029/2018WR023615

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.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership