SIAM/ASA J. Uncertainty Quantification
Vol. 3, pp. 954–968

Impact of Data Assimilation on Cost-Accuracy Tradeoff in Multifidelity Models*

Michael Sinsbeck† and Daniel M. Tartakovsky‡

Abstract. Observable phenomena can often be described by alternative models with different degrees of fidelity. Such models typically contain uncertain parameters and forcings, rendering predictions of the state variables uncertain as well. Within the probabilistic framework, solutions of these models are given in terms of their probability density functions (PDFs). In the presence of data, the latter can be treated as prior distributions. Uncertainty and assimilation of measurements into model predictions, e.g., via Bayesian updating of solution PDFs, pose a question of model selection: Given a significant difference in computational cost, is a lower-fidelity model preferable to its higher-fidelity counterpart? We investigate this question in the context of multiphase flow in heterogeneous porous media whose hydraulic properties are uncertain. While low-fidelity (reduced-complexity) models introduce a model error, their moderate computational cost makes it possible to generate more realizations, which reduces the (e.g., Monte Carlo) sampling error. These two errors determine the model with the smallest total error. Our analysis suggests that assimilation of measurements of a quantity of interest (a medium's saturation, in our example) influences both types of errors, increasing the probability that the predictive accuracy of a reduced-complexity model exceeds that of its higher-fidelity counterpart.

Key words. uncertainty quantification, reduced complexity, stochastic, subsurface, porous media, unsaturated

AMS subject classifications. 60H30, 35Q86, 76S05, 86A05

DOI. 10.1137/141001743

1. Introduction.

Every physical, biological, and chemical phenomenon can be described by alternative mathematical models that differ in their degree of fidelity. Fine-scale models (e.g., molecular dynamics simulations) typically rely on fewer assumptions but are computationally prohibitive at a scale of practical interest. Their coarse-scale counterparts (e.g., reaction-diffusion equations) are orders of magnitude faster to compute but rest on a number of foundational assumptions whose veracity might be hard to ascertain and lead to model errors. The standard choice between different fidelity models is a compromise between representational accuracy and computational expediency.

Uncertainty in models' parameterizations and forcings (e.g., initial/boundary conditions and sources) complicates the model selection. When quantified probabilistically, this uncertainty gives rise to multiple model predictions, whose likelihood of occurrence is expressed in

*Received by the editors December 29, 2014; accepted for publication (in revised form) August 3, 2015; published electronically September 30, 2015.
http://www.siam.org/journals/juq/3/100174.html

†Institute for Modeling Hydraulic and Environmental Systems (LS3), SimTech, University of Stuttgart, Stuttgart, 70569, Germany (michael.sinsbeck@iws.uni-stuttgart.de). This author's work was supported by the German Research Foundation (DFG grant 805/3-1) and the Cluster of Excellence in Simulation Technology (EXC 310/1) at the University of Stuttgart.

‡Department of Mechanical and Aerospace Engineering, University of California, San Diego, La Jolla, CA 92037 (dmt@ucsd.edu). This author's work was supported in part by the Air Force Office of Scientific Research under grant FA9550-12-1-0185 and by the National Science Foundation under grant EAR-1246315.

terms of the probability density function (PDF) of a quantity of interest (QoI) or its statistical moments (e.g., ensemble mean and variance). A significant computational cost of each deterministic solve of a high-fidelity model implies that the PDF of its solution must be estimated from only a few realizations. This gives rise to a sampling error.

Given a fixed amount of time for the calculations, a reduced-complexity model enables one to compute more realizations, decreasing the sampling error in estimation of the solution PDF. The sampling error can be eliminated completely if a reduced-complexity model allows derivation of a closed-form PDF equation [23, 27, 28]. If the sampling error of a high-fidelity model dominates the model error of a low-fidelity model, then, other factors being equal, the low-fidelity model is preferable [15]. The balance between model error and sampling error triggers a model-selection problem: Given limited computational resources, which model yields the smallest *total* error?

Availability and assimilation of QoI measurements add another facet to the model-selection problem. When used within a Bayesian framework, data assimilation would treat the PDFs obtained with multifidelity models as priors. If Bayesian data assimilation were to be robust, and as more data become available, the significance of the choice of a prior is expected to diminish. In other words, the solution PDF computed with a reduced-complexity model might lead to a posterior distribution that is "close" to that calculated with the solution PDF from its high-fidelity counterpart.

Moreover, the calculation of a posterior distribution becomes increasingly sampling intensive as the amount of available data increases. As more data become available, the percentage of realizations that match the data decreases, and therefore larger sample sizes are required.

In other words, one can expect data assimilation to decrease the model error of a reduced-complexity model and to increase the sampling error of all calculations. These two effects change the balance between the model and sampling errors and, therefore, affect the optimal model selection.

From the outset, it is worthwhile contrasting these aspects of the optimal model selection with much of the existing literature on model selection and model averaging. The field of model selection is well established [2, 3, 4]; it impacted application areas as diverse as psychology [30], hydrology [10], ecology [1], and sociology [19]. The main focus of such studies is to identify a model with highest accuracy or best predictive power, without considering computational costs. We pose different questions: Given pervasive parametric uncertainty, is the use of high-fidelity, computationally expensive models justified? Given computational constraints, what model should be used? These are questions of *efficiency* rather than accuracy. This aspect of model selection also lies outside the scope of Bayesian model averaging (BMA) (see, e.g., [8, 13, 18]) and multilevel Monte Carlo (MMC) simulations (see, e.g., [11, 5]). BMA assigns a weight to each model, which is equal to its posterior probability of predicting given data, without considering the cost-accuracy tradeoff. In our analysis the fidelity of individual models is known a priori. MMC uses multiple models in a hierarchical way to increase the computational efficiency of stochastic calculations.

We investigate the impact of cost-accuracy tradeoff on model selection in the context of multiphase flow in heterogeneous porous media whose hydraulic properties are uncertain. More specifically, we consider an infiltration process, which is described alternatively by the Richards (nonlinear diffusion-advection) equation [29] and a Green–Ampt (Laplacian growth)

model [28]. The former is treated as a high-fidelity reference model and the latter as its reduced-complexity approximation. Both models use the same physical quantities as input and predict the infiltration depth as output. Uncertain properties of a porous medium are modeled as random fields. Measurements of water content, acquired, e.g., with moisture sensors, are assimilated into model predictions to update the alternative predictions of the infiltration depth.

The paper is structured as follows. The two alternative models are described in section 2. Detailed descriptions of the model for moisture measurements and the data assimilation procedure are given in section 3. Simulation results are presented in section 4. Section 5 discusses potential generalizations of our findings, and section 6 presents general conclusions drawn from the results.

2. Alternative infiltration models.

We consider infiltration of water into a two-dimensional heterogeneous soil, whose initial state is characterized by water content θ_{init}. Infiltration is driven by a constant pressure head of ponding water, ψ_0, prescribed at the soil's surface ($z = 0$). Our QoIs are the wetting depth $z_{\text{f}}(t; x)$ and the total amount of infiltrated water $Q(t)$.

2.1. Richards equation.

At any point $\mathbf{x} = (x, z)^\top$ in the flow domain $\mathcal{D}$, the temporal evolution of water content $\theta(\mathbf{x}, t) : \mathcal{D} \times \mathbb{R}^+ \to [\theta_{\text{i}}, \phi]$ and pressure head $\psi(\mathbf{x}, t) : \mathcal{D} \times \mathbb{R}^+ \to \mathbb{R}^-$ are described by the Richards equation [29]

$$\frac{\partial \theta}{\partial t} = \nabla \cdot (K \nabla \psi) - \frac{\partial K}{\partial z}, \qquad \mathbf{x} \in \mathcal{D}, \qquad t > 0, \tag{2.1}$$

where θ_{i} is the irreducible water content, ϕ is the porosity, $K(\mathbf{x}, \theta)$ is the soil hydraulic conductivity, and z denotes depths. This equation is supplemented by two constitutive relations $K = K_{\text{s}}(\mathbf{x}) K_{\text{r}}(\mathbf{x}, \psi)$ and $\theta = f(\psi)$, where K_{s} and K_{r} are the saturated and relative hydraulic conductivities, respectively. We employ the van Genuchten constitutive model [29],

$$K_{\text{r}} = \frac{\left[1 - \psi_{\text{d}}^{mn} \left(1 + \psi_{\text{d}}^{n}\right)^{-m}\right]^2}{\left(1 + \psi_{\text{d}}^{n}\right)^{m/2}}, \quad \frac{\theta - \theta_{\text{i}}}{\phi - \theta_{\text{i}}} = \frac{1}{(1 + \psi_{\text{d}}^{n})^{m}}, \quad \psi_{\text{d}} = \alpha \left|\psi\right|, \quad m = 1 - \frac{1}{n}. \tag{2.2}$$

The shape parameters $\alpha > 0$ and $n > 0$ may vary in space, reflecting the soil heterogeneity. Equations (2.1) and (2.2) are defined on domain $\mathcal{D} = \{\mathbf{x} : -L \le x \le L, 0 \le z \le \infty\}$, subject to initial and boundary conditions

$$\theta(\mathbf{x}, 0) = \theta_{\text{init}}, \quad \psi(x, z = 0, t) = \psi_0, \quad \theta(x, z \to \infty, t) = \theta_{\text{init}}, \quad \frac{\partial \psi}{\partial x}(x = \pm L, z, t) = 0. \tag{2.3}$$

Probabilistic model parameterization. Among all the model parameters, saturated hydraulic conductivity K_{s} and soil parameter α vary most and exhibit the highest degree of uncertainty (see, e.g., [21, 24] and the references therein). In line with this observation, we treat $K_{\text{s}}(\mathbf{x})$ and $\alpha(\mathbf{x})$ as random fields, while assuming the remaining parameters (ϕ, θ_{i}, and n) to be constant and known with certainty. Following the standard practice (ibid.), we assume that the random fields K_{s} and α are statistically independent, log-normal, and second-order stationary (statistically homogeneous). The latter means that they have constant means and variances

and autocovariance functions that depend only on the distance between two points. The soil data analyzed in [21] and various subsequent studies suggest the use of an anisotropic exponential covariance function

$$C(d_x, d_z) = \sigma^2 \mathrm{e}^{-s}, \qquad s = \sqrt{(d_x/\lambda_x)^2 + (d_z/\lambda_z)^2}, \tag{2.4}$$

where σ^2 is the variance, d_x and d_z are the horizontal and vertical distances between two points, and λ_x and λ_z denote the horizontal and vertical correlation lengths.

Computation of QoIs. Solving (2.1)–(2.3) yields realizations of the state variable $\theta(x, z, t)$. Realizations of the QoIs, infiltration depths $z_\mathrm{f}(t; x)$ and the amount of infiltrated water $Q(t)$, are then computed as

$$z_\mathrm{f}(t; x) = \int_0^\infty \frac{\theta(x, z, t) - \theta_\mathrm{init}}{\phi - \theta_\mathrm{init}} \, \mathrm{d}z \quad \text{and} \quad Q(t) = (\phi - \theta_\mathrm{init}) \int_{-L}^{L} z_\mathrm{f}(t; x) \, \mathrm{d}x. \tag{2.5}$$

2.2. Green–Ampt model. The Green–Ampt model provides a simplified description of infiltration. It assumes (i) homogeneity of the soil parameters in the z direction; (ii) one-dimensional vertical flow from the soil surface downwards; and (iii) the existence of a sharp wetting front $z_\mathrm{f}(t)$, which separates the dry soil ($\theta = \theta_\mathrm{init}$) ahead of the front from the wet ($\theta = \phi$) behind it, such that

$$\theta(z, t) = \begin{cases} \theta_\mathrm{wet} = \phi & \text{for } z < z_\mathrm{f}(t), \\ \theta_\mathrm{dry} = \theta_\mathrm{init} & \text{for } z \geq z_\mathrm{f}(t). \end{cases} \tag{2.6}$$

For the problem under consideration, the Green–Ampt model yields an implicit solution for the infiltration front $z_\mathrm{f}(t)$ [17, 28, 29],

$$z_\mathrm{f} - (\psi_0 - \psi_\mathrm{f}) \ln\left(1 + \frac{z_\mathrm{f}}{\psi_0 - \psi_\mathrm{f}}\right) = \frac{K_\mathrm{s}}{\phi - \theta_\mathrm{i}} t, \tag{2.7}$$

where the pressure head at the infiltration front, ψ_f, is set to a capillary drive [17, 28],

$$\psi_\mathrm{f} = -\int_{\psi_\mathrm{i}}^{0} K_\mathrm{r}(\psi) \, \mathrm{d}\psi. \tag{2.8}$$

The pressure head in the dry soil, ψ_i, is related to the corresponding irreducible water content θ_i by (2.2).

Probabilistic model parameterization. The foundational assumptions of the Green–Ampt solution (2.7) preclude the direct use of the input soil parameters described in section 2.1. The vertical flow assumption replaces the two-dimensional flow field with a collection of one-dimensional isolated flow tubes labeled by x, in a manner consistent with the Dagan–Bresler parameterization [7]. The vertical homogeneity assumption requires one to average out the vertical variability of the soil properties. These spatial averages must be computed over the a priori unknown interval $0 \leq z \leq z_\mathrm{f}(t; x)$.

To estimate the averaging intervals at a point x, we sample K_s and α from their respective PDFs and insert them into the Green–Ampt solution (2.7). The resulting ensemble of infiltration depths $z_\mathrm{f}(t; x)$ is then used to construct an empirical cumulative distribution function

(CDF) $F(z) = \mathbb{P}(z_\mathrm{f} < z)$ for any z. Finally, the parameters are averaged using a weighted mean with weights proportional to $1-F(z)$. The conductivity K_s is averaged with a harmonic mean (analogous to multiple resistors in a series connection) [25]. The shape parameter α is averaged arithmetically.

3. Data assimilation. We use the solution PDFs computed with the two alternative models to assimilate the state variable's measurements into model predictions. We adopt a Bayesian updating strategy, in which these PDFs serve as prior distributions and the data-informed (improved) model predictions are given by posterior PDFs.

3.1. Data acquisition and processing. Soil-moisture sensors provide pointwise measurements of the water content $\theta(x,z,t)$. While these data can be assimilated into predictions of $\theta(x,z,t)$ obtained with the Richards equation (section 2.1), the Green–Ampt model (section 2.2) predicts only the QoI $z_\mathrm{f}(t;x)$. To guarantee a meaningful comparison between the two models, and to isolate the model error's impact on posterior distributions, we compare the probabilistic predictions of the QoI $z_\mathrm{f}(t;x)$ obtained with the two models described in section 2. This requires one to convert measurements of the state variable $\theta(x,z,t)$ into "measurements" of the QoI $z_\mathrm{f}(t;x)$.

Let (x,z) denote a sensor's position. We model the sensor's output as

$$s(x,z) = \begin{cases} \text{dry} & \text{if } z_\mathrm{f}(x) < \varepsilon\, z, \\ \text{wet} & \text{otherwise,} \end{cases} \tag{3.1}$$

where ε represents a (small) measurement error. It is assumed to have a log-normal distribution, such that $\ln\varepsilon \sim \mathcal{N}(0,\sigma_\varepsilon)$. This allows the measurements to be wrong in some cases, especially if the sensor is very close to the wetting front ($z \approx z_\mathrm{f}(x)$). In the numerical examples presented below, we set $\sigma_\varepsilon = 0.1$.

Solving for ε and inserting the result into the normal CDF (`normcdf`), we express the conditional probability of getting the measurement "wet" in terms of the sensor's position (x,z), given a wetting front z_f, as

$$\mathbb{P}[s(x,z) = \text{wet} \mid z_\mathrm{f}] = \texttt{normcdf}\left(\sigma_\varepsilon^{-1} \ln\frac{z_\mathrm{f}(x)}{z}\right). \tag{3.2a}$$

The conditional probability of measuring "dry" is, of course, the complement

$$\mathbb{P}[s(x,z) = \text{dry} \mid z_\mathrm{f}] = 1 - \mathbb{P}[s(x,z) = \text{wet} \mid z_\mathrm{f}]. \tag{3.2b}$$

In the presence of multiple sensors, we assume their measurement errors to be independent. That means that the noise ε is different for each sensor. The probability of measuring values at multiple sensors is then the product of the probabilities of the individual measurements. For a number of sensor positions $(x_1,z_1),\dots,(x_n,z_n)$ and measurement values $d_1,\dots,d_n$, with $d_i \in \{\text{dry},\text{wet}\}$, we obtain

$$\mathbb{P}\left[\mathbf{s} = \mathbf{d} \mid z_\mathrm{f}\right] = \mathbb{P}\left[s(x_1,z_1) = d_1,\dots,s(x_n,z_n) = d_n \mid z_\mathrm{f}\right] = \prod_{i=1}^{n} \mathbb{P}\left[s(x_i,z_i) = d_i \mid z_\mathrm{f}\right]. \tag{3.3}$$

3.2. Bayesian filtering. We use the Smith and Gelfand implementation [22] of Bayes' theorem to assimilate a set of binary data into model predictions. The representation of all uncertain quantities is sample-based. After running Monte Carlo simulations for N realizations of the soil parameters, we obtain N wetting fronts $\{z_{\mathrm{f}}^{(1)}(t), \dots, z_{\mathrm{f}}^{(N)}(t)\}$. This (unweighted) sample is a representation of the prior distribution. A prior sample of the total water content $\{Q^{(1)}(t), \dots, Q^{(N)}(t)\}$ is obtained by inserting each realization into (2.5).

To generate the posterior, each realization is assigned a weight proportional to its likelihood of measuring the data. The likelihood $l^{(i)}$ of the ith realization is

$$l^{(i)} := \mathbb{P}[\mathbf{s} = \mathbf{d} \mid z_{\mathrm{f}}^{(i)}], \tag{3.4}$$

where $\mathbb{P}[\mathbf{s} = \mathbf{d} \mid z_{\mathrm{f}}^{(i)}]$ is given by (3.3). The weights of the posterior sample are computed as

$$w^{(i)} = \frac{l^{(i)}}{\sum_{j=1}^{N} l^{(j)}}. \tag{3.5}$$

The posterior distribution is then represented by a weighted sample $\{z_{\mathrm{f}}^{(1)}, w^{(1)}; \dots; z_{\mathrm{f}}^{(N)}, w^{(N)}\}$ and $\{Q^{(1)}, w^{(1)}; \dots; Q^{(N)}, w^{(N)}\}$. More details are given in [22].

The above implementation of Bayes' theorem is slower than other methods, such as Markov Chain Monte Carlo methods [9], but it has an important advantage. Sampling independently from the data allows one to control which realizations of the input are inserted into the models. This ensures that both the high-fidelity and the reduced-complexity models are run with the same input. If we define a finite sample of the input parameters as the reference, then all of the discrepancy in the output is due to the model errors. In that case, there is no sampling (Monte Carlo) error, even with a finite sample size.

3.3. Statistical distance. The Monte Carlo simulations and Bayesian updating described above yield the prior and posterior distributions of the total water content. To compare the distributions obtained with the two alternative models, we employ the Earth Mover's distance (EMD) [20]. If a distribution is thought of as a pile of earth, the EMD between two distributions is the minimal work required to turn one distribution into the other one. In one dimension, the EMD is computed as the area between the two CDFs $F_X(x)$ and $F_Y(y)$ of random variables X and Y [6],

$$D(F_X, F_Y) = \int |F_X(x) - F_Y(x)| \, \mathrm{d}x. \tag{3.6}$$

The EMD is preferred over other error measures, such as the Kullback–Leibler divergence [14] or Hellinger distance [26], because it does not require a PDF estimate.

4. Simulation results. In the numerical experiments reported below, all soil properties are taken from [21]: statistical properties of the uncertain soil properties K_{s} and α are summarized in Table 1; the remaining parameters are set to $n = 1.81$, $\phi = 0.42$, $\theta_{\mathrm{i}} = 0.13$, and $\psi_0 = 0.01\,\mathrm{m}$. The initial soil moisture content is set to $\theta_{\mathrm{init}} = 0.2$. This value is different from θ_{i} to avoid numerical instabilities.

The simulations are carried out on a $2.0\,\mathrm{m} \times 2.0\,\mathrm{m}$ rectangular domain. The horizontal length of the domain was chosen to exceed $2\lambda_x^\star$, where $\lambda_x^\star$ is the largest of the horizontal correlation lengths in Table 1. This ensures that the considered domain is representative of the full variability in the soil. The vertical size of the domain was determined after preliminary simulations to ensure that no water leaves the domain at the bottom boundary during the considered simulation time.

A solution of the Richards equation is treated as a reference model, which does not introduce any model error. An error of the Green–Ampt model is reported with respect to solutions of the Richards equation.

The stochastic reference is a sample of $10,000$ realizations of the soil properties, called the *base sample*. Using them to parameterize the Richards equation yields the overall reference solution, which is assumed to have no model error and no sampling error.

Moisture sensors are located at depth $z_\mathrm{s} = 0.1\,\mathrm{m}$; virtual measurements are taken at time $t = 30\,\mathrm{min}$. For that, one realization of the Richards equation is selected as the virtual truth and inserted into (2.5), and then noise is added according to (3.1). Both the horizontal position and the number of sensors differ from experiment to experiment. Whenever a quantity depends on the number of moisture sensors, we equip it with an index for the number of sensors, e.g., T_0 for a certain time in the prior and T_7 for the same quantity in the posterior with data from seven sensors.

Table 1

Statistical properties of K_s and α: mean μ, variance σ^2, and correlation lengths λ_x, λ_z [21, Table 3a].

	μ	σ^2	λ_x[m]	λ_z[m]
$\ln K_\mathrm{s}$	–3.58	0.89	0.7840	0.2123
$\ln \alpha$	–3.01	0.63	0.2554	0.1117

4.1. Notation. To emphasize the difference between the model error and sampling error, all errors are written in the form $D^N_{\alpha,n}$. The subscript $\alpha \in \{\mathrm{R}, \mathrm{G}\}$ denotes either the Richards equation ($\alpha = \mathrm{R}$) or the Green–Ampt model ($\alpha = \mathrm{G}$). The subscript $n = 0, 1, \ldots$ denotes the number of soil moisture sensors used ($n = 0$ for the prior). Finally, N is the number of realizations used in the Monte Carlo simulations. In the convergence analysis, N is varied from 1 to $1,000,000$ and samples are drawn with replacement. More details on this approach are given in section 4.4. A special case is simulations with all $10,000$ realizations from the base sample (without replacement): These calculations do not have a sampling error, so the error represents the model error. Such cases are marked with the superscript star *, e.g., $D^\star_{\mathrm{G},0}$ for the prior model error of the Green–Ampt model.

4.2. Numerical implementation. The Richards equation is solved using the USGS software package `vs2dt`. Horizontally, the domain is discretized with 50 equally spaced cells. Vertically, it is discretized with 30 cells and a grid refinement towards the top boundary.

The Green–Ampt model is solved using the MATLAB function `fzero`, which uses a combination of bisection, secant, and inverse quadratic interpolation methods. One simulation run solves (2.7) for all 50 soil columns defined by the discretization of the Richards equation.

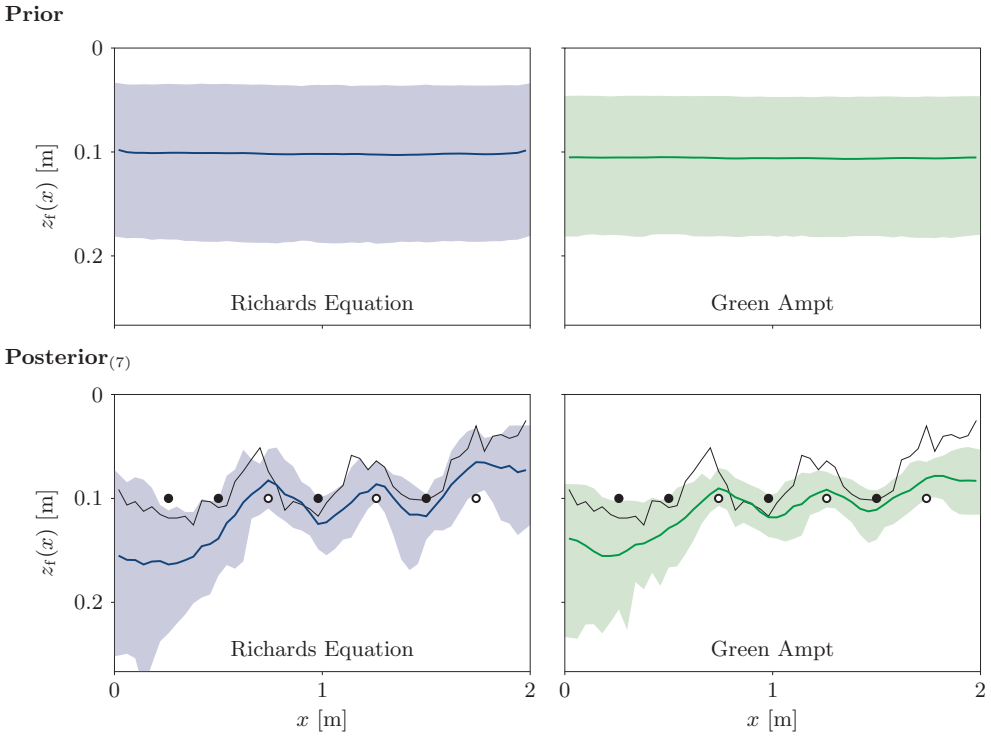

Figure 1. *Prior and posterior statistics of the infiltration depth $z_f(x)$ at $t = 30$. Thick lines: ensemble mean. Light colored areas: pointwise 0.1 and 0.9 percentiles. Black thin line: virtual truth. Circles: moisture sensors, wet (black), dry (white). Both z_f and x are displayed in* [m]. *All diagrams show the same part of the domain.*

4.3. Infiltration depth. Figure 1 shows the spatial variability of the prior and posterior distributions of the infiltration depth $z_f(t;x)$. The posterior was calculated with data from seven equidistant moisture sensors. The figure also shows the virtual truth from which the data are generated via the measurement model; see (3.1). The plots were generated using the full base sample. Therefore, there is no sampling error in these calculations, and all discrepancies between the two results are due to the model error.

The prior computed with the Green–Ampt model overestimates the infiltration depths on average by 0.005 m, which corresponds to the relative error of about 5%. The distribution's width is slightly underestimated. In the posterior distributions, the Green–Ampt model again underestimates the distribution's width: The shaded area is smaller than its reference. The virtual truth leaves the shaded area in about one third of the domain.

Figure 2 exhibits density estimates (using the MATLAB function `ksdensity`) of the total amount of infiltrated water Q; see (2.5). Again, the full base sample was used. This plot confirms the previous observations. In the prior, the reduced model slightly overestimates the water content. In the posterior, the means of both models almost align. The model errors of the Green–Ampt model are $D^\star_{\mathrm{G},0} = 1.87 \cdot 10^{-3}\,\mathrm{m}^2$ and $D^\star_{\mathrm{G},7} = 1.60 \cdot 10^{-3}\,\mathrm{m}^2$.

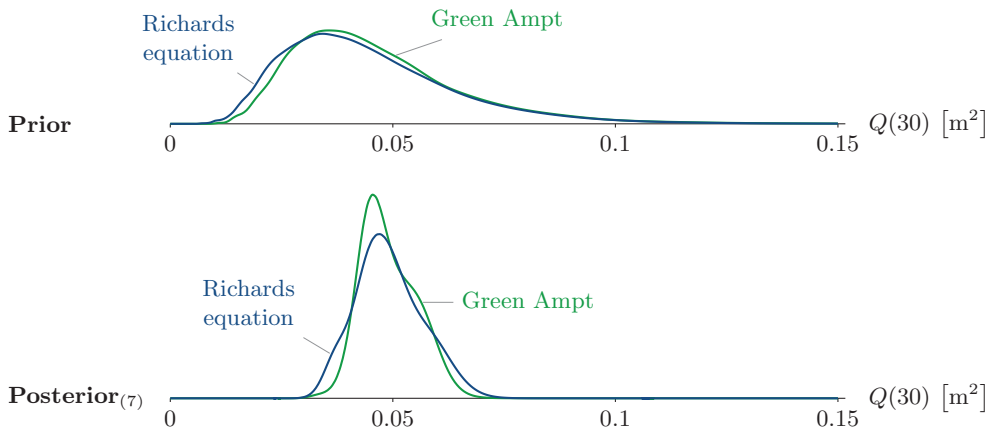

Figure 2. *Prior and posterior density estimates of the total amount of infiltrated water $Q(30)$ for the two models.*

4.4. Simulation times and convergence rates. We now return to the model-selection problem: Given limited computational resources, which model has the smallest total error? We compare the total errors (model error plus sampling error) of both models as a function of calculation time.

In varying the sample size, we follow a bootstrapping approach. This means that the samples are drawn from the initial base sample with replacement. This procedure allows one to extend the analysis to sample sizes larger than the base sample and still obtain the typical Monte Carlo convergence rate of $\mathcal{O}(N^{-1/2})$; see, e.g., [16]. Additionally, for each data point the procedure is repeated 250 times with different random samples and averaged to ensure that the results are robust against sampling artifacts.

We employ the setup with seven soil moisture sensors and use the same measurements as in the previous section. Figure 3 shows the convergence plot of $D^N_{\mathrm{G},n}$ for sample sizes between $N = 1$ (first data point of each line) and $N = 1,000,000$ (last data point). One realization of the Richards equation takes about 56 s to compute, while one realization of the Green–Ampt model takes 0.087 s. This is a ratio of more than $600 : 1$.

Since the Richards equation represents the reference model, its error converges to zero, $D^\star_{\mathrm{R},n} = 0$, by definition. The convergence rate is $\mathcal{O}(N^{-1/2})$, as expected. The error of the Green–Ampt model solution does not converge to zero, but to $D^\star_{\mathrm{G},0}$ and $D^\star_{\mathrm{G},7}$, respectively.

The model-selection problem can be solved directly from the convergence plot. For both the prior and the posterior there exists a computation time threshold T, which marks the time after which the models should be switched. If the modeler has less than this time available, the Green–Ampt model should be used; otherwise the Richards equation yields better results. For the prior this threshold is $T_0 = 1.1 \cdot 10^4\,\mathrm{s} \approx 3\,\mathrm{h}$; for the posterior it is $T_7 = 6.3 \cdot 10^5\,\mathrm{s} \approx 174\,\mathrm{h}$. This means that if the available computation time is between T_0 and T_7, then the model selection depends on whether the prior or the posterior is to be calculated. The availability of data favors the use of the reduced model.

A comparison of the two convergence plots in Figure 3 suggests that this result could be

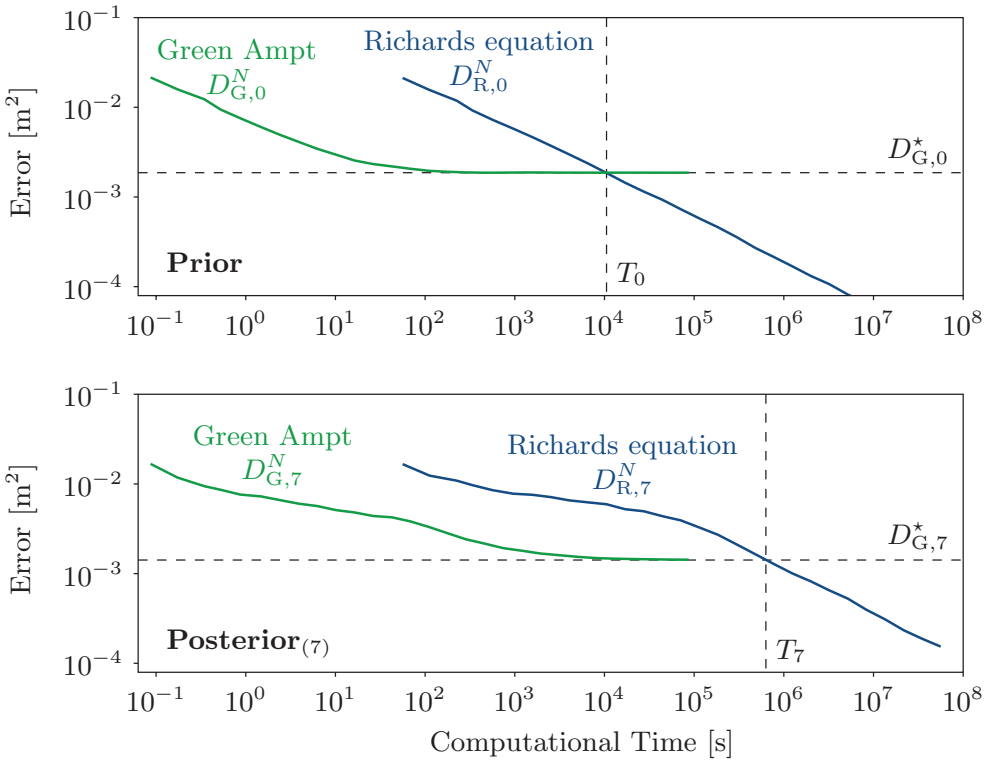

Figure 3. *Convergence of the prior and posterior distributions of the total amount of infiltrated water. Dashed horizontal lines: model errors $D^*_{\mathrm{G},0}$ and $D^*_{\mathrm{G},7}$. Dashed vertical lines: time thresholds T_0 and T_7.*

caused by two effects:

1. Available data reduce the model error, $D^\star_{\mathrm{G},7} < D^\star_{\mathrm{G},0}$.
2. Available data increase the sampling error. While the asymptotic convergence behavior of the Richards equation is $C/\sqrt{n}$ in both cases, the posterior convergence starts with a larger multiplicative constant C and therefore reaches the same accuracy later than the prior does.

In the following two sections we investigate these two effects in more detail.

4.5. Impact of data on the model error. To check the extent to which measurements can reduce the model error, we vary the sampling density (i.e., the number of sensors n) and calculate the model error $D^\star_{\mathrm{G},n}$. One would expect the influence of the prior to diminish and the model error to decrease as the number of sensors increases.

Horizontal positions of the sensors (the experimental designs) are shown in Figure 4(left). Each row represents one design, and each design is created by adding one more sensor to the previous design (the newly added sensors are shown in red). The sensors are spread out equidistantly as much as possible. This is achieved by using the Hammersley sampling [12].

Figure 4(right) shows the model error $D^\star_{\mathrm{G},n}$ as a function of the number of sensors $n =$

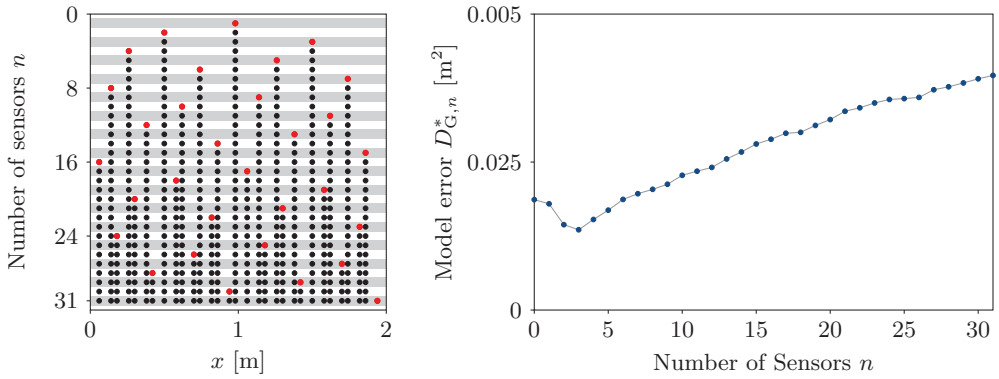

Figure 4. *Left: experimental designs (horizontal positions of the sensors). Each row shows one design, and each design contains the previous design and one additional sensor (shown in red). Right: model error as a function of the number of sensors.*

$1, \ldots, 31$. The error is averaged over 100 repetitions, in which a different realization was selected to represent the reality for generating the measurement data. One can see that the model error decreases at first until a minimum is reached with three sensors. Then the error gradually increases. The error with seven sensors is on average larger than the error without measurements. The decrease observed in section 4.4 was specific to the precise data used in that section and cannot be expected on average.

The increase in the model error for a large number of sensors shows that, among the base sample of $10,000$ Green–Ampt solutions, there are no realizations that fully resemble the true wetting front. In the situation with data from more than five sensors, the model complexity of the Green–Ampt model is too low to keep up with the increasing sampling density. An exact point of the minimal model error is, of course, problem dependent. A sampling density of 31 sensors on a domain of 2 m long would not be practical.

We conclude that the initial conjecture was incorrect: Additional measurements do not, in general, lower the model error.

4.6. Impact of data on the sampling error. Finally, we investigate the extent to which sampling density affects the sampling error. Figure 5(left) shows the convergence of the Monte Carlo solution of the Richards equation with 0 to 16 moisture sensors. At the right end of the plot, where the asymptotic convergence behavior is attained, the individual data lines are perfectly ordered according to the number of sensors.

Figure 5(right) shows the rightmost data points in Figure 5(left) (the data points for sample sizes of $1,000,000$) as a function of the number of sensors. This figure confirms the previous observation that the sampling error increases with the sampling density.

5. An approach to model selection. In this section, we recap the findings from the previous section and formulate a possible approach to the model-selection problem. Solving the model-selection problem is a matter of determining the time threshold T. Once it is known, the modeler can decide which model to use.

The convergence behavior shown in Figure 3 gives rise to two observations.

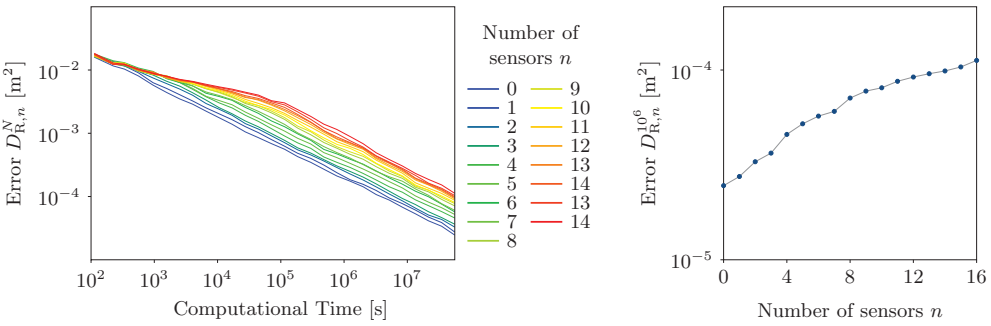

Figure 5. *Left: convergence of the Monte Carlo simulations of the Richards equation for different numbers of sensors. Since the Richards equation is the reference, the error shown is a pure sampling error. Right: the last data point from the left plot ($D^{10^6}_{\mathrm{R},n}$), shown as a function of the number of sensors.*

1. If a reduced-complexity model is much faster to solve than its high-fidelity counterpart, then, given sufficient simulation time T, the sampling error in the solution of the reduced-complexity model is negligible relative to its model error. In other words, the total error of the reduced-complexity model at the time threshold T is constant and equal to $D := D^{\star}_{\mathrm{G},n}$.
2. The simulation time T is sufficient to enable the high-fidelity model to reach the asymptotic convergence behavior of the form $C/\sqrt{N}$ (see, e.g., [16]).

Let t_{c} denote the simulation time necessary to solve one realization of the high-fidelity model. Then, the time threshold T is found by equating the two errors, $D = C/\sqrt{N}$, which yields

$$T = t_{\mathrm{c}} N = t_{\mathrm{c}} \left(\frac{C}{D}\right)^2. \tag{5.1}$$

This general result holds for any QoI and any error measure, as long as the QoI estimate converges with the rate of $\mathcal{O}(N^{-1/2})$.

Figure 6 exhibits the dependence of the simulation time threshold T estimated with (5.1) on the number of sensors n. The increase in the sampling error, quantified by the factor C (Figure 5), outweighs the increase in the model error D (Figure 4), such that the time threshold T increases with the sampling density. These results are averaged over the data from 100 different "virtual truths." Therefore, the effect is not as strong as in the example given in section 4.4, which represented a single realization of the "ground truth."

Equation (5.1) reveals the difficulty in solving the model-selection problem a priori. To do so, one needs to determine both the model error of the reduced model, D, and the multiplicative constant C in the convergence behavior of the complex model. These two quantities depend on the amount of available measurements, as shown in the previous two sections.

The constant C could be estimated using the reduced model if one assumes that the reduced model converges to its limit with the same constant as the complex model does (in terms of number of realizations, not in terms of computer time). We are not able to provide a general approach for estimation of D: While in the absence of measurements one could compare a small number of realizations of both models to get an estimate of D, the presence

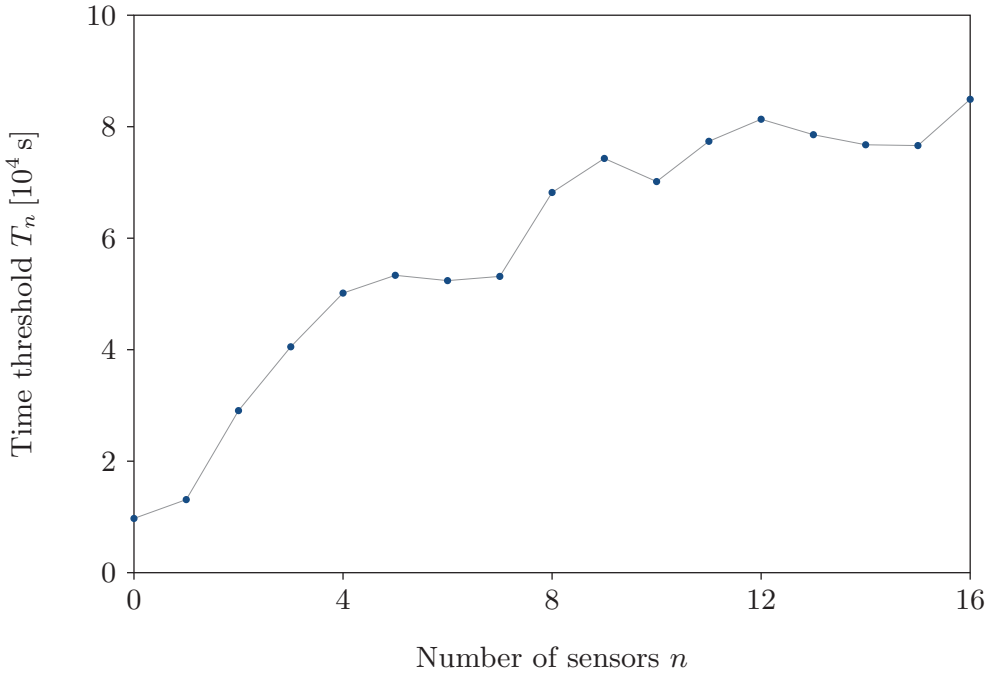

Figure 6. *Dependence of the time threshold T_n, predicted with (5.1), on the number of sensors n.*

of QoI measurements changes the model error; this change is nonmonotonic in the amount of available data (Figure 4). This makes it difficult, if not impossible, to get an a priori estimate of the model error.

6. Conclusions. We investigated the impact of data assimilation on model selection in the presence of uncertainty. Two models with different degrees of fidelity were considered in the context of infiltration into heterogeneous porous media with uncertain hydraulic properties. We found that Bayesian assimilation of data (water content measurements) changes both the posterior representational (model) error of the reduced-complexity (low-fidelity) model relative to its prior counterpart and the sampling error of the high-fidelity model. These changes in the two errors shift the optimal model selection towards the reduced-complexity model. There are situations in which the best result for the prior is calculated using the high-fidelity model, while for the posterior the best result is obtained with the reduced-complexity model.

This effect becomes most apparent when considering the limited amount of computational resources. The latter is expressed in terms of the simulation time threshold T. The use of the high-fidelity model is beneficial only if the available simulation time exceeds T. In our study, the time threshold T increased almost monotonically with the amount of QoI measurements. This increase is rather drastic, changing by a factor of 8.

The study elucidates the relationship between measurements of state variables and the model-selection problem. The two factors that influence the model selection are (i) the model error of the low-fidelity model and (ii) the sampling error of the high-fidelity model. Both

factors strongly depend on the amount of available data. Our analysis suggests that the availability of data favors the use of reduced models. The generality of this finding for other phenomena remains an open question.

REFERENCES

[1] K. Aho, D. Derryberry, and T. Peterson, *Model selection for ecologists: The worldviews of AIC and BIC*, Ecology, 95 (2014), pp. 631–636.

[2] S. T. Buckland, K. P. Burnham, and N. H. Augustin, *Model selection: An integral part of inference*, Biometrics, 53 (1997), pp. 603–618.

[3] K. P. Burnham and D. R. Anderson, *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd ed., Springer-Verlag, New York, 2002.

[4] G. Claeskens and N. L. Hjort, *Model Selection and Model Averaging*, Cambridge University Press, Cambridge, UK, 2008.

[5] K. A. Cliffe, M. B. Giles, R. Scheichl, and A. L. Teckentrup, *Multilevel Monte Carlo methods and applications to elliptic PDEs with random coefficients*, Comput. Vis. Sci., 14 (2011), pp. 3–15.

[6] S. D. Cohen and L. J. Guibas, *The Earth Mover's Distance: Lower Bounds and Invariance under Translation*, Tech. report, Computer Science Department, Stanford University, Stanford, CA, 1997.

[7] G. Dagan and E. Bresler, *Unsaturated flow in spatially variable fields: 1. Derivation of models of infiltration and redistribution*, Water Resour. Res., 19 (1983), pp. 413–420.

[8] D. Draper, *Assessment and propagation of model uncertainty*, J. Roy. Statist. Soc. Ser. B, 57 (1995), pp. 45–97.

[9] D. Gamerman and H. F. Lopes, *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*, CRC Press, Boca Raton, FL, 2006.

[10] T. Y. Gan, E. M. Dlamini, and G. F. Biftu, *Effects of model complexity and structure, data quality, and objective functions on hydrologic modeling*, J. Hydrol., 192 (1997), pp. 81–103.

[11] M. Giles, *Multilevel Monte Carlo path simulation*, Oper. Res., 56 (2008), pp. 607–617.

[12] J. M. Hammersley, *Monte Carlo methods for solving multivariable problems*, Ann. New York Acad. Sci., 86 (1960), pp. 844–874.

[13] J. Hoeting, D. Madigan, A. Raftery, and C. Volinsky, *Bayesian model averaging: A tutorial*, Statist. Sci., 14 (1999), pp. 382–417.

[14] S. Kullback and R. A. Leibler, *On information and sufficiency*, Ann. Math. Statistics, 22 (1951), pp. 79–86.

[15] P. C. Leube, F. P. J. de Barros, W. Nowak, and R. Rajagopal, *Towards optimal allocation of computer resources: Trade-offs between uncertainty quantification, discretization and model reduction*, Environ. Model. Software, 50 (2013), pp. 97–107.

[16] J. S. Liu, *Monte Carlo Strategies in Scientific Computing*, Springer, New York, 2008.

[17] S. P. Neuman, *Wetting front pressure head in the infiltration model of Green and Ampt*, Water Resour. Res., 12 (1976), pp. 564–566.

[18] S. P. Neuman, *Maximum likelihood Bayesian averaging of uncertain model predictions*, Stoch. Environ. Res. Risk Assess., 17 (2003), pp. 291–305.

[19] A. E. Raftery, *Bayesian model selection in social research*, Sociol. Methodol., 25 (1995), pp. 111–163.

[20] Y. Rubner, C. Tomasi, and L. J. Guibas, *A metric for distributions with applications to image databases*, in Proceedings of the Sixth IEEE International Conference on Computer Vision, 1998, pp. 59–66.

[21] D. Russo and M. Bouton, *Statistical analysis of spatial variability in unsaturated flow parameters*, Water Resour. Res., 28 (1992), pp. 1911–1925.

[22] A. F. M. Smith and A. E. Gelfand, *Bayesian statistics without tears: A sampling-resampling perspective*, Amer. Statist., 46 (1992), pp. 84–88.

[23] D. M. Tartakovsky, M. Dentz, and P. C. Lichtner, *Probability density functions for advective-reactive transport in porous media with uncertain reaction rates*, Water Resour. Res., 45 (2009), W07414.

[24] D. M. Tartakovsky, Z. Lu, A. Guadagnini, and A. M. Tartakovsky, *Unsaturated flow in heterogeneous soils with spatially distributed uncertain hydraulic parameters*, J. Hydrol., 275 (2003), pp. 182–193.
[25] D. M. Tartakovsky and S. P. Neuman, *Transient effective hydraulic conductivities under slowly and rapidly varying mean gradients in bounded three-dimensional random media*, Water Resour. Res., 34 (1998), pp. 21–32.
[26] A. W. Van der Vaart, *Asymptotic Statistics*, Cambridge University Press, Cambridge, UK, 2000.
[27] D. Venturi, D. M. Tartakovsky, A. M. Tartakovsky, and G. E. Karniadakis, *Exact PDF equations and closure approximations for advective-reactive transport*, J. Comput. Phys., 243 (2013), pp. 323–343.
[28] P. Wang and D. M. Tartakovsky, *Reduced complexity models for probabilistic forecasting of infiltration rates*, Adv. Water Resour., 34 (2011), pp. 375–382.
[29] A. W. Warrick, *Soil Water Dynamics*, Oxford University Press, New York, 2003.
[30] W. Zucchini, *An introduction to model selection*, J. Math. Psychol., 44 (2000), pp. 41–61.