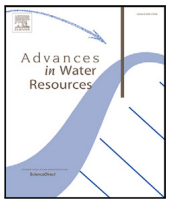
Advances in Water Resources 193 (2024) 104836

Contents lists available at ScienceDirect

Advances in Water Resources

journal homepage: www.elsevier.com/locate/advwatres



Computationally efficient and error aware surrogate construction for numerical solutions of subsurface flow through porous media

Aleksei G. Sorokin [a,d,*], Aleksandra Pachalieva [a,b], Daniel O'Malley [b], James M. Hyman [a], Fred J. Hickernell [d], Nicolas W. Hengartner [c]

[a] Center for Non-Linear Studies, Los Alamos National Laboratory, Los Alamos, 87545, NM, United States
[b] Energy and Natural Resources Security Group (EES-16), Earth and Environmental Sciences Division, Los Alamos National Laboratory, Los Alamos, 87545, NM, United States
[c] Theoretical Biology and Biophysics Group (T-6), Theoretical Division, Los Alamos National Laboratory, Los Alamos, 87545, NM, United States
[d] Illinois Institute of Technology, Department of Applied Mathematics, 10 W 35th Street, Chicago, 60616, IL, United States

ARTICLE INFO

Keywords:
Partial differential equations
Darcy's equation
Random coefficients
Surrogate model
Gaussian process regression

ABSTRACT

Limiting the injection rate to restrict the pressure below a threshold at a critical location can be an important goal of simulations that model the subsurface pressure between injection and extraction wells. The pressure is approximated by the solution of Darcy's partial differential equation for a given permeability field. The subsurface permeability is modeled as a random field since it is known only up to statistical properties. This induces uncertainty in the computed pressure. Solving the partial differential equation for an ensemble of random permeability simulations enables estimating a probability distribution for the pressure at the critical location. These simulations are computationally expensive, and practitioners often need rapid online guidance for real-time pressure management. An ensemble of numerical partial differential equation solutions is used to construct a Gaussian process regression model that can quickly predict the pressure at the critical location as a function of the extraction rate and permeability realization. The Gaussian process surrogate analyzes the ensemble of numerical pressure solutions at the critical location as noisy observations of the true pressure solution, enabling robust inference using the conditional Gaussian process distribution.

Our first novel contribution is to identify a sampling methodology for the random environment and matching kernel technology for which fitting the Gaussian process regression model scales as $\mathcal{O}(n \log n)$ instead of the typical $\mathcal{O}(n^3)$ rate in the number of samples n used to fit the surrogate. The surrogate model allows almost instantaneous predictions for the pressure at the critical location as a function of the extraction rate and permeability realization. Our second contribution is a novel algorithm to calibrate the uncertainty in the surrogate model to the discrepancy between the true pressure solution of Darcy's equation and the numerical solution. Although our method is derived for building a surrogate for the solution of Darcy's equation with a random permeability field, the framework broadly applies to solutions of other partial differential equations with random coefficients.

1. Introduction

Pressure management strategies are essential to prevent overpressurization in the subsurface caused by resource extraction/injection such as wastewater injection and carbon sequestration (Viswanathan et al., 2008; Benson and Cole, 2008; Birkholzer and Zhou, 2009; Stauffer et al., 2011; Middleton et al., 2012; Gholami et al., 2021). Failure to accomplish this goal can lead to induced seismicity (Majer et al., 2007; Zoback, 2012; Keranen et al., 2014; McNamara et al., 2015), leakage of sequestered resources (wastewater or CO_2) (Buscheck et al., 2011; Cihan et al., 2015; Harp et al., 2017; Chen et al., 2018; Chen and Pawar, 2019) and along abandoned wellbores (Pruess, 2008; Watson and Bachu, 2009; Nordbotten et al., 2009; Carey et al., 2010; Huerta et al., 2013; Jordan et al., 2015; Harp et al., 2016; Yonkofski et al., 2019; Lackey et al., 2019; Mehana et al., 2022), and potential contamination of water aquifers (Keating et al., 2010; Little and Jackson, 2010; Trautz et al., 2013; Navarre-Sitchler et al., 2013; Keating et al., 2016; Bacon et al., 2016; Xiao et al., 2020). These events erode public trust, increase economic costs and financial risk, create obstacles to deployment of future projects (Bielicki et al., 2016; Gholami et al., 2021), and can even result in project cancellation (Palmgren et al.,

* Corresponding author at: Illinois Institute of Technology, Department of Applied Mathematics, 10 W 35th Street, Chicago, 60616, IL, United States.
E-mail address: asorokin@hawk.iit.edu (A.G. Sorokin).

https://doi.org/10.1016/j.advwatres.2024.104836
Received 25 June 2024; Received in revised form 4 October 2024; Accepted 10 October 2024
Available online 19 October 2024

2004; Curry et al., 2005; Miller et al., 2007; Wilson et al., 2008; Court et al., 2012; Tcvetkov et al., 2019; Whitmarsh et al., 2019). The key to minimizing the risk of such events is to develop successful reservoir pressure management strategies for choosing well-suited reservoir sites, that are robust against failure and require minimal cost.

Reservoir management operators will benefit from computationally efficient risk analysis and uncertainty quantification tools to support pressure management in heterogeneous subsurface flow fields. To address these issues, complex physics models must be solved with sufficient fidelity and enough realizations to reduce the inherent uncertainties in the heterogeneous subsurface, as well as in the GCS site characterization and operations (Ben-Haim, 2006; O'Malley and Vesselinov, 2015; Chen et al., 2020; Vasylkivska et al., 2021).

Existing pressure management models are often costly to fit and do not account for the discretization error in numerical simulations. To overcome these challenges, our surrogate model for the pressure exploits strategically selected sampling locations and a matching covariance kernel to enable fast model fitting and evaluation. Computations that would typically cost $\mathcal{O}(n^3)$ can be done in $\mathcal{O}(n \log n)$ using our technology of matching n samples to a nicely structured kernel. Moreover, the discretization error in our numerical simulations is systematically encoded into the model through noisy observations and noise variance calibration. The modeled discretization error accounts for both the error in the finite dimensional representation of the random permeability field and the error in using a computational mesh. We emphasize that while our method is applicable to the fine computational mesh discretizations used in practical applications, the large errors incurred by course discretizations are sufficient to build a surrogate with higher uncertainty.

When modeling subsurface flow between injection and extraction wells, there are scenarios where it is essential to ensure that the pressure at a critical location is below a given threshold value. Critical locations would be chosen based on geological, hydrogeological, and operational considerations. For example, they might be chosen to be near preexisting boreholes that could cause leakage or faults/fractures that could cause induced seismicity. To simplify the presentation of our results, we consider the situation where fluid is injected down a single well at a fixed rate into a heterogeneous porous media which increases subsurface pressure, pushing other fluids out at a single extraction well. The method is applicable in more general settings where multiple critical locations are used, and in general the results will depend strongly on the locations of the critical points. To calculate the subsurface pressure, one must solve Darcy's partial differential equation.

We demonstrate how the extraction rate can be controlled so there is a high probability that pressure at a critical location in the subsurface is below a given threshold. For this, we simulate an ensemble of both porous media realizations and extraction rates at which we solve Darcy's equation numerically at several fidelities. These numerical solutions are treated as noisy observations of the unknown analytic solutions. Tens of thousands of simulations can be required to characterize the distribution of solutions for the extraction rate and permeability field inputs. Even though a single computer simulation might only require a few minutes of computer time, an ensemble of thousands of simulations can take hours or days. Because practitioners often need rapid online guidance for real-time control of the injection-extraction process, we use a database of past simulations to create a surrogate model that can provide this guidance in a few seconds.

The application of Machine Learning (ML) models in geosciences has seen increasing popularity in recent years. ML offers new opportunities to learn from big data and fill knowledge gaps in geoscientific models. Data-driven ML models such as deep neural networks have been applied to a wide range of problems within the geosciences such as subsurface characterization (Misra et al., 2019; Chang and Zhang, 2019; Shi and Wang, 2022; Mishra et al., 2021), reservoir modeling (Schuetter et al., 2018; Holdaway, 2014; Mohaghegh, 2017), precipitation prediction (Pan et al., 2022; Shi et al., 2015), and water quality and groundwater levels (Lin et al., 2022; Varadharajan et al., 2022), just to name a few. However, these methods, though useful, have limitations such as the need for large amounts of data (often unavailable), model quality being strongly dependent on the data quality, and model interpretability.

One way to overcome the lack of data and interpretability is by adding physics to the ML models, which gave rise to Physics-Informed Neural Networks (PINNs) (Cai et al., 2021). PINN methods have been used to make predictions of unknown parameters and/or states in the subsurface (Tartakovsky et al., 2020; He et al., 2020; Wu and Qiao, 2021), reservoir management and production forecast by monitoring the injection/extraction rates (Pachalieva et al., 2022; Harp et al., 2021; Gross et al., 2021; Mudunuru et al., 2020), and CO_2 storage predicting saturation and pressure (Shokouhi et al., 2021; Chu et al., 2022), among numerous other applications. PINNs are usually trained so the relationship between derivatives of the neural network matches the governing equations of the PDE. Often PINNs are used to model PDEs with deterministic coefficients (Tang et al., 2020), but their application to PDEs with random coefficients has seen increased interest in recent years (Pachalieva et al., 2022). For example, physics-informed convolutional neural networks have been used for simulating two-phase Darcy flows in heterogeneous media (Rabczuk et al., 2022). Similarly, stochastic deep collocation methods that incorporates neural architecture search and transfer learning for modeling flow in heterogeneous porous media (Zhang et al., 2023). A main limitation of neural networks is their lack of prediction uncertainty.

One proposed method to overcome the lack of predictive uncertainty suffered by neural networks is to use a polynomial chaos expansion (PCE) model. The PCE model has been studied in the context of wave problems in El Mocayd et al. (2021) and for computational hydraulics problems in Al-Ghosoun et al. (2021), Alghosoun et al. (2022). These works also explore encoding the numerical discretization error into the computational model. Following these papers, we use a Karhunen-Loève expansion, also called a proper orthogonal decomposition, to efficiently reduce the dimension of the underlying random field. However, we use a Gaussian process regression model instead of a PCE model as the former may be fit more efficiently by pairing quasi-random sampling locations with matching kernels. Specifically, PCE costs $\mathcal{O}(n^3)$ to fit to n samples, while Gaussian process regression can be performed in $\mathcal{O}(n \log n)$ using efficiently computational pairings.

We have chosen Gaussian process regression (GPR) (Williams and Rasmussen, 2006) to build a surrogate between the inputs of both the extraction rate and permeability field and the output pressure at the critical location. The primary motivation for choosing a GPR model is the ability to encode the discretization error into the surrogate systematically. The discretization error in numerically solving the PDE comes from both choosing a lower fidelity approximation of the permeability field and using a numerical solver on a discrete mesh. We treat the numerical PDE solutions at some target fidelity as noisy observations of the analytical PDE solution. The resulting surrogate approximates the analytic PDE solution based on these noisy numerical solutions. Note that our GPR approach accounts for numerical error but not measurement error. Here, we do not do data assimilation on measurements, so this type of error is outside the scope of our work.

GPR models naturally support noisy observations where the noise is assumed to be zero mean Gaussian with a common variance for each observation. We automatically calibrate a well-adjusted noise variance through hyperparameter optimization on a restricted domain. A conservative starting value for this variance optimization is derived from an approximate upper bound on the error between the analytic PDE solution and the numerical PDE solution. This upper bound is approximated by tracking the decay of solution differences as the problem fidelity increases. A lower bound for the noise variance optimization is derived from the variance of the difference between high-fidelity numerical PDE solves and those at the target fidelity used to fit the GPR.

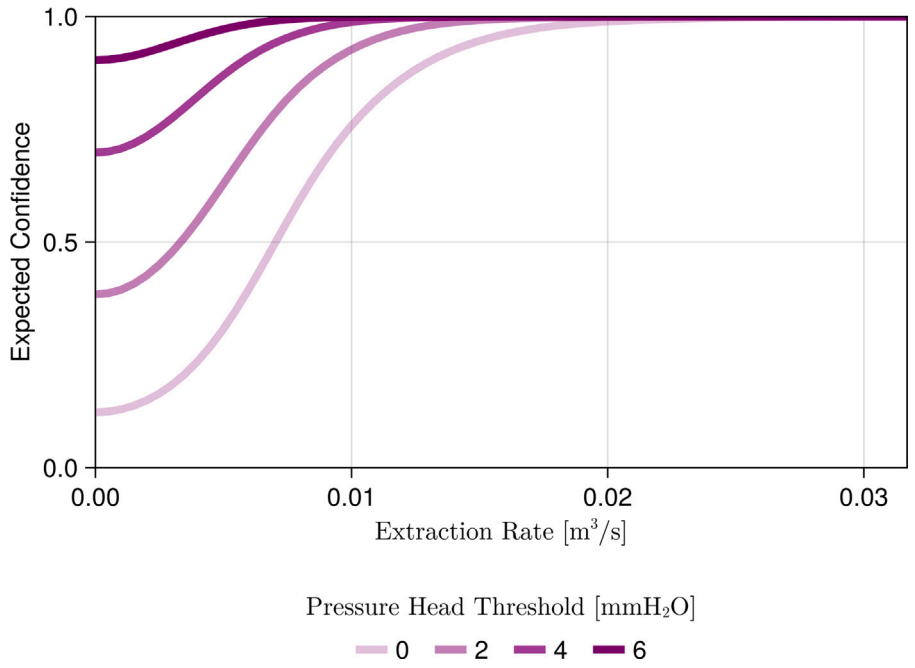


Fig. 1. The expected confidence in maintaining pressure below a threshold as a function of extraction rate [m^3/s]. Fig. 8 extends this plot to a continuum of thresholds.

GPR models enable robust uncertainty quantification in the output predictions. Unlike most machine learning models, the GPR model provides a probability distribution over a broad class of possible analytic PDE solutions. For example, while neural networks may provide predictions of equal quality, a Gaussian process provides confidence levels associated with the predictions based on similarity to already seen data. Specifically, GPR provides easily computed confidence intervals for point predictions or credible intervals for linear functionals of the GPR model. Such predictions under uncertainty are crucial for pressure management systems and would allow operators at underground reservoir sites to make informed decisions for the injection/extraction rates to minimize the risk of overpressurization while maximizing the amount of injected fluid (e.g. reservoir's performance).

Another advantage is that GPR surrogates can be quickly fit when one controls the design of experiments, as is the case here. Fitting a standard GPR surrogate typically costs $\mathcal{O}(n^3)$ where n is the number of data points. By strategically matching sampling locations and covariance kernels, we reduced the cost to $\mathcal{O}(n \log n)$. This reduced cost includes optimizing the noise variance among other hyperparameters, which makes this computationally intensive heterogeneous subsurface problem feasible.

Fig. 1 shows the GPR model prediction for the expected confidence that the pressure at the critical location will be below a given threshold. For example, suppose the threshold pressure is 2 mmH_2O (millimeters of water). In this case, the extraction rate must be at least 0.01 m^3/s (cubic meters per second) to have 90% confidence that the pressure at the critical location will be below the threshold. Notice the confidence computed using the GPR surrogate increases in both the extraction rate and threshold, which matches the physics of the simulation.

Fig. 2 shows the workflow of our approach:

1. We sample the feasibility space by generating a uniform quasi-random (low-discrepancy) sample for the extraction rates and permeability fields. Then, for each extraction-permeability pairing, we solve Darcy's equation numerically at a sequence of increasing fidelities using the DPFEHM software package (O'Malley, 2023). DPFEHM supports a variety of solvers including direct solvers and preconditioned iterative solvers. For this work, we used the default solver, which is a conjugate gradient iterative solver preconditioned with algebraic multigrid. The simulation errors depend on the fidelity of the permeability field and the fidelity of the finite volume numerical solver. The decay of differences between simulations at increasing fidelities informs an approximate upper bound on the noise variance. The differences between simulations at the highest fidelity and the target fidelity inform an approximate lower bound on the noise variance. These bounds are used in the next step for optimizing the noise variance. Once these bounds have been found, we may choose to sample more at the target fidelity as these numerical solutions will be used to build the GPR model.
2. With numerical solutions in hand, the observations at the target fidelity are used to fit a fast GPR model. We emphasize that the target fidelity is not necessarily the maximum among the increasing sequence of fidelities at which we numerically solve the PDE. The target fidelity is chosen based on budgetary restrictions and the GPR surrogate's noise will adapt to this choice. The GPR model may be fit quickly since the chosen quasi-random sampling locations and matching kernel have induced a circulant kernel matrix structure. This fast-fitting cost includes optimizing hyperparameters such as the noise variance. The noise variance optimization is initialized to the approximate upper bound and restricted to be above the approximate lower bound derived in the previous step.
3. The GPR surrogate fit to the target fidelity observations can then be used to select an optimal extraction rate in real-time for any pressure threshold. The key is that the GPR surrogate is much faster to evaluate than the high-fidelity numerical PDE solver and the GPR models the true analytic PDE solution, not the numerical one. This enables rapid real-time analysis for a variety of objectives from an error-aware model.

We emphasize that our proposed workflow readily generalizes from the example we provide in this paper to other PDEs with random coefficients. Our method makes the inclusive assumption that the PDE of interest has a random coefficient which can be approximated at a sequence of increasing fidelities. This paper explores random log-normal permeability fields in two or three dimensions and uses the classic Karhunen-Loève expansion to get lower fidelity approximations. Other random fields in more than three dimensions are also supported

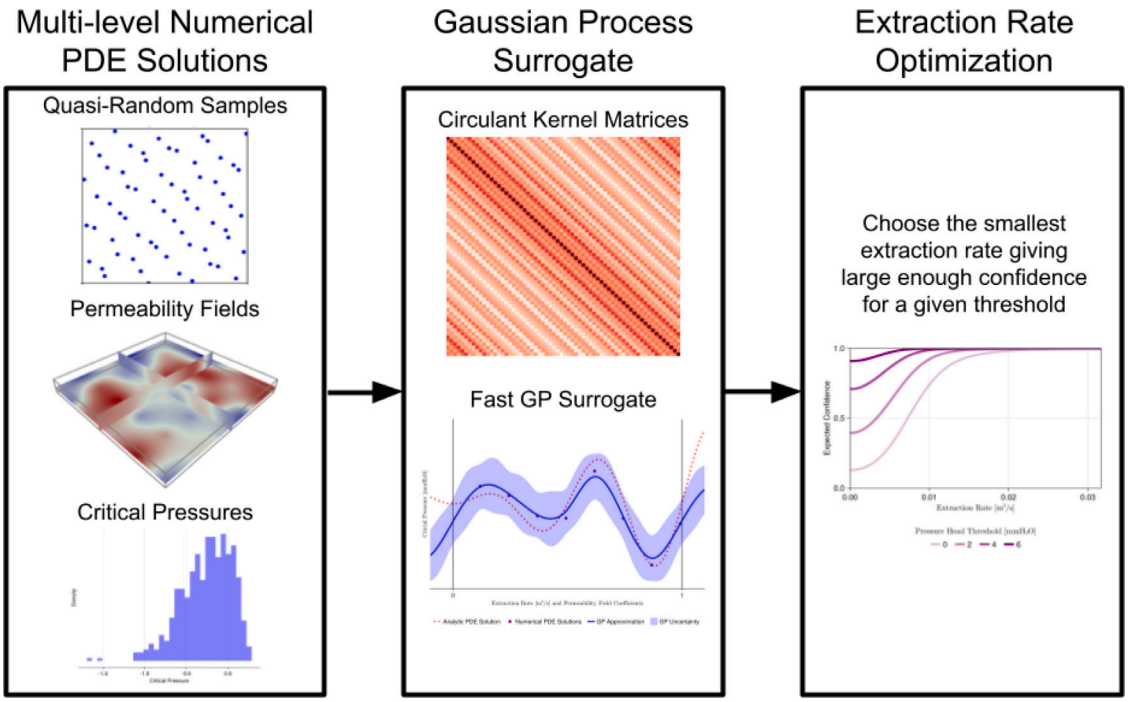


Fig. 2. Workflow diagram visualizing the three stages of our method. First, the possible extraction rate and permeability field realizations are sampled with a low-discrepancy quasi-random uniform distribution. Each pair is input to the numerical partial differential equation solver, which returns an approximation for the pressure at a critical location. Next, a GPR model is optimized to the data relating extraction rate and permeability to pressure at the critical location. The optimization for n samples is done at $\mathcal{O}(n \log n)$ cost by exploiting structure in the Gram kernel matrix induced by using quasi-random samples and matching kernels. The trained GPR model can be quickly evaluated to identify the lowest extraction rate, which rarely overpressurizes a critical location.

when nice parameterizations exist. Examples of methods in geophysics which decouple a permeability field into a series of random variables include optimized principal component analysis methods (Liu, 2017), discrete cosine transforms (Wang et al., 2023), or probability perturbation methods (Grana et al., 2012).

Moreover, numerical PDE solvers exist for more challenging problems including higher dimensional PDEs and those on more intricate domains. These alternative solvers may be plugged into our method without change. The simple two-dimensional Darcy flow we solve with a finite volume method should be viewed as a proof of concept for our generally applicable method.

Section 2 introduces the modeling equations, notation for the problem formulation, and the existing methods we build upon later in the paper. This includes details on numerical solutions of Darcy's equation and an overview of GPR modeling. Our novel theoretical contributions are detailed in Section 3 where we first discuss a method for calibrating the GPR noise to the numerical PDE solution error and then discuss details on fitting a fast Gaussian process at $\mathcal{O}(n \log n)$ cost. Section 4 discusses details of our numerical simulations, exemplifies the use of the trained GPR model for real-time pressure management, and explores the efficacy of both the Gaussian noise assumption and calibration routine. This section concludes by applying our method to the Darcy problem with a three-dimensional subsurface, emphasizing the generality of our algorithm. Finally, Section 5 ends with a brief conclusion and discussion of future work.

2. Methods

This section describes the problem and the model equations of interest. We start by formulating Darcy's equation and describing our quantity of interest: the confidence (probability) that the pressure at a critical location stays below a given threshold. Using a Gaussian process regression (GPR) surrogate gives a distribution over possible pressures at the critical location leading to a random confidence. Next, we discuss the numerical solution of Darcy's equation including the permeability field discretization using the Karhunen-Loève expansion and the two-point flux finite volume method. Finally, we describe how the GPR surrogate views the numerical PDE solutions as noisy evaluations of the true pressure solutions in order to fit a distribution over admissible pressure solutions.

2.1. Problem formulation

Consider a pressure management problem of a single-phase fluid in a heterogeneous permeability field. Darcy's partial differential equation can model the pressure throughout the subsurface

$$\nabla \cdot (G(x) \cdot \nabla H(x)) = f(x), \tag{1}$$

when the subsurface permeability field is known. Darcy's equation describes the pressure head $H(x)$ in the subsurface over a domain $D \subset \mathbb{R}^2$ with permeability field $G(x)$ and external forcing function $f(x)$. The steady-state Darcy equation (1) allows us to evaluate the long-term impact of the injection and extraction on the pressure head. For pressure management, the forcing function f is composed of an injection rate $w \geq 0$ at $x_{\text{injection}}$ and an extraction rate $-r \leq 0$ at $x_{\text{extraction}}$. Following Pachalieva et al. (2022), we write

$$f(x; r) := \begin{cases} w, & x = x_{\text{injection}} \\ -r, & x = x_{\text{extraction}} \\ 0, & x \in D \backslash \{x_{\text{injection}}, x_{\text{extraction}}\}. \end{cases} \tag{2}$$

Throughout this paper, we treat the *injection rate* w as fixed and focus on optimizing the *extraction rate* r, assuming r does not exceed w. Note that in general, these could both be time-dependent. We treat that as a constant because operators may be constrained by the maximum rates at which they can inject/extract or permitting requirements. The use of a constant also simplifies the design problem and gives an indication of the long-term operation at the site that could be sustained.

The details of the permeability field G are rarely known in practice. Instead, one is often given statistical properties of that field, such as

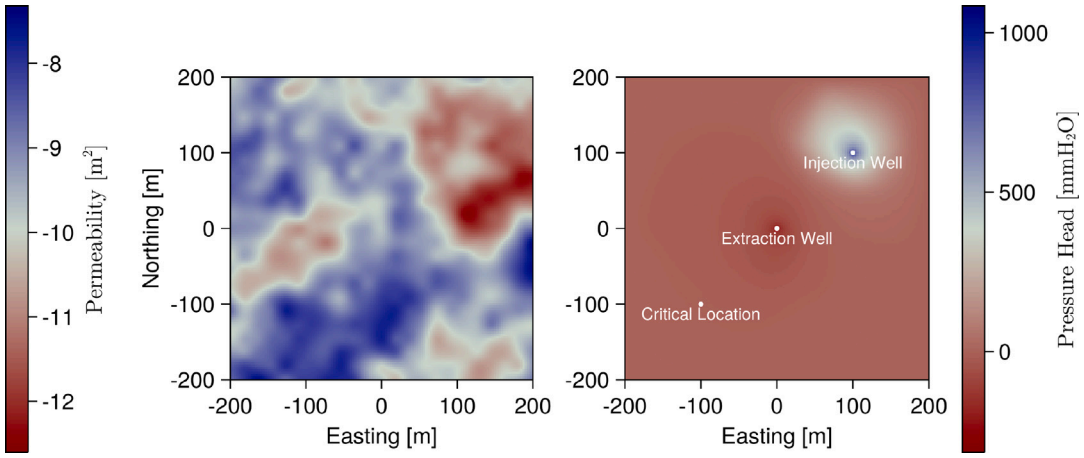


Fig. 3. The left plot shows a realization of the log-permeability field $\log G(x)[\mathrm{m}^2]$ and the right plot shows the resulting pressure $H(x; r, G)$ for some extraction rate r $[\mathrm{m}^3/\mathrm{s}]$. The fixed injection, extraction, and critical locations in our two-dimensional setup are also shown.

the mean permeability and spatial correlations of variations of the permeability around its mean. Given these descriptors, it is convenient to model the permeability G as a *random log-normal* field. Solving Darcy's Eq. (1) using a stochastic permeability field G induces randomness in the pressure H.

Our goal is to quickly estimate the probability that the pressure at the *critical location* x_{critical} remains below a desired *threshold* $\bar{h}$ as a function of the extraction rate r. This enables practitioners to implement control policies to manage pressure at critical locations in real-time. Let $H^c(r, G) := H(x_{\text{critical}}; r, G)$ denote the pressure at the critical location, which is a function of the extraction rate, r, and permeability field, G. For a fixed upper bound $\bar{h}$, we seek to evaluate the *confidence*,

$$c(r) := P_G(H^c(r, G) \le \bar{h}), \tag{3}$$

where the probability P_G is taken over the distribution permeability field G. Fig. 3 illustrates the described setup.

Although the confidence, $c(r)$, cannot be computed explicitly, it can be approximated for each fixed extraction rate r by numerically solving for critical pressure $H^c(r, G)$ in Darcy's Eq. (1) for many realizations of G. Unfortunately, this method is biased as the numerical critical pressure is only approximated the critical pressure $H^c(r, G)$. Also, the associated cost of solving the PDE multiple times is impractical for practitioners desiring fast online inference.

Our approach provides rapid solutions and error estimates for the confidence $c(r)$ by building a surrogate model for the critical pressure $H^c(r, G)$. This statistical approach treats the numerically computed critical pressures as noisy observations of the analytic critical pressures. GPR is a natural and efficient approach for this framework and can provide immediate online estimates for $c(r)$ as a function of the extraction rate.

Given n numerical critical pressure observations, the GPR surrogate $H_n^c(r, G)$ estimates the critical pressure $H^c(r, G)$. We plug this estimate into (3) and get the *conditional confidence*

$$\hat{C}_n(r) := P_G(H_n^c(r, G) \le \bar{h} | H_n^c). \tag{4}$$

The *expected conditional confidence* is a natural estimate for $c(r)$ denoted by

$$c_n(r) := \mathbb{E}_{H_n^c}\left[\hat{C}_n(r)\right] = P_{(G, H_n^c)}(H_n^c(r, G) \le \bar{h}). \tag{5}$$

We approximate the unknown analytic solution $c(r)$ by the computationally tractable $c_n(r)$, which only uses the surrogate model. After interchanging expectations, the above equation can be efficiently computed with (Quasi-)Monte Carlo.

2.2. Numerical solution of Darcy's equation

To solve Darcy's Eq. (1), we apply a standard two-point flux finite volume method in the square domain D on a discrete mesh. A truncated Karhunen-Loève expansion represents the log of the log-normal permeability field G over D. This enables us to draw samples of G, which can be evaluated at a mesh grid of any fidelity.

We say the physical domain has discretization dimension d when the finite volume mesh has $d + 1$ mesh points in each dimension of D. The choice of $d = 2^m$ creates nested mesh grids. While there is no restriction on the mesh grids with an equal number of points in each dimension, this reduces the number of parameters we must consider when approximating the numerical error later in this section.

The Karhunen-Loève expansion (Karhunen, 1947) of the permeability field may be used to find a good finite-dimensional approximation of G. Specifically, we may write

$$\log G(x) = \sum_{j=1}^{\infty} \sqrt{\lambda_j}\varphi_j(x) Z_j \tag{6}$$

where φ_j are deterministic and orthonormal and $Z_1, Z_2, \dots$ are independent standard Gaussian random variables. Ordering $\lambda_1 \ge \lambda_2 \ge \lambda_3 \ge \dots$, we approximate G by

$$\log G^s(x) := \sum_{j=1}^{s} \sqrt{\lambda_j}\varphi_j(x) Z_j \tag{7}$$

which optimally compacts the variance into earlier terms. Fig. 4 shows different pairs of s and d for a common realization of $Z_1, Z_2, \dots$. Notice the greater detail in G^s as s increases and the finer mesh over D as d increases. We will often call the pair (s, d) the fidelity of the numerical solution.

We let $H_{s,d}^c(r, \boldsymbol{Z}^s)$ denote the *numerical critical pressure* computed by solving the PDE (1) with domain discretization dimension d and permeability discretization dimension s. Here $\boldsymbol{Z}^s = (Z_1, \dots, Z_s)$ is a vector of independent standard Gaussians, which uniquely determine the approximate permeability field G^s. In our implementation, we use the `GaussianRandomFields.jl` package (Robbe, 2023) to simulate permeability fields G^s and solve the PDE numerically with the `DPFEHM.jl` (O'Malley, 2023).

2.3. A probabilistic GPR surrogate

We model the relationship between the inputs of extraction rate r and permeability field G and the output critical pressure $H^c(r, G)$. The model is built on observed numerical critical pressures $Y^n :=$

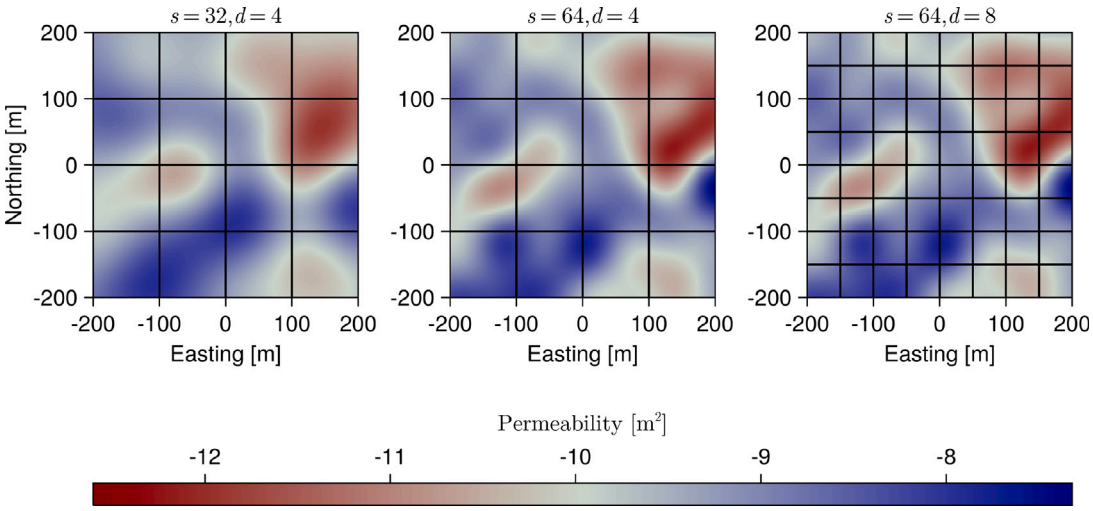


Fig. 4. The same realization of the log-permeability field as in Fig. 3 but for various choices of permeability discretization dimension s and domain discretization dimension d. In the left plot, a small s and d are chosen, corresponding to a lack of small-scale changes in the permeability realization and a coarse mesh grid over the domain, respectively. Moving from the left to center plot maintains the same mesh grid while increasing s to yield more small-scale changes in the realization. Moving from the center to the right plot keeps the same realization while increasing d to yield a finer mesh over D.

$\{H^c_{s,d}(r_i, \boldsymbol{Z}^s_i)\}_{i=1}^n$ at strategically chosen sampling locations $(r_i, G^s_i)_{i=1}^n$. Our GPR surrogate views Y^n as noisy observations of H^c with

$$H^c_{s,d}(r, \boldsymbol{Z}^s) = H^c(r, G) + \varepsilon_{s,d}. \tag{8}$$

We model the noise $\varepsilon_{s,d}$, which encodes the discretization error, as a random variable. That random variable is assumed to be independent of the sampling location $(r, \boldsymbol{Z}^s)$ but dependent on the fidelity (s, d). We further assume that the errors are zero mean Gaussians with variance $\zeta_{s,d}$, i.e.

$$\varepsilon_{s,d} \sim \mathcal{N}(0, \zeta_{s,d}). \tag{9}$$

Assuming homogeneous variances enables us to exploit numerical tricks that lead to fast computations of the posterior expectations. In Section 3.1 we discuss a method for approximating bounds on the noise variance $\zeta_{s,d}$, which are then used during hyperparameter optimization.

GPR assumes H^c is a Gaussian process, and therefore, the conditional distribution of H^c given Y^n is also a Gaussian process (Williams and Rasmussen, 2006). We use this conditional, or posterior, distribution on H^c as an error-aware surrogate. The conditional mean and covariance functions determining the posterior Gaussian process are

$$m_n(t) := \mathbb{E}\left[H^c(t)|Y^n\right] \quad \text{and} \tag{10}$$

$$k_n(t, t') := \mathrm{Cov}\left[H^c(t), H^c(t')|Y^n\right] \tag{11}$$

respectively where $t := (r, \boldsymbol{Z}^s)$ is the $1+s$ dimensional input to the GPR. The conditional variance is written as

$$\sigma_n^2(t) := \mathrm{Var}\left[H^c(t)|Y^n\right] = k_n(t, t). \tag{12}$$

Building upon the above notation, we denote the posterior Gaussian process by

$$H^c_n := H^c|Y_n \sim \mathrm{GP}(m_n, k_n). \tag{13}$$

Fig. 5 illustrates an example posterior Gaussian process. While the figure assumes t is one-dimensional, which is impossible for our problem, GPR extends naturally to arbitrarily large dimensions. We emphasize that the posterior Gaussian process is a surrogate for the critical pressure H^c, not the numerical critical pressure $H^c_{s,d}$ whose evaluations are used for fitting. Our GPR surrogate provides a distribution on H^c whose expectation can be taken as a point estimate for the analytic critical pressure solution.

3. Theory

This section describes in detail the novel theoretical contributions of this work. First, we approximate upper and lower bounds on the variance of the error between the numerical PDE solutions used to fit the GPR and the true PDE solutions. When optimizing the GPR noise variance $\zeta_{s,d}$, the approximate upper bound is used as a starting value and the optimization is restricted to search above the approximate lower bound. Second, we discuss how the fitting of the GPR model to n data points can be accelerated from $\mathcal{O}(n^3)$ to $\mathcal{O}(n \log n)$ using an intelligent design of experiments and matching GPR kernel. This speedup technology is more generally applicable to surrogate modeling when one has control over the design of experiments.

3.1. Approximate bounds on GPR noise variance

Recall the GPR model with zero mean Gaussian noise assumes the numerical solution is unbiased for the analytic solution. Therefore, the noise variance $\zeta_{s,d}$ is the Mean Squared Error (MSE) between the analytic PDE solution and the numerical PDE solutions. This section derives approximate upper and lower bounds on the Root Mean Squared Error (RMSE) $\sqrt{\zeta_{s,d}}$. The upper bound is used as a starting point to calibrate $\zeta_{s,d}$ when performing hyperparameter optimization of the GPR model. This bound is derived by tracking the decay in average solution differences as the problem fidelity is increased. The lower bound is used to restrict the search domain for this hyperparameter optimization. This heuristic lower bound is derived by looking at the MSE of differences between solutions at the maximum and target fidelities.

We start by approximating an upper bound on $\zeta_{s,d}$. First, notice the assumption of zero mean Gaussian noise in (9) implies that the standard deviation of the GPR noise may be written as the RMSE

$$\begin{aligned}\sqrt{\zeta_{s,d}} &= \left\| H^c(R, G) - H^c_{s,d}(R, \boldsymbol{Z}^s) \right\| \\ &= \sqrt{\mathbb{E}_{(R,\boldsymbol{Z}^s)}\left[H^c(R, G) - H^c_{s,d}(R, \boldsymbol{Z}^s)\right]^2} \\ &=: \mathrm{RMSE}_{s,d}.\end{aligned} \tag{14}$$

Here the extraction rate R is assume to be uniformly distributed between 0 and w, i.e. $R \sim \mathcal{U}[0, w]$. Moreover, the extraction rate is assumed to be independent of $\boldsymbol{Z}^s$.

Following ideas from Multi-level Monte Carlo (Giles, 2008; Robbe et al., 2017), we choose strictly increasing sequences $(s_j)_{j\geq 0}$ and $(d_j)_{j\geq 0}$

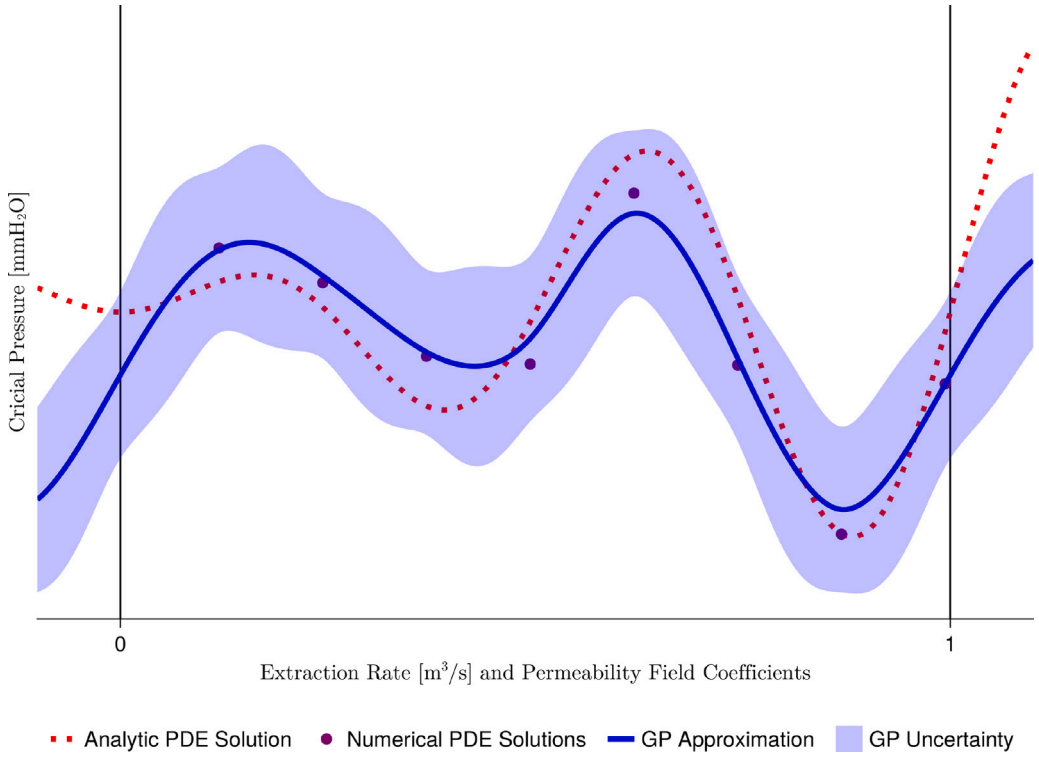


Fig. 5. Cartoon Gaussian process model. Suppose we are interested in recovering the analytic PDE solution H^c from numerical PDE solutions, which we treat as noisy observations of H^c. The posterior Gaussian process (GP) for H^c is visualized through its posterior mean approximation and 99% confidence interval uncertainty at each point. While the cartoon shows a one-dimensional input, our actual model takes a $1+s$ dimensional input: 1 for the extraction rate and s for the random coefficients in the Karhunen-Loève expansion of the permeability field.

with $s = s_T$ and $d = d_T$ and rewrite (14) as the telescoping sum

$$\mathrm{RMSE}_{s_T,d_T} = \left\| \sum_{j=T+1}^{\infty} \left[\Delta_{s_j}(R, Z^{s_j}) + \Delta_{d_j}(R, Z^{s_j}) \right] \right\| \tag{15}$$

where

$$\Delta_{s_{j+1}}(R, Z^{s_{j+1}}) = H^c_{s_{j+1},d_j}(R, Z^{s_{j+1}}) - H^c_{s_j,d_j}(R, Z^{s_j}),$$
$$\Delta_{d_{j+1}}(R, Z^{s_{j+1}}) = H^c_{s_{j+1},d_{j+1}}(R, Z^{s_{j+1}}) - H^c_{s_{j+1},d_j}(R, Z^{s_{j+1}}).$$

Here $T \geq 0$ is the index of the target fidelity whose observations will be used to fit the GPR model. Let us assume that

$$\|\Delta_{s_j}(R, Z^{s_j})\| = 2^{b_s} s_j^{a_s} \quad \text{and} \quad \|\Delta_{d_j}(R, Z^{s_j})\| = 2^{b_d} d_j^{a_d}. \tag{16}$$

The parameters (a_s, b_s) and (a_d, b_d) will be fit using linear regression in the log–log domain. Let $s_j = v_s 2^j$ and $d_j = v_d 2^j$ where v_s and v_d are the respective initial values chosen by the user. Applying the triangle inequality to (15) gives

$$\begin{aligned} \mathrm{RMSE}_{s_T,d_T} &\leq \sum_{j=T+1}^{\infty} \left[2^{b_s} v_s^{a_s} (2^{a_s})^j + 2^{b_d} v_d^{a_d} \left(v_d^{a_d}\right)^j \right] \\ &= 2^{b_s} v_s^{a_s} \frac{2^{(T+1)a_s}}{1-2^{a_s}} + 2^{b_d} v_d^{a_d} \frac{2^{(T+1)a_d}}{1-2^{a_d}} \\ &=: \overline{\mathrm{RMSE}}_{s_T,d_T} \end{aligned} \tag{17}$$

using the expression for the sum of a geometric series.

Fig. 6 illustrates the above idea for both a two dimensional subsurface (above plot) and three dimensional subsurface (below plot). First, we pick an $M \geq T$ so that (s_M, d_M) is the maximum fidelity at which we will numerically solve the PDE and (s_T, d_T) is the target fidelity whose numerical solutions will be used to fit the GPR surrogate. We emphasize that $M \geq T$ since we do not require the GPR surrogate to be built on maximum fidelity solves. Now, at every fidelity (s_j, d_j) with $1 \leq j \leq M$ we solve the PDE at the same $(R_i, Z_i^{s_M})_{i=1}^m$ points to get $\{H^c_{s_j,d_j}(R_i, Z_i^{s_j})\}_{i=1}^m$. Here $Z_i^{s_j}$ is the first s_j element of $Z_i^{s_M}$. For $1 \leq j \leq M$ we make the approximations

$$\begin{aligned} \|\Delta_{s_j}(R, Z^{s_j})\| &\approx \sqrt{\frac{1}{m} \sum_{i=1}^{m} \Delta^2_{s_j}(R_i, Z^{s_j})}, \\ \|\Delta_{d_j}(R, Z^{s_j})\| &\approx \sqrt{\frac{1}{m} \sum_{i=1}^{m} \Delta^2_{d_j}(R_i, Z^{s_j})} \end{aligned} \tag{18}$$

corresponding to the plotted red dots and blue squares, respectively. The slope intercept pairings (a_s, b_s) and (a_d, b_d) from (16) are fit to the values in (18) with lines in the respective colors. The upper bounds $\overline{\mathrm{RMSE}}_{s_j,d_j}$ from (17) are visualized by the purple stars. Notice that the model will find $\overline{\mathrm{RMSE}}_{s_T,d_T}$ for any target fidelity (s_T, d_T) we choose. As d_T increases, the mesh size shrinks, and the PDE becomes more expensive to solve numerically. As s_T increases, the input dimension to the GPR model grows, and more PDE solves are required to build an accurate model. Notice that the RMSE is dominated by the error in the permeability field discretization rather than the error in the domain discretization.

We now approximate a lower bound on $\zeta_{s,d} = \zeta_{s_T,d_T}$. Recall that ζ_{s_T,d_T} is MSE between analytic solutions and the numerical solutions at target fidelity (s_T, d_T). Practically speaking, we expect this to be at least as large as the MSE between the solutions at the maximum fidelity (s_M, d_M) and target fidelity (s_T, d_T) i.e.

$$\begin{aligned} \sqrt{\zeta_{s_T,d_T}} &= \sqrt{\mathbb{E}_{(R,Z^{s_M})} \left[H^c(R, G) - H^c_{s_M,d_M}(R, Z^{s_M}) \right]^2} \\ &\gtrapprox \sqrt{\mathbb{E}_{(R,Z^{s_M})} \left[H^c_{s_M,d_M}(R, Z^{s_M}) - H^c_{s_T,d_T}(R, Z^{s_T}) \right]^2} \\ &=: \underline{\mathrm{RMSE}}_{s_T,d_T} \end{aligned}$$

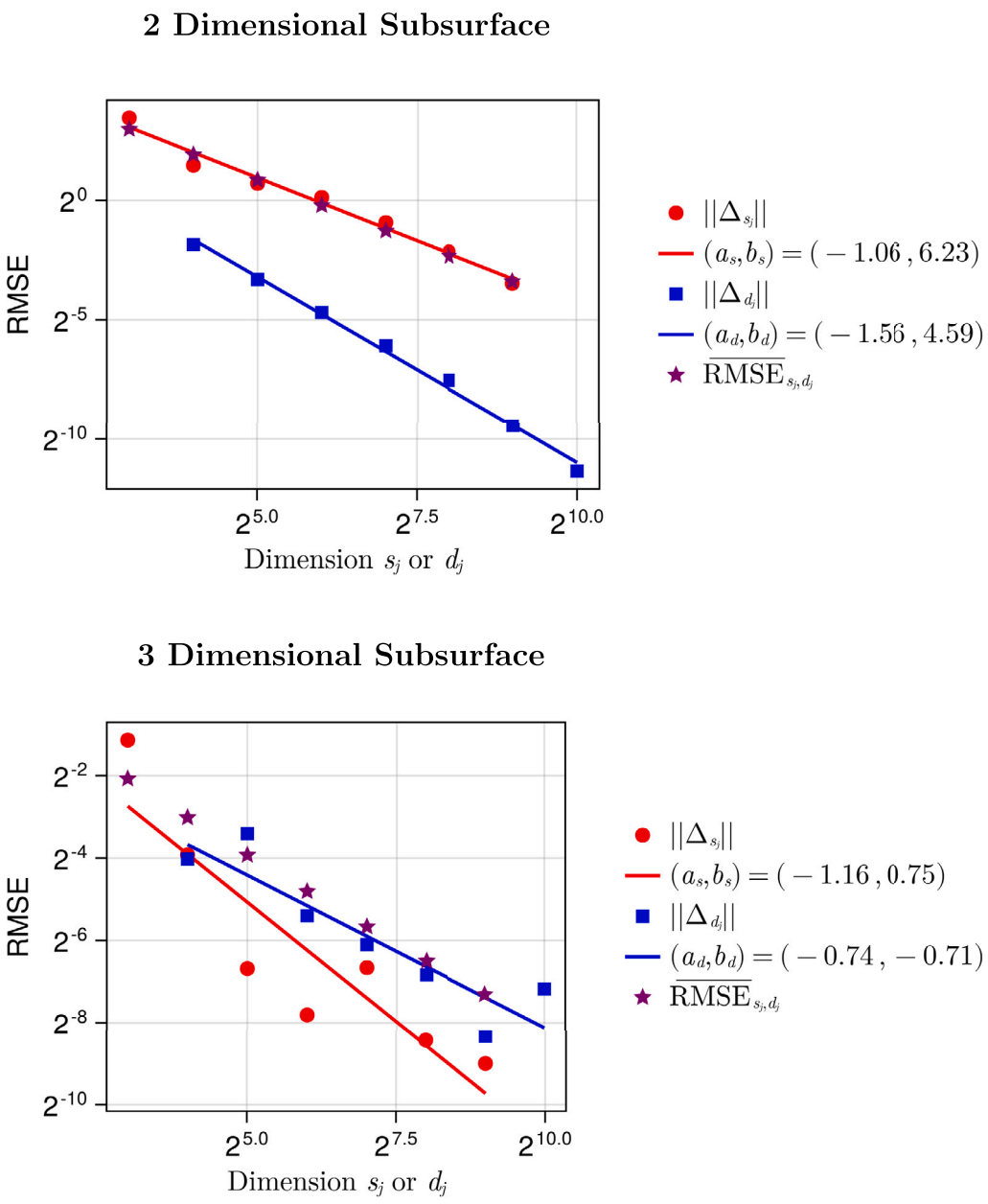


Fig. 6. The Karhunen-Loève expansion of the permeability field is approximated by the sum of the first s_j terms. The physical domain D is discretized into a grid with mesh width $1/d_j$ in each dimension. The RMSE of the difference between numerical critical pressures with discretizations $(s_j, d_j/2)$ and $(s_j/2, d_j/2)$ is scatter plotted as $\|\Delta_{s_j}\|$. The RMSE of the difference between numerical critical pressures with discretizations (s_j, d_j) and $(s_j, d_j/2)$ is scatter plotted as $\|\Delta_{d_j}\|$. Simple regression models are fit to these two scatter trends where a and b are the slope and intercept, respectively, of the plotted lines. These models are extrapolated through an infinite telescoping sum to derive the approximate upper bound on the RMSE of the difference between the numerical critical pressure with discretization (s_j, d_j) and the target critical pressure.

Similar to (18), we approximate this heuristic lower bound by the sample RMSE:

$$\underline{\text{RMSE}}_{s_T,d_T} \approx \sqrt{\frac{1}{m}\sum_{i=1}^{m} \tilde{\Delta}^2_{M,T,i}}. \tag{19}$$

where

$$\tilde{\Delta}_{M,T,i} = H^c_{(s_M,d_M)}(R_i, Z_i^{s_M}) - H^c_{s_T,d_T}(R_i, Z_i^{s_T}). \tag{20}$$

3.2. Fast GPR

Gaussian processes regression models are defined with a positive-definite covariance kernel k, which assumes $\text{Cov}[H^c(t), H^c(t')] = k(t,t')$. The posterior mean m_n and posterior covariance k_n given the noisy data Y^n requires solving the linear system $\tilde{K}a = b$ for $a \in \mathbb{R}^n$ given $b \in \mathbb{R}^n$ where $\tilde{K} = K + \zeta_{s,d} I_n$ is the $n \times n$ noisy kernel matrix. Here $K = (k(t_i, t_j))_{i,j=1}^n$ is the kernel matrix of pairwise evaluations at sampling locations $(t_i)_{i=1}^n$ and I_n is the $n \times n$ identity matrix. The cost of solving the system $\tilde{K}a = b$ for a is $\mathcal{O}(n^3)$ in the general case where $\tilde{K}$ is dense and unstructured. This computational cost limits the sample size used to build the surrogate.

For some structured $\tilde{K}$, the linear system can be solved in $\mathcal{O}(n \log n)$. For example, when K is circulant or block Toeplitz (Gray et al., 2006) the same structure is induced in $\tilde{K}$ and the linear system can be solved using fast Fourier transforms. We can induce such structure in K by strategically choosing the sampling locations $(t_i)_{i=1}^n$ and matching covariance kernel k. Two available flavors are:

Table 1
Comparison of construction and evaluation costs between the unstructured and structured Gaussian processes in terms of the number of samples n used to fit the Gaussian process. Construction costs include hyperparameter optimization (tuning) and computing constants for evaluating the posterior mean and covariance.

	Construction		Evaluation	
	Tuning	Constants	Mean	Covariance
Unstructured GP	$\mathcal{O}(n^3)$	$\mathcal{O}(n^3)$	$\mathcal{O}(n)$	$\mathcal{O}(n^2)$
Structured GP	$\mathcal{O}(n \log n)$	$\mathcal{O}(n \log n)$	$\mathcal{O}(n)$	$\mathcal{O}(n \log n)$

- Lattice sequence $(t_i)_{i=1}^n$ and periodic shift invariant k produce circulant K.
- Digital sequence $(t_i)_{i=1}^n$ and digitally shift invariant k produce block Toeplitz K.

These methods for inducing circulant and block Toeplitz structure in K were previously used for Bayesian cubature in Jagadeeswaran and Hickernell (2019) and Jagadeeswaran and Hickernell (2022) respectively with a unifying thesis on fast Bayesian cubature available in Rathinavel (2019). Parallel developments for kernel interpolation in reproducing kernel Hilbert space are described in Kaarnioja et al. (2022, 2023), Kuo et al. (2023). This paper differs from the methods in these references in two ways. First, our confidence quantity of interest is a non-linear functional of the GPR surrogate as opposed to the linear mean studied in the cubature context. Second, we add support for noisy observations, including methods to optimize the noise variance $\zeta_{s,d}$.

Fig. 7 plots lattice and digital sequences $(t_i)_{i=1}^n$ alongside matching kernels and the induced kernel matrices. Notice that lattice and digital sequences lie in the unit cube and have low discrepancy with the standard uniform distribution (Owen, 2013, Chapters 15 and 16). The periodicity in the lattice sequence kernels and discontinuities in the digital sequence kernels induce the same features in the posterior mean. For example, Fig. 5 showed a GPR surrogate with lattice sampling locations and matching kernel.

Table 1 compares construction and evaluation costs between the unstructured and structured Gaussian processes described in this section. The construction costs include the tuning parameters to maximize the marginal likelihood of the Gaussian process and the computing constants to be used during evaluation. Evaluation costs for both the posterior mean and covariance occur after construction.

Recall that the noise variance $\zeta_{s,d}$ encodes the approximation error and acts as a regularization of the posterior Gaussian processes, which honors observations Y^n less as $\zeta_{s,d}$ grows. We choose $\zeta_{s,d}$ to optimize the marginal likelihood of the Gaussian process. The upper bound in (17) is used as a conservative initial guess for the noise variance $\zeta_{s,d}$ while the optimization is constrained to search above the lower bound in (19).

4. Numerical experiments

This section discusses the numerical experiments used to test our method. We stress that the experiments are meant to be a proof of concepts which is easily generalized to more realistic problems rather than an exhaustive real-life application. This section begins by describing implementation specifics such as transformations required to make the fast GPR framework compatible with Darcy's problem and estimation techniques for the expected conditional confidence. We then describe two numerical experiments used to evaluate our method. The first is a straightforward application of our trained GPR model for real-time analysis of the relationship between extraction rate and confidence in maintaining a low enough pressure at the critical location. The second experiment visualizes the calibration procedure for the GPR noise variance and explores the requisite assumption of Gaussian noise. Finally, we emphasize the generality of our algorithm by applying it to the same Darcy problem but with a three-dimensional subsurface.

4.1. Implementation considerations

To enable the reproducibility of our results, we describe the specifics for approximating the confidence $c(r)$ in (3) by the expected conditional confidence $c_n(r)$ in (5). Recall from Section 2.3 that lattice sequences are defined in the unit cube $[0,1]^{1+s}$. Moreover, the posterior mean of a GPR surrogate fit with lattice sampling locations, and a matching covariance kernel has a periodic posterior mean. We first transform the extraction rate r and independent Gaussians $\boldsymbol{Z}^s$, which defines the permeability field G^s, to the unit cube domain. We then periodize the critical pressure using the baker transform.

To transform our problem to the unit cube $[0,1]^{1+s}$, recall the assumption that the extraction rate no larger than the injection rate i.e. $r \in [0, w]$. Let Φ^{-1} denote the inverse distribution function of a standard Gaussian random variable. For any $r_u \in [0,1]$ and $u_1, u_2, \ldots \in (0,1)$ we may use (6) to write

$$M^c(r_u, u_1, u_2, \ldots) := H^c\left(wr_u, \exp\left(\sum_{j\geq 1} \sqrt{\lambda_j}\varphi_j(x)\Phi^{-1}(u_j)\right)\right). \tag{21}$$

For independent standard uniform random variables $U_1, U_2, \ldots \overset{\text{iid}}{\sim} \mathcal{U}[0,1]$ we have

$$P_G(H^c(r, G) \leq \bar{h}) = P_{(U_1,U_2,\ldots)}(M^c(r/w, U_1, U_2, \ldots) \leq \bar{h}).$$

To periodize $M^c(r_u, u_1, u_2, \ldots)$, define the baker transform (Owen, 2013, Chapter 16)

$$b(u) = 1 - 2\left|u - \frac{1}{2}\right| = \begin{cases} 2u, & 0 \leq u \leq 1/2 \\ 2(1-u) & 1/2 \leq u \leq 1 \end{cases} \tag{22}$$

so that

$$\mathring{M}^c(r_u, u_1, u_2, \ldots) := M^c(b(r_u), b(u_1), b(u_2), \ldots). \tag{23}$$

Since $b(U) \sim \mathcal{U}[0,1]$ when $U \sim \mathcal{U}[0,1]$, we have

$$\begin{aligned} c(r) &= P_{(U_1,U_2,\ldots)}\left(M^c(r/w, U_1, U_2, \ldots) \leq \bar{h}\right) \\ &= P_{(U_1,U_2,\ldots)}\left(\mathring{M}^c(r/(2w), U_1, U_2, \ldots) \leq \bar{h}\right). \end{aligned} \tag{24}$$

Let $\mathring{M}_n^c(r_u, u_1, \ldots, u_s)$ denote the posterior Gaussian process for $\mathring{M}^c(r_u, u_1, u_2, \ldots)$. Substituting $\mathring{M}_n^c(r_u, u_1, \ldots, u_s)$ into (24) and taking the expectation gives

$$\begin{aligned} \mathring{c}_n(r) &:= P_{(\mathring{M}_n^c, U_1, \ldots, U_s)}(\mathring{M}_n^c(r/(2w), U_1, \ldots, U_s) \leq \bar{h}) \\ &= \mathbb{E}_{(U_1,\ldots,U_s)}\left[\Phi\left(\frac{\bar{h} - \mathring{m}_n(r, U_1, \ldots, U_s)}{\mathring{\sigma}_n(r, U_1, \ldots, U_s)}\right)\right]. \end{aligned} \tag{25}$$

Eq. (24) motivates us approximating $c(r)$ by $\mathring{c}_n(r)$. Here the final inequality follows from Fubini's theorem (Fubini, 1907) and $\mathring{m}_n$ and $\mathring{\sigma}_n$ are the posterior mean and standard deviation of $\mathring{M}_n^c$ when plugged into (10) and (12) respectively.

We use QMCGenerators.jl (Sorokin, 2023) to generate lattice (or digital) sequences. The lattice or digital sequences can also efficiently approximate (25) using Quasi-Monte Carlo (Niederreiter, 1992; Owen, 2013). Specifically, we use the Quasi-Monte Carlo estimate

$$\mathring{c}_n(r) \approx \frac{1}{N}\sum_{i=1}^{N} \Phi\left(\frac{\bar{h} - \mathring{m}_n(r, U_{i1}, \ldots, U_{is})}{\mathring{\sigma}_n(r, U_{i1}, \ldots, U_{is})}\right) \tag{26}$$

where $(U_{i1}, \ldots, U_{is})_{i=1}^N$ are low-discrepancy points e.g. the first N points of a lattice or digital sequence.

4.2. Two-dimensional experiments and analysis

The domain of our subsurface is square with side lengths of 200 m with the injection well, extraction well, and critical location shown in Fig. 3. We set the injection rate to 0.031688 m^3/s (equivalent to 1 million metric tons per year [MMT/y]) and test extraction rates between $-0.031688\,\text{m}^3/s$ and $0.0\,\text{m}^3$/s. We use a zero mean log-normal

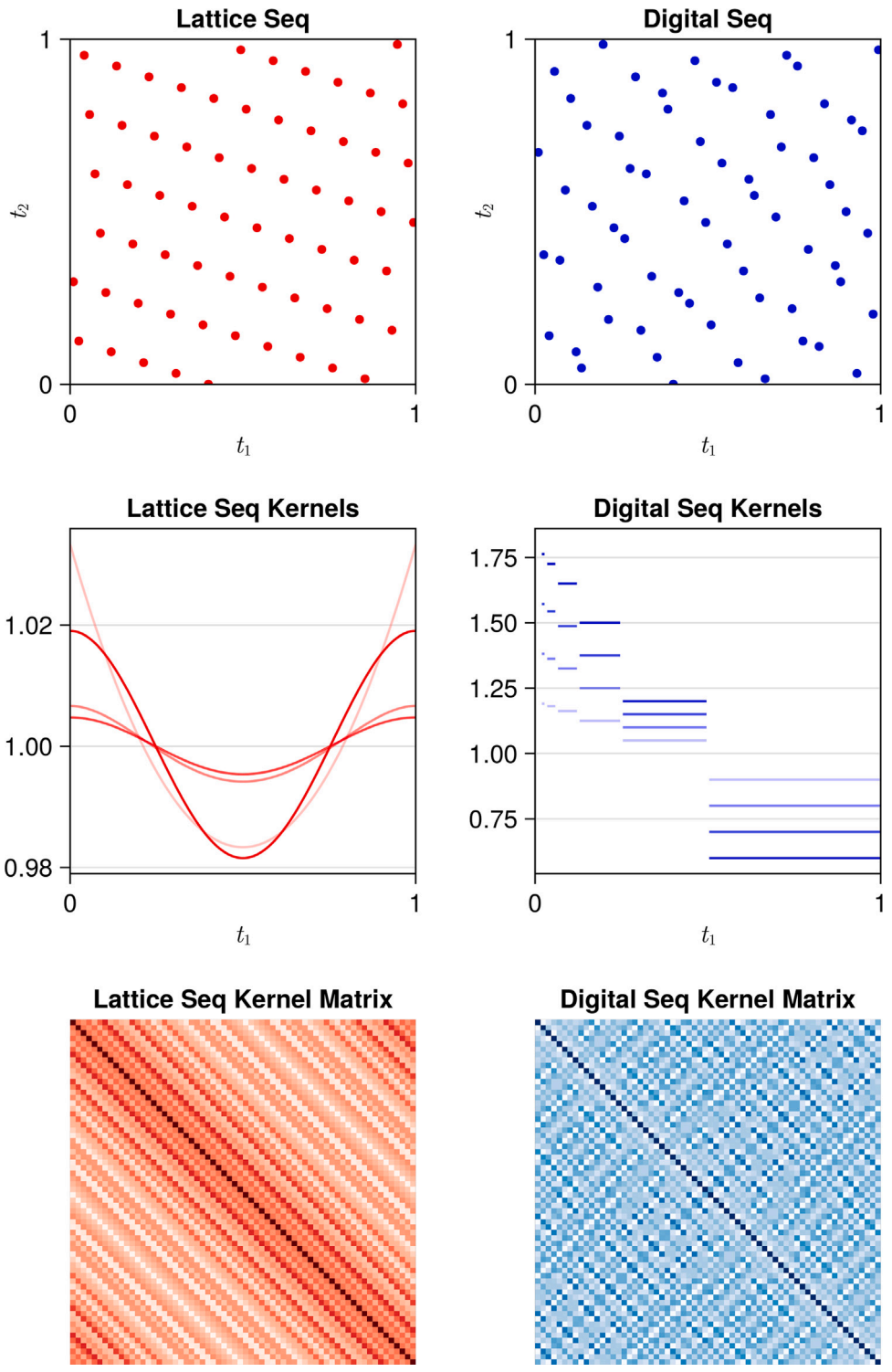


Fig. 7. The first column shows the first n points of a lattice sequence and some matching periodic kernels that yield circulant kernel matrices. The second column shows the first n points of a digital sequence and some matching step function kernels that yield block Toeplitz kernel matrices. Features of the kernel functions are induced in the prediction function. For instance, using the lattice sequence and matching kernels will yield a periodic prediction, while using digital sequences and matching kernels will yield a discontinuous step-function prediction.

permeability field with a Matérn covariance kernel having a correlation length 50 m.

Our GPR surrogate is fit to numerical experiments with fidelity $(s, d) = (64, 128)$ i.e. 64 were terms kept in the KL expansion and the mesh width for the finite volume solver was $1/128$ in both dimensions. The sequence of fidelities used to find upper and lower bounds for noise variance tuning were $(s_j)_{j=0}^{M} = (4, 8, 16, 32, 64, 128, 256, 512)$ and $(d_j)_{j=0}^{M} = (8, 16, 32, 64, 128, 256, 512, 1024)$ as shown in Fig. 6 for the two dimensional subsurface. At each fidelity, the PDE was solved numerically at $m = 128$ extraction-permeability pairings, and $n = 1024$ solves at fidelity $(s_T, d_T) = (64, 128)$ were used to fit the GPR. The Quasi-Monte Carlo approximation in (26) was performed using $N = 1024$ randomly shifted lattice points. The CPU time required for each step of experimentation are given in Table 2. The condition number of the

Table 2
CPU time for different stages of the proposed method on the Darcy problem in two and three dimensions. First, the KL expansion is performed on a fine grid. Then the finite volume method from DPFEHM (O'Malley, 2023) is used solve Darcy's equation on common permeability realizations at different fidelities. The Gaussian process regression (GPR) model is then fit at the target fidelity, and GPR inference may be performed at a fraction of the cost compared to the a traditional solver. Note that GPR fitting includes FV The fidelity parameters are the number of samples m, the number of KL terms/number of input dimensions to the GPR s, and the domain discretization fidelity d. For the two dimensional Darcy problem, the fidelity d indicates a $d \times d$ computational mesh while for the three dimensional Darcy problem the fidelity d indicates a $d \times d \times 9$ computational mesh. Note that the cost of KL is independent of the number of samples, so these entries are left blank.

Step	Darcy 2D					Darcy 3D				
	CPU time [sec]		m	s	d	CPU time [sec]		m	s	d
KL on fine grid	229			512	1024	2578			512	1024
		47	128	4	8		320	128	4	8
		76	128	8	8		341	128	8	8
		82	128	8	16		340	128	8	16
		137	128	16	16		384	128	16	16
		220	128	16	32		383	128	16	32
FV solves	7857	263	128	32	32	46657	468	128	32	32
		317	128	32	64		532	128	32	64
		336	128	64	64		663	128	64	64
		340	128	64	128		745	128	64	128
		375	128	128	128		1077	128	128	128
		389	128	128	256		1645	128	128	256
		464	128	256	256		2309	128	256	256
		527	128	256	512		5502	128	256	512
		671	128	512	512		6849	128	512	512
		908	128	512	1024		19182	128	512	1024
		2704	1024	64	128		5916	1024	64	128
GPR fitting	3.5		1024	64	128	2.0		1024	64	128
GPR inference	3.0		1024	64	128	2.8		1024	64	128

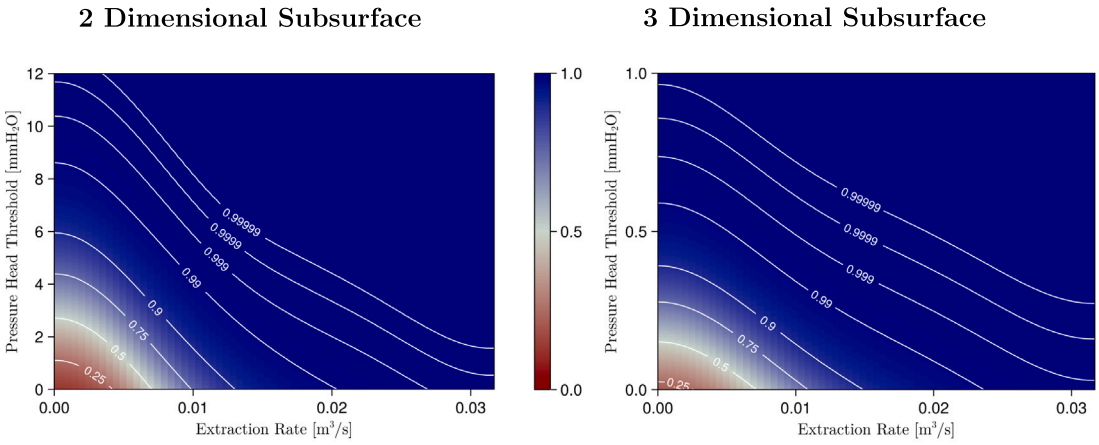


Fig. 8. Approximate expected posterior confidence from (26) by extraction rate r and pressure threshold $\bar{h}$.

noisy circulant kernel matrix is 387. As the condition number is the ratio between the largest and smallest eigenvalues, one may decrease the condition number by raising the lower bound on the GP noise variance.

In Fig. 8 for the two dimensional subsurface, we plot the approximate expected conditional confidence in (26) for a range of extraction rates r and pressure thresholds $\bar{h}$. While the surrogate is not constrained to be monotonically increasing in both extraction rate r and threshold $\bar{h}$, the expected confidence appears to have this qualitative behavior. This reassures us that our surrogate captures the physics in the model. Fig. 1 may be viewed as slices of the left plot of Fig. 8 at fixed pressure threshold $\bar{h}$. For a fixed pressure threshold $\bar{h}$, numerous methods exist to use the GPR model to find an extraction rate which yields a desired confidence. We emphasize this can all be done in real-time using only evaluations of the GPR surrogate.

We now analyze our assumption of Gaussian noise for the GPR model and our method of optimization. Fig. 9 illustrates the critical distributions considered for noise variance fitting. A frequency plot of the errors between the maximum and target fidelities is plotted i.e. a frequency plot of $\tilde{\Delta}_{M,T,i}$ from (20). The sample MSE of these errors is used an approximate lower bound on the noise variance for optimization. The initial noise variance for optimization was set to the upper bound approximated the decay of differences in numerical solves at a sequence of increasing fidelities, see (17). Section 3.1 contains details on both these approximate bounds.

For the two dimensional subsurface, Fig. 9 shows the fitted GPR noise variance essentially matches the lower bound. In fact, if we do not lower bound the noise variance, our optimization to maximize the marginal likelihood will choose an optimal noise variance orders of magnitude smaller than the lower bound. In practice, we observed the GPR based confidence estimate is robust to the choice of observation noise.

It may also be observed in the left plot of Fig. 9 that the distribution of errors between the target and maximum fidelity numerical solutions does not appear Gaussian but instead appears to have heavier tails. The distribution of these errors should be close to the distribution of errors between the target fidelity numerical solutions and analytic solutions. This later error is what is modeled by our GPR noise. The use of a GPR necessitates the assumption of Gaussian noise, but our results suggest this assumption may not hold in practice.

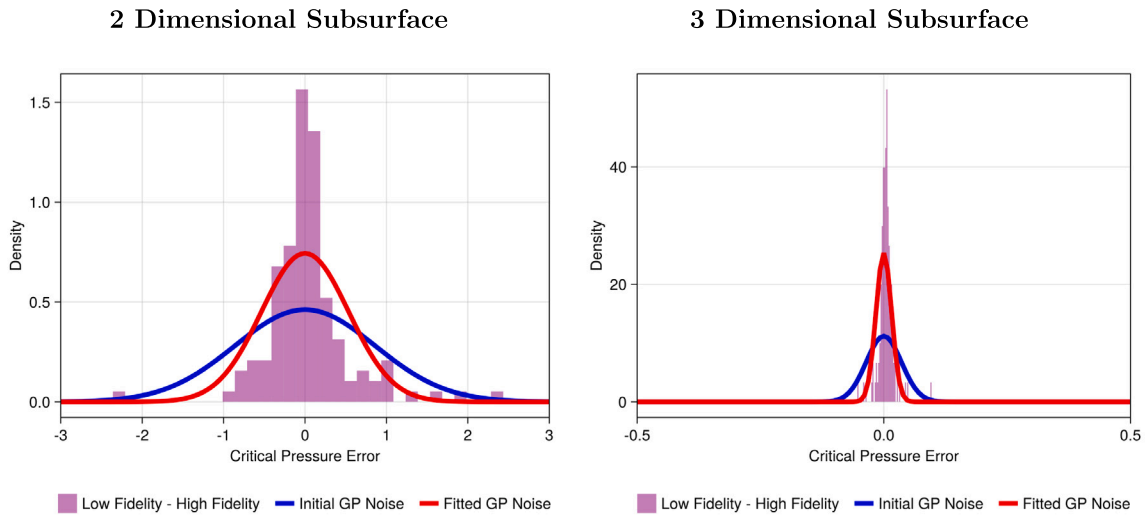


Fig. 9. Analysis of optimization for GPR noise. The histogram shows frequency of errors between target and maximum fidelity observations i.e. frequency of $\tilde{\Delta}_{M,T,i}$ from (20). The mean square error of these $\tilde{\Delta}_{M,T,i}$ is used as a lower bound when optimizing the Gaussian process's noise variance. The initial noise variance for optimization, corresponding to the blue curve, is set to an upper bound approximated from differences in numerically solves of the PDE at a sequence of increasing fidelities, see (17). The noise distribution after optimization is the red curve, which is indistinguishable from the lower bound distribution (not plotted). Our use of GPR modeling necessitates the assumption of Gaussian noise which prohibits a better fit.

4.3. Three-dimensional experiments and analysis

To emphasize the generality of our method, we applied our algorithm to the Darcy flow problem to a three-dimensional domain. The experimental setup is the same as in the two-dimensional case, except now we set the subsurface to have a depth of 20 m while the injection, extraction, and critical locations are all set at a depth of 10 m. Also, the mesh grid for the finite volume solver had $(d_j + 1) \times (d_j + 1) \times 9$ mesh points in each dimension so the mesh width in each dimension at fidelity j is $1/d_j, 1/d_j, 1/8$. The CPU time required for each step of experimentation are given in Table 2. The condition number of the noisy circulant kernel matrix is 613.

Fig. 6 shows the convergence of telescoping sums used to derive an upper bound on the noise variance for the three dimensional subsurface. The coefficients of determination are 0.81 and 0.83 for the s and d trends respectively. These are lower than the 0.98 and 0.99 respective values for Darcy's problem with a two-dimensional subsurface, but still large enough to justify the linear fits.

Fig. 8 shows, for the three dimensional subsurface, the confidence in maintaining a low enough pressure at the critical location as a function of both pressure threshold and extraction rate. The pressure at the critical location is generally much lower in this three-dimensional setup than in the two-dimensional one. For instance, at an extraction rate of $0\,m^3/s$ the two-dimensional setup gives a confidence of around 25% that the pressure at the critical location is below $1\,mmH_2O$, while the three-dimensional setup has almost a 100% confidence for the same extraction rate and threshold. Again, the monotonicity in both extraction rate and pressure threshold computed from the surrogate match our physical intuition for this three-dimensional subsurface problem.

Finally, 9 shows the noise calibration process for the three dimensional subsurface, specifically the initial upper bound and final optimized bound. We again observe the heavier tails in the distribution of differences between target and maximum fidelities when compared to the assumed Gaussian distribution. There also appears to be a slight skew to the right in the distribution of these differences, indicating a potential bias in low-fidelity approximation. Again, we defer remedies to future work but discuss ideas in the next section.

5. Discussion and conclusions

We fit a GPR surrogate model to a subsurface pressure management problem with a random log-normal permeability field. We solve the Darcy single-phase steady-state equation that examines the long-term impact of the injection/extraction on the reservoir. We consider that the pressure at the critical location is influenced strongly by the random permeability field during injection/extraction. Our GPR model predicts the pressure at a fixed critical location in the subsurface from an extraction rate and (truncated) permeability field realization. After we train the GPR surrogate model offline, it is used online to quickly determine the smallest extraction rate required to preserve a pressure at the critical location below a threshold with high probability.

Two discretizations must be made to solve the problem. First, we truncate the Karhunen-Loève expansion of the log-permeability field to a finite sum. The random coefficients in this sum determining the (truncated) permeability field are inputs to the GPR model alongside the extraction rate. Second, the domain must be meshed in order to apply a finite volume method to solve the PDE numerically. Each of these discretizations induces a numerical error, and these errors are often neglected in subsurface flow problems. By contrast, our GPR modeling approach accounts for these errors in the uncertainty analysis.

Our novel contributions are as follows. First, we use ideas in multilevel Monte Carlo to derive an approximate upper bound on the root mean squared error between the discretized numerical solution and the analytic PDE solution. Then, this upper bound is used as an initial guess for the noise variance in our GPR model before hyperparameter optimization. A lower bound for optimization is also derived based on the differences in numerical PDE solves at our maximum tested fidelity and the target fidelity used to fit the GPR model. Finally, we use a quasi-random design of experiments and matching covariance kernel to accelerate GPR model fitting and hyperparameter optimization from the classic $\mathcal{O}(n^3)$ rate to $\mathcal{O}(n \log n)$.

These ideas enable error-aware GPR modeling that can scale to tens or even hundreds of thousands of observations for an accurate fit in high dimensions. Moreover, the GPR predictions come with a notion of uncertainty. In fact, the GPR surrogate is a distribution over possible functions mapping the extraction rate and permeability field to pressure at the critical location.

In conclusion, we would like to summarize our findings:

- The GPR fitting and optimization scales like $\mathcal{O}(n \log n)$ in the number of numerical PDE pressure solutions.
- The GPR model is error-aware by calibrating surrogate noise to numerical errors in solving the subsurface flow problem.

- The surrogate model quantifies the uncertainties in the predicted pressures by providing a probability distribution over a broad class of possible solutions.
- In addition to subsurface flow problems, our approach can be directly applied to a variety of other problems that consider PDEs with random coefficients.

There is a delicate trade-off to consider when selecting good values for s and d. Recall that s is the fidelity of the permeability field; specifically s is the number of terms to keep in the KL expansion and $1+s$ is the number of input dimensions to the GPR surrogate. Increasing s has the drawback of increasing the dimension of the GPR surrogate and making it harder to fit by the curse of dimensionality. So if s is increased one should also increase the number of numerical PDE solves to attain a surrogate of equal quality. Moreover, d determines the fidelity of the PDE solver; specifically, the mesh width is $1/d$ in each dimension for the finite volume method. Increasing d has the drawback of making each numerical PDE solve more expensive. Therefore, increasing s and d necessitates more numerical PDE solves, each of which is more expensive. The advantage of our method is the ability to accommodate smaller values for s and d by simply encoding a larger error into the model. Valuable future work should explore the trade-offs in increasing s and d while taking advantage of our error-aware modeling methodology.

Future work may also look into more accurate models for the noise. Fig. 9 suggest the true noise distributions have heavier tails than a Gaussian for these problems. While this is not immediately accommodated by our GPR modeling framework, alternative models and methods may utilize a more accurate distribution for the noise in order to attain a better fit.

Our primary focus here is solving problems in the context of pressure management to prevent overpressurization in the subsurface due to climate mitigation operations such as injecting wastewater or CO_2 sequestration. To allow for CO_2 sequestration applications, a more complex multiphase flow model would be needed, but the process for applying the GPR would remain the same. The key common ground is the existence of PDEs with random coefficients, which is very common in subsurface applications where the random coefficients are used to represent subsurface heterogeneity, e.g., in permeability fields in subsurface flow or velocity fields in seismic problems.

CRediT authorship contribution statement

Aleksei G. Sorokin: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Aleksandra Pachalieva:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **Daniel O'Malley:** Writing – review & editing, Writing – original draft, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization. **James M. Hyman:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **Fred J. Hickernell:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Methodology, Funding acquisition, Formal analysis, Conceptualization. **Nicolas W. Hengartner:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Aleksei Sorokin, Aleksandra Pachalieva, Nicolas Hengartner and James Hyman acknowledge the Center for Nonlinear Studies at Los Alamos National Laboratory. The work was supported by the U.S. Department of Energy, United States through the Los Alamos National Laboratory. 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). Fred Hickernell and Aleksei Sorokin acknowledge support from the U.S. National Science Foundation grant DMS-2316011. Nicolas Hengartner acknowledges support from LDRD, United States grant 20210043DR. The views expressed herein do not necessarily represent the views of the U.S. National Science Foundation, the U.S. Department of Energy, or the United States Government. Work by Daniel O'Malley was supported by the US Department of Energy Office of Science Energy Earthshot Initiative as part of the "Learning reduced models under extreme data conditions for design and rapid decision-making in complex systems" project under award number LANLE31D.

Data availability

No data was used for the research described in the article.

References

Al-Ghosoun, A., El Moçayd, N., Seaid, M., 2021. A surrogate model for efficient quantification of uncertainties in multilayer shallow water flows. Environ. Model. Softw. 144, 105176.

Alghosoun, A., Mocayd, N.E., Seaid, M., 2022. A nonintrusive reduced-order model for uncertainty quantification in numerical solution of one-dimensional free-surface water flows over stochastic beds. Int. J. Comput. Methods 19 (04), 2150073.

Bacon, D.H., Qafoku, N.P., Dai, Z., Keating, E.H., Brown, C.F., 2016. Modeling the impact of carbon dioxide leakage into an unconfined, oxidizing carbonate aquifer. Int. J. Greenh. Gas Control 44, 290–299.

Ben-Haim, Y., 2006. Info-Gap Decision Theory: Decisions Under Severe Uncertainty. Elsevier.

Benson, S.M., Cole, D.R., 2008. CO_2 sequestration in deep sedimentary formations. Elements 4 (5), 325–331.

Bielicki, J.M., Pollak, M.F., Deng, H., Wilson, E.J., Fitts, J.P., Peters, C.A., 2016. The leakage risk monetization model for geologic CO_2 storage. Environ. Sci. Technol. 50 (10), 4923–4931.

Birkholzer, J.T., Zhou, Q., 2009. Basin-scale hydrogeologic impacts of CO_2 storage: Capacity and regulatory implications. Int. J. Greenh. Gas Control 3 (6), 745–756.

Buscheck, T.A., Sun, Y., Hao, Y., Wolery, T.J., Bourcier, W., Tompson, A.F., Jones, E.D., Friedmann, S.J., Aines, R.D., 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.

Cai, S., Mao, Z., Wang, Z., Yin, M., Karniadakis, G.E., 2021. Physics-informed neural networks (PINNs) for fluid mechanics: A review. Acta Mech. Sin. 37 (12), 1727–1738.

Carey, J.W., Svec, R., Grigg, R., Zhang, J., Crow, W., 2010. Experimental investigation of wellbore integrity and CO_2–brine flow along the casing–cement microannulus. Int. J. Greenh. Gas Control 4 (2), 272–282.

Chang, H., Zhang, D., 2019. Machine learning subsurface flow equations from data. Comput. Geosci. 23, 895–910.

Chen, B., Harp, D.R., Lin, Y., Keating, E.H., Pawar, R.J., 2018. Geologic CO_2 sequestration monitoring design: A machine learning and uncertainty quantification based approach. Appl. Energy 225, 332–345.

Chen, B., Harp, D.R., Lu, Z., Pawar, R.J., 2020. Reducing uncertainty in geologic CO_2 sequestration risk assessment by assimilating monitoring data. Int. J. Greenh. Gas Control 94, 102926.

Chen, B., Pawar, R.J., 2019. Characterization of CO_2 storage and enhanced oil recovery in residual oil zones. Energy 183, 291–304.

Chu, A.K., Benson, S.M., Wen, G., 2022. Deep-learning-based flow prediction for CO_2 storage in shale–sandstone formations. Energies 16 (1), 246.

Cihan, A., Birkholzer, J.T., Bianchi, M., 2015. Optimal well placement and brine extraction for pressure management during CO_2 sequestration. Int. J. Greenh. Gas Control 42, 175–187.

Court, B., Elliot, T.R., Dammel, J., Buscheck, T.A., Rohmer, J., Celia, M.A., 2012. Promising synergies to address water, sequestration, legal, and public acceptance issues associated with large-scale implementation of CO_2 sequestration. Mitig. Adapt. Strateg. Glob. Chang. 17, 569–599.

Curry, T., Reiner, D.M., Ansolabehere, S., Herzog, H.J., 2005. How aware is the public of carbon capture and storage? Greenh. Gas Control Technol. 7, 1001–1009.
El Mocayd, N., Mohamed, M.S., Seaid, M., 2021. Non-intrusive polynomial chaos methods for uncertainty quantification in wave problems at high frequencies. J. Comput. Sci. 53, 101344.
Fubini, G., 1907. Sugli integrali multipli. Rend. Acc. Naz. Lincei 16, 608–614.
Gholami, R., Raza, A., Iglauer, S., 2021. Leakage risk assessment of a CO_2 storage site: A review. Earth-Sci. Rev. 223, 103849.
Giles, M.B., 2008. Multilevel Monte Carlo path simulation. Oper. Res. 56 (3), 607–617.
Grana, D., Mukerji, T., Dvorkin, J., Mavko, G., 2012. Stochastic inversion of facies from seismic data based on sequential simulations and probability perturbation method. Geophysics 77 (4), M53–M72.
Gray, R.M., et al., 2006. Toeplitz and circulant matrices: A review. Found. Trends Commun. Inf. Theory 2 (3), 155–239.
Gross, M.R., Hyman, J.D., Srinivasan, S., O'Malley, D., Karra, S., Mudunuru, M.K., Sweeney, M., Frash, L., Carey, B., Guthrie, G.D., et al., 2021. A physics-informed machine learning workflow to forecast production in a fractured marcellus shale reservoir. In: Unconventional Resources Technology Conference, 26–28 July 2021. Unconventional Resources Technology Conference (URTeC), pp. 3641–3648.
Harp, D.R., O'Malley, D., Yan, B., Pawar, R., 2021. On the feasibility of using physics-informed machine learning for underground reservoir pressure management. Expert Syst. Appl. 178, 115006.
Harp, D.R., Pawar, R., Carey, J.W., Gable, C.W., 2016. Reduced order models of transient CO_2 and brine leakage along abandoned wellbores from geologic carbon sequestration reservoirs. Int. J. Greenh. Gas Control 45, 150–162.
Harp, D.R., Stauffer, P.H., O'Malley, D., Jiao, Z., Egenolf, E.P., Miller, T.A., Martinez, D., Hunter, K.A., Middleton, R.S., Bielicki, J.M., et al., 2017. Development of robust pressure management strategies for geologic CO_2 sequestration. Int. J. Greenh. Gas Control 64, 43–59.
He, Q., Barajas-Solano, D., Tartakovsky, G., Tartakovsky, A.M., 2020. Physics-informed neural networks for multiphysics data assimilation with application to subsurface transport. Adv. Water Resour. 141, 103610.
Holdaway, K.R., 2014. Harness Oil and Gas Big Data with Analytics: Optimize Exploration and Production with Data-Driven Models. John Wiley & Sons.
Huerta, N.J., Hesse, M.A., Bryant, S.L., Strazisar, B.R., Lopano, C.L., 2013. Experimental evidence for self-limiting reactive flow through a fractured cement core: Implications for time-dependent wellbore leakage. Environ. Sci. Technol. 47 (1), 269–275.
Jagadeeswaran, R., Hickernell, F.J., 2019. Fast automatic Bayesian cubature using lattice sampling. Stat. Comput. 29 (6), 1215–1229. http://dx.doi.org/10.1007/s11222-019-09895-9, URL http://dx.doi.org/10.1007/s11222-019-09895-9.
Jagadeeswaran, R., Hickernell, F.J., 2022. Fast automatic Bayesian Cubature using sobol' sampling. In: Advances in Modeling and Simulation: Festschrift for Pierre L'Ecuyer. Springer, pp. 301–318.
Jordan, A.B., Stauffer, P.H., Harp, D., Carey, J.W., Pawar, R.J., 2015. A response surface model to predict CO_2 and brine leakage along cemented wellbores. Int. J. Greenh. Gas Control 33, 27–39.
Kaarnioja, V., Kazashi, Y., Kuo, F.Y., Nobile, F., Sloan, I.H., 2022. Fast approximation by periodic kernel-based lattice-point interpolation with application in uncertainty quantification. Numer. Math. 1–45.
Kaarnioja, V., Kuo, F.Y., Sloan, I.H., 2023. Lattice-based kernel approximation and serendipitous weights for parametric PDEs in very high dimensions. arXiv preprint arXiv:2303.17755.
Karhunen, K., 1947. Under lineare methoden in der wahr scheinlichkeitsrechnung. Ann. Acad. Sci. Fennicae Ser. A1: Math. Phys. 47.
Keating, E.H., Fessenden, J., Kanjorski, N., Koning, D.J., Pawar, R., 2010. The impact of CO_2 on shallow groundwater chemistry: observations at a natural analog site and implications for carbon sequestration. Environ. Earth Sci. 60, 521–536.
Keating, E.H., Harp, D.H., Dai, Z., Pawar, R.J., 2016. Reduced order models for assessing CO_2 impacts in shallow unconfined aquifers. Int. J. Greenh. Gas Control 46, 187–196.
Keranen, K.M., Weingarten, M., Abers, G.A., Bekins, B.A., Ge, S., 2014. Sharp increase in central Oklahoma seismicity since 2008 induced by massive wastewater injection. Science 345 (6195), 448–451.
Kuo, F.Y., Mo, W., Nuyens, D., Sloan, I.H., Srikumar, A., 2023. Comparison of two search criteria for lattice-based kernel approximation. arXiv preprint arXiv:2304.01685.
Lackey, G., Vasylkivska, V.S., Huerta, N.J., King, S., Dilmore, R.M., 2019. Managing well leakage risks at a geologic carbon storage site with many wells. Int. J. Greenh. Gas Control 88, 182–194.
Lin, G.-Y., Chen, H.-W., Chen, B.-J., Yang, Y.-C., 2022. Characterization of temporal PM2.5, nitrate, and sulfate using deep learning techniques. Atmospheric Pollut. Res. 13 (1), 101260.
Little, M.G., Jackson, R.B., 2010. Potential impacts of leakage from deep CO_2 geosequestration on overlying freshwater aquifers. Environ. Sci. Technol. 44 (23), 9225–9232.
Liu, Y., 2017. Multilevel strategy for O-PCA-based history matching using mesh adaptive direct search (Ph.D. thesis). Stanford University Stanford, California.
Majer, E.L., Baria, R., Stark, M., Oates, S., Bommer, J., Smith, B., Asanuma, H., 2007. Induced seismicity associated with enhanced geothermal systems. Geothermics 36 (3), 185–222.
McNamara, D.E., Benz, H.M., Herrmann, R.B., Bergman, E.A., Earle, P., Holland, A., Baldwin, R., Gassner, A., 2015. Earthquake hypocenters and focal mechanisms in central Oklahoma reveal a complex system of reactivated subsurface strike-slip faulting. Geophys. Res. Lett. 42 (8), 2742–2749.
Mehana, M., Chen, B., Pawar, R., 2022. Reduced-order models for wellbore leakage from depleted reservoirs. In: SPE/AAPG/SEG Unconventional Resources Technology Conference. URTEC, D031S053R003.
Middleton, R.S., Keating, G.N., Stauffer, P.H., Jordan, A.B., Viswanathan, H.S., Kang, Q.J., Carey, J.W., Mulkey, M.L., Sullivan, E.J., Chu, S.P., et al., 2012. The cross-scale science of CO_2 capture and storage: from pore scale to regional scale. Energy Environ. Sci. 5 (6), 7328–7345.
Miller, E., Bell, L., Buys, E., 2007. Public understanding of carbon sequestration in Australia: socio-demographic predictors of knowledge, engagement and trust. Int. J. Emerg. Technol. Soc. 5 (1), 15–33.
Mishra, S., Schuetter, J., Datta-Gupta, A., Bromhal, G., 2021. Robust data-driven machine-learning models for subsurface applications: are we there yet? J. Pet. Technol. 73 (03), 25–30.
Misra, S., Li, H., He, J., 2019. Machine Learning for Subsurface Characterization. Gulf Professional Publishing.
Mohaghegh, S.D., 2017, Data-Driven Reservoir Modeling, Society of Petroleum Engineers, https://doi.org/10.2118/9781613995600.
Mudunuru, M.K., O'Malley, D., Srinivasan, S., Hyman, J.D., Sweeney, M.R., Frash, L., Carey, B., Gross, M.R., Welch, N.J., Karra, S., et al., 2020. Physics-informed machine learning for real-time unconventional reservoir management. In: CEUR Workshop Proceedings. pp. 1–10.
Navarre-Sitchler, A.K., Maxwell, R.M., Siirila, E.R., Hammond, G.E., Lichtner, P.C., 2013. Elucidating geochemical response of shallow heterogeneous aquifers to CO_2 leakage using high-performance computing: Implications for monitoring of CO_2 sequestration. Adv. Water Resour. 53, 45–55.
Niederreiter, H., 1992. Random Number Generation and Quasi-Monte Carlo Methods. SIAM.
Nordbotten, J.M., Kavetski, D., Celia, M.A., Bachu, S., 2009. Model for CO_2 leakage including multiple geological layers and multiple leaky wells. Environ. Sci. Technol. 43 (3), 743–749.
O'Malley, D., 2023. DPFEHM.jl: A differentiable subsurface physics simulator. URL https://github.com/OrchardLANL/DPFEHM.jl.
O'Malley, D., Vesselinov, V., 2015. Bayesian-information-gap decision theory with an application to CO_2 sequestration. Water Resour. Res. 51 (9), 7080–7089.
Owen, A.B., 2013. Monte Carlo Theory, Methods and Examples. URL https://artowen.su.domains/mc/.
Pachalieva, A., O'Malley, D., Harp, D.R., Viswanathan, H., 2022. Physics-informed machine learning with differentiable programming for heterogeneous underground reservoir pressure management. Sci. Rep. 12 (1), 18734.
Palmgren, C.R., Morgan, M.G., Bruine de Bruin, W., Keith, D.W., 2004. Initial Public Perceptions of Deep Geological and Oceanic Disposal of Carbon Dioxide. ACS Publications.
Pan, B., Anderson, G.J., Goncalves, A., Lucas, D.D., Bonfils, C.J., Lee, J., 2022. Improving seasonal forecast using probabilistic deep learning. J. Adv. Modelling Earth Syst. 14 (3), e2021MS002766.
Pruess, K., 2008. On CO_2 fluid flow and heat transfer behavior in the subsurface, following leakage from a geologic storage reservoir. Environ. Geol. 54, 1677–1686.
Rabczuk, T., Guo, H., Zhuang, X., Chen, P., Alajlan, N., 2022. Stochastic deep collocation method based on neural architecture search and transfer learning for heterogeneous porous media.
Rathinavel, J., 2019. Fast Automatic Bayesian Cubature Using Matching Kernels and Designs. Illinois Institute of Technology.
Robbe, P., 2023. GaussianRandomFields.jl. URL https://github.com/PieterjanRobbe/GaussianRandomFields.jl.
Robbe, P., Nuyens, D., Vandewalle, S., 2017. A multi-index quasi–Monte Carlo algorithm for lognormal diffusion problems. SIAM J. Sci. Comput. 39 (5), S851–S872.
Schuetter, J., Mishra, S., Zhong, M., LaFollette, R., 2018. A data-analytics tutorial: Building predictive models for oil production in an unconventional shale reservoir. SPE J. 23 (04), 1075–1089.
Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., Woo, W.-C., 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. Adv. Neural Inf. Process. Syst. 28.
Shi, C., Wang, Y., 2022. Data-driven construction of three-dimensional subsurface geological models from limited site-specific boreholes and prior geological knowledge for underground digital twin. Tunn. Undergr. Space Technol. 126, 104493.
Shokouhi, P., Kumar, V., Prathipati, S., Hosseini, S.A., Giles, C.L., Kifer, D., 2021. Physics-informed deep learning for prediction of CO_2 storage site response. J. Contam. Hydrol. 241, 103835.
Sorokin, A.G., 2023. QMCGenerators.jl. URL https://github.com/alegresor/QMCGenerators.jl.
Stauffer, P.H., Keating, G.N., Middleton, R.S., Viswanathan, H.S., Berchtold, K.A., Singh, R.P., Pawar, R.J., Mancino, A., 2011. Greening Coal: Breakthroughs and Challenges in Carbon Capture and Storage. ACS Publications.
Tang, M., Liu, Y., Durlofsky, L.J., 2020. A deep-learning-based surrogate model for data assimilation in dynamic subsurface flow problems. J. Comput. Phys. 413, 109456.

Tartakovsky, A.M., Marrero, C.O., Perdikaris, P., Tartakovsky, G.D., Barajas-Solano, D., 2020. Physics-informed deep neural networks for learning parameters and constitutive relationships in subsurface flow problems. Water Resour. Res. 56 (5), e2019WR026731.

Tcvetkov, P., Cherepovitsyn, A., Fedoseev, S., 2019. Public perception of carbon capture and storage: A state-of-the-art overview. Heliyon 5 (12).

Trautz, R.C., Pugh, J.D., Varadharajan, C., Zheng, L., Bianchi, M., Nico, P.S., Spycher, N.F., Newell, D.L., Esposito, R.A., Wu, Y., et al., 2013. Effect of dissolved CO_2 on a shallow groundwater system: a controlled release field experiment. Environ. Sci. Technol. 47 (1), 298–305.

Varadharajan, C., Appling, A.P., Arora, B., Christianson, D.S., Hendrix, V.C., Kumar, V., Lima, A.R., Müller, J., Oliver, S., Ombadi, M., et al., 2022. Can machine learning accelerate process understanding and decision-relevant predictions of river water quality? Hydrol. Process. 36 (4), e14565.

Vasylkivska, V., Dilmore, R., Lackey, G., Zhang, Y., King, S., Bacon, D., Chen, B., Mansoor, K., Harp, D., 2021. NRAP-open-IAM: A flexible open-source integrated-assessment-model for geologic carbon storage risk assessment and management. Environ. Model. Softw. 143, 105114.

Viswanathan, H.S., Pawar, R.J., Stauffer, P.H., Kaszuba, J.P., Carey, J.W., Olsen, S.C., Keating, G.N., Kavetski, D., Guthrie, G.D., 2008. Development of a hybrid process and system model for the assessment of wellbore leakage at a geologic CO_2 sequestration site. Environ. Sci. Technol. 42 (19), 7280–7286.

Wang, Z., Wang, H., Yan, L., Chi, F., 2023. Random-field generation method based on discrete cosine transform and application to landslide analysis. Eur. J. Environ. Civ. Eng. 27 (7), 2435–2446.

Watson, T.L., Bachu, S., 2009. Evaluation of the potential for gas and CO_2 leakage along wellbores. SPE Drill. & Complet. 24 (01), 115–126.

Whitmarsh, L., Xenias, D., Jones, C.R., 2019. Framing effects on public support for carbon capture and storage. Palgrave Commun. 5 (1).

Williams, C.K., Rasmussen, C.E., 2006. Gaussian Processes for Machine Learning, vol. 2, (3), MIT press Cambridge, MA.

Wilson, E.J., Morgan, M.G., Apt, J., Bonner, M., Bunting, C., Gode, J., Haszeldine, R.S., Jaeger, C.C., Keith, D.W., McCoy, S.T., et al., 2008. Regulating the Geological Sequestration of CO_2. ACS Publications.

Wu, H., Qiao, R., 2021. Physics-constrained deep learning for data assimilation of subsurface transport. Energy and AI 3, 100044.

Xiao, T., McPherson, B., Esser, R., Jia, W., Dai, Z., Chu, S., Pan, F., Viswanathan, H., 2020. Chemical impacts of potential CO_2 and brine leakage on groundwater quality with quantitative risk assessment: A case study of the farnsworth unit. Energies 13 (24), 6574.

Yonkofski, C., Tartakovsky, G., Huerta, N., Wentworth, A., 2019. Risk-based monitoring designs for detecting CO_2 leakage through abandoned wellbores: An application of NRAP's WLAT and DREAM tools. Int. J. Greenh. Gas Control 91, 102807.

Zhang, Z., Yan, X., Liu, P., Zhang, K., Han, R., Wang, S., 2023. A physics-informed convolutional neural network for the simulation and prediction of two-phase Darcy flows in heterogeneous porous media. J. Comput. Phys. 477, 111919.

Zoback, M.D., 2012. Managing the seismic risk posed by wastewater disposal. Earth 57 (4), 38–43.