GEM - International Journal on Geomathematics (2023) 14:30
https://doi.org/10.1007/s13137-023-00240-x

ORIGINAL PAPER

Solute transport prediction in heterogeneous porous media using random walks and machine learning

Lazaro J. Perez[1] · George Bebis[2] · Sean A. McKenna[1] · Rishi Parashar[1]

Received: 20 May 2022 / Accepted: 22 September 2023 / Published online: 27 October 2023

Abstract

Solute transport processes in heterogeneous porous media have been traditionally studied through the parameterization of macroscale properties using upscaling approaches over a representative elementary volume. As a result, our ability to accurately model solute transport at fine-scale is limited. Combining multiple transport and geometrical observations from the pore scale in a multiphysics framework can enhance the understanding of transport mechanisms that manifest at larger scales. In this paper, we predict conservative solute transport in three sandstone geometries (Castlegate, Bentheimer, and sandpack) that range across different degrees of heterogeneity using a machine learning approach. Our approach, which is based on the random forests (RF) algorithm, performs simulated transport predictions such as solute breakthrough curves. The RF algorithm used in our workflow is a tree-based ensemble method, which builds several different decision tree models independently and then computes a final prediction by combining the outputs of the individual trees. We employ observations, such as solute arrival time and distance traveled, as input to train the predictive model using random walk particle tracking (RWPT) simulations in the sandstones. We employ Bayesian optimization techniques to select the hyperparameter values controlling the structure of the RF model in order to avoid overfitting. Results of our workflow show accurate RF predictions of the RWPT breakthrough curves demonstrating the ability of the RF algorithm to capture the critical flow and transport properties of porous media. We also examine the sensitivity to geometrical sample effects in the training data, which can impact machine learning predictions. The RF algorithm used is able to provide accurate results in real rock samples spanning from unconsolidated granular to consolidated media, highlighting the ability of the model to generalize solute transport problems in porous media.

✉ Lazaro J. Perez
lazaro.perez@dri.edu

[1] Division of Hydrologic Sciences, Desert Research Institute, 2215 Raggio Pkwy, Reno, NV 89512, USA

[2] Department of Computer Science and Engineering, University of Nevada, Reno, 1664 N. Virginia Street, Reno, NV 89557, USA

Keywords Machine learning · Random forest algorithm · Random walks model · Solute transport · Anomalous transport · Heterogeneous porous media

Mathematics Subject Classification 76Rxx Diffusion and convection · 76Sxx Flows in porous media; filtration; seepage · 68Txx Artificial intelligence

1 Introduction

Understanding solute transport phenomena in sedimentary rocks is key in a range of scientific and engineering applications, such as groundwater management (Al-Salamah et al. 2011; Swanson et al. 2015), contaminant transport (Vesper 2019; Brusseau et al. 2020; Guo et al. 2020a), oil recovery and soil carbon storage (Popova et al. 2012; Poffenbarger et al. 2023). The heterogeneous porous structure of geological media, ranging from pore to field scales, leads to anomalous solute transport that cannot be adequately described by an effective advection-dispersion equation. Such anomalous solute transport is ubiquitous in hydrological settings and has been observed in sandstone aquifers (Edmunds and Smedley 2000; Cortis and Berkowitz 2004) and fractured porous media environments (Haggerty et al. 2000; Aquino et al. 2015). Solute transport typically consists of a broad range of behaviors across spatial and temporal scales that must be incorporated in any modeling framework to successfully reproduce transport characteristics.

Various modeling methodologies could, in principle, be employed to quantify anomalous transport in porous media. Some of the most widely used include multirate mass transfer (Guo et al. 2020b), continuous time random walks (CTRW) (Kim and Kang 2020; Engdahl and Aquino 2022; Ben-Noah et al. 2023), and fractional advection dispersion equations (fADE) (Qiao et al. 2020; Sharma et al. 2022). Despite their effectiveness in reproducing actual observations in diverse hydrological scenarios, these models have their limitations. CTRW models, for instance, typically depend on fitting parameters that do not correspond to the physical characteristics of the system (Bolster et al. 2019; Kurotori et al. 2020; Gouze et al. 2023). On the other hand, fADE models can be resource-intensive computationally when used to forecast solute transport in intricate environments (Sun et al. 2020).

Recently, machine learning methods have been introduced as a modeling framework for learning from observational data of physical phenomena and predict variables of interest. These novel methods benefit from benchmark datasets and the capabilities of surrogate models that serve as effective approximations for complex problems (Schilders et al. 2008; De Lucia et al. 2017; Tang et al. 2020). Various machine learning techniques (e.g., ensemble methods, kernel-based methods, neural networks, etc.) have been successfully applied to reactive transport applications such as fluid mixing estimation (Ahmad et al. 2019; Ahmmed et al. 2021; Li et al. 2021), chemical equilibria computations (Leal et al. 2020; Li et al. 2021), and fluid flow and solute dispersion prediction (Santos et al. 2020; Kamrava et al. 2020; He et al. 2020; He and Tartakovsky 2021; Kamrava et al. 2021). These approaches substantially improve our capabilities to develop fast, accurate, and robust predictions of contaminant fate and transport under natural conditions.

Among various machine learning methods, the random forest (RF) algorithm (Breiman 2001; Breiman and Cutler 2004) has shown promising results for classification and regression problems. The RF algorithm is a tree-based ensemble method that builds several different decision tree models independently and then computes a final prediction by combining the outputs of the individual trees (Breiman et al. 2017). RF presents several advantages over other machine learning techniques as it is highly efficient with large datasets, is less sensitive to noise or over-fitting (Zhou et al. 2016; Hong and Lynn 2020), and employs fewer parameters compared to neural networks or support vector machines (Lee et al. 2005). The RF classification framework has been shown to successfully describe particle trajectory characteristics (Kowalek et al. 2019; Muñoz-Gil et al. 2020), soil and rock physical properties (Al-Farisi et al. 2019; Zhang and Cai 2021), and groundwater pollution and water quality (Rodriguez-Galiano et al. 2014; Singh et al. 2017; Naghibi et al. 2017). To the best of our knowledge, only a few studies have reported the use of the RF regression framework for modeling flow and transport processes in heterogeneous porous media applications (Wang et al. 2015; Shiri 2018; Lange and Sippel 2020). While these approaches provide dynamic frameworks to model large-scale transport features, their tree parameterization, selection of transport-independent parameters, and predictive power are still open questions.

In this work, spatial and temporal behaviors of solute particle transport in two natural sandstones and a sandpack are investigated numerically. The aim of the study is to provide accurate and less computationally expensive simulated transport predictions such as solute breakthrough curves (BTC) using a random forest algorithm. To this end, we follow a workflow that uses direct measurements from particle tracking models in synthetic and real rock geometries to accurately train the machine learning algorithm and predict the transport features in the Bentheimer (one of the sandstone) geometry. The methodology chosen benefits from working with large datasets, being computationally efficient, and having an automated optimized parameter selection. Our approach provides the flexibility to extend the study by assimilating other types of variables and physical laws to predict transport features in complex engineered and natural porous media.

This paper is organized as follows. Section 2 describes the machine learning algorithm used, the training and test data, the simulation framework for conservative transport in the 2D geometries studied, and feature extraction. The performance and transport prediction of the approach used, including estimation errors and limitations, are given in Sect. 3. Concluding remarks and future directions of our work are given in Sect. 4.

2 Methodology

In this section, we present the Machine Learning (ML) approach adopted in our work which is based on observed transport features and its application to subsurface solute transport. The ML approach is a regression model based on random forests (RFs) (Breiman 2001) that predicts individual particle transport time from a feature set calculated using geometric and statistical attributes of individual particle trajectories.

We also present the numerical methods for simulating the flow and transport physics to generate the training data.

2.1 Random forest regression

In this work, we use RFs for regression to predict particle transport time (i.e., a quantitative outcome and the dependent variable) from statistical and geometric features (i.e., predictors/independent variables), which are computed from observed particle trajectories. RFs is a powerful, state-of-the-art technique for both classification and regression problems, usually outperforming more sophisticated models as shown in a thorough comparison study (Breiman et al. 2017). They represent an extension of single classification and regression trees (CART) (Breiman et al. 1984) coupled with an effective methodology for building CART ensembles of high variance (i.e., minimize the correlation among CART members in the ensemble). To introduce RFs, we first introduce CART followed by bagging (Breiman 1996) that is used to build CART ensembles, a simpler version of an RF.

Traditional regression techniques, such as logistic and linear regression, rely on a mathematical formula for data classification or regression. CART, on the other hand, does not develop a predictor equation. Instead, it develops a decision tree by iteratively partitioning the data along the predictor axes into subsets. The decision tree is a set of conditions or restrictions hierarchically organized and successively applied to the predictors from the root to any leaf of the tree where each tree node represents a subset of the data made as homogeneous as possible with respect to the dependent variable. This is typically performed by minimizing the weighted average of mean square errors over the subsets: $MSE = \sum_i (n_i/n) MSE_i$ where n is the size of the data set, n_i is the size of the i-th subset, $MSE_i = (1/n_i) \sum_j (\hat{y}_i - y_j)^2$ is the mean square error of the i-th subset, $\hat{y}_i$ is the average of the $i - th$ subset, and y_j is the j-th sample of the i-th subset. The model prediction is the average value $\hat{y}_i$ of the dependent variable in each subset.

To build an ensemble of regression models it is necessary to resample the data, both predictors and target variable, multiple times. Bagging builds an ensemble of regression models using bootstrap sampling (Breiman 1996). Assuming n samples in the data set, bootstrap samples of size n are generated from the original data through sampling with replacement and used to create a regression tree for each bootstrap sample. The individual bootstrap predictions can then be combined into a single prediction, for example, by averaging the outputs of the decision trees in the ensemble. The optimal number of bootstrap samples is problem dependent.

RFs are an extension of CART bagging to address the issue of highly correlated bootstrap samples that can lead to similarity between regression trees and reduce the mitigating effect of bagging. To improve the variance of the regression trees in the ensemble, RFs split the data at each node of a regression tree using a subset of the predictors only. Assuming $\mathcal{P}$ predictors, a subset p is randomly chosen to split the data at a given regression tree node. This random selection of p predictors reduces the similarity of the regression trees built even when the bootstrap samples are similar. In short, RF is an ensemble-based approach that combines a large set of relatively

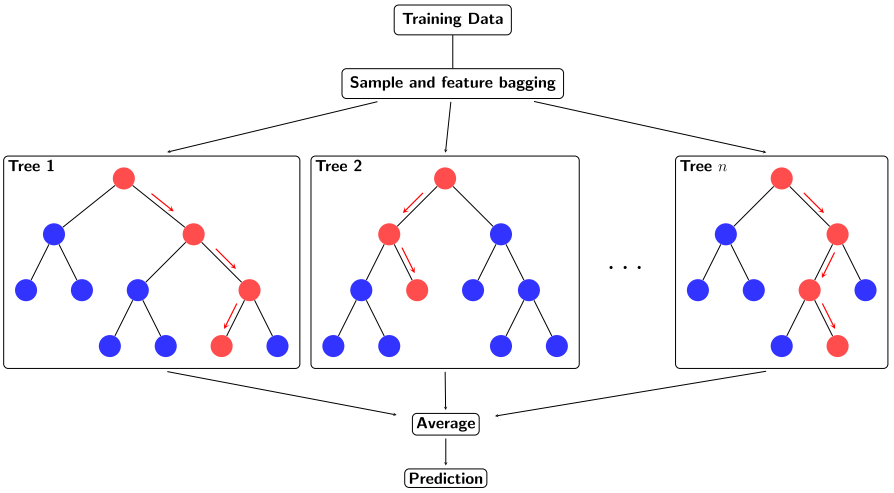

Fig. 1 Illustration of the Random Forest algorithm. After sample and feature bagging, the algorithm grows a forest of n trees. At each node, the algorithm selects m variables randomly out of $\mathcal{M}$ possible variables and finds the best split for the selected m variables. The trees are then grown to a maximum depth followed by averaging of the trees to obtain new predictions

uncorrelated regression trees to reduce over-fitting and improve predictions. In this work, we used Matlab's function *fitrensemble.m* where the number of randomly chosen predictors p for each split is equal to one third of the original number of predictors P.

A common challenge with the application of RFs is determining the appropriate number of decision trees and their size in the ensemble. We consider the minimum leaf size ($\mathcal{L}_s$), which controls the depth/size of the tree, and number of decision trees ($\mathcal{M}$) as unknown hyperparameters and couple Bayesian optimization with a surrogate function to select their optimum values for robust predictions (Snoek et al. 2012; Wu et al. 2019). Here, $h(x)$ represents an unknown function that characterizes the performance of the model given the hyperparameters x. Optimizing $h(x)$ requires evaluating $h(x)$ over a large number of hyperparameter values which can be computationally expensive. In the case of RFs, one would need to create many different ensembles, each with different hyperparameter values, test each one of them, and choose the ensemble that performs best.

The Bayesian optimization algorithm, used attempts to minimize the scalar objective function $h(x)$, is described in detail in Snoek et al. (2012). The underlying probabilistic model for the objective function $h(x)$ is a Gaussian process (GP). Using Bayesian inference, the GP is updated iteratively with new sample points to more closely approximate $h(x)$. Through this process, we are able to find the best combination for hyperparameters, that is minimum leaf size ($\mathcal{L}_s$) and number of learners ($\mathcal{P}$) determines the minimum mean squared error using the training data (Fig. 1).

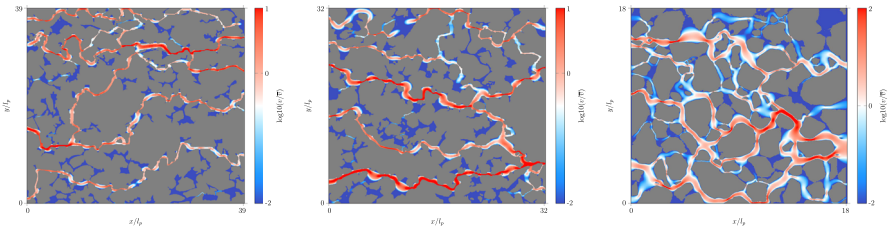

Fig. 2 Pore geometry and flow velocity field for the Castlegate (**a**), Bentheimer (**b**), and Sandpack (**c**). Warmer colors correspond to higher velocities, and solid grains are shown in gray

2.2 Training and test datasets

Three different porous media geometries are used in our numerical experiments where the RF model is trained on two of the geometries and then tested on the third one. Therefore, we perform three different tests, one for each geometry. The geometries include two natural sandstones: Bentheimer and Castlegate and an engineered sandpack. The datasets used to train the RF model are built from random walk particle simulations for each geometry.

2.2.1 Geometries

The natural sandstones studied in this work are from the Castlegate Formation (Cretaceous Mesa Verde Group, Utah) and from the Bentheimer shallow marine formation deposited during the Lower Cretaceous in the Netherlands and Germany. The first is commonly used in experimental rock mechanics studies as an analogue reservoir rock (DiGiovanni et al. 2000), while the latter is considered to be an ideal sedimentary rock for reservoir studies due to its lateral continuity and homogeneous block-scale nature (Peksa et al. 2015).

Both Castlegate and Bentheimer geometries were obtained from 3D samples archived at the digital rock portal (https://www.digital-rocksportal.org/), from which a slice from each 3D sample was selected and processed to increase its original grid resolution 1002×1000 by a factor of two. The Castlegate sandstone, with porosity $\phi = 0.2$ and average pore length $l_p = 5.77 \times 10^{-5}$m, is discretized in a regular grid that consists of 2004×2000 pixels in the x and y dimensions, respectively. Each pixel is a square with a size of 1.125 μm. The Bentheimer sandstone, with dimensions $x \times y = 0.00225 \times 0.00225$ m and a pixel size of 1.125 μm ($\Delta x = \Delta y$), shows a porosity of $\phi = 0.23$ and $l_p = 6.95 \times 10^{-5}$ m. Lastly, the constructed sandpack is a close packing of irregular quartz grains of different size that aims to replicate aquifer material (e.g. alluvial (Di Palma et al. 2019)). The discretization level selected for the sandpack sandstone is similar to the natural sandstones with $\phi = 0.37$ and $l_p = 1.2 \times 10^{-4}$ m. The geometries resulting from the three media are illustrated together with the flow field in Fig. 2. All media differ in the distributions of pore sizes and connectivity, which leads to different degree of flow heterogeneity as quantified by the variance σ_v^2 of the logarithm of the flow speed $v = \log_{10}(\mathbf{v}(\mathbf{x}))$. The variance

of the log-speed for the Castlegate geometry is $\sigma_v^2 = 6.71$, while for Bentheimer and Sandpack σ_v^2 equals 4.75 and 2.17, respectively. The significant high value of the variance of the flow speed for the Castlegate sandstone reflects higher flow heterogeneity in comparison to Bentheimer (the middle case) and Sandpack (the less heterogeneous case).

2.2.2 Flow

In order to simulate solute transport for each geometry, we first compute the 2-D flow field by solving the Navier-Stokes equations:

$$\nabla \cdot \mathbf{v}(\mathbf{x}) = 0, \tag{1a}$$

$$\nabla^2 \mathbf{v}(\mathbf{x}) = -\frac{1}{\mu} \nabla P(\mathbf{x}), \tag{1b}$$

where $\mathbf{v}$ is the velocity vector (m/s), P (kg/m/s^2) is the fluid pressure, and the viscosity $\mu = 10^{-3}$ Pa s. The flow field is solved using the steady-state solver for incompressible flow simpleFOAM that belongs to the open-source code OpenFOAM (Weller et al. 1998), which uses a semi-implicit method for pressure linked equations (Icardi et al. 2014). We apply a constant pressure boundary condition at the inlet and outlet faces of the image. On other solid faces, including the void-rock interface, we apply no-slip boundary conditions. After convergence, that is once the residual of the pressure and flow fields between two consecutive numerical iterations are smaller than a user defined criterion, $\epsilon = 10^{-9}$, we extract the complete velocity field. The Reynolds number used in simulations ranges from 2.7×10^{-4} in Castlegate to 10^{-3} in Sandpack.

2.2.3 Transport

The transport problem can be formulated in a Lagrangian modeling framework based on the equivalence of the advection–diffusion equation

$$\frac{\partial c(\mathbf{x}, t)}{\partial t} + \mathbf{v}(\mathbf{x}) \cdot \nabla c(\mathbf{x}, t) - D \nabla^2 c(\mathbf{x}, t) = 0 \tag{2}$$

where D is the molecular diffusion coefficient, $\mathbf{v}(\mathbf{x})$ is the flow velocity, and $c(\mathbf{x}, t)$ is the concentration of a scalar particle, with the Langevin equations (Perez et al. 2019; Yoon and Kang 2021)

$$\mathbf{x}(t + \Delta t) = \mathbf{x}(t) + \mathbf{v}\left[\mathbf{x}(t)\right] \Delta t + \sqrt{2 D_m \Delta t}\, \xi(t), \tag{3}$$

where $\mathbf{x}(t)$ is the position of the particles at time t, $D_m = 3.5 \times 10^{-10}$ m^2/s is the molecular diffusion coefficient used in all geometries, and $\xi(t)$ are independently distributed Gaussian random variables with $\mathbf{0}$ mean and unit variance. The advective step during a time interval $\Delta t = 0.05$ s, requires the interpolation of the flow velocities that are defined at the faces of the finite voxels. We use a quadratic velocity interpolation (Mostaghimi et al. 2012; Puyguiraud et al. 2019; Perez et al. 2021b) as this

approximation respects the no-slip boundary condition at the void-solid interface in contrast to the linear interpolation that has been traditionally used in particle tracking models (Pollock 1988).

We consider a solute line pulse injection perpendicular to the mean flow direction with initial particle positions assigned using a flux-weighted approach. Particle transport in the simulations is dominated by advection defined by Peclet dimensionless number $Pe = \bar{v} l_p / 2D_m = 100$, where $\bar{v}$ is the mean flow velocity. The mean flow velocity ranges from 4.7×10^{-6} m/s in the Castlegate to 2.56×10^{-4} m/s in the Sandpack. Particle trajectories are simulated until they exit the medium.

2.3 Feature extraction

In this section we describe the features used for training and testing the RF regression algorithm. There are $N_0 = 7.5 \times 10^5$ particles simulated through each geometry. For each of the N_0 particles, a total of five features, described in the subsections below, and one target variable, the travel time, are extracted.

2.3.1 Cumulative sum of displacements

The path of each particle is divided into N segments. The distance traveled within each segment is determined by the local fluid velocity and the time step length. We use the total distance traveled by an individual particle in the geometries,

$$\kappa = \sum_{i=1}^{N-1} \mid \mathbf{x}_{i\Delta t} - \mathbf{x}_{(i-1)\Delta t} \mid, \tag{4}$$

as a feature that provides an indirect informative observable about the time of arrival of a particle to the outlet. Figure 3 shows the bivariate distribution of κ / l_p against time t normalized by the characteristic time $\tau_c = l_p / \bar{v}$. Larger values of κ / l_p for a particle generally translate to late arrival time, which have been related to anomalous transport features (Comolli et al. 2019; Hidalgo et al. 2021; Perez et al. 2021a).

2.3.2 Variance

The variance of the displacements of a particle relate to their individual dispersion and their spreading in the geometries (Perez et al. 2019; Puyguiraud et al. 2020), which provide information about their path length,

$$\sigma^2 = \langle x^2 \rangle - \langle x \rangle^2, \tag{5}$$

where the angular brackets denote the average position. Higher particle σ^2 reflects longer path lengths and late time arrivals, thus informing the learning algorithm about tailing effects on breakthrough curves.

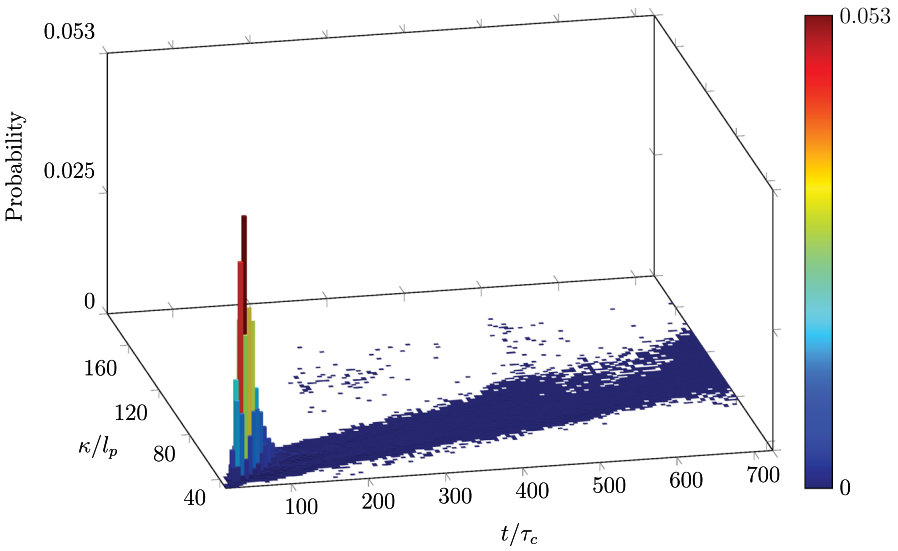

Fig. 3 Bivariate distribution of cumulative sum of displacements and particle arrival times in the Bentheimer sandstone. The colorbars shows the κ values. Particles that travel longer distances (high values of κ) take more time to arrive at the outlet

2.3.3 Straightness

Straightness is a measure of the average direction change between subsequent steps and it is essentially the inverse of tortuosity of the particle path (Sherman et al. 2020; Puyguiraud et al. 2021) defined as ratio between the net displacement of the particle from the start $\mathbf{x_0}$ to the end point $\mathbf{x_f}$, $\mid \mathbf{x_f} - \mathbf{x_0} \mid$, and the sum of step lengths as

$$S = \frac{\mid \mathbf{x_f} - \mathbf{x_0} \mid}{\sum_{i=1}^{N-1} \mid \mathbf{x}_i - \mathbf{x}_{i-1} \mid}. \tag{6}$$

Particles that show high S values provide transport information about tortuous paths and thus late time arrivals.

2.3.4 Mean velocity

The mean Lagrangian velocity across all segments along a particle path provides information for fast particles trajectories and early time of arrivals,

$$\bar{v} = \frac{1}{N} \sum_{i=1}^{N} v_i, \tag{7}$$

where v_i is the particle's velocity.

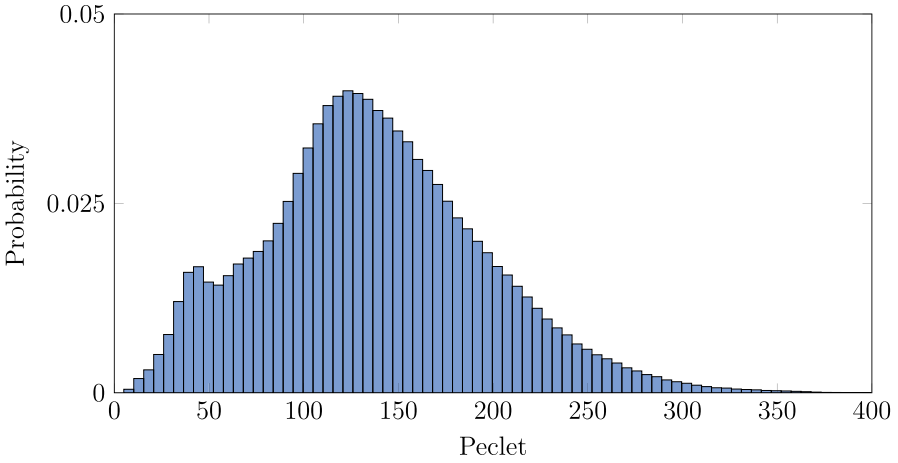

Fig. 4 Particle Peclet distribution in the Bentheimer sandstone shows a broad Pe distribution where fast particles which have higher Pe values arrive earlier than particles with lower Pe

2.3.5 Mean Peclet number

The mean Peclet number connected to the transport of an individual particle affects tailing behavior in case of low particle Peclet numbers and describes fast arriving particles in case of high Peclet numbers

$$Pe = \frac{\bar{v} l_p}{2D}, \tag{8}$$

where $\bar{v}$ is the mean particle velocity. The distribution of particle Pe for the Bentheimer sandstone is shown in Fig. 4, which shows a broad distribution of particle's Pe. Fast particles that will arrive earlier at the outlet have higher Pe values, while particles with lower Pe inform the RF algorithm about tailing.

3 Results

In this section, we first discuss the accuracy of the RF algorithm in the prediction of solute concentration breakthrough curves, and later we analyze the model's robustness and its sensitivity to data scarcity.

3.1 Random forest prediction

Simulation results from two of the geometries are used to train the RF model, next the trained random forests model is applied to the third geometry. The data structure used for the RF development is a matrix where each set of predictors and target variable represent a single row. As transport observations from two domains are used to predict results in the third domain, the input data matrix has $2N_0$ rows where each row has

Table 1 Best hyperparameters selected using Bayesian optimization that yielded the minimum mean squared error between estimated and observed data

Hyperparameters			
Geometry	Number of iterations	Minimum leaf size	Number of learners
Castlegate	30	5	40
Bentheimer	30	3	47
Sandpack	30	7	38

six columns: the five features and the single target variable. This data matrix provides the input to the Bayesian hyperparameter estimation to determine the structure of the RF model.

Table 1 shows the best hyperparameters selected for each geometry. The selected $\mathcal{L}_s$ and $\mathcal{P}$ for the Bentheimer are 3 & 47 respectively, while for the Castlegate and Sandpack geometries prediction, $\mathcal{L}_s$ and $\mathcal{P}$ values are 5 and 40, and 7 and 38.

We used the square root of the mean squared error (RMSE) as the accuracy measurement between RF predictions and RWPT simulations. The RMSE is the average squared difference between each true data point (y_i) and its corresponding predicted value ($\widehat{y_i}$), defined as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}. \tag{9}$$

Figure 5 shows the comparison between the prediction of the RF algorithm and the actual breakthrough curves from the RWPT numerical simulations for the three geometries. Results for the Castlegate (Fig. 5a) show a root mean squared error (RMSE) of 0.0781 between predicted and measured data. The random forest predicts early arrivals that do not match the observed RWPT simulation results. Later, at intermediate times, the RF prediction underestimates the RWPT concentration, while at late times the RF prediction matches RWPT results. Here, the proposed RF estimation suffers from the lower degree of heterogeneity inherent in the training data. Recall that the Castlegate prediction is carried out by training the random forest algorithm on the Bentheimer and Sandpack geometries. These geometries are less heterogeneous compared to the Castlegate geometry, which affects the predictions. We attribute this prediction mismatch to geometrical sample effects, which refer to training a model on one type of data and then testing it on another type of data that is structurally different. More specifically, this difference means that the training data may not adequately represent all the structural complexities present in the test data. This geometrical sample effect is responsible for the earlier and intermediate mismatch between the RF prediction and the simulation results.

In the Bentheimer geometry (Fig. 5b), RMSE = 0.0314, the RF algorithm matches the observed results from the numerical random walk simulations. The training data used by the random forest prediction in this case is composed of particle information from higher (Castlegate) and lower (Sandpack) degrees of heterogeneity. This

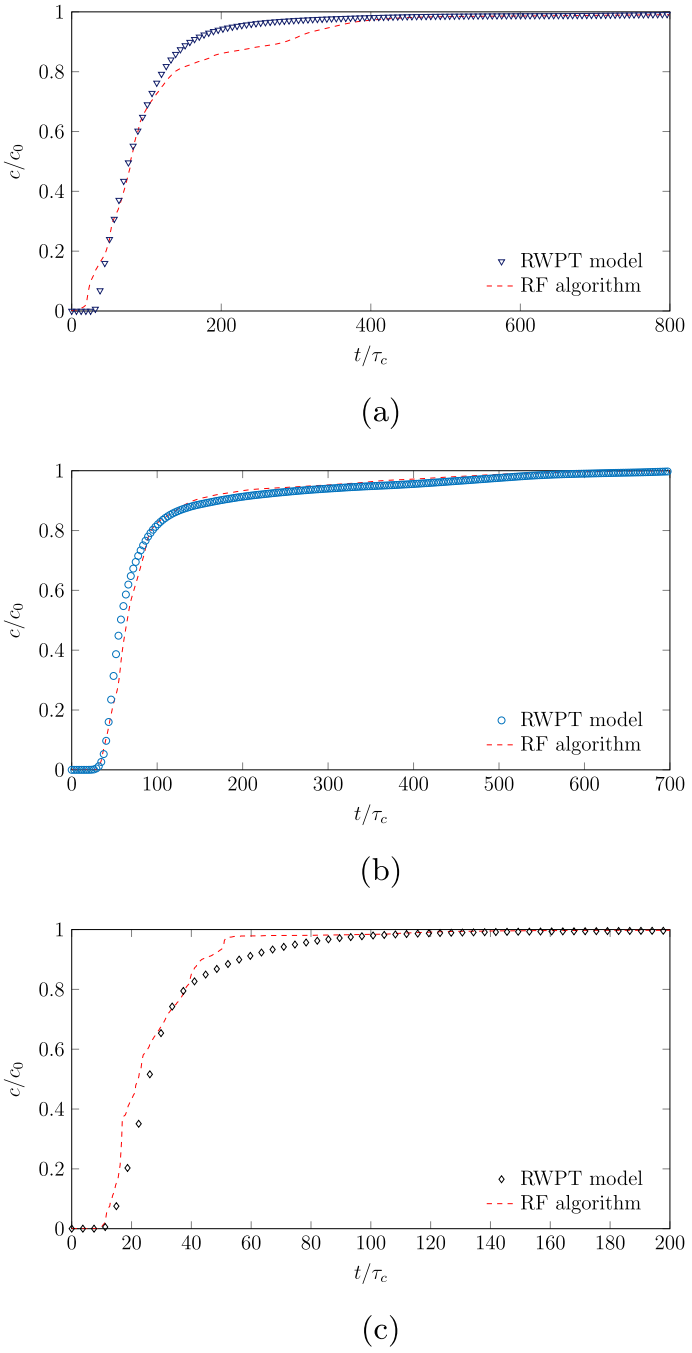

Fig. 5 Measured and predicted breakthrough curve from RWPT model (symbols) and random forest prediction (red dashed line) in the Castlegate (top), Bentheimer (middle), and Sandpack (bottom) geometries (color figure online)

decreases the accuracy of the prediction and result in early concentration overestimation and intermediate underestimation in the case of the Castlegate RF prediction. Figure 5c shows the machine learning prediction for the sandpack. A slight overprediction in concentration arrival is observed at early and intermediate times. The RF prediction reflects that geometrical sample effects play a role as the training data used by the machine learning algorithm comes from more heterogeneous geometries. Here the computed RMSE between predicted and measured data equals 0.0427. Our results suggest that solute transport predictions by machine learning algorithms matches observations when training data used comprises data from higher and lower heterogeneous cases. On the other hand, mismatch between learning algorithms and actual data results is observed when concentration estimations are made in highly heterogeneous samples while the learning algorithm is trained with data from lower heterogeneous samples, or vice-versa.

Features of anomalous transport are observed in all geometries. Stronger non-Fickian behaviors are confirmed in the case of the Castlegate geometry due to higher value of the variance σ_ν^2 of the logarithm $\nu = \log 10(\mathbf{v}(\mathbf{x}))$ of the flow velocities that reflects the presence of preferential flow paths and stagnant zones. For the Castlegate sandstone, $\sigma_\nu^2 = 3.7$, while for Bentheimer and sandpack geometries the σ_ν^2 values were 3.2 and 2.1 respectively.

3.2 Sensitivity to data scarcity

Overcoming data scarcity in machine learning approaches is critical when developing robust models. Additionally, in the case of surrogate models, such as random forests, data-driven optimization improves the speed and computational cost of the numerical workflow (Alizadeh et al. 2020). Here we show the efficiency of the learning algorithm by computing the root mean squared error (RMSE) between the solute breakthrough prediction using different proportions of data and the measured RWPT breakthrough using all data. Later, we show how models based on limited data compares to the RWPT and machine learning prediction using all data available.

Figure 6 shows the error estimation of the random forest trained with limited data. Higher RMSE values are observed in the Castlegate geometry (top), which is consistent with the geometrical sample observations discussed above. Since the RF algorithm is trained with data from less heterogeneous samples, RMSE values are higher as predictions are less able to match the RWPT results. The RF algorithm trained with data from less heterogeneous samples understandably are not able to fully capture the transport dynamics observed in samples with higher heterogeneity. In contrast, the lower RMSE values observed in Bentheimer prediction shows that the machine learning model prediction is accurate if training data cover a broad range of heterogeneities. Note that the RMSE values in all samples reaches a constant value at 0.7 fraction of data indicating that robust predictions can be achieved using this amount of data and thus reducing the computational cost associated with RWPT simulations.

Figure 7 shows the probability of the arrival time obtained from the RWPT simulations, and RF algorithm prediction using 100% and 10% of training data. Results show that the RF prediction using 100% of training data matches well the RWPT model

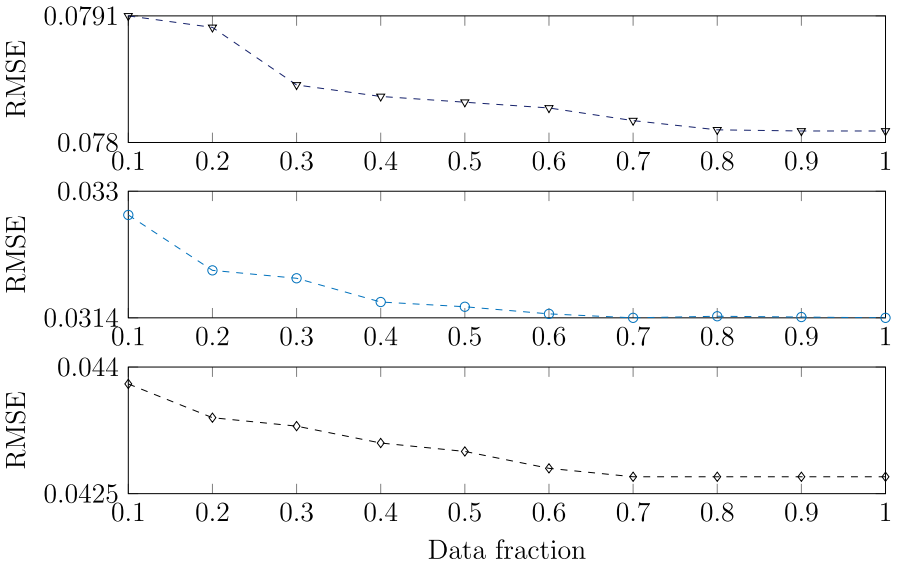

Fig. 6 Root mean squared error between the measured breakthrough curves from RWPT model and random forest prediction as a function of different fractions of training data in Castlegate (top), Bentheimer (middle), and sandpack (bottom) geometries

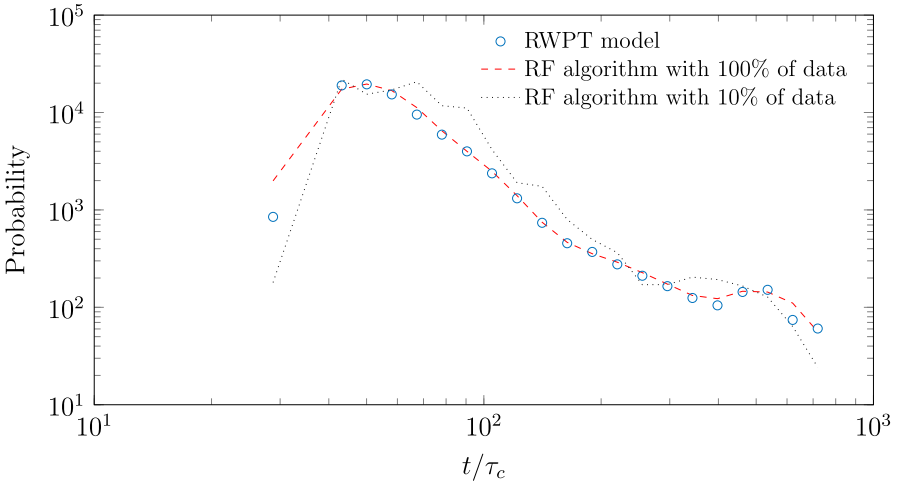

Fig. 7 Breakthrough curve from RWPT model (blue symbols) and random forest prediction using all available data (red dashed line) and limiting data to only 10% (black dotted line) in the Bentheimer geometry (color figure online)

Table 2 Comparison of key indicators of breakthrough time statistics between the RWPT predictions, and RF predictions using all available data (100%) and using limited data (10%) obtained for the Bentheimer geometry

Breakthrough time statistics			
Time metrics	RWPT	RF at 100% data	RF at 10% data
Peak arrival time	50.05	50.05 (9.9×10^{-6})	43.10 (0.1389)
25% breakthrough time	235.50	235.70 (8.5×10^{-4})	236.98 (0.0063)
50% breakthrough time	401.80	401.00 (0.002)	397.70 (0.01)
75% breakthrough time	562.70	561.55 (0.002)	551.80 (0.019)

The relative error for the RF prediction with respect to the RWPT simulation is shown in parenthesis

while noisy results are observed when the amount of training data is restricted to 10%. Nevertheless, RF prediction is robust and visually captures the BTC peak and tailing fairly well. To quantitatively evaluate the performance of RF models, and examine the effect on RF predictions in case of limited training data, we compare some key indicators of breakthrough time statistics obtained from RWPT simulations and RF models using 100% and 10% of training data. Table 2 shows that for the example case of Bentheimer geometry, the peak arrival time as well as the breakthrough time for various quantiles, matches very well between RWPT simulations and RF predictions when all available data is used for training of the model. When only 10% of data is used to train the RF model, the breakthrough statistics still show a high degree of match (within 2% error) with the RWPT values for all quantile measurements, though the error in estimation of the peak arrival time (which occurs at a relatively early time) is more noticeable.

4 Conclusions

Computing the solute breakthrough curves in flow through porous media, a fundamental characteristic of transport in geological formations, is a time-consuming task due to modeling limitations. By applying random forest algorithms trained with data from random walk particle tracking simulations, we predict solute BTCs in 2D geometries extracted from images of two natural sandstones and a sandpack. Using Bayesian optimization, we selected the best hyperparameters for predictions, avoiding data overfitting and bias of the machine learning predictions. The accuracy of the RF predictions and the actual RWPT workflow built carefully for this study demonstrates the ability of the random forests algorithm in capturing the critical flow and transport properties of porous media for new input data that other generalized auto-machine learning tools may not effectively capture.

Our analysis shows that the random forests algorithm accurately predicts the transport behaviors, and computational cost can be reduced when training data cover broad range of heterogeneities. The output of RF algorithm quantitatively compare very well to key indicators of breakthrough time statistics produced using RWPT simulations,

and moreover the impact on model performance is minimal when the amount of training data is reduced to only 10%. However, results are sensitive to the properties of the training data if geometrical sample effects are present. We find that RF predictions made in the Castlegate sandstone, the highest heterogeneous sample studied, underestimate peak concentration due to geometrical sample effects in the training data, highlighting the shortcoming of the random forest algorithm when the representative heterogeneities supporting the training data has a limited range.

Our work highlights the potential benefits of the random forest algorithms for predicting transport behaviors in porous media when using limited data, while also drawing attention to the need for careful management of the algorithm's training data, particularly with respect to the range of heterogeneities represented. Work in this direction is in progress. The development of these methods and other machine learning architectures may help avoid the time-consuming procedure in solute transport predictions while increasing the accuracy of the problems described in this work.

Funding L.J.P. and R.P. acknowledges the support of the Desert Research Institute (DRI) through the Internal Project Assignment award (PG20133).

Data availability All datasets and codes generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflicts of 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.

References

Ahmad, I., Ilyas, H., Urooj, A., et al.: Novel applications of intelligent computing paradigms for the analysis of nonlinear reactive transport model of the fluid in soft tissues and microvessels. Neural Comput. Appl. **31**(12), 9041–9059 (2019)

Ahmmed, B., Mudunuru, M.K., Karra, S., et al.: A comparative study of machine learning models for predicting the state of reactive mixing. J. Comput. Phys. **432**(110), 147 (2021)

Al-Farisi, O., Zhang, H., Raza, A., et al.: Machine learning for 3D image recognition to determine porosity and lithology of heterogeneous carbonate rock. In: SPE Reservoir Characterisation and Simulation Conference and Exhibition, OnePetro (2019)

Al-Salamah, I.S., Ghazaw, Y.M., Ghumman, A.R.: Groundwater modeling of Saq Aquifer Buraydah Al Qassim for better water management strategies. Environ. Monit. Assess. **173**(1), 851–860 (2011)

Alizadeh, R., Allen, J.K., Mistree, F.: Managing computational complexity using surrogate models: a critical review. Res. Eng. Des. **31**(3), 275–298 (2020)

Aquino, T., Aubeneau, A., Bolster, D.: Peak and tail scaling of breakthrough curves in hydrologic tracer tests. Adv. Water Resour. **78**, 1–8 (2015)

Ben-Noah, I., Hidalgo, J.J., Jimenez-Martinez, J., et al.: Solute trapping and the mechanisms of non-Fickian transport in partially saturated porous media. Water Resources Res. 59(2):e2022WR033,613 (2023)

Bolster, D., Roche, K.R., Morales, V.L.: Recent advances in anomalous transport models for predicting contaminants in natural groundwater systems. Curr. Opin. Chem. Eng. **26**, 72–80 (2019)

Breiman, L.: Bagging predictors. Mach. Learn. **24**, 123–140 (1996)

Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)

Breiman, L., Cutler, A.: Random forest-manual (2004). http://www.statberkeleyedu/breiman/RandomForests/cc_manual.htm

Breiman, L., Friedman, J.H., Olshen, R.A., et al.: Classification and Regression Trees. Brooks. Wadsworth and Brooks, Monterey (1984)
Breiman, L., Friedman, J.H., Olshen, R.A., et al.: Classification and Regression Trees. Routledge, London (2017)
Brusseau, M.L., Anderson, R.H., Guo, B.: PFAS concentrations in soils: background levels versus contaminated sites. Sci. Total Environ. **740**(140), 017 (2020)
Comolli, A., Hakoun, V., Dentz, M.: Mechanisms, upscaling, and prediction of anomalous dispersion in heterogeneous porous media. Water Resour. Res. **55**(10), 8197–8222 (2019)
Cortis, A., Berkowitz, B.: Anomalous transport in "classical" soil and sand columns. Soil Sci. Soc. Am. J. **68**(5), 1539–1548 (2004)
De Lucia, M., Kempka, T., Jatnieks, J., et al.: Integrating surrogate models into subsurface simulation framework allows computation of complex reactive transport scenarios. Energy Procedia **125**, 580–587 (2017)
Di Palma, P.R., Guyennon, N., Parmigiani, A., et al.: Impact of synthetic porous medium geometric properties on solute transport using direct 3d pore-scale simulations. Geofluids 2019 (2019)
DiGiovanni, A.A., Fredrich, J.T., Holcomb, D.J., et al. Micromechanics of compaction in an analogue reservoir sandstone. In: 4th North American Rock Mechanics Symposium, OnePetro (2000)
Edmunds, W., Smedley, P.: Residence time indicators in groundwater: the east midlands triassic sandstone aquifer. Appl. Geochem. **15**(6), 737–752 (2000)
Engdahl, N.B., Aquino, T.: Upscaled models for time-varying solute transport: transient spatial-Markov dynamics. Adv. Water Resour. **166**(104), 271 (2022)
Gouze, P., Puyguiraud, A., Roubinet, D., et al.: Pore-scale transport in rocks of different complexity modeled by random walk methods. Transp. Porous Med. **146**(1–2), 139–158 (2023)
Guo, B., Zeng, J., Brusseau, M.L.: A mathematical model for the release, transport, and retention of per-and polyfluoroalkyl substances (PFAS) in the vadose zone. Water Resour. Res. **56**(2), e2019WR026667 (2020a)
Guo, Z., Henri, C.V., Fogg, G.E., et al.: Adaptive multirate mass transfer (aMMT) model: a new approach to upscale regional-scale transport under transient flow conditions. Water Resour. Res. **56**(2), e2019WR026000 (2020b)
Haggerty, R., McKenna, S.A., Meigs, L.C.: On the late-time behavior of tracer test breakthrough curves. Water Resour. Res. **36**(12), 3467–3479 (2000)
He, Q., Tartakovsky, A.M.: Physics-informed neural network method for forward and backward advection-dispersion equations. Water Resour. Res. **57**(7), e2020WR029479 (2021)
He, Q., Barajas-Solano, D., Tartakovsky, G., et al.: Physics-informed neural networks for multiphysics data assimilation with application to subsurface transport. Adv. Water Resour. **141**(103), 610 (2020)
Hidalgo, J.J., Neuweiler, I., Dentz, M.: Transport under advective trapping. J. Fluid Mech. **907**, A36 (2021)
Hong, S., Lynn, H.S.: Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction. BMC Med. Res. Methodol. **20**(1), 1–12 (2020)
Icardi, M., Boccardo, G., Marchisio, D.L., et al.: Pore-scale simulation of fluid flow and solute dispersion in three-dimensional porous media. Phys. Rev. E **90**(1), 013032 (2014)
Kamrava, S., Tahmasebi, P., Sahimi, M.: Linking morphology of porous media to their macroscopic permeability by deep learning. Transp. Porous Med. **131**(2), 427–448 (2020)
Kamrava, S., Im, J., de Barros, F.P.J., et al.: Estimating dispersion coefficient in flow through heterogeneous porous media by a deep convolutional neural network. Geophys. Res. Lett. **48**(18), e2021GL094443 (2021). https://doi.org/10.1029/2021GL094443
Kim, J.S., Kang, P.K.: Anomalous transport through free-flow-porous media interface: pore-scale simulation and predictive modeling. Adv. Water Resour. **135**(103), 467 (2020)
Kowalek, P., Loch-Olszewska, H., Szwabiński, J.: Classification of diffusion modes in single-particle tracking data: feature-based versus deep-learning approach. Phys. Rev. E **100**(3), 032410 (2019)
Kurotori, T., Zahasky, C., Benson, S.M., et al.: Description of chemical transport in laboratory rock cores using the continuous random walk formalism. Water Resour. Res. **56**(9), e2020WR027511 (2020)
Lange, H., Sippel, S.: Machine learning applications in hydrology. In: Forest-Water Interactions, pp. 233–257. Springer, Cham (2020)
Leal, A.M., Kyas, S., Kulik, D.A., et al.: Accelerating reactive transport modeling: on-demand machine learning algorithm for chemical equilibrium calculations. Transp. Porous Med. **133**(2), 161–204 (2020)
Lee, J.W., Lee, J.B., Park, M., et al.: An extensive comparison of recent classification tools applied to microarray data. Comput. Stat. Data Anal. **48**(4), 869–885 (2005)

Li, Y., Lu, P., Zhang, G.: An artificial-neural-network-based surrogate modeling workflow for reactive transport modeling. Pet. Res. **7**(1), 13–20 (2021)
Mostaghimi, P., Bijeljic, B., Blunt, M.J.: Simulation of flow and dispersion on pore-space images. SPE J. **17**(04), 1131–1141 (2012)
Muñoz-Gil, G., Garcia-March, M.A., Manzo, C., et al.: Single trajectory characterization via machine learning. New J. Phys. **22**(1), 013010 (2020)
Naghibi, S.A., Ahmadi, K., Daneshi, A.: Application of support vector machine, random forest, and genetic algorithm optimized random forest models in groundwater potential mapping. Water Resour. Manag. **31**(9), 2761–2775 (2017)
Peksa, A.E., Wolf, K.H.A., Zitha, P.L.: Bentheimer sandstone revisited for experimental purposes. Mar. Pet. Geol. **67**, 701–719 (2015)
Perez, L.J., Hidalgo, J.J., Dentz, M.: Upscaling of mixing-limited bimolecular chemical reactions in Poiseuille flow. Water Resour. Res. **55**(1), 249–269 (2019). https://doi.org/10.1029/2018WR022730
Perez, L.J., Parashar, R., Plymale, A., et al.: Contributions of biofilm-induced flow heterogeneities to solute retention and anomalous transport features in porous media. Water Res. **209**, 117896 (2021a). https://doi.org/10.1016/j.watres.2021.117896
Perez, L.J., Puyguiraud, A., Hidalgo, J.J., et al.: Upscaling mixing-controlled reactions in unsaturated porous media. Transp. Porous Med. **146**, 177–196 (2021b)
Poffenbarger, H., Castellano, M., Egli, D., et al.: Contributions of plant breeding to soil carbon storage: retrospect and prospects. Crop Sci. **63**, 990–1018 (2023)
Pollock, D.W.: Semianalytical computation of path lines for finite-difference models. Ground Water **26**(6), 743–750 (1988). https://doi.org/10.1111/j.1745-6584.1988.tb00425.x
Popova, O.H., Small, M.J., McCoy, S.T., et al.: Comparative analysis of carbon dioxide storage resource assessment methodologies. Environ. Geosci. **19**(3), 105–124 (2012)
Puyguiraud, A., Gouze, P., Dentz, M.: Stochastic dynamics of Lagrangian pore-scale velocities in three-dimensional porous media. Water Resour. Res. **55**(2), 1196–1217 (2019). https://doi.org/10.1029/2018WR023702
Puyguiraud, A., Perez, L.J., Hidalgo, J.J., et al.: Effective dispersion coefficients for the upscaling of pore-scale mixing and reaction. Adv. Water Resour. **146**(103), 782 (2020)
Puyguiraud, A., Gouze, P., Dentz, M.: Pore-scale mixing and the evolution of hydrodynamic dispersion in porous media. Phys. Rev. Lett. **126**(16), 164501 (2021)
Qiao, C., Xu, Y., Zhao, W., et al.: Fractional derivative modeling on solute non-Fickian transport in a single vertical fracture. Front. Phys. **8**, 378 (2020)
Rodriguez-Galiano, V., Mendes, M.P., Garcia-Soldado, M.J., et al.: Predictive modeling of groundwater nitrate pollution using random forest and multisource variables related to intrinsic and specific vulnerability: a case study in an agricultural setting (southern Spain). Sci. Total Environ. **476**, 189–206 (2014)
Santos, J.E., Xu, D., Jo, H., et al.: Poreflow-net: a 3d convolutional neural network to predict fluid flow through porous media. Adv. Water Resour. **138**(103), 539 (2020)
Schilders, W.H., Van der Vorst, H.A., Rommes, J.: Model Order Reduction: Theory, Research Aspects and Applications, vol. 13. Springer, Cham (2008)
Sharma, P.K., Agarwal, P., Mehdinejadiani, B.: Study on non-Fickian behavior for solute transport through porous media. ISH J. Hydraul. Eng. **28**(sup1), 171–179 (2022)
Sherman, T., Engdahl, N.B., Porta, G., et al.: A review of spatial Markov models for predicting pre-asymptotic and anomalous transport in porous and fractured media. J. Contam. Hydrol. **236**, 103734 (2020)
Shiri, J.: Improving the performance of the mass transfer-based reference evapotranspiration estimation approaches through a coupled wavelet-random forest methodology. J. Hydrol. **561**, 737–750 (2018)
Singh, B., Sihag, P., Singh, K.: Modelling of impact of water quality on infiltration rate of soil by random forest regression. Model. Earth Syst. Environ. **3**(3), 999–1004 (2017)
Snoek J, Larochelle H, Adams RP (2012) Practical bayesian optimization of machine learning algorithms. In: Advances in neural information processing systems, vol. 25
Sun, L., Qiu, H., Wu, C., et al.: A review of applications of fractional advection-dispersion equations for anomalous solute transport in surface and subsurface water. Wiley Interdiscip. Rev. Water **7**(4), e1448 (2020)

Swanson, R.D., Binley, A., Keating, K., et al.: Anomalous solute transport in saturated porous media: relating transport model parameters to electrical and nuclear magnetic resonance properties. Water Resour. Res. **51**(2), 1264–1283 (2015)

Tang, M., Liu, Y., Durlofsky, L.J.: A deep-learning-based surrogate model for data assimilation in dynamic subsurface flow problems. J. Comput. Phys. **413**(109), 456 (2020)

Vesper, D.J.: Contamination of cave waters by heavy metals. In: Encyclopedia of Caves, pp. 320–325. Elsevier, Netherlands (2019)

Wang, Z., Lai, C., Chen, X., et al.: Flood hazard risk assessment model based on random forest. J. Hydrol. **527**, 1130–1141 (2015)

Weller, H.G., Tabor, G., Jasak, H., et al.: A tensorial approach to computational continuum mechanics using object-oriented techniques. Comput. Phys. **12**(6), 620–631 (1998)

Wu, J., Chen, X.Y., Zhang, H., et al.: Hyperparameter optimization for machine learning models based on Bayesian optimization. J. Electron. Sci. Technol. **17**(1), 26–40 (2019)

Yoon, S., Kang, P.K.: Mixing-induced bimolecular reactive transport in rough channel flows: pore-scale simulation and stochastic upscaling. Transp. Porous Med. **146**, 329–350 (2021)

Zhang, Z., Cai, Z.: Permeability prediction of carbonate rocks based on digital image analysis and rock typing using random forest algorithm. Energy Fuels **35**(14), 11271–11284 (2021)

Zhou, X., Zhu, X., Dong, Z., et al.: Estimation of biomass in wheat using random forest regression algorithm and remote sensing data. Crop J. **4**(3), 212–219 (2016)