Water Resources Research

RESEARCH ARTICLE
10.1029/2017WR021993

Information-Theoretic Approach to Bidirectional Scaling

Francesca Boso[1] **and Daniel M. Tartakovsky**[1]

[1] Department of Energy Resources Engineering, Stanford University, Stanford, CA, USA

Key Points:
- Information theory is used to up- and down-scale conductivity of heterogeneous formations
- Information theory recovers classic results for effective conductivity of uniform and convergent flows
- Information theory also yields statistical parameterization of effective properties

Correspondence to:
D. M. Tartakovsky,
tartakovsky@stanford.edu

Citation:
Boso, F., & Tartakovsky, D. M. (2018). Information-theoretic approach to bidirectional scaling. *Water Resources Research*, *54*, 4916–4928. https://doi.org/10.1029/2017WR021993

Received 2 OCT 2017
Accepted 10 MAY 2018
Accepted article online 29 MAY 2018
Published online 23 JUL 2018

Abstract We present an information-theoretic approach for integration of multiresolution data into multiscale simulations. This general framework is used to upscale and downscale equations of fluid flow in heterogeneous porous media. Fine-scale information can comprise observational data and/or simulation results related to both system states and system parameters. It is aggregated into its coarse-scale representation by setting a probabilistic equivalence between the two scales, with parameters that are determined via minimization of observable error and mutual information across scales. The same quantities facilitate the use of coarse-scale data to constrain compatible fine-scale distributions.

Plain Language Summary Information-theoretic arguments are used to construct probabilistic descriptions of effective hydraulic conductivity for mean uniform and convergent flows in heterogeneous formations. For classical uniform and convergent flow scenarios where the comparison with Monte Carlo is available, different coarse-scale parameter distributions are identified, and an overall quantification of uncertainty in the coarse-scale estimates is obtained.

1. Introduction

Fusion of multiresolution data with multiscale modeling and simulations is an essential component of modern science-based predictions. It poses a number of questions, such as how do experimental and/or simulation data collected at one scale inform model predictions at another scale? How much information is lost/gained in the process of moving upscale or downscale? and How does structural and/or parametric uncertainty propagate between scales? Considerable advancements have been made in relating smaller-scale models and their parameters to their larger-scale counterparts: various deterministic and stochastic upscaling strategies were deployed to relate pore-scale models to Darcy-scale models (e.g., Korneev & Battiato, 2016, and the references therein) and Darcy-scale models to field-scale ones (e.g., Neuman & Tartakovsky, 2009, among many others); to compute, analytically (e.g., Lichtner & Tartakovsky, 2003) or numerically (e.g., Efendiev & Pankov, 2004), effective coarse-scale parameters from their spatially varying fine-scale counterparts; and to propagate parametric uncertainty from the pore scale to the Darcy scale (e.g., Um et al., 2018) and from the Darcy scale to the field scale (e.g., de Barros & Rubin, 2011).

Progress in the opposite direction, from a large scale to a smaller one, is significantly more modest, in large part because of the lack of uniqueness (Wigmosta & Prasad, 2005). Downscaling aims to identify plausible fine-scale scenarios that are compatible with coarse-scale observations and/or numerical simulations. Given its intrinsic ill-posedness, downscaling employs mostly probabilistic techniques (such as refinement of spatial resolution, spatial interpolation/Kriging, and disaggregation of coarse-scale information) and, typically, requires auxiliary information (Blöschl, 2005), for example, knowledge of the spatial correlation of properties of interest. Downscaling is routinely used in atmospheric sciences and climatology (e.g., Fowler et al., 2007), for instance, to disaggregate remote sensing and precipitation data (e.g., Ferraris et al., 2003). Applications of downscaling techniques in subsurface modeling are appreciably more scarce; for example, downscaling procedures have been used as an inverse modeling tool (Hassane & Ackerer, 2017) and to estimate groundwater recharge (Allen et al., 2010).

While downscaling is data driven, upscaling is largely physics based with fine-scale measurements (e.g., of hydraulic conductivity) typically used to parameterize fine-scale models (e.g., a groundwater flow equation). Within the latter framework, measurements of system states (e.g., hydraulic head) on a given scale serve to either validate or calibrate a same-scale model. If explicit mappings between the fine- and coarse-scale models and between their corresponding parameters are available, for example, from homogenization, then data

assimilation techniques (various flavors of Kalman filters, Bayesian updating, etc.) allow one to pass information between the scales (Montzka et al., 2012; Zhu et al., 2017). Regardless of the data assimilation strategy adopted, system states and system parameters are treated as random fields/processes, which are specified in terms of their respective probability density functions (PDFs).

We posit that information theory (IT) provides a unifying framework for bidirectional (upscale and downscale) transfer of information (experimental or simulated data) and allows one to study questions similar to those raised in the opening paragraph of this article. IT uses probabilistic concepts to quantify the amount of information present in uncertain quantities. Specifically, information content at a given scale is quantified in terms of differential entropy, whose definition relies on the probabilistic description of a system in terms of (joint) PDFs of its system states and system parameters. These PDFs, which codify uncertain knowledge of the system behavior, are constructed from experimental measurements, numerical simulations, or both. They constitute a *training set* on which the *learning process* is based. Relevant information from the training set is then propagated to a different scale to obtain conditional probabilistic predictions. Relative entropy and mutual information are used to quantify the difference in information content between scales, while scaling is directed by the minimization of an average measure of discrepancy between estimates of quantities of interest (QoI) at both scales. Similar information-theoretic techniques have been used, mostly for upscaling, in computer science (e.g., Rose, 1998) with the purpose of reducing storage requirements, and in biology and material science to obtain bulk descriptions of molecular ensembles (e.g., Schöberl et al., 2017; Shell, 2008) or material properties (Koutsourelakis, 2007).

We employ information-theoretic tools to upscale and downscale flow fields in heterogeneous media, either in terms of state variables (i.e., hydraulic head) or parameters (i.e., hydraulic conductivity). Our goal is to identify probability distributions of effective variables (upscaling) and fine-scale quantities (downscaling) informed by data collected at the opposite scale. Section 2 contains a brief overview of information-theoretic concepts used to quantify information content at each scale and information transfer between scales; in this section we also deploy those concepts to minimize the discrepancy between predictions at the two scales. Our information-theoretic approach to bidirectional scaling and introduction of the novel discrepancy measure represent the main contributions of this work. In section 3, we formulate two examples of fluid flow in heterogeneous porous media with uncertain hydraulic conductivity, which are used to illustrate this general information-theoretic approach. The first deals with flow induced by externally imposed hydraulic head gradient, while the second concerns groundwater flow toward a pumping well. The upscaling and downscaling results for both problems are presented in sections 4 and 5, respectively. Major conclusions drawn from this study are summarized in section 6.

2. Information Theory and Bidirectional Scaling

We use lower- and upper-case letters to designate physical quantities on a fine scale and coarse scale, respectively. Thus, $\tilde{\mathbf{s}}$ and $\tilde{\mathbf{S}}$ denote state variables (e.g., fluid pressure or solute concentration at each element of a numerical mesh) at these two scales; their behavior is described by corresponding models m and M, with respective sets of parameters $\mathbf{p}$ and $\mathbf{P}$, such that

$$m(\tilde{\mathbf{s}};\mathbf{p}) = 0 \qquad \text{and} \qquad M(\tilde{\mathbf{S}};\mathbf{P}) = 0. \tag{1}$$

The models m and M can take the form of either an identical differential operator (e.g., the groundwater flow equation defined, respectively, on the Darcy and field scales) or distinct differential operators (e.g., Stokes equations defined on the pore scale, and a groundwater flow equation defined on the field scale). In the former case the parameters $\mathbf{p}$ and $\mathbf{P}$ represent upscaled/downscaled versions of each other (e.g., hydraulic conductivities defined on the Darcy and field scales), while in the latter case they are distinct sets (e.g., pore geometry used in the pore-scale simulations and hydraulic conductivity used in the field-scale model). Data scarcity and spatial heterogeneity are ubiquitous at any scale, rendering both model parameters, $\mathbf{p}$ and $\mathbf{P}$, and model predictions, $\tilde{\mathbf{s}}$ and $\tilde{\mathbf{S}}$, uncertain. We model this uncertainty in probabilistic terms, such that PDFs of $\mathbf{p}$ and $\mathbf{P}$ (denoted by $f_{\mathbf{p}}$ and $f_{\mathbf{P}}$) are inferred from either data or expert knowledge, while PDFs of $\tilde{\mathbf{s}}$ and $\tilde{\mathbf{S}}$ (denoted by $f_{\tilde{\mathbf{s}}}$ and $f_{\tilde{\mathbf{S}}}$) are computed by solving (1) with, for example, Monte Carlo simulations or more efficient uncertainty quantification techniques (Tartakovsky, 2016). Alternatively, both (or one of) $f_{\tilde{\mathbf{s}}}$ and (or) $f_{\tilde{\mathbf{S}}}$ can be inferred directly from measurements of the system states at their corresponding scale.

2.1. Basic Concepts of Information Theory

Let $\mathbf{s} = (\tilde{\mathbf{s}}, \mathbf{p})$ and $\mathbf{S} = (\tilde{\mathbf{S}}, \mathbf{P})$ denote complete sets of random variables (consisting of both the system parameters and the system states) at the fine and coarse scales, respectively. These so-called *augmented states* are characterized by corresponding joint PDFs $f_{\mathbf{s}}(\mathbf{s}') \equiv f_{\tilde{\mathbf{s}},\mathbf{p}}(\tilde{\mathbf{s}}', \mathbf{p}')$ and $f_{\mathbf{S}}(\mathbf{S}') \equiv f_{\tilde{\mathbf{S}},\mathbf{P}}(\tilde{\mathbf{S}}', \mathbf{P}')$, where the primes are used to indicate deterministic arguments (e.g., $\mathbf{s}'$) of the PDFs of the corresponding random variables (e.g., $\mathbf{s}$). *Differential entropy*, $\mathcal{H}(\mathbf{s})$, provides a measure of the fine-scale information content (e.g., Cover & Thomas, 2012). It is defined as

$$\mathcal{H}(\mathbf{s}) = -\int_{\Omega_{\mathbf{s}}} f_{\mathbf{s}}(\mathbf{s}') \ln\left[f_{\mathbf{s}}(\mathbf{s}')\right] \mathrm{d}\mathbf{s}', \tag{2}$$

with an analogous definition for coarse-scale differential entropy $\mathcal{H}(\mathbf{S})$, given in terms of PDF $f_{\mathbf{S}}$. The integration is over the domain of definition of the fine- and coarse-scale augmented states, $\Omega_{\mathbf{s}}$ and $\Omega_{\mathbf{S}}$, respectively. For example, if $\mathbf{p}$ and $\tilde{\mathbf{s}}$ represent two N-dimensional arrays of values of hydraulic conductivity and fluid pressure in N elements of a fine-scale numerical mesh, then $\Omega_{\mathbf{s}} = \mathbb{R}^{2N}$.

Let $f_{\mathbf{s},\mathbf{S}}(\mathbf{s}', \mathbf{S}')$ denote a joint PDF of the augmented states on the fine and coarse scales. Then the amount of information contained at both scales is quantified in terms of *mutual information* (e.g., Cover & Thomas, 2012),

$$\mathcal{I}(\mathbf{s}, \mathbf{S}) = \int_{\Omega_{\mathbf{s}}}\int_{\Omega_{\mathbf{S}}} f_{\mathbf{s},\mathbf{S}}(\mathbf{s}', \mathbf{S}') \ln\left[\frac{f_{\mathbf{s},\mathbf{S}}(\mathbf{s}', \mathbf{S}')}{f_{\mathbf{s}}(\mathbf{s}')f_{\mathbf{S}}(\mathbf{S}')}\right] \mathrm{d}\mathbf{s}'\mathrm{d}\mathbf{S}'. \tag{3}$$

If the fine- and coarse-scale augmented states $\mathbf{s}$ and $\mathbf{S}$ are statistically independent, then $f_{\mathbf{s},\mathbf{S}} = f_{\mathbf{s}}f_{\mathbf{S}}$ and (3) yields $\mathcal{I}(\mathbf{s}, \mathbf{S}) = 0$. In other words, if the fine- and coarse-scale models (augmented states) are fully decoupled, that is, statistically independent, then the mutual information between the scales is zero.

2.2. An Information-Theoretic Approach to Bidirectional Scaling

Multiscale modeling and data assimilation rely on the ability to estimate the relation between model predictions at the fine (m) and coarse (M) scales. In the probabilistic context, these predictions are expressed in terms of PDFs of the augmented system states $f_{\mathbf{s}}(\mathbf{s}')$ and $f_{\mathbf{S}}(\mathbf{S}')$, so this question translates into a question of how to quantify a distance between two functions. The task of defining such a distance is by no means trivial (or unique), and in our case it is complicated further by the fact that $f_{\mathbf{s}}(\mathbf{s}')$ and $f_{\mathbf{S}}(\mathbf{S}')$ are defined on different domains $\Omega_{\mathbf{s}}$ and $\Omega_{\mathbf{S}}$.

One possibility is to employ the Kullback-Leibler divergence D_{KL} as a measure of distance between two PDFs (see Appendix A for detail). This approach has been used in, for example, multiscale simulations of molecular dynamics (e.g., Schöberl et al., 2017; Shell, 2008). It requires knowledge of a map, $\hat{\mathbf{S}} = \mathcal{R}(\mathbf{s})$ with $\hat{\mathbf{S}} \in \Omega_{\mathbf{S}}$, between the states at different scales, which replaces $f_{\mathbf{s}}(\mathbf{s}')$ with $f_{\hat{\mathbf{S}}}(\hat{\mathbf{S}}')$ and renders $D_{\mathrm{KL}}(f_{\hat{\mathbf{S}}}, f_{\mathbf{S}})$ meaningful. If one were to deal only with state variables, then the operational definition of $\mathcal{R}$ is straightforward; for example, if $\mathbf{s}$ represents pore-scale concentration $c(\mathbf{x})$, then $\mathcal{R}$ represents a volume-averaging operator which yields Darcy-scale concentration $\hat{C} = \phi|\Delta|^{-1}\int_{\Delta} c(\mathbf{x})\mathrm{d}\mathbf{x}$, where ϕ is porosity and $|\Delta|$, the volume of Δ, defines the coarse scale. When information about parameter states is of interest and hence forms part of the augmented states, the mapping $\mathcal{R}$ and its inverse $\mathcal{R}^{-1}$ are unknown; finding them is a purpose of bidirectional scaling (upscaling and downscaling).

We, therefore, pursue an alternative strategy, which focuses on the difference (distance), $d(\hat{\boldsymbol{\mathcal{A}}}, \boldsymbol{\mathcal{A}}) = \|\hat{\boldsymbol{\mathcal{A}}}(\mathbf{s}) - \boldsymbol{\mathcal{A}}(\mathbf{S})\|$, between QoIs predicted with the fine-scale model, $\hat{\boldsymbol{\mathcal{A}}}(\mathbf{s})$, and its coarse-scale counterpart, $\boldsymbol{\mathcal{A}}(\mathbf{S})$. As usual, the choice of the norm $\|\cdot\|$ is somewhat arbitrary (Marošević, 1996). Clustering algorithms, whose goal is to reduce the size of data sets, employ a *squared error distortion* norm (Parekh et al., 2015), while Koutsourelakis (2007) used the ℓ^2 norm, $d(\hat{\boldsymbol{\mathcal{A}}}, \boldsymbol{\mathcal{A}}) = \sum_{i=1}^{M}(\hat{\mathcal{A}}_i - \mathcal{A}_i)^2$ for M point-wise QoIs $\boldsymbol{\mathcal{A}} = (\mathcal{A}_i, \ldots, \mathcal{A}_M)$, to upscale descriptive material properties. We generalize the latter approach by assigning weights w_i ($i = 1, \ldots, M$) to individual point-wise QoIs,

$$d_w(\hat{\boldsymbol{\mathcal{A}}}, \boldsymbol{\mathcal{A}}) = \sum_{i=1}^{M} w_i[\hat{\mathcal{A}}_i(\mathbf{s}) - \mathcal{A}_i(\mathbf{S})]^2. \tag{4a}$$

These weights can be defined, for example, in terms of the means, $\langle\hat{\mathcal{A}}_i\rangle$ or $\langle\mathcal{A}_i\rangle$, or the standard deviations, $\sigma_{\hat{\mathcal{A}}}$ or $\sigma_{\mathcal{A}}$. We choose a dependence on the means, such that either

$$w_i = \frac{1}{\langle\hat{\mathcal{A}}_i\rangle^2} \quad \text{with} \quad \langle\hat{\mathcal{A}}_i\rangle^2 = \int \hat{\mathcal{A}}_i(\mathbf{s}')f_{\mathbf{s}}(\mathbf{s}')\mathrm{d}\mathbf{s}' \tag{4b}$$

or

$$w_i = \frac{1}{\langle \mathcal{A}_i \rangle^2} \quad \text{with} \quad \langle \mathcal{A}_i \rangle^2 = \int \mathcal{A}_i(\mathbf{S}') f_{\mathbf{S}}(\mathbf{S}') \mathrm{d}\mathbf{S}', \tag{4c}$$

depending on the scale at which the data are available, that is, on whether upscaling or downscaling is to be performed.

A goal of bidirectional scaling (upscaling or downscaling) can then be thought of as finding (joint) PDFs of $\mathbf{s}$ and $\mathbf{S}$, which minimize the expected value (mean) of the distance,

$$\langle d_w(\hat{\mathcal{A}}, \mathcal{A}) \rangle = \sum_{i=1}^{M} w_i \int_{\Omega_{\mathbf{S}}} \int_{\Omega_{\mathbf{s}}} [\hat{\mathcal{A}}_i(\mathbf{s}') - \mathcal{A}_i(\mathbf{S}')]^2 f_{\mathbf{s},\mathbf{S}}(\mathbf{s}', \mathbf{S}') \, \mathrm{d}\mathbf{s}' \mathrm{d}\mathbf{S}'. \tag{5}$$

Functional minimization problems of this kind are notoriously challenging because of the dependence of d_w on noisy Monte Carlo estimates. Following the standard practice (Rose, 1998), we regularize this problem by replacing the goal of finding the minimum of $\langle d_w \rangle$ in (5) with an optimization problem

$$\underset{f_{\mathbf{s},\mathbf{S}}}{\operatorname{argmin}} \, \mathcal{J}(\mathbf{s}, \mathbf{S}), \qquad \mathcal{J} \equiv \mathcal{I}(\mathbf{s}, \mathbf{S}) + \lambda \left\{ \langle d_w[\hat{\mathcal{A}}(\mathbf{s}), \mathcal{A}(\mathbf{S})] \rangle - d_0 \right\}, \tag{6}$$

where λ is a Lagrange multiplier, d_0 is a prescribed mean discrepancy or scaling error, and $\mathcal{I}(\mathbf{s}, \mathbf{S})$ is the mutual information defined in (3). In upscaling applications, the fine-scale PDF $f_{\mathbf{s}}$ is known and, since $f_{\mathbf{s},\mathbf{S}} = f_{\mathbf{S}|\mathbf{s}} f_{\mathbf{s}}$, the optimization is over the conditional PDF $f_{\mathbf{S}|\mathbf{s}}$. The situation is reversed for downscaling, wherein $f_{\mathbf{S}}$ is accessible and the optimization is over the conditional PDF $f_{\mathbf{s}|\mathbf{S}}$. In either case, the goal is to minimize the average discrepancy between the QoIs inferred from the fine- and coarse-scale models, while keeping the information transfer across the scales (as quantified by the mutual information) to a minimum.

For upscaling, we set to 0 the functional derivative of $\mathcal{J}$ with respect to $f_{\mathbf{S}|\mathbf{s}}(\mathbf{S}'; \mathbf{s})$, that is, the PDF for the augmented coarse-scale state $\mathbf{S}$ conditioned on the augmented fine-scale state $\mathbf{s}$. This yields a solution of (6) in the form of a Gibbs distribution (Rose, 1998),

$$f_{\mathbf{S}|\mathbf{s}}(\mathbf{S}'; \mathbf{s}) = \frac{f_{\mathbf{S}} \, \mathrm{e}^{-\lambda d_w[\hat{\mathcal{A}}(\mathbf{s}), \mathcal{A}(\mathbf{S}')]}}{\int_{\Omega_{\mathbf{S}}} f_{\mathbf{S}}(\mathbf{S}') \, \mathrm{e}^{-\lambda d_w} \mathrm{d}\mathbf{S}'}, \tag{7}$$

where $f_{\mathbf{S}}(\mathbf{S}')$ is an unknown prior coarse-scale distribution. This formal solution holds for any distance $d(\hat{\mathcal{A}}, \mathcal{A})$, including d_w given by (4). We use deterministic annealing (see Appendix B for detail) to compute $f_{\mathbf{S}}$ and, hence, $f_{\mathbf{S}|\mathbf{s}}(\mathbf{S}'; \mathbf{s})$ in (7). This procedure carries high computational cost but yields a nonparametric estimation of $f_{\mathbf{S}|\mathbf{s}}(\mathbf{S}'; \mathbf{s})$, that is, it does not require any assumptions about its functional form. Another attractive feature of this approach (as opposed to, say, the postprocessing of Monte Carlo realizations) is its ability to directly recover (joint) PDFs of the coarse-scale parameters regardless of the existence of an analytical relation between the QoIs and the effective parameters.

For downscaling, we minimize (6) with respect to $f_{\mathbf{s}|\mathbf{S}}$, which gives an expression for this conditional PDF in the form of a Gibbs distribution,

$$f_{\mathbf{s}|\mathbf{S}}(\mathbf{s}'; \mathbf{S}) = \frac{f_{\mathbf{s}} \, \mathrm{e}^{-\lambda d_w[\hat{\mathcal{A}}(\mathbf{s}), \mathcal{A}(\mathbf{S}')]}}{\int_{\Omega_{\mathbf{s}}} f_{\mathbf{s}}(\mathbf{s}') \, \mathrm{e}^{-\lambda d_w} \mathrm{d}\mathbf{s}'}, \tag{8}$$

where $f_{\mathbf{s}}(\mathbf{s}')$ is an unknown prior fine-scale distribution. Substituting (8) into (6) yields

$$\mathcal{J}_{\mathrm{ds}}(\mathbf{s}, \mathbf{S}) = -\frac{1}{\lambda} \int_{\Omega_{\mathbf{S}}} f_{\mathbf{S}}(\mathbf{S}') \ln \left(\int_{\Omega_{\mathbf{s}}} f_{\mathbf{s}}(\mathbf{s}') \, \mathrm{e}^{-\lambda d_w(\hat{\mathcal{A}}(\mathbf{s}'), \mathcal{A}(\mathbf{S}'))} \mathrm{d}\mathbf{s}' \right) \mathrm{d}\mathbf{S}'. \tag{9}$$

We use parametric estimation to identify $f_{\mathbf{s}|\mathbf{S}}(\mathbf{s}', \mathbf{S}')$, thus assigning to the prior $f_{\mathbf{s}}(\mathbf{s})$ a functional form and identifying its parameters via minimization of (9). This strategy enables us to reduce the computational cost of evaluating the quadrature over $\Omega_{\mathbf{s}}$. Expert knowledge and/or site-specific information can be used to guide the selection of $f_{\mathbf{s}}$ (Blöschl, 2005); for example, it is common in subsurface hydrology to assume that hydraulic conductivity is multivariate lognormal and hydraulic head is multivariate normal (PDFs of these and other physicochemical QoIs are collated in Table 1 in Tartakovsky, 2013).

3. Formulation of Computational Examples

We assume that fluid flow models at fine and coarse scales, m and M in (1), are the same and consist of steady-state two-dimensional groundwater flow equations,

$$\nabla \cdot (k\nabla h) = \rho(\mathbf{x}) \qquad \text{and} \qquad \nabla \cdot (K\nabla H) = R(\mathbf{X}). \tag{10}$$

Here the spatial coordinates $\mathbf{x}$ and $\mathbf{X}$ represent centroids of the volumes of a heterogeneous porous medium corresponding to the fine and coarse scales, respectively; $k(\mathbf{x})$, $h(\mathbf{x})$, and $\rho(\mathbf{x})$ denote fine-scale hydraulic conductivity, hydraulic head, and sources/sinks; and $K(\mathbf{X})$, $H(\mathbf{X})$, and $R(\mathbf{X})$ stand for their coarse-scale counterparts. Fine- and coarse-scale Darcy velocities are computed as $\mathbf{u}(\mathbf{x}) = -k(\mathbf{x})\nabla h(\mathbf{x})$ and $\mathbf{U}(\mathbf{X}) = -K(\mathbf{X})\nabla H(\mathbf{X})$.

Hydraulic conductivity at both scales is uncertain. We define the coarse scale as a scale on which a single conductivity value K is assigned over the whole computational domain. Since this value is uncertain, and treated as random, it is characterized by a PDF f_K. When used in the upscaling mode, the fine-scale conductivity field $k(\mathbf{x})$ is assumed to be lognormal, such that log-conductivity $Y = \ln k$ has mean $\bar{Y} = 0.0$, variance $\sigma_Y^2 = 1.0$, and an isotropic exponential autocorrelation function $\rho_Y(\mathbf{x}, \mathbf{x}') = \exp\left(-\sqrt{(x_1 - x_1')^2 + (x_2 - x_2')^2}/\ell_Y\right)$ with correlation length $\ell_Y = 0.05$ (these and other physical quantities are defined in consistent units). This choice of the statistical properties of $k(\mathbf{x})$ is for illustrative purposes only and can be replaced with other statistical models. We use `Hydro_gen` (Bellin & Rubin, 1996) to generate $N_r = 100$ realizations of the fine-scale conductivity field $k(\mathbf{x})$ on a discretized staggered grid of size $\Delta x < \ell_Y$, and `MODFLOW` (Harbaugh et al., 2000) to solve the fine-scale flow equation in (10). Representative realizations of fine-scale hydraulic conductivity $k(\mathbf{x})$ and hydraulic head $h(\mathbf{x})$ for two flow scenarios considered below are shown in Figure 1.

3.1. Mean Uniform Flow

Flow takes place in a square domain $\mathcal{D} = \{\mathbf{x} = (x_1, x_2)^\top : 0 \le x_1 \le L, 0 \le x_2 \le L\}$ with $L = 1.0$ in the absence of sources $\rho = R = 0$. Hydraulic heads H_1 and H_2 imposed at the vertical Dirichlet boundaries, $x_1 = 0$ and $x_1 = L$, combined with the impermeable horizontal boundaries, $x_2 = 0$ and $x_2 = L$, result in a mean uniform hydraulic gradient of $(H_2 - H_1)/L = 0.01$ in the longitudinal direction. The top row in Figure 1 shows a representative realization of the fine-scale hydraulic conductivity $k(\mathbf{x})$ and the resulting hydraulic head $h(\mathbf{x})$ computed on a uniform square mesh of size $\Delta x = 1/64$. The coarse-scale model with uniform, albeit uncertain, hydraulic conductivity K admits an analytical solution,

$$H(\mathbf{X}) = -\frac{H_2 - H_1}{L} X_1, \tag{11}$$

leading to the constant coarse-scale flow velocity in the horizontal direction, $U = K(H_1 - H_2)/L$, and the volumetric flux $\mathcal{Q} = UL$.

3.2. Radially Convergent Flow

Flow takes place in a circular domain $\mathcal{D} = \{\mathbf{x} = (x_1, x_2)^\top : r^2 \equiv x_1^2 + x_2^2 \le r_e^2\}$ with $r_e = 0.5$. It is driven by groundwater withdrawal with rate $Q = -1.0$ from a well located at the center of the domain; constant hydraulic head H_e is imposed at the external boundary, $h(r = r_e) = H_e = 1.0$. The bottom row in Figure 1 shows a representative realization of the fine-scale hydraulic conductivity $k(\mathbf{x})$ and the resulting hydraulic head $h(\mathbf{x})$ computed on a uniform square mesh of size $\Delta x = 1/101$. The coarse-scale model with uniform, albeit uncertain, hydraulic conductivity K admits an analytical solution,

$$H(\mathbf{X}) = H_e - \frac{Q}{4\pi K^2} \ln\left(\frac{X_1^2 + X_2^2}{r_e^2}\right). \tag{12}$$

Upscaling of convergent flows in randomly heterogeneous porous media has been extensively investigated (e.g., Sanchez-Vila & Tartakovsky, 2007, and the literature therein). A key quantity of interest in such studies is well productivity $\mathcal{W}$, defined as (Durlofsky, 2000)

$$\mathcal{W} \equiv -\frac{Q}{H_e - H_w} = \frac{2\pi K}{\ln(r_e/r_w)}. \tag{13}$$

Here H_w and r_w are the hydraulic head in the well and the well radius, respectively. Following Peaceman (1978), we set $r_w = 0.2\Delta x$.

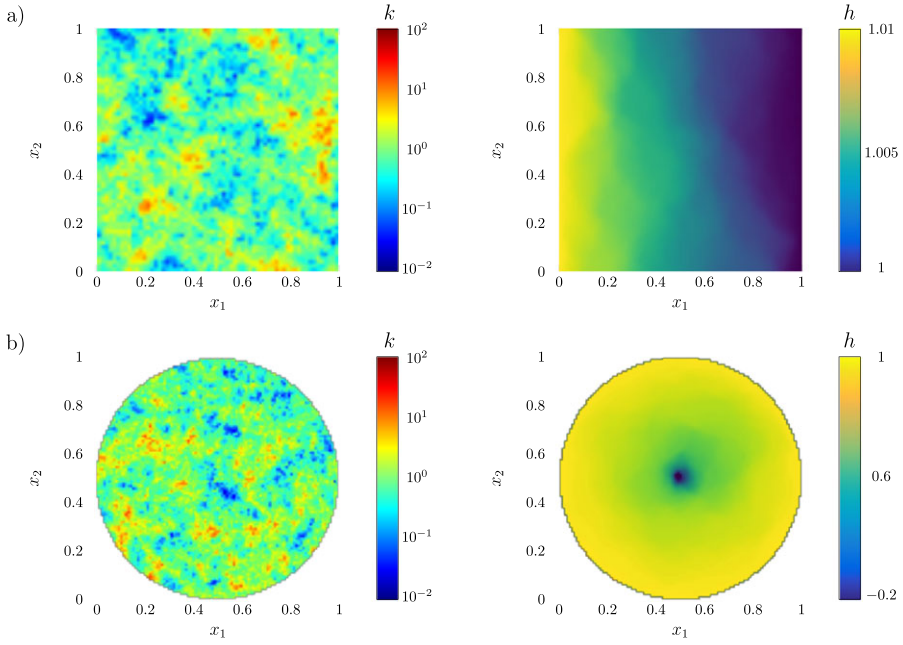

Figure 1. Typical realizations of the fine-scale hydraulic conductivity $k(\mathbf{x})$ (left column) and hydraulic head $h(\mathbf{x})$ (right column) for (a) the mean uniform (top row) and (b) convergent (bottom row) flow regimes.

4. Information-Theoretic Approach to Upscaling

For both flow scenarios, our goal is to upscale the fine-scale hydraulic conductivity $k(\mathbf{x})$ to the coarse scale, that is, to estimate f_K, the PDF of the coarse-scale conductivity K, from its fine-scale counterpart. The discretization of the flow domains $\mathcal{D}$ into N_{el} square elements translates into the fine-scale parameter set $\mathbf{p} = \{k_1, \dots, k_{N_{\text{el}}}\}$; the coarse-scale parameter set consists of a single member, $\mathbf{P} = \{K\}$. The mean uniform flow example has one QoI (the total flux through the domain), while the radially convergent flow example has two (the well productivity and the volume of the depression cone around the well).

4.1. Mean Uniform Flow

For the flow conditions described in section 3.1, our QoI is the total volumetric flux through the flow domain $\mathcal{D}$. Its coarse-scale prediction is given by $Q = K(H_1 - H_2)$ with uncertain K, while its fine-scale estimate is computed as $\hat{Q} = \int_0^L u_1(L, x_2)\mathrm{d}x_2$. Our goal is to obtain a probabilistic description of unknown K, that is, its PDF f_K, via minimization of (6) wherein $d_w = (\hat{Q} - Q)^2$.

We start by generating fine-scale raw data consisting of N_r solutions of the fine-scale model (10) with different realization of the parameter $\mathbf{p}$, and computing the corresponding values of $\hat{Q}$. Next, we use deterministic annealing (see Appendix B) to solve the optimization problem (6). For the discrepancy level set to $d_0 = 1.5 \times 10^{-7}$, this procedure results in a discrete approximation of $f_K(K')$, which is shown by the solid line in Figure 2a. This value of d_0 is achieved by discretizing $f_K(K')$ with $N_{\text{dis}} = 35$ points. One can decrease the computational burden of deterministic annealing by reducing N_{dis} at the cost of increasing the discrepancy level; for example, using $N_{\text{dis}} = 19$ discretization points results in $d_0 = 1.6 \times 10^{-6}$ and leads to the estimate of $f_K(K')$ exhibited in Figure 2b.

Mean uniform flow provides a rare example of subsurface models in which an analytical (and invertible) map between the coarse-scale QoI Q and the coarse-scale parameter K is available; in our example, they are linearly related! This map enables one to represent $F_K(K')$ as a histogram of K values computed from individual realizations of the QoI $\hat{Q}(k)$ by setting $\hat{Q} = Q$. Figure 2 demonstrates the consistency between the resulting histogram and our information-theoretic estimate of $F_K(K')$. Also shown in this figure is a histogram of the geometric means $K_g = \exp(\sum_{i=1}^{N_{\text{el}}} k_i)$ computed for each realization of the fine-scale conductivity field $k(\mathbf{x})$, discretized into N_{el} square elements. The theoretical result of $K = K_g$ is strictly valid for two-dimensional flow domains of infinite extent (Matheron, 1967) or, as in the present test case, when $\ell_Y \leq L$ so that the presence of boundaries has a negligible effect on K (Paleologos et al., 1996).

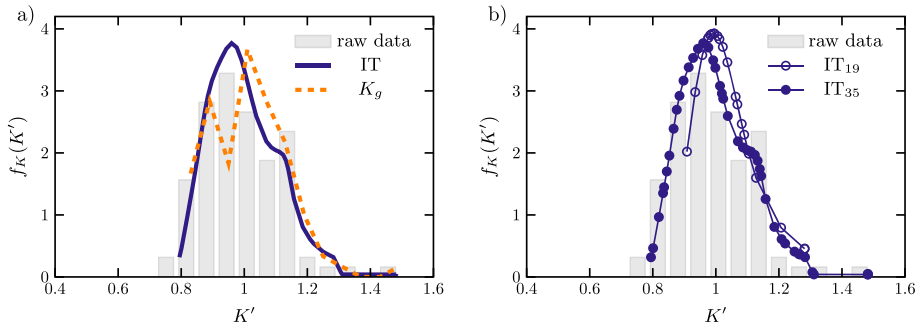

Figure 2. Probability density functions, $f_K(K')$, of the effective (coarse scale) hydraulic conductivity K for the mean uniform flow. (a) The estimates of $f_K(K')$ predicted with the information-theoretic approach (solid line), histogram of K's resulting from $N_r = 100$ fine-scale flow simulations (bars), and histogram of geometric mean (K_g, dashed line) estimated from each realization of $k(\mathbf{x})$. (b) Comparison of the estimates of $f_K(K')$ computed via deterministic annealing with $N_{\text{dis}} = 35$ and 19 discretization points, whose locations are shown by the filled and open circles, respectively.

4.2. Radially Convergent Flow

For the flow conditions described in section 3.2, we have two QoIs, $\boldsymbol{\mathcal{A}} = \{\mathcal{A}_1, \mathcal{A}_2\}$. The first is the well productivity $\mathcal{A}_1 \equiv \mathcal{W}$, which on the coarse scale is computed from (12) as $\mathcal{W} = 2\pi K/\ln(r_e/r_w)$, and on the fine scale as $\hat{\mathcal{W}} = -Q/(H_e - h_w)$ where the hydraulic head at the well walls h_w is computed by solving (10). The second QoI is the volume of the depression cone around the well, $\mathcal{A}_2 \equiv \mathcal{V}$, whose fine- and coarse-scale representations are

$$\hat{\mathcal{V}} = \int_D H_e - h(\mathbf{x})\mathrm{d}\mathbf{x} \quad \text{and} \quad \mathcal{V} = 2\pi \int_{r_w}^{r_e} \frac{Qr}{2\pi K} \ln\left(r/r_e\right) \mathrm{d}r \approx -\frac{Qr_e^2}{4K}, \tag{14}$$

respectively.

Estimates of the PDF $f_K(K')$, obtained with our information-theoretic approach for one ($\mathcal{A} = \mathcal{W}$) and two ($\mathcal{A}_1 = \mathcal{W}$ and $\mathcal{A}_2 = \mathcal{V}$) QoIs, are shown in Figures 3a and 3b, respectively. In this flow regime, the reliance on two QoIs slightly reduces the predictive uncertainty about K: the PDF $f_K(K')$ in Figure 3a, computed with our IT approach, is wider than that in Figure 3b. This result also demonstrates a well-known observation that effective (upscaled) parameters depend not only on their fine-scale counterparts and flow regime but also on QoIs used to establish equivalency between scales.

The radial flow regime is another example where explicit maps between QoIs and coarse-scale parameter(s) are available. These are constructed by equating their fine- and coarse-scale estimates, $\hat{\mathcal{V}} = \mathcal{V}$ and $\hat{\mathcal{W}} = \mathcal{W}$. This yields realizations of K corresponding directly to realizations of the fine-scale solutions of the flow equation (raw data); the histogram of these K's closely matches the information-theoretic estimate of $f_K(K')$ in the case of one QoI (Figure 3a) but is qualitatively different when two QoIs are present (Figure 3b).

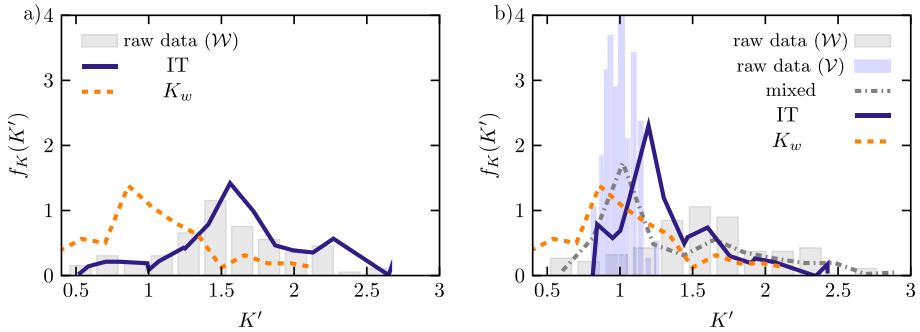

Figure 3. Probability density function, $f_K(K')$, of the effective (coarse scale) hydraulic conductivity K for the radially convergent flow in the case of one (a) and two (b) quantities of interests. The estimates of $f_K(K')$ are alternatively obtained with the information-theoretic approach (IT, solid line), histograms of the raw data (samples of K) resulting from $N_r = 100$ fine-scale flow simulations (bars), a mixture of the histograms for two quantities of interest (dash-dotted line), and histograms of the weighted averages estimated from each realization of $k(\mathbf{x})$ (K_w, dashed line).

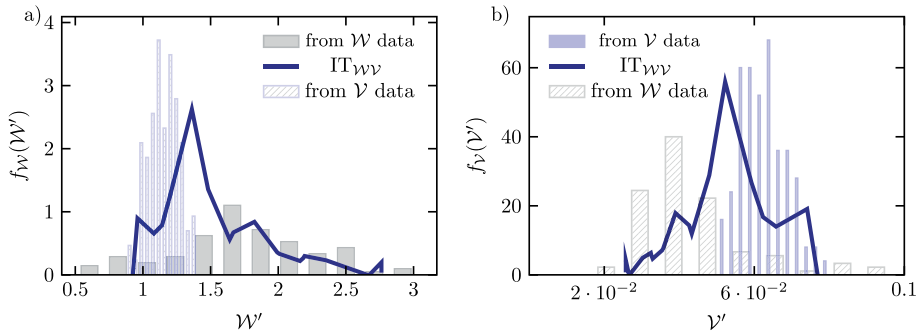

Figure 4. (a) Frequency distributions for the well productivity $\mathcal{W}$ obtained directly from the fine-scale simulations (from $\mathcal{W}$ data) and inferred from the depression cone volume data $\mathcal{V}$ (from $\mathcal{V}$ data), as well as the probability density function $f_{\mathcal{W}}(\mathcal{W}')$ estimated with the information-theoretic approach for $\mathcal{A} = \{\mathcal{W}, \mathcal{V}\}$. (b) Frequency distributions for the depression cone volume $\mathcal{V}$ obtained directly from the fine-scale simulations (from $\mathcal{V}$ data) and inferred from the well productivity data $\mathcal{W}$ (from $\mathcal{W}$ data), as well as the probability density function $f_{\mathcal{V}}(\mathcal{V}')$ estimated with the information-theoretic approach for $\mathcal{A} = \{\mathcal{W}, \mathcal{V}\}$.

That is because the Monte Carlo-like approach used to generate these histograms aims to minimize the discrepancy with one QoI at a time, resulting in a large discrepancy with the other QoI. One can alleviate this problem by using a weighted sum of the two histograms (the dash-dotted line in Figure 3b with both weights equal to 1/2), which has an effect of looking for a compromise between the two discrepancies, just as the information-theoretic approach does.

Empirical studies (Desbarats, 1992) suggest the use of a weighted spatial average K_w,

$$\ln K_w = \frac{1}{|\mathcal{D}|}\int_{\mathcal{D}} \frac{\ln k(\mathbf{x})}{x_1^2 + x_2^2}\,\mathrm{d}\mathbf{x}, \qquad |\mathcal{D}| = \int_{\mathcal{D}} \frac{\mathrm{d}\mathbf{x}}{x_1^2 + x_2^2}, \tag{15}$$

as an effective (coarse-scale) representation of the fine-scale conductivity $k(\mathbf{x})$. The frequency distribution of K_w estimates from $N_r = 100$ realizations of $k(\mathbf{x})$ is shown by the dashed line in Figure 3b. The performance of this empirical mapping improves when both QoIs are taken into account. The effective conductivity K_w accounting solely for the well productivity $\mathcal{W}$ is inadequate because of the local nature of the latter quantity, which depends on the hydraulic head at the well; K_w becomes more representative of the whole formation when the depression cone volume $\mathcal{V}$ is accounted for as well.

The estimates of $f_K(K')$ shown in Figure 3 can be used to compute PDFs of the QoIs via the coarse-scale model (10). These PDFs quantify uncertainty in estimates of the QoIs obtained from the coarse-scale model, taking into account all the available information. We compare the PDFs $f_{\mathcal{W}}(\mathcal{W}')$ and $f_{\mathcal{V}}(\mathcal{V}')$ predicted by our information-theoretic approach, (4) and (6), with the raw data for the well productivity $\mathcal{W}$ and the volume of the depression cone $\mathcal{V}$ (histograms with filled bars in Figure 4). A Monte Carlo-like estimate of the PDF of one QoI, for example, $f_{\mathcal{W}}(\mathcal{W}')$, is appreciably off when using the raw data of the other QoI, for example, $\mathcal{V}$, (histograms with patterned bars in Figure 4), with our PDF predictions striking a balance between the two.

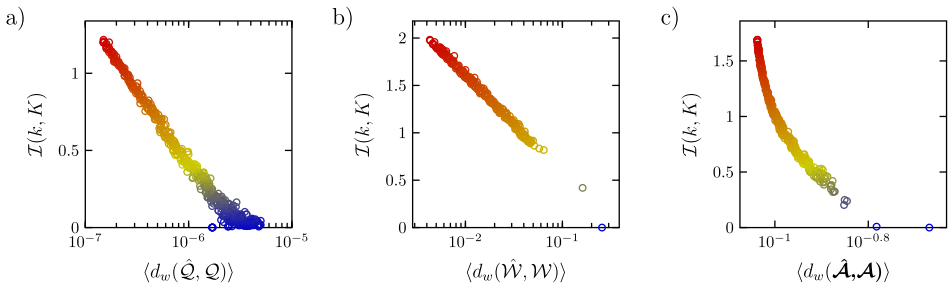

Figure 5. Scatter plots of the average distortion rates $\langle d_w \rangle$ and the mutual information $\mathcal{I}(k, K)$ for (a) the mean uniform flow, $d_w = d_w(\hat{\mathcal{Q}}, \mathcal{Q})$; (b) the radial flow with one quantities of interest, $d_w = d_w(\hat{\mathcal{W}}, \mathcal{W})$; and (c) the radial flow with two quantities of interests, $d_w = d_w(\hat{\mathcal{A}}, \mathcal{A})$ where $\mathcal{A} = \{\mathcal{W}, \mathcal{V}\}$. Iteration of the deterministic annealing progresses from blue to red, achieving better accuracy (i.e., smaller $\langle d_w \rangle$) and approximating the posterior probability density functions f_K with a larger number of discretization points N_{dis}.

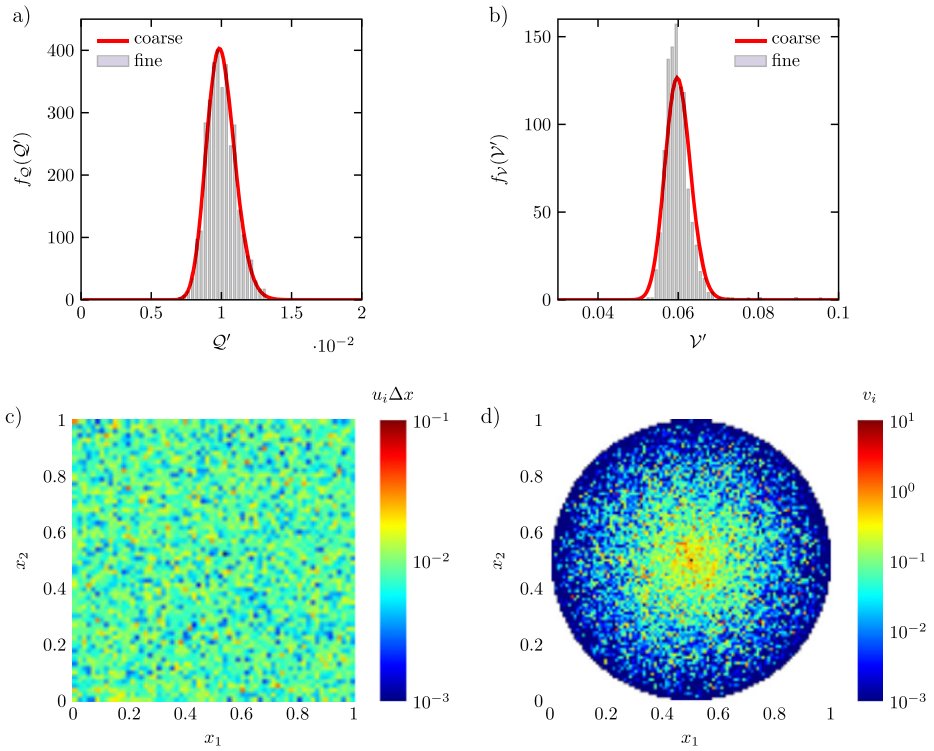

Figure 6. *Top:* Probability density functions of the fine- and coarse-scale predictions of the quantities of interests for the mean uniform flow (left) and the convergent flow (right). At the coarse scale, the distribution is provided by either simulations or measurements (solid line). The histograms representing the quantities of interests are obtained by generating one fine-scale realization with the statistical properties inferred by the information-theoretic downscaling procedure. *Bottom:* Representative spatial distributions of $u_i \Delta x$ (left) and v_i (right).

The magnitude of the discrepancy level d_0 is controlled by the number of discretization points N_{dis} used in deterministic annealing (Figure 2b). In general, the average discrepancy $\langle d_w(\hat{\boldsymbol{\mathcal{A}}}, \boldsymbol{\mathcal{A}}) \rangle$ decreases and the mutual information $\mathcal{I}(k, K)$ increases as N_{dis} becomes larger. This finding is illustrated by the scatter plots in Figure 5, wherein the colors become warmer as the iteration of deterministic annealing progresses (N_{dis} increases, i.e., d_0 decreases). The increase in $\mathcal{I}$ indicates that larger amount of information is passed between scales in order to achieve better overall accuracy of the coarse-scale model. At convergence, $\mathcal{I}$ quantifies the amount of information, which is transferred from the fine scale to the coarse scale upon upscaling.

5. Information-Theoretic Approach to Downscaling

Stochastic downscaling aims to identify the joint PDF of the fine-scale state variables and/or parameters that are compatible with available coarse-scale data. Coarse-scale (experimental or simulation) data seldom include coarse-scale (effective) parameters, since those often cannot be directly observed. That is because their physical meaning is conferred on them by a coarse-scale model (e.g., hydraulic conductivity, as opposed to fine-scale properties such as pore geometry that can be measured directly). Hence, we focus on downscaling of state variables.

5.1. Mean Uniform Flow

A coarse-scale model of the mean uniform flow regime is described in section 3.1. The coarse-scale quantity U_1, the longitudinal Darcy velocity across the domain $\mathcal{D}$, is measured to have a lognormal PDF $f_{U_1}(U_1')$ with mean $\mu_{\ln U_1}$ and variance $\sigma^2_{\ln U_1}$, giving rise to the longitudinal volumetric flux by $\mathcal{Q} = U_1 L$ whose PDF is shown in Figure 6a (solid line). It is compatible with a coarse-scale hydraulic conductivity K, a constant whose uncertain value is assigned a lognormal PDF $f_K(K')$. Our goal is to identify a PDF of the fine-scale Darcy velocity, $u_1(L, x_2)$, at the outlet $x_1 = L$, which provides a fine-scale estimate of the volumetric flux through the outlet, $\hat{\mathcal{Q}} = \int_0^L u_1(L, x_2) \mathrm{d}x_2$, and, by continuity, in all transverse cross sections of $\mathcal{D}$.

Using a numerical solution of the fine-scale flow field, $\hat{Q} = \Delta x \sum_{i=1}^{N_2} u_{1,i}$ where $N_2 = L/\Delta x$ is the number of cells in the x_2 direction along the outflow boundary, and $u_{1,i}$ is the longitudinal velocity in the ith cell. We assume $u_{1,i}$ to be identical and independently distributed (i.i.d.) lognormal random variables (RVs). Then the goal is identify their mean and variance, μ_i and σ_i^2, which minimize (9). (Alternatively, one could assume these RVs to be correlated and treat the correlation length as another optimization variable.)

For $d_w(\hat{Q}, Q) = (\ln \hat{Q} - \ln Q)^2$, we employ the Fenton-Wilkinson approximation (e.g., Cobb et al., 2012) for the summation of i.i.d. lognormal RVs to obtain a closed-form solution of this optimization problem,

$$\mu_i = \mu_{\ln Q} - \frac{1}{2}(\sigma_i^2 - \sigma_{\ln Q}^2) - \ln(\Delta x), \qquad \sigma_i^2 = \ln\left[-N_2(1 - e^{\sigma_{\ln Q}^2}) + 1\right]. \tag{16}$$

Generating fine-scale samples of u_i with these statistical properties yields a frequency distribution for $\hat{Q}$ that is in agreement with the coarse-scale input data (histogram in Figure 6a). N_2 realizations of u_i constitute one realization of a random field; generating a random field with the same statistical properties for each transverse section, we obtain a random field of the fine-scale longitudinal fluxes in $\mathcal{D}$ (Figure 6c).

5.2. Convergent Flow

For the convergent flow regime described in section 3.2, we consider downscaling based on the depression-cone volume $\mathcal{V}$ whose PDF is estimated from coarse-scale measurements or simulations. For a homogeneous coarse-scale hydraulic conductivity K with lognormal PDF, both hydraulic head at the well, H_w, and depression-cone volume $\mathcal{V}$ have lognormal PDFs (solid line in Figure 6b). Our goal is to identify a compatible PDF of the fine-scale hydraulic heads within the flow domain $\mathcal{D}$ via minimization of (9).

The volume of the depression cone is approximated from the fine scale by

$$\hat{\mathcal{V}} = \Delta x^2 \sum_{i=1}^{N} v_i, \tag{17}$$

where $v_i = H_e - h_i$ is the drawdown in each of the N cells with area Δx^2. As before, we treat v_i as independent. RVs and approximate the sum of independent. lognormal RVs with the Fenton-Wilkinson formula. Then, the optimization problem (9) admits an analytical solution for the uniform variance of the lognormal PDF of v_i,

$$\sigma_i^2 = \frac{1}{N}\left[\exp(\sigma_{\ln \mathcal{V}}^2) - 1\right]\left[\exp\left(\mu_{\ln \mathcal{V}} - \ln(\Delta x^2) + \frac{\sigma_{\ln \mathcal{V}}^2}{2}\right)\right]^2, \tag{18}$$

whose mean is given by the coarse-scale solution; and, hence a lognormal PDF of h_i. Here $\mu_{\ln \mathcal{V}}$ and $\sigma_{\ln \mathcal{V}}^2$ are the prescribed the mean and variance of the coarse-scale PDF of $\mathcal{V}$ in Figure 6b. This figure demonstrates an agreement between this PDF (solid line) and that of $\hat{\mathcal{V}}$ in (17)–(18) (histogram). Figure 6d exhibits a fine-scale lognormal field v_i whose statistics are given by (18). The fact that thus inferred $v(\mathbf{x})$ (Figure 6d) displays stronger variability than $u(\mathbf{x})$ (Figure 6c) reveals how the choice of observables determines the uncertainty level in predictions of downscaled quantities.

6. Summary and Conclusions

We present a probabilistic framework, which uses information-theoretic concepts to transfer information from a fine scale to a coarse scale (upscaling) or from a coarse scale to a fine scale (downscaling). The bidirectional scaling is guided by the functional minimization of a modified cost function, where mutual information between scales is minimized subject to a constrain imposed by a weighted-average discrepancy measure between the fine- and coarse-scale model predictions for a set of QoIs.

The optimization yields conditional PDFs for system states and/or system parameters at either scale, taking into account the available information at the opposite scale. The procedure enables one to evaluate the accuracy of the scaled predictions in terms of the average discrepancy and PDF of the observables. It also allows one to quantify both the loss of information upon upscaling and the redistribution of information upon downscaling: mutual information measures the difference in information content between the scaled description of (augmented) states and their probabilistic description if they were independent.

The performance and challenges of our approach are discussed in the context of two examples of groundwater flow in heterogeneous formations: mean uniform flow and radial convergent flow. These examples have been used in the past to derive theoretical and empirical expressions for the effective conductivity, respectively. They provide familiar settings to study the assimilation of conductivity measurements collected

at different scales and the quantification of differences in the information content provided by the fine- and coarse-scale models. Future work will explore applications of this approach to dynamical systems and to the inference (learning) of coarse-scale models.

Appendix A: Kullback-Leibler Divergence and Distance Between Two PDFs

A distance between any two PDFs $f_1(\mathbf{s})$ and $f_2(\mathbf{s})$, both defined on the same domain $\Omega_{\mathbf{s}}$, can be quantified in terms of the *Kullback-Leibler (KL) divergence*

$$D_{\mathrm{KL}}(f_1, f_2) \equiv \int_{\Omega_{\mathbf{s}}} f_1(\mathbf{s}') \ln\left[\frac{f_1(\mathbf{s}')}{f_2(\mathbf{s}')}\right] \mathrm{d}\mathbf{s}'. \tag{A1}$$

It is also referred to as *relative entropy* between two PDFs, $\mathcal{S}_{\mathrm{rel}} = D_{\mathrm{KL}}(f_1, f_2)$. In the context of bidirectional scaling, we deal with $f_{\mathbf{s}}(\mathbf{s}')$ and $f_{\mathbf{S}}(\mathbf{S}')$, PDFs of the fine- and coarse-scale states $\mathbf{s}$ and $\mathbf{S}$, respectively. The KL divergence cannot be used to compare these PDFs, since $\mathbf{s}$ and $\mathbf{S}$ have different dimensions and, correspondingly, $f_{\mathbf{s}}(\mathbf{s}')$ and $f_{\mathbf{S}}(\mathbf{S}')$ have different domains of definition $\Omega_{\mathbf{s}}$ and $\Omega_{\mathbf{S}}$.

Instead, given the knowledge of a deterministic mapping, $\hat{\mathbf{S}} \equiv \mathcal{R}(\mathbf{s})$, between the fine- and coarse-scale states, (A1) is used to compare $f_1 \equiv f_{\mathbf{s}}(\mathbf{s}')$ and $f_2 \equiv f_{\hat{\mathbf{S}}}(\mathbf{s}; \boldsymbol{\phi})$ (Schöberl et al., 2017),

$$f_{\hat{\mathbf{S}}}(\mathbf{s}') = \frac{f_{\mathbf{S}}[\hat{\mathbf{S}}(\mathbf{s}')[}{Z_{\hat{\mathbf{S}}}[\hat{\mathbf{S}}(\mathbf{s}')]}, \qquad Z_{\hat{\mathbf{S}}}(\mathbf{s}) = \int_{\Omega_{\mathbf{s}}} \delta(\mathbf{S} - \hat{\mathbf{S}}(\mathbf{s}'))\mathrm{d}\mathbf{s}', \tag{A2}$$

where $f_{\mathbf{S}}(\mathbf{S}')$ is the prior coarse-scale PDF. Minimization of (A1) with respect to the PDF parameters $\boldsymbol{\phi}$ yields a parameterized $f_{\hat{\mathbf{S}}}(\mathbf{s}')$ as close as possible to $f_{\mathbf{s}}(\mathbf{s}')$. This approach is used in chemistry as a fundamental quantity for multiscale analysis of molecular dynamics (Shell, 2008). This procedure does not yield dimensionality reduction because the PDFs are defined on the same space $\Omega_{\mathbf{s}}$. A posteriori estimates of QoIs in the form of their ensemble averages are given by $\langle \hat{\mathcal{A}}(\mathbf{s}) \rangle = \int_{\Omega_{\mathbf{s}}} \hat{\mathcal{A}}(\mathbf{s}') f_{\hat{\mathbf{S}}}(\mathbf{s}')\mathrm{d}\mathbf{s}'$. A coarse-to-fine probabilistic mapping can be used to estimate PDFs of the observables (Schöberl et al., 2017).

Dimensional reduction and upscaling can be achieved by setting $f_1 \equiv f_{\hat{\mathcal{A}}}(\mathcal{A}')$ and $f_2 \equiv f_{\mathcal{A}}(\mathcal{A}'; \boldsymbol{\phi})$ in (A1) (Arnst & Ghanem, 2008). Here $\hat{\mathcal{A}}(\mathbf{s})$ and $\mathcal{A}(\mathbf{S})$ are defined in section 2 as the QoIs obtained from the fine-scale and the coarse-scale, respectively. Next, (A1) is minimized with respect to the parameters $\boldsymbol{\phi}$ of $f_{\mathcal{A}}(\mathcal{A}'; \boldsymbol{\phi})$ to obtain an optimal probabilistic description of the QoIs at the coarse scale, which are equivalent to those at the fine scale.

Appendix B: Deterministic Annealing

We employ deterministic annealing (Rose, 1998) to minimize (6) with respect to $f_{\mathbf{S}|\mathbf{s}}(\mathbf{S}'|\mathbf{s})$. Deterministic annealing replaces the continuous $f_{\mathbf{S}}(\mathbf{S}')$ in (7) with its discrete version composed of N_{dis} pairs $(\mathbf{S}_k, f_k)$ with $k = 1, \ldots, N_{\mathrm{dis}}$, such that $\sum_{k=1}^{N_k} f_k = 1$. Upon substituting (7) into (6) this minimization problem is replaced with

$$\underset{\{\mathbf{S}_k, f_k\}_{k=1}^{N_{\mathrm{dis}}}}{\operatorname{argmin}} \ \mathcal{J}_D, \qquad \mathcal{J}_D = -\frac{1}{\lambda} \int_{\Omega_{\mathbf{s}}} f_{\mathbf{s}}(\mathbf{s}') \ln\left(\sum_{k=1}^{N_{\mathrm{dis}}} f_k \mathrm{e}^{-\lambda d_w(\mathbf{s}', \mathbf{S}_k)}\right) \mathrm{d}\mathbf{s}' + \lambda_q \left(\sum_{k=1}^{N_{\mathrm{dis}}} f_k - 1\right), \tag{B1}$$

where λ_q is a Lagrange multiplier. This is advantageous because no assumption on the parametric form of $f_{\mathbf{S}|\mathbf{s}}$ is needed. Gradient-based minimization of (B1) with respect to $\mathbf{S}_k$ and f_k is performed by solving $\nabla_{\mathbf{S}_k} \mathcal{J}_D = \mathbf{0}$ and $\partial \mathcal{J}_D / \partial f_k = 0$. Recalling (7) and making use of the definition $f_k = \int f_{\mathbf{s}}(\mathbf{s}') f_{\mathbf{S}_k|\mathbf{s}}(\mathbf{S}'_k|\mathbf{s}')\mathrm{d}\mathbf{s}'$, these yield

$$\int_{\Omega_{\mathbf{s}}} f_{\mathbf{s}}(\mathbf{s}') f_{\mathbf{S}|\mathbf{s}}(\mathbf{S}'_k|\mathbf{s}) \frac{\partial d_w(\mathbf{s}', \mathbf{S}_k)}{\partial S_k^i} \mathrm{d}\mathbf{s}' = 0, \quad i = 1, \ldots, \dim(\mathbf{S}); \quad k = 1, N_k \tag{B2}$$

and

$$-\frac{1}{\lambda} \int_{\Omega_{\mathbf{s}}} \frac{1}{\mathcal{W}(\mathbf{s}', \lambda)} \exp\left(-\lambda d_w(\mathbf{s}', \mathbf{S}_k)\right) \mathrm{d}\mathbf{s}' + \lambda_q = 0, \tag{B3}$$

respectively. Equation (B3) is solved exactly by $\lambda = 1/\lambda_q$, while the solution of (B2) proceeds iteratively as follows:

1. Initialization of λ, $\mathbf{S}_k$ and f_k with $N_k = 1$;
2. $f_{\mathbf{S}|\mathbf{s}}$ and f_k are updated using the definitions:

$$f_{\mathbf{S}_k|\mathbf{s}}(\mathbf{S}'_k|\mathbf{s}) = \frac{f_k \exp\left(-\lambda d_w(\mathbf{s}, \mathbf{S}_k)\right)}{\sum_{l=1}^{N_k} f_l \exp\left(-\lambda d_w(\mathbf{s}, \mathbf{S}_l)\right)} \qquad k = 1, \dots, N_k \tag{B4}$$

$$f_k = \int_{\Omega_\mathbf{s}} f_\mathbf{s}(\tilde{\mathbf{s}}') f_{\mathbf{S}_k|\mathbf{s}}(\mathbf{S}'_k|\mathbf{s}') \mathrm{d}\mathbf{s}'. \tag{B5}$$

   The integral in $\Omega_\mathbf{s}$ is computed approximately from the training set (i.e., the finite set of fine-scale realizations that inform upscaling).
3. The expression for $\partial d_w / \partial S^i_k$ ($i = 1, \dots, \dim(\mathbf{S})$ and $k = 1, \dots, N_k$) in (B2) depends on the choice of $d_w(\mathbf{s}, \mathbf{S})$. The weighted ℓ^2 norm introduced in (4) yields a system

$$\sum_j w_j \mathcal{A}_j(\mathbf{S}) \frac{\partial \mathcal{A}_j(\mathbf{S})}{\partial S^i_k} f_k = \sum_j w_j \frac{\partial \mathcal{A}_j(\mathbf{S})}{\partial S^i_k} \int_{\Omega_\mathbf{s}} f_\mathbf{s}(\mathbf{s}') f_{\mathbf{S}|\tilde{\mathbf{s}}}(\mathbf{S}|\mathbf{s}') \hat{\mathcal{A}}_j(\mathbf{s}') \mathrm{d}\mathbf{s}', \tag{B6}$$

   whose solution $\mathbf{S}_k$ guarantees the minimization of (B1) for the chosen λ and N_k; in case of multiple solutions, we retain the one that provides the smallest $\langle d_w \rangle$.
4. The updated average discrepancy is computed and compared with the convergence tolerance; if convergence is not reached, iteration is repeated. At each iteration, λ is increased and, if the components of $\mathbf{S}_k$ are distinct enough (i.e., the critical annealing of the objective function is reached), N_k is artificially doubled.

Convergence proceeds along a rate distortion curve that is characterized by a decreasing average discrepancy and increasing mutual information between scales, progressively taking into account larger exchange of information between levels.

Acknowledgments
This work was supported in part by the National Science Foundation under grant DMS-1620103 and by the Air Force Office of Scientific Research. There are no data sharing issues since all of the numerical information is provided in the figures produced by solving the equations in the paper.

References

Allen, D. M., Cannon, A. J., Toews, M. W., & Scibek, J. (2010). Variability in simulated recharge using different GCMs. *Water Resources Research*, *46*, W00F03. https://doi.org/10.1029/2009WR008932

Arnst, M., & Ghanem, R. (2008). Probabilistic equivalence and stochastic model reduction in multiscale analysis. *Computer Methods in Applied Mechanics and Engineering*, *197*(43-44), 3584–3592.

Bellin, A., & Rubin, Y. (1996). HYDRO_GEN: A spatially distributed random field generator for correlated properties. *Stochastic Hydrology and Hydraulics*, *10*(4), 253–278.

Blöschl, G. (2005). *Statistical upscaling and downscaling in hydrology*. New York, NY: Wiley Online Library.

Cobb, B. R., Rumi, R., & Salmerón, A. (2012). Approximating the distribution of a sum of log-normal random variables. *Statistics and Computing*, *16*(3), 293–308.

Cover, T. M., & Thomas, J. A. (2012). *Elements of information theory* (2nd ed.). New York, NY: John Wiley.

de Barros, F. P. J., & Rubin, Y. (2011). Modelling of block-scale macrodispersion as a random function. *Journal of Fluid Mechanics*, *676*, 514–545.

Desbarats, A. J. (1992). Spatial averaging of hydraulic conductivity in three-dimensional heterogeneous porous media. *Mathematical Geology*, *24*(3), 249–267.

Durlofsky, L. J. (2000). An approximate model for well productivity in heterogeneous porous media. *Mathematical Geology*, *32*(4), 421–438.

Efendiev, Y., & Pankov, A. (2004). Numerical homogenization of nonlinear random parabolic operators. *Multiscale Modeling and Simulation*, *2*(2), 237–268.

Ferraris, L., Gabellani, S., Rebora, N., & Provenzale, A. (2003). A comparison of stochastic models for spatial rainfall downscaling. *Water Resources Research*, *39*(12), 1368. https://doi.org/10.1029/2003WR002504

Fowler, H. J., Blenkinsop, S., & Tebaldi, C. (2007). Linking climate change modelling to impacts studies: Recent advances in downscaling techniques for hydrological modelling. *International Journal of Climatology*, *27*(12), 1547–1578.

Harbaugh, A. W., Banta, E. R., Hill, M. C., & McDonald, M. G. (2000). *MODFLOW–2000, The U. S. Geological Survey Modular Ground-Water Model—User guide to modularization concepts and the ground-water flow process*. Reston, VA: U. S. Geological Survey.

Hassane, M. M. F. Z., & Ackerer, P. (2017). Groundwater flow parameter estimation using refinement and coarsening indicators for adaptive downscaling parameterization. *Advances in Water Resources*, *100*, 139–152.

Korneev, S., & Battiato, I. (2016). Sequential homogenization of reactive transport in polydisperse porous media. *Multiscale Modeling and Simulation*, *14*(4), 1301–1318.

Koutsourelakis, P.-S. (2007). Stochastic upscaling in solid mechanics: An excercise in machine learning. *Journal of Computational Physics*, *226*(1), 301–325.

Lichtner, P. C., & Tartakovsky, D. M. (2003). Stochastic analysis of effective rate constant for heterogeneous reactions. *Stochastic Environmental Research and Risk Assessment*, *17*(6), 419–429.

Marošević, T. (1996). A choice of norm in discrete approximation. *Mathematical Communications*, *1*(2), 147–152.

Matheron, G. (1967). *Eléments Pour Une Théorie des Milieux Poreux*. Paris, France: Masson.

Montzka, C., Pauwels, V. R. N., Franssen, H.-J. H., Han, X., & Vereecken, H. (2012). Multivariate and multiscale data assimilation in terrestrial systems: A review. *Sensors*, *12*(12), 16,291–16,333.
Neuman, S. P., & Tartakovsky, D. M. (2009). Perspective on theories of anomalous transport in heterogeneous media. *Advances in Water Resources*, *32*(5), 670–680. https://doi.org/10.1016/j.advwatres.2008.08.005
Paleologos, E. K., Neuman, S. P., & Tartakovsky, D. (1996). Effective hydraulic conductivity of bounded, strongly heterogeneous porous media. *Water Resources Research*, *32*(5), 1333–1341.
Parekh, P. M., Katselis, D., Beck, C. L., & Salapaka, S. M. (2015). Deterministic annealing for clustering: Tutorial and computational aspects. In *American Control Conference (ACC), 2015* (pp. 2906–2911). Chicago, IL: IEEE.
Peaceman, D. W. (1978). Interpretation of well-block pressures in numerical reservoir simulation (includes associated paper 6988). *Society of Petroleum Engineers Journal*, *18*(03), 183–194.
Rose, K. (1998). Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. *Proceedings of the IEEE*, *86*(11), 2210–2239.
Sanchez-Vila, X., & Tartakovsky, D. M. (2007). Ergodicity of pumping tests. *Water Resources Research*, *43*, W03414. https://doi.org/10.1029/2006WR005241
Schöberl, M., Zabaras, N., & Koutsourelakis, P.-S. (2017). Predictive coarse-graining. *Journal of Computational Physics*, *333*, 49–77.
Shell, M. S. (2008). The relative entropy is fundamental to multiscale and inverse thermodynamic problems. *Journal of Chemical Physics*, *129*(14), 144108.
Tartakovsky, D. M. (2013). Assessment and management of risk in subsurface hydrology: A review and perspective. *Advances in Water Resources*, *51*, 247–260. https://doi.org/10.1016/j.advwatres.2012.04.007
Tartakovsky, D. (2016). Uncertainty quantification in subsurface modeling (3rd ed.). In D. Tartakovsky & J. Cushman (Eds.), *The handbook of groundwater engineering* (pp. 625–640). Boca Raton, FL: CRC Press.
Um, K., Zhang, X., Katsoulakis, M., Plechac, P., & Tartakovsky, D. M. (2018). Global sensitivity analysis of multiscale properties of porous materials. *Journal of Applied Physics*, *123*, 075103. https://doi.org/10.1063/1.5009691
Wigmosta, M., & Prasad, R. (2005). Upscaling and downscaling–dynamic models, Encyclopedia of hydrological sciences.
Zhu, P., Shi, L., Zhu, Y., Zhang, Q., Huang, K., & Williams, M. (2017). Data assimilation of soil water flow via ensemble Kalman filter: Infusing soil moisture data at different scales. *Journal of Hydrology*, *555*, 912–925.