Computational Geosciences (2018) 22:1093–1113
https://doi.org/10.1007/s10596-018-9740-x

ORIGINAL PAPER

Machine-learning-based modeling of coarse-scale error, with application to uncertainty quantification

Sumeet Trehan[1] · **Louis J. Durlofsky**[1]

Received: 19 February 2017 / Accepted: 27 March 2018 / Published online: 11 May 2018

Abstract

The use of upscaled models is attractive in many-query applications that require a large number of simulation runs, such as uncertainty quantification and optimization. Highly coarsened models often display error in output quantities of interest, e.g., phase production and injection rates, so the direct use of these results for quantitative evaluations and decision making may not be appropriate. In this work, we introduce a machine-learning-based post-processing framework for modeling the error in coarse-model results in the context of uncertainty quantification. Coarse-scale models are constructed using an accurate global single-phase transmissibility upscaling procedure. The framework entails the use of high-dimensional regression (random forest in this work) to model error based on a number of error indicators or features. Many of these features are derived from approximations of the subgrid effects neglected in the coarse-scale saturation equation. These features are identified through volume averaging, and they are generated by solving a fine-scale saturation equation with a constant-in-time velocity field. Our approach eliminates the need for the user to hand-design a small number of informative (relevant) features. The training step requires the simulation of some number of fine and coarse models (in this work we perform either 10 or 30 training simulations), followed by construction of a regression model for each well. Classification is also applied for production wells. The methodology then provides a correction at each time step, and for each well, in the phase production and injection rates. Results are presented for two- and three-dimensional oil–water systems. The corrected coarse-scale solutions show significantly better accuracy than the uncorrected solutions, both in terms of realization-by-realization predictions for oil and water production rates, and for statistical quantities important for uncertainty quantification, such as P10, P50, and P90 predictions.

Keywords Upscaling · Machine learning · Uncertainty quantification · Reservoir simulation · Classification · Error modeling · Random forest · Surrogate model

1 Introduction

Field measurements in subsurface flow applications are almost always sparse, so there commonly remains a high degree of uncertainty in the geological description. For this reason, in a comprehensive simulation study, a large number of plausible realizations, generated such that they are consistent with measured data and the prior geological concept(s), are generated. Because geological features at a variety of scales can impact subsurface flow behavior, these models typically contain a high degree of resolution. The simulation of a large number of high-fidelity models can be very expensive, however, and for this reason, upscaling (or coarsening) procedures are commonly applied. The flow responses from upscaled models generally contain some error relative to the responses for the corresponding fine-scale models. Thus, if coarse-scale models (CSMs) are to be used for practical decision making and uncertainty quantification (UQ), these errors must be quantified.

Our goal in this paper is to apply a recently developed machine-learning framework to model the time-instantaneous errors in a specified set of quantities of interest, or QoI. In this work, the QoI correspond to the time-varying injection and phase production rates for all wells in the model. The general framework, referred to as error modeling using machine learning (EMML), was described

✉ Sumeet Trehan
strehan@alumni.stanford.edu

[1] Department of Energy Resources Engineering, Stanford University, Stanford, CA 94305-2220, USA

in an earlier study [52]. In that work, approximate results were computed using a reduced-order modeling approach based on proper orthogonal decomposition and trajectory piecewise linearization [7, 23, 53], and QoI error using this method was successfully modeled using the EMML framework.

The EMML methodology requires a set of error indicators or features, which are used in the statistical regression (random forest [5] in this work) applied for error modeling. In this paper we use the term "features" to refer to quantities used as error indicators. These are distinct from geological features, though of course error is impacted by aspects of the geological model. Various treatments, including the modeling of both absolute and relative errors, were considered in [52], and the best-performing approaches were identified. In this work, a key intent is to provide features that characterize the subgrid effects neglected in the CSM. An important strength of our procedure is that a large number of features can be considered, i.e., the user is not required to select a small number of the most informative features. As is the case with all regression-based methods, some amount of training, which entails the simulation of high-fidelity models (HFMs) and the construction of the regression-based error model, is required in a preprocessing (offline) step.

A variety of upscaling methods have been developed for use in reservoir simulation problems (see, e.g., [15] for a detailed discussion). In highly coarsened models, accurate results can be achieved if both permeability and relative permeability are upscaled, though the latter requires substantial upscaling computation. Therefore, in this work, we upscale only single-phase parameters (absolute permeability and porosity). We apply a global single-phase transmissibility upscaling method, as this approach provides (essentially) the most accurate coarse-scale models [9, 10, 55], along with the global fine-scale velocity field (which we reuse in the error modeling procedure). Although this upscaling technique requires the global solution of the steady-state single-phase pressure equation, it is still less expensive computationally than relative permeability upscaling.

CSMs have been used in previous studies to solve computationally demanding many-query problems. For example, CSMs were applied for UQ [3, 8, 11, 30, 35, 45], in inverse modeling [2, 16, 26, 33, 34], and for optimization [1, 29]. The ensemble-level upscaling procedure, introduced in [8] and applied for UQ, shares some similarities with the EMML-based framework applied here. In this approach, upscaled relative permeability curves for the coarse blocks in a few training realizations are computed in a preprocessing step. An appropriate curve is then assigned statistically to each grid block in new coarse-scale models based on efficiently computed features. The transmissibilities in the CSMs are constructed using a single-phase upscaling procedure. The method has since been extended to three-dimensional well-driven flow systems [11] and to oil-gas compositional problems [30]. Lødøen et al. [35] also used CSMs constructed using local or local-global upscaling techniques, along with upscaling error models, for UQ. In subsequent work, they used CSMs, generated using harmonic averaging of permeability, in conjunction with error models, for inverse modeling [33, 34]. In these studies, the error in the QoI was modeled using a linear regression model, which was constructed with the coarse-scale QoI as the independent variable. See Omre and Lødøen [39] for more details.

Scheidt et al. [45] modeled upscaling error in the QoI using a distance-based clustering procedure. In this approach, the coarse-scale QoI prediction is first assigned to a particular cluster, where the error in the QoI is known at the cluster center (this error is computed in an offline step by performing both fine- and coarse-scale simulation). During the online stage, the error in the QoI is assumed to be constant over the cluster. Similarly, Glimm et al. [20] used a Gaussian process (kriging) to model the error introduced due to coarsening. Other statistical techniques, such as a functional principal-component-analysis-based approach, as presented by Josset et al. [25] in the context of a proxy model, could also be employed. However, as demonstrated recently by [12, 38], the direct modeling of QoI error (as performed in the above-mentioned studies) may be problematic if there is a complex dependency, such as an oscillatory or highly nonlinear relationship, between the error and the CSM prediction.

In this work, we apply a high-dimensional regression technique, random forest, to model the upscaling error, for each well at each time step, based on a set of features. As discussed by [13, 17, 18], CSM error results from neglected subgrid effects. Using the methods described in these and related papers, it is possible to construct improved CSMs. Our intent here, however, is to instead compute features that represent the neglected effects and to use these features within the EMML framework. This is accomplished by solving an approximate fine-scale saturation equation, with a constant-in-time velocity field, to provide estimates of subgrid (fluctuating) quantities. The EMML training dataset is generated by simulating both the HFM and the corresponding CSM, obtained using upscaling, for a fraction of the realizations in the ensemble. An error model is then constructed from the EMML training dataset based on a prescribed set of features. Classification is also applied to distinguish pre- versus post-water-breakthrough behavior. For new (test) realizations, we estimate the error from the regression model and the features, and then use this estimate to correct the coarse-scale QoI. We demonstrate the application of this framework for UQ in two- and three-dimensional

oil–water problems. The approaches presented here should also enable the use of CSMs in optimization or data assimilation problems.

This paper proceeds as follows. In Section 2, we present the governing equations for both the fine- and coarse-scale models. The subgrid terms neglected in the CSM are then discussed. We next present the EMML framework as specialized for the oil–water upscaling problem. This includes using estimates of the neglected subgrid effects as features. Results demonstrating the application of the EMML framework for two systems are presented in Section 3. Concluding remarks are provided in Section 4. The Appendix provides a brief overview of random-forest regression.

2 Problem description and error modeling

In this section, we describe the HFM and CSM representations for two-phase oil–water systems. We discuss our estimates of the subgrid effects, which are used as features, and we then describe the overall EMML framework. For full details on the EMML methodology, see [51, 52].

2.1 Fine-scale equations

We consider an isothermal two-phase oil–water system with no mass transfer between the phases. Neglecting capillary pressure and (for now) gravitational effects, the governing equations for phase j, where $j = o, w$ denotes the oil and water phase, are

$$\frac{\partial}{\partial t}\left(\phi\rho_j S_j\right) + \nabla\cdot\left(\rho_j \mathbf{v}_j\right) + \rho_j\tilde{q}_j = 0, \tag{1a}$$

$$\mathbf{v}_j = -\lambda_j \mathbf{k}\cdot\nabla p. \tag{1b}$$

Here t is time, ϕ is porosity, ρ_j, S_j and $\mathbf{v}_j$ are phase density, saturation and velocity, $\tilde{q}_j$ denotes the well phase flow rate per unit volume (positive for production wells), $\mathbf{k}$ is the absolute permeability tensor, $\lambda_j = k_{rj}/\mu_j$ is the phase mobility, with $k_{rj}(S_j)$ the relative permeability and μ_j the phase viscosity, and p is pressure. From hereon, we denote S_w by S.

In our simulations, well bottom-hole pressures (BHPs) are specified in all cases. Thus, the output QoI correspond to the time-varying volumetric phase flow rates q_j (units of volume/time) at the production and injection wells. If rates were specified, the QoI would include BHPs.

For a well located in grid block $d \in \mathcal{D} := \{d_1, \ldots, d_{N_w}\}$, where N_w is the total number of wells, the volumetric phase flow rate is given by the usual Peaceman [40] well model:

$$(q_j)_d = WI_d\left(\lambda_j\right)_d\left(p_d - p_w\right). \tag{2}$$

Here, WI_d denotes the well index, which depends on well-block properties and the wellbore radius, p_d is the well-block pressure, and p_w is the wellbore pressure. Note that $(q_j)_d = (\tilde{q}_j)_d\Omega_d$, where Ω_d is the volume of well block d. The set of production wells is denoted by $\mathcal{D}_P$ and the set of injection wells by $\mathcal{D}_I$. Thus, $\mathcal{D} = \mathcal{D}_P \cup \mathcal{D}_I$, and the total number of wells is $N_w = |\mathcal{D}| = |\mathcal{D}_P| + |\mathcal{D}_I|$, where $|\cdot|$ denotes the cardinality of the set.

Standard finite-volume discretizations are applied in this work (including the representation of source terms using the well model in Eq. 2). The state vector $\mathbf{x}^f$ for the fine-scale model at a particular time step, computed by solving the discretized system, is denoted by

$$\mathbf{x}^f = \left[p_1\ S_1\ \cdots\ p_{N_c}\ S_{N_c}\right]^T \in \mathbb{R}^{2N_c}, \tag{3}$$

where N_c is the number of grid blocks in the HFM.

Assuming incompressible fluid and rock, we can rearrange Eq. 1a into so-called pressure and saturation equations. This representation, which will be useful when we apply volume averaging in the development below, yields the following equations

$$\nabla\cdot\left[\lambda_t\mathbf{k}\cdot\nabla p\right] = \tilde{q}_o + \tilde{q}_w, \tag{4a}$$

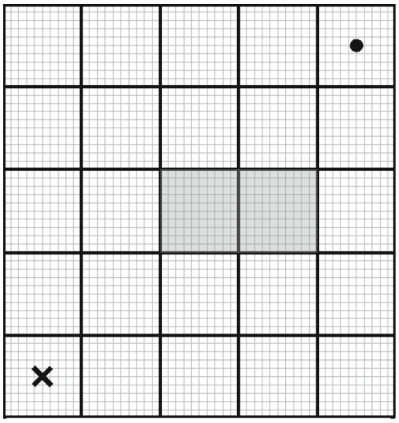
$$\phi\frac{\partial S}{\partial t} + \nabla\cdot\left[\mathbf{v}_t f(S)\right] = -\tilde{q}_w, \tag{4b}$$

where $f(S) = \lambda_w/\lambda_t$ is the Buckley-Leverett flux function, $\lambda_t = \lambda_o + \lambda_w$ is the total mobility, and $\mathbf{v}_t = \mathbf{v}_o + \mathbf{v}_w = -\lambda_t\mathbf{k}\cdot\nabla p$ is the total Darcy velocity.

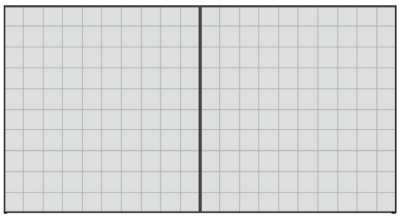
2.2 Coarse-scale equations

The coarse-scale governing equations involve replacing the permeability tensor $\mathbf{k}$ with the effective (equivalent or upscaled) grid block permeability $\mathbf{k}^*$ for each of the N_c^* coarse blocks. In our actual implementation, we use upscaled transmissibility T^* rather than $\mathbf{k}^*$, although we continue to express the coarse-scale equations in terms of $\mathbf{k}^*$. We also upscale the porosity ϕ to ϕ^* (in a manner that conserves pore volume from the set of fine-scale blocks to the corresponding coarse-scale block), and the well index WI to WI^*. Figure 1 shows a schematic of the grid blocks corresponding to the HFM and the CSM. The coarse-scale grid is represented by thick black lines, while the fine-scale grid is represented by thin gray lines.

In this study, we apply global single-phase transmissibility upscaling (SPTU) [10, 30, 55] to compute T^* for each interface and WI^* for each well block. This approach requires the solution of the steady-state single-phase fine-scale pressure equation, but it avoids the two-phase flow computations required for two-phase upscaling. Further, among the different single-phase upscaling techniques, global single-phase upscaling appears to provide the most accurate coarse models [10, 15, 55]. For more details on upscaling and a discussion of different upscaling techniques, see [14, 15].



(a) Global domain



(b) Local region showing coarse blocks i and $i+1$ and coarse interface $i+1/2$

Fig. 1 Schematic showing **a** fine-scale (50×50) and coarse-scale (5×5) grids, and **b** a local two-coarse-block region. In (a), × indicates a production well and • indicates an injection well

When using the single-phase upscaled parameters in the coarse-scale two-phase model, the equations are taken to be of the same form as Eq. 4a. Specifically, the coarse-scale equations can now be expressed as

$$\nabla \cdot \left[\lambda_t(S^c)\mathbf{k}^* \cdot \nabla p^c\right] = \tilde{q}_o^c + \tilde{q}_w^c, \tag{5a}$$

$$\phi^* \frac{\partial S^c}{\partial t} + \nabla \cdot \left[\mathbf{v}_t^c f(S^c)\right] = -\tilde{q}_w^c, \tag{5b}$$

where the superscript $*$ designates a precomputed coarse-scale parameter, and superscript c indicates a coarse-scale variable. The relative permeability function $k_{rj} = k_{rj}(S^c)$ and viscosity μ_j are the same as on the fine scale. The total velocity, $\mathbf{v}_t^c$, is now given by $\mathbf{v}_t^c = \mathbf{v}_o^c + \mathbf{v}_w^c = -\lambda_t(S^c)\mathbf{k}^* \cdot \nabla p^c$. For a well in the coarse well block d, the volumetric phase flow rate is given by

$$(q_j)_d^c = WI_d^* \left(\lambda_j\right)_d \left(p_d^c - p_w\right), \tag{6}$$

where $(q_j)_d^c = (\tilde{q}_j)_d^c \Omega_d^c$, with Ω_d^c the volume of coarse well block d. The state vector $\mathbf{x}^c$, computed by simulating the CSM, is denoted by

$$\mathbf{x}^c = \left[p_1^c \; S_1^c \; \cdots \; p_{N_c^*}^c \; S_{N_c^*}^c\right]^T \in \mathbb{R}^{2N_c^*}. \tag{7}$$

For a given geological realization, to compute the upscaled transmissibility T^* for each coarse block-to-block interface and the upscaled well index WI^* for each coarse block containing a well, we solve the fine-scale incompressible single-phase pressure equation, with the flow driven by the wells:

$$\nabla \cdot (\mathbf{k} \cdot \nabla p) = q. \tag{8}$$

The upscaled transmissibility is then computed as

$$T_{i+\frac{1}{2}}^* = \frac{q^c}{\overline{p}_i - \overline{p}_{i+1}}, \tag{9}$$

where $q^c = \sum_{j \in i+\frac{1}{2}} q_j$ is the sum of the fine-scale flux contributions q_j through coarse interface $i+\frac{1}{2}$, and $\overline{p}_i$ and $\overline{p}_{i+1}$ are the bulk-volume average of the fine-scale pressure over coarse blocks i and $i+1$, respectively. These quantities are defined with reference to Fig. 1b. Similarly, we compute the upscaled well index WI^* as in [10]. Using the solution of Eq. 8, we also compute the constant-in-time single-phase velocity field, designated $\mathbf{v}_{1p}$ and given by $\mathbf{v}_{1p} = -\mathbf{k} \cdot \nabla p$. This quantity will be required for the construction of some of the features used in the EMML framework, as we will see in Section 2.4.

2.3 Volume averaging for modeling of subgrid effects

We apply the volume averaging procedure described in [13] to characterize the upscaling error that results from the neglected subgrid effects. This averaging provides an equation for the neglected quantities, which in turn enables us to identify some potential features (error indicators) for use in the EMML framework. For any variable $\psi(x, y)$ defined on the fine scale, we can write

$$\psi(x, y) = \overline{\psi}(x^c, y^c) + \psi'(x, y), \tag{10}$$

where $\overline{\psi}(x^c, y^c) = \frac{1}{\Omega^c} \int_{\Omega^c} \psi(x, y) d\Omega^c$ is the volume average (coarse-scale) quantity, and $\psi'(x, y)$ is the spatially varying fluctuating (subgrid) quantity. In a single coarse block, $\overline{\psi}(x^c, y^c)$ is constant, and $\psi'(x, y)$ varies from fine cell to fine cell, with an average of zero.

Our focus for now will be on the saturation Eq. 4b, since subgrid effects can have a large impact on phase transport. Using the representation in Eq. 10, we write

$$S(x, y) = \bar{S}(x^c, y^c) + S'(x, y), \tag{11a}$$

$$f\left(S(x, y)\right) = \bar{f}(x^c, y^c) + f'\left(S(x, y)\right), \tag{11b}$$

$$\mathbf{v}_t(x, y) = \overline{\mathbf{v}_t}(x^c, y^c) + \mathbf{v}_t'(x, y). \tag{11c}$$

Substituting these expressions into Eq. 4b, we have

$$\phi^* \frac{\partial(\bar{S} + S')}{\partial t} + \nabla \cdot \left[(\overline{\mathbf{v}_t} + \mathbf{v}_t')(\bar{f} + f')\right] = 0, \tag{12}$$

where, for simplicity, we have omitted the source term. Averaging Eq. 12 over the volume Ω^c, and using the fact that $\overline{\overline{\psi}'} = 0$ and $\overline{\overline{\psi}} = \overline{\psi}$, we have

$$\phi^* \frac{\partial \overline{S}}{\partial t} + \nabla \cdot \left[\overline{\mathbf{v}}_t \overline{f} + \overline{\mathbf{v}_t' f'} \right] = 0. \tag{13}$$

Recall that the overbar denotes the volume average over all fine-scale blocks inside a coarse block. By comparing Eqs. 5b and 13, we observe that, assuming $S^c \approx \overline{S}$ and $\mathbf{v}_t^c \approx \overline{\mathbf{v}}_t$, the unmodeled subgrid effects in Eq. 5b derive from neglecting $\overline{\mathbf{v}_t' f'}$ and from the fact that $\overline{f} \neq f$ when f is a nonlinear function of S. It was shown in [13] that these terms depend, to leading order (i.e., neglecting terms that are third-order in fluctuating quantities), on $\overline{\mathbf{v}_t' S'}$ and $\overline{S' S'}$. It is therefore appropriate to use $\overline{\mathbf{v}_t' S'}$ and $\overline{S' S'}$, and the quantities on which they depend, as features in the EMML framework, since they are directly related to the upscaling error.

Towards this goal, we next present equations for the fluctuating quantities S', $\overline{\mathbf{v}_t' S'}$ and $\overline{S' S'}$. Subtracting Eq. 13 from Eq. 12, we obtain the equation for the fluctuating quantity S' as

$$\phi^* \frac{\partial S'}{\partial t} + \nabla \cdot \left[\overline{\mathbf{v}}_t f' + \mathbf{v}_t' \overline{f} \right] = 0. \tag{14}$$

Following [13], the transport equations for $\overline{\mathbf{v}_t' S'}$ and $\overline{S' S'}$ are derived by multiplying Eq. 14 with $\mathbf{v}_t'$ and S', respectively, followed by volume averaging. This gives

$$\begin{aligned} \phi^* \frac{\partial (\overline{\mathbf{v}_t' S'})}{\partial t} + & \ \overline{\mathbf{v}_t' \nabla f'} \cdot \overline{\mathbf{v}}_t + \overline{f' \mathbf{v}_t'} \nabla \cdot \overline{\mathbf{v}}_t + \\ & \overline{f} \overline{\mathbf{v}_t' \nabla \cdot \mathbf{v}_t'} + \overline{\mathbf{v}_t' \mathbf{v}_t'} \nabla \overline{f} = 0, \end{aligned} \tag{15a}$$

$$\begin{aligned} \frac{1}{2} \phi^* \frac{\partial \overline{(S' S')}}{\partial t} + & \ \overline{S' \nabla f'} \cdot \overline{\mathbf{v}}_t + \overline{S' f'} \nabla \cdot \overline{\mathbf{v}}_t + \\ & \overline{f} \overline{S' \nabla \cdot \mathbf{v}_t'} + \overline{S' \mathbf{v}_t'} \cdot \nabla \overline{f} = 0. \end{aligned} \tag{15b}$$

The neglected terms described above are directly relevant to the saturation equation. Efendiev and Durlofsky [18] additionally estimated the subgrid effects neglected in the coarse-scale pressure equation given in Eq. 5a. For this equation, subgrid effects involve terms such as $f_S \lambda_S(\overline{S}) \overline{\mathbf{v}_t' S'}$ and $f_S(\overline{S}) \lambda_S(\overline{S}) \overline{\mathbf{v}}_t \overline{S' S'}$, where $f_S = df/dS$ and $\lambda_S = d\lambda_t/dS$. These terms can also be estimated in terms of the features defined below.

2.4 Error modeling for upscaling

Our intent in this work is to model the actual error in the output QoI at each time step,

$$(\delta_q^n(\mathbf{m}))_d := (q_f^n)_d - (q_c^n)_d, \ n = 1, \ldots, N_t, d \in \mathcal{D}. \tag{16}$$

Here, $(\delta_q^n(\mathbf{m}))_d \in \mathbb{R}$, d denotes the coarse well block, subscript q denotes the QoI, i.e., $q = o, w, inj$, with o the oil production rate, w the water production rate and inj the water injection rate, superscript n denotes the time step, N_t is the total number of time steps, $\mathbf{m}$ indicates the geological parameters corresponding to a particular HFM, $(q_f^n)_d = (q_f^n(\mathbf{m}))_d$ is the QoI from the fine-scale simulation, and $(q_c^n)_d = (q_c^n(\mathbf{m}))_d$ is the QoI from the coarse-scale simulation. This error originates from the fact that we perform the simulation on the coarse-scale model (obtained using the SPTU technique), instead of the fine-scale model.

From Eq. 6, we see that $(q_c^n)_d$ depends directly on the coarse well-block pressure and saturation. Since the errors in these quantities are correlated to the subgrid effects, it follows that $(\delta_q^n)_d$ is correlated to the neglected subgrid effects in the coarse well block d. These subgrid effects are described by Eqs. 13, 14 and 15a, b, and involve the terms with fluctuating quantities. However, to compute these fluctuating quantities during the online computations, we require the fine-scale solution, which is unavailable since our intent is to perform only coarse-scale simulations during the online stage.

To obtain an estimate of these fluctuating quantities, we instead solve a simplified (and computationally efficient) fine-scale transport equation:

$$\phi \frac{\partial S_{\mathrm{a}}}{\partial t} + \nabla \cdot \left[\mathbf{v}_{1\mathrm{p}} f_{\mathrm{a}} \right] = -q_w/\rho_w. \tag{17}$$

Here, $\mathbf{v}_{1\mathrm{p}}$ is the constant-in-time velocity field computed while performing global single-phase upscaling (as described in Section 2.2), S_{a} is the saturation associated with this equation, $f_{\mathrm{a}} = f(S_{\mathrm{a}})$, where $f(S)$ is the same nonlinear flux function as in Eq. 4b, and all other variables are as previously defined. Solving Eq. 17 is relatively inexpensive since it involves a constant-in-time velocity field. Recall that in (fine-scale) Eq. 4b, $\mathbf{v}_t$ changes at each time step and must be computed through the solution of the pressure equation. Using the solution of Eq. 17, we estimate the fluctuating quantities at the well blocks $d \in \mathcal{D}$. These estimates are then used as features.

For example, for a well located in the coarse well block d, the QoI error $(\delta_q^n)_d$ may be correlated to the quantity $\overline{(S' S')^n}$, which appears in Eq. 15b. Thus, using the solution of Eq. 17, we approximate $\overline{(S' S')^n}$ in well block d as $\overline{\left(S_{\mathrm{a}}'\right)_d^i \left(S_{\mathrm{a}}'\right)_d^i}$, where i is the time index such that $\overline{(S_{\mathrm{a}})_d^i} \approx (S^c)_d^n$. Similarly, we identify $\overline{\left(\mathbf{v}_t' f'\right)^n}$, which appears in Eq. 13, as a potential feature and estimate it (for well block d) as $\overline{(\mathbf{v}_{1\mathrm{p}}')_d \left(f_{\mathrm{a}}'\right)_d^i}$. We also include in the features global quantities such as the total simulated time t^n and the pore volume injected (PVI^n). Other realization properties such as the well-pair volumes (WPV) and the well-allocation factors (WAF) (described in more detail in Section 2.6) corresponding to both the CSM and the HFM are also included.

Some of the features used in this work are shown in Table 1. In the table, $\boldsymbol{f}^n \in \mathbb{R}^{1\times N_f}$, $n = 1, \ldots, N_t$, $(\cdot)^n_d$ refers to a coarse-scale quantity corresponding to the coarse well block $d \in \mathcal{D}$ at time step n, and $(\cdot)^i_d$ refers to the fine-scale quantity, obtained by solving Eq. 17, for all the fine-scale cells in the coarse well block d at time step i. Recall that the time step i is chosen such that $\overline{(S_\mathrm{a})^i_d} \approx (S^c)^n_d$, which acts to "align" the coarse-scale and approximate fine-scale solutions. The features $\boldsymbol{f}^n$ are computed using the solutions of Eq. 5a, b and 17.

Note that some of the features shown in the table, and many additional features that are included in the set but are not displayed in the table, involve various combinations (products) of terms. Although many of these combinations do not appear directly in the governing equations, we expect that their inclusion in the feature set will act to capture higher-moment type effects and will thus enrich the representation. In this way we specify a fairly exhaustive set of features. Since some features in $\mathbb{F}$ will be highly correlated among themselves, as described in [52], we remove such highly correlated features in a preprocessing step by computing the feature-feature Pearson correlation coefficient for all pairs. Refer to [21] for further details on correlation-criteria-based feature selection. Because we remove redundant features in this way, we eliminate the need to prescribe (a priori) a small number of 'most-representative' features.

Next, we apply the EMML framework, specifically Method 2 as described in [51, 52], to model $(\delta^n_q)_d, n = 1, \ldots, N_t, d \in \mathcal{D}$. Although this approach was presented in [52], we describe it below while specializing it for the upscaling application considered here.

2.5 Modeling of the upscaling errors

We use a high-dimensional regression technique (random forest, which is described later) to correlate the QoI error $(\delta^n_q(\mathbf{m}))_d$ to the features $\boldsymbol{f}^n$. Since the QoI error can exhibit a wide range of values, we define a relative error $(\mathring{\delta}^n_q(\mathbf{m}))_d \in \mathbb{R}$ as

$$(\mathring{\delta}^n_q(\mathbf{m}))_d := \frac{(\delta^n_q)_d}{(q^n_f)_d},\ n = 1, \ldots, N_t, d \in \mathcal{D}. \tag{18}$$

From hereon, we drop the subscript d for notational simplicity. We then assume that

$$\mathring{\delta}^n_q = \mathring{r}_q(\boldsymbol{f}^n) + \varepsilon^n,\ n = 1, \ldots, N_t, \tag{19}$$

where $\mathring{\delta}^n_q = \mathring{\delta}^n_q(\mathbf{m})$, $\boldsymbol{f}^n = \boldsymbol{f}^n(\mathbf{m}) \in \mathbb{R}^{1\times N_f}$ corresponding to a well in coarse block d, and $\mathring{r}_q : \mathbb{R}^{1\times N_f} \to \mathbb{R}$ denotes an unknown function. The noise ε^n, which we assume to be a random variable with zero mean, represents the information due to missing features, sampling variability and unaccounted for nonlinear effects. While the noise ε^n may depend on the features, i.e., $\varepsilon^n = \varepsilon(\boldsymbol{f}^n)$, we neglect this dependence in this study.

To estimate the regression model $\mathring{r}_q$, we construct $\hat{\mathring{r}}_q : \mathbb{R}^{1\times N_f} \to \mathbb{R}$, such that $\hat{\mathring{r}}_q(\boldsymbol{f}^n) \approx \mathring{r}_q(\boldsymbol{f}^n)$. The error model $\hat{\mathring{r}}_q$ is constructed here using random forest. Refer to [52] for a succinct description of random forest in the context of error modeling, and to [5] for a more detailed development. A brief description of random forest is given in the Appendix. Use of a high-dimensional regression technique allows the estimated error $\hat{\mathring{\delta}}^n_q \in \mathbb{R}$ to be expressed as

$$\hat{\mathring{\delta}}^n_q = \hat{\mathring{r}}_q(\boldsymbol{f}^n), \quad n = 1, \ldots, N_t. \tag{20}$$

Table 1 Features ($\boldsymbol{f}^n$) used in EMML

No.	Feature	No.	Feature
1.	$\overline{(S'_\mathrm{a})^i_d\,(S'_\mathrm{a})^i_d}$	2.	$\overline{(S_\mathrm{a})^i_d}/(S^c)^n_d$
3.	$\left(\overline{(\mathbf{v}'_{1\mathrm{p}})_d\,(f'_\mathrm{a})^i_d}\right)_\ell$	4.	$\left(\overline{(\mathbf{v}'_{1\mathrm{p}})_d\,(S'_\mathrm{a})^i_d}\right)_\ell$
5.	$\overline{(f'_\mathrm{a})^i_d\,(f'_\mathrm{a})^i_d}$	6.	$\overline{(f_\mathrm{a})^i_d}/(f^c)^n_d$
7.	$(f^c)^n_d(S^c)^n_d\left(\overline{(\mathbf{v}'_{1\mathrm{p}})_d\,(S'_\mathrm{a})^i_d}\right)_\ell$	8.	$\overline{\left(\mathbf{v}'_{1\mathrm{p}})_d\right)_\ell\left(\nabla\,(f'_\mathrm{a})^i_d\right)_\ell}\left(\overline{(\mathbf{v}_{1\mathrm{p}})_d}\right)_\ell$
9.	t^n	10.	$\left((S^c)^n_d - (S^c)^{n-1}_d\right)/\Delta t^n$
11.	$(p^c)^n_d$	12.	$\left((\mathbf{v}^c)^n_d\right)_\ell - \left(\overline{(\mathbf{v}'_{1\mathrm{p}})_d}\right)_\ell$
13.	PVI^n	14.	$\overline{(f'_\mathrm{a})^i_d(S'_\mathrm{a})^i_d\,\nabla\cdot(\mathbf{v}'_{1\mathrm{p}})_d}$
15.	$\left(\overline{(\mathbf{v}_{1\mathrm{p}})_d(f_\mathrm{a})^i_d} - (\mathbf{v}^c f^c)^n_d\right)_\ell$	16.	$\left(\overline{(\mathbf{v}_{1\mathrm{p}})_d(S_\mathrm{a})^i_d}\right)_\ell / \left((\mathbf{v}^c S^c)^n_d\right)_\ell$
17.	$\overline{(S'_\mathrm{a})^i_d\,(S'_\mathrm{a})^i_d\,\left((\mathbf{v}^c)^n_d\right)_\ell}$	18.	$\overline{\left((\mathbf{v}'_{1\mathrm{p}})_d\right)_\ell\left(\nabla\,(f'_\mathrm{a})^i_d\right)_\ell}$
19.	$\overline{\left((S'_\mathrm{a})^i_d\,(\mathbf{v}'_{1\mathrm{p}})_d\right)_\ell\left(\nabla\,(f'_\mathrm{a})^i_d\right)_\ell}$	20.	$\left(\overline{(S'_\mathrm{a})^i_d\,\nabla\,(f'_\mathrm{a})^i_d}\right)_\ell\left(\overline{(\mathbf{v}_{1\mathrm{p}})_d}\right)_\ell$
21.	$\left(\overline{(\mathbf{v}'_{1\mathrm{p}})_d\,(S'_\mathrm{a})^i_d}\right)_\ell / \left(\overline{(\mathbf{v}_{1\mathrm{p}})_d}\right)_\ell$	22.	$(\mathrm{WPV})_{\mathrm{CSM,HFM}}$, $(\mathrm{WAF})_{\mathrm{CSM,HFM}}$

$\ell = 1, 2$ (for 2D problems) or $\ell = 1, 2, 3$ (for 3D problems) indicates vector component

From Eq. 18, it follows that

$$q_f^n = \frac{q_c^n}{1 - \mathring{\delta}_q^n}, \quad n = 1, \ldots, N_t, \tag{21}$$

which enables us to express the QoI error δ_q^n in terms of relative error $\mathring{\delta}_q^n$ as

$$\delta_q^n = q_c^n \times \left(\frac{\mathring{\delta}_q^n}{1 - \mathring{\delta}_q^n} \right), \quad n = 1, \ldots, N_t. \tag{22}$$

Using our estimate for the error $\hat{\delta}_q^n$ in Eq. 22, we arrive at

$$\hat{\delta}_q^n = q_c^n \times \left(\frac{\hat{\mathring{\delta}}_q^n}{1 - \hat{\mathring{\delta}}_q^n} \right), \quad n = 1, \ldots, N_t. \tag{23}$$

In this modeling approach, we implicitly assume that the samples (each corresponding to a time step) are independent and identically distributed (i.i.d.). This enables us to frame the construction of $\hat{\mathring{r}}_q$ as a prediction problem rather than as a time-series problem. This framing is appropriate because, in our assessments, we perform coarse-scale simulation for different models $\mathbf{m}$, but over the same time interval.

2.6 Realization selection for EMML training dataset

To construct the mapping $\hat{\mathring{r}}_q$, $q = o, w, inj$, we first collect the EMML training dataset. The construction of the EMML training dataset involves simulating both the HFM and the CSM for a fraction of the realizations in the ensemble. In the context of uncertainty quantification, if the ensemble contains N_{real} different geological realizations, we define the EMML training set as $\mathcal{T}_{\text{train}} := \{\mathbf{m}^1, \mathbf{m}^2, \ldots, \mathbf{m}^{N_{\text{train}}}\}$, where N_{train} denotes the number of realizations in the EMML training dataset ($N_{\text{train}} < N_{\text{real}}$). The remaining $N_{\text{real}} - N_{\text{train}}$ realizations in the ensemble comprise the EMML test dataset $\mathcal{T}_{\text{test}} := \{\mathbf{m}_{\text{test}}^1, \mathbf{m}_{\text{test}}^2, \ldots, \mathbf{m}_{\text{test}}^{N_{\text{real}}-N_{\text{train}}}\}$. Therefore, the ensemble can be described as $\mathcal{T}_{\text{train}} \cup \mathcal{T}_{\text{test}}$. The EMML training dataset inform the relationship between the response δ_q^n and the features described in Table 1.

To select the N_{train} realizations that constitute $\mathcal{T}_{\text{train}}$ from the full ensemble of N_{real} realizations, we cluster the N_{real} realizations, based on the well-pair volumes (WPVs) and the well-allocation factors (WAFs), into N_{train} clusters. We then choose the realizations that are closest to the cluster centers. To compute the WPV and WAF for a realization, we solve the stationary time-of-flight (TOF) and the stationary tracer equation at the fine scale using MRST [28, 32]. See [37, 46] for more details. The WPV represents the pore volume associated with an injector–producer pair, and the WAF indicates the fraction of the flux at a production well that can be attributed to a given injection well. As shown by [37, 48], the TOF and the tracer information provide a dynamic measure of heterogeneity, and can thus be used to cluster the realizations. We note that clustering can also be performed based on other measures, as discussed in [44, 47, 50], or by ranking the realizations [19, 27, 36, 43] based on particular metrics and then selecting representative realizations.

2.7 Feature-space partitioning and error prediction

As an alternative to constructing one global regression model $\mathring{r}_q$ for a given well over the entire time frame of the EMML training dataset, we instead partition the EMML training dataset into categories or groups. These categories are designed to capture the different stages that typically exist in water-flooding problems. We then construct a "local" error model for each category, with the intent of improving prediction accuracy. As described in [52], we can use classification or clustering to perform feature-space partitioning. In this work, we focus on classification as it was shown to perform well in [52].

Classification, a supervised machine learning technique [22], involves the construction of a statistical model that identifies the relationship between category membership $L \in \mathbb{R}$ and categorization criteria, referred as classification features. The statistical model is constructed based on the samples for which the category membership and the classification features are known. Following [52], we perform feature-space partitioning only for production wells $d_p \in \mathcal{D}_P$, since injection-well behavior does not display clear stages. For the production wells, we construct unique local models (in each partition) for each QoI, i.e., oil and water production rates.

The first step in classification is to define the category L for all samples in the EMML training dataset. These categories are defined based on HFM and CSM water breakthrough and water cut (fraction of water in the produced fluid) characteristics. The specific categorization, with reference to Fig. 2, is as follows. Samples for which water breakthrough has not occurred in either the CSM or

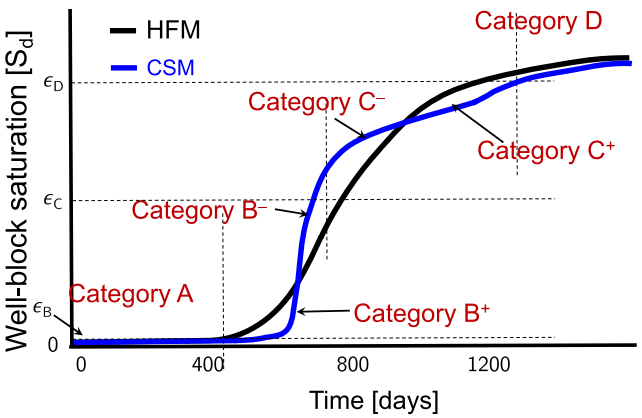


Fig. 2 Category assignment for a production well for the EMML training dataset

the HFM, i.e., $S_d^c \leq \epsilon_B$ and $S_d^{\text{HFM}} \leq \epsilon_B$, are assigned to category A. Samples around water breakthrough, i.e., $\epsilon_B < S_d^c \leq \epsilon_C$, are assigned to category B^+ (if $S_d^c < S_d^{\text{HFM}}$) or B^- (if $S_d^c \geq S_d^{\text{HFM}}$). Samples beyond the breakthrough transition, i.e., $\epsilon_C < S_d^c \leq \epsilon_D$, are assigned to category C^+ (if $S_d^c < S_d^{\text{HFM}}$) or C^- (if $S_d^c \geq S_d^{\text{HFM}}$). Finally, later-time samples with $S_d^c > \epsilon_D$ are assigned to category D. For each production well d_p, we therefore have $L = A, B^+, B^-, C^+, C^-$ or D. In this study, we set $\epsilon_B = 0.01$, $\epsilon_C = 0.25$ and $\epsilon_D = 0.5$.

Here, we take the classification features to be the same as the regression features, $\boldsymbol{f}^n,\ n = 1, \ldots, N_t$, defined above. Given the category membership L^n and the features for all samples in the EMML training dataset, we use random forest to construct the statistical model for the classification. For a sample in $\mathcal{T}_{\text{train}}$, we have available the features described in Table 1, the true error $\delta_q^n,\ q = o, w, inj$, in the CSM prediction for the QoI, and the category assignment L (A, B^+, B^-, C^+, C^- or D). For a well located in grid block d, this information is represented as

$$\mathbb{F} = \begin{bmatrix} (\boldsymbol{f}^1)^1 \\ (\boldsymbol{f}^2)^1 \\ \vdots \\ (\boldsymbol{f}^{N_t})^1 \\ (\boldsymbol{f}^1)^2 \\ (\boldsymbol{f}^2)^2 \\ \vdots \\ (\boldsymbol{f}^{N_t})^2 \\ \vdots \\ (\boldsymbol{f}^{N_t})^{N_{\text{train}}} \end{bmatrix}, \underline{\boldsymbol{\delta}}_q = \begin{bmatrix} \left(\delta_q^1\right)^1 \\ \left(\delta_q^2\right)^1 \\ \vdots \\ \left(\delta_q^{N_t}\right)^1 \\ \left(\delta_q^1\right)^2 \\ \left(\delta_q^2\right)^2 \\ \vdots \\ \left(\delta_q^{N_t}\right)^2 \\ \vdots \\ \left(\delta_q^{N_t}\right)^{N_{\text{train}}} \end{bmatrix}, \mathbb{L} = \begin{bmatrix} (L^1)^1 \\ (L^2)^1 \\ \vdots \\ (L^{N_t})^1 \\ (L^1)^2 \\ (L^2)^2 \\ \vdots \\ (L^{N_t})^2 \\ \vdots \\ (L^{N_t})^{N_{\text{train}}} \end{bmatrix},$$

where $q = o, w, inj$, $\mathbb{F} \in \mathbb{R}^{N_{\text{train}} N_t \times N_f}$, $\underline{\boldsymbol{\delta}}_q \in \mathbb{R}^{N_{\text{train}} N_t \times 1}$, $\mathbb{L} \in \mathbb{Z}^{+^{N_{\text{train}} N_t \times 1}}$, and the superscript on the parenthesis is a counter for the fine-scale simulation in $\mathcal{T}_{\text{train}}$. For the test realizations $\mathbf{m}_{\text{test}} \in \mathcal{T}_{\text{test}}$, we only have information for features $\boldsymbol{f}^n$.

Finally, we note that for samples in categories B^- and C^-, relative error $(\mathring{\delta}_q^n)_d$ can potentially be large. This is because, in the absence of water breakthrough in the fine-scale simulation, the denominator $(q_f^n)_d$ in the definition of relative error in Eq. 18 can be very small. Therefore, for scaling purposes, for all samples in categories B^- and C^-, we redefine the relative error so it is in terms of $(q_c^n)_d$ rather than $(q_f^n)_d$. This gives

$$(\mathring{\delta}_q^n(\mathbf{m}))_d := \frac{(\delta_q^n)_d}{(q_c^n)_d}, \quad n = 1, \ldots, N_t, d \in \mathcal{D}. \tag{24}$$

Following the development in Section 2.5, we now arrive at

$$\hat{\delta}_q^n = q_c^n \times \hat{\mathring{\delta}}_q^n,\ n = 1, \ldots, N_t. \tag{25}$$

3 Numerical results

In this section, we present numerical results demonstrating the quality of the estimated error $(\hat{\delta}_q^n)_d$ obtained using the EMML framework. We perform numerical experiments on two sets of synthetic models. The first set corresponds to a bimodal channelized system. The HFMs contain 100×100 grid blocks, and these are upscaled to 10×10 CSMs. The second set of models involves three-dimensional Gaussian permeability fields defined on grids containing $50 \times 50 \times 20$ blocks. The corresponding upscaled models are of dimensions $10 \times 10 \times 2$. In both cases, the permeability field on the fine scale is isotropic, and the porosity is constant and set to 0.2. Similarly, in both cases, the relative permeability curves are given by $k_{rw}(S) = S^2$ and $k_{ro}(S) = (1 - S)^2$, on both the fine and coarse scales. The fluids in both cases are slightly compressible. The oil and water viscosities are 3 cp and 1 cp, and the oil and water densities are 641 kg/m^3 and 997 kg/m^3. Capillary pressure effects are neglected. MRST [28, 32] is used for both the CSM and the HFM simulations.

3.1 Error modeling for 2D channelized models

The two-dimensional bimodal channelized system contains four production wells (P_1–P_4) and one injection well (I_1) arranged in a five-spot pattern. The fine-grid cells are of dimensions 32 ft × 32 ft × 30 ft, while the coarse-grid cells are of dimensions 320 ft × 320 ft × 30 ft. The initial reservoir pressure is 4700 psi and the initial water saturation is 0.001. Geological uncertainty is treated by considering an ensemble of 250 unconditioned realizations. Figure 3 shows six realizations from the ensemble: three of these belong to $\mathcal{T}_{\text{train}}$, while the other three belong to $\mathcal{T}_{\text{test}}$. These realizations were generated using SNESIM [49], with a channelized training image and Gaussian permeability distributions in each facies introduced using the cookie-cutter approach. For more details on the models, refer to [54]. The wells in all of the simulations are BHP controlled, with the BHPs for all of the production wells set to 2200 psi, and the BHP for the injection well set to 9800 psi. As BHPs are prescribed in this case, the QoI are the injection and production rates.

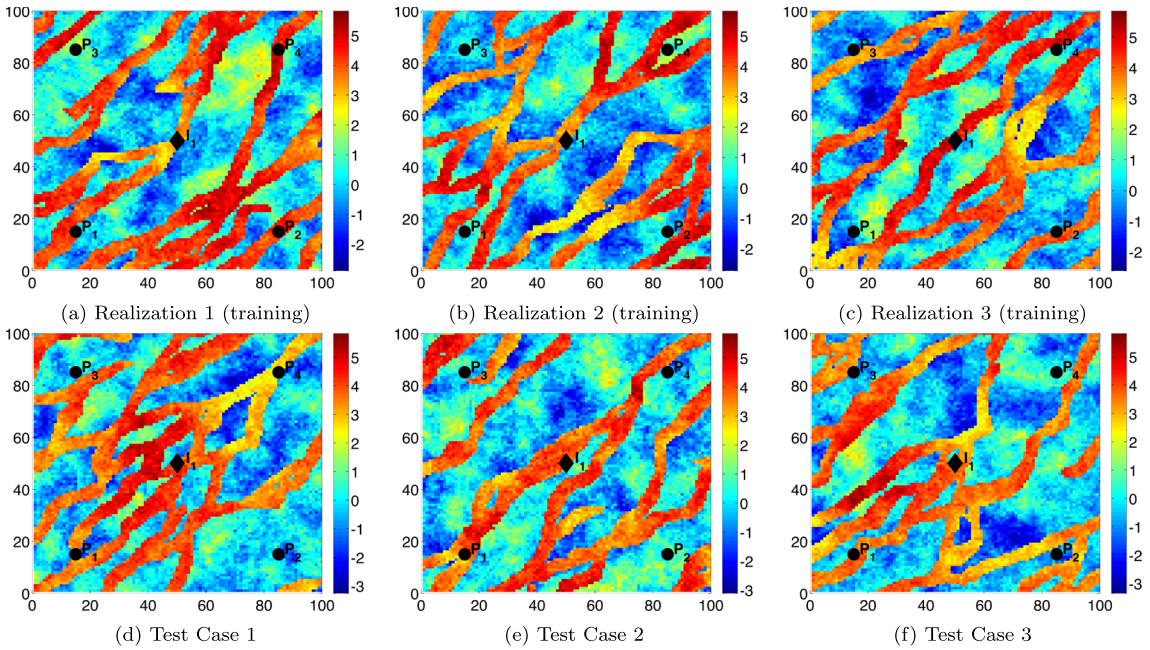


Fig. 3 Realizations of the permeability field ($\log k$) from the ensemble of two-dimensional bimodal channelized models. Well locations are also shown

3.1.1 Upscaling error in channelized model

We first illustrate the performance of the SPTU upscaling method for single-phase flow for a particular realization. The log transmissibility in the y-direction for the HFM and the CSM, corresponding to Test Case 4, are shown in Fig. 4. Well-by-well flow rates for the CSM and HFM are shown in Fig. 5. There we see that, despite the fact that the channels are not evident in the CSM (Fig. 4b) due to the high degree of coarsening, the upscaling procedure nonetheless provides accurate production and injection rates for all wells. A high degree of accuracy in these quantities is typically observed using this global upscaling method.

Highly coarsened models such as that shown in Fig. 4b are usually less accurate in terms of breakthrough and fluid distributions in two-phase flow problems. This is evident in Fig. 6, where we plot fieldwide oil cut for the waterflood problem of interest, for various coarsening levels. We observe that the error in the two-phase flow results increases with increasing coarsening, as expected, with the 10×10 CSM solution (blue curve) showing significant error relative to the HFM solution (black curve). The increase in error with degree of coarsening is accompanied, however, by a corresponding decrease in computational effort, so there is a clear tradeoff between accuracy and cost. Our intent in applying the EMML methodology is to correct the error associated with the 10×10 CSM solution while retaining much of the efficiency inherent in coarse-scale simulation.

In our evaluations of the EMML framework, we choose the level of upscaling such that the CSM provides significant speedup, but at the same time retains enough resolution to capture the basic flow response. This balance is achieved through application of SPTU and by retaining a

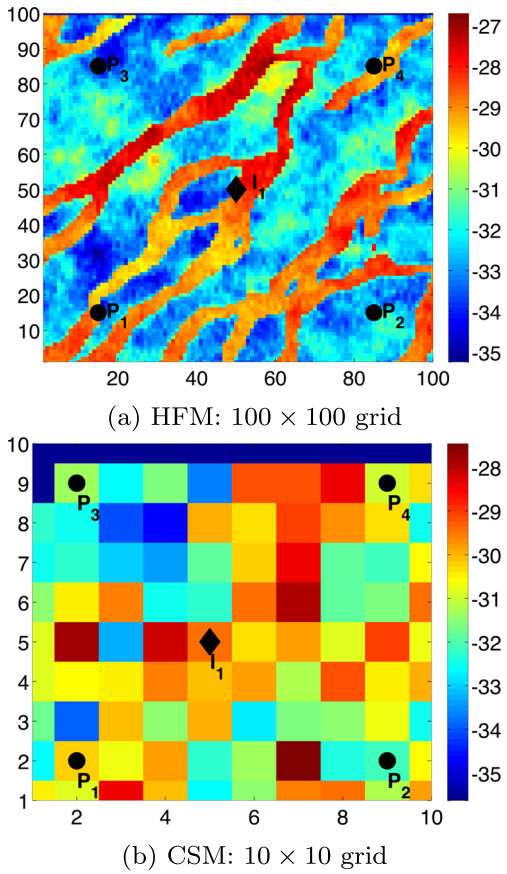


Fig. 4 Log transmissibility (with T in units of m^3) in the y-direction corresponding to Test Case 4 (two-dimensional bimodal channelized system)

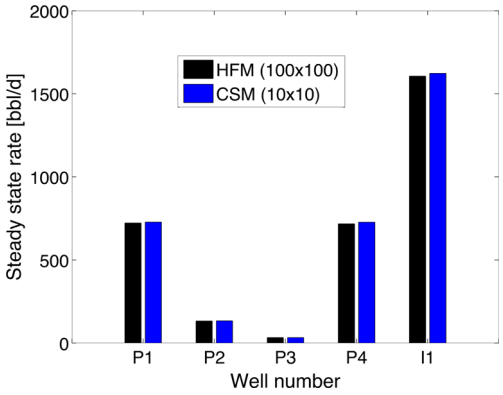


Fig. 5 Upscaling results for single-phase flow for Test Case 4 (two-dimensional bimodal channelized system)

reasonable number of grid blocks, for example five, between injector–producer pairs in the CSM. It may be possible to use even coarser models, but then the error model would have to be capable of accurately capturing large effects, which may require extension of the methodology.

To construct the error model $\hat{\hat{r}}_q$, we first generate the EMML training dataset, which involves performing 30 fine-scale and corresponding coarse-scale simulations (i.e., $N_{\text{train}} = 30$). To reduce the elapsed (wall-clock) time, these simulations are performed in parallel, using 12 cores. Here $\mathcal{T}_{\text{train}}$ contains 30 realizations, while $\mathcal{T}_{\text{test}}$ contains 220 realizations (the full ensemble contains a total of 250 unconditioned realizations). The error model, $\hat{\hat{r}}_q$, is constructed by performing classification on the EMML training dataset followed by regression. To estimate $(\delta_q^n)_d$ for a test sample, we first predict the category L^n, and we then compute the QoI error $\hat{\delta}_q^n$ accordingly.

Misclassification error, defined as the ratio of the number of EMML test samples misclassified to the total number of EMML test samples over all realizations in $\mathcal{T}_{\text{test}}$, may be incurred when predicting the category L. The percentage of misclassified EMML test samples is about 17% for this case. For the two-dimensional model considered here, the total number of features per sample (N_f), after removing the highly correlated features, is on average about 110 for production wells and 30 for the injection well. Note that each QoI may retain a different subset of features. Finally, we reiterate that none of the test cases is used in the construction of the EMML training dataset.

3.1.2 QoI error modeling for 2D model: Test Case 1

We first present results for Test Case 1, represented by $\mathbf{m}_{\text{test}}^1 \in \mathcal{T}_{\text{test}}$ and shown in Fig. 3d. Figure 7 displays the results for the production and injection rates for wells P_1 and I_1, and Fig. 8 shows the fieldwide oil cut and cumulative fieldwide production profiles. We present results for well P_1 in Fig. 7a, b since it has the highest cumulative liquid production. In all figures, the black curves denote the fine-scale solution, the blue curves indicate the coarse-scale solution (10×10 model), and the red curves show the corrected solution. The error in the CSM relative to the HFM is clearly evident in Figs. 7 and 8. The high

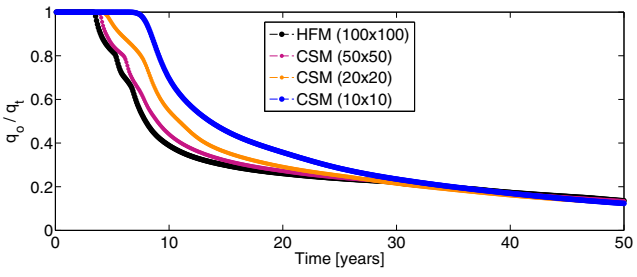
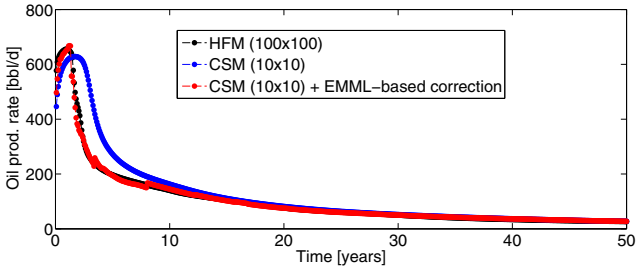
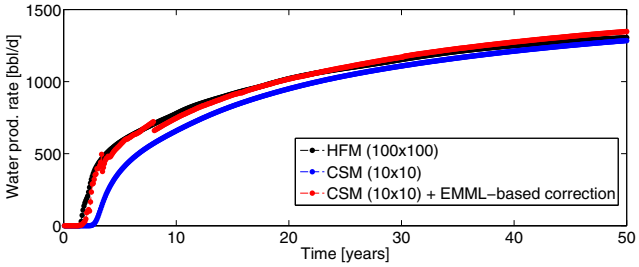


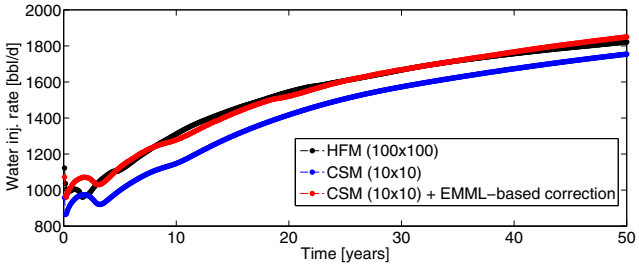
Fig. 6 Fieldwide oil cut with varying coarsening levels for Test Case 4 (two-dimensional bimodal channelized model)



(a) Oil production rate

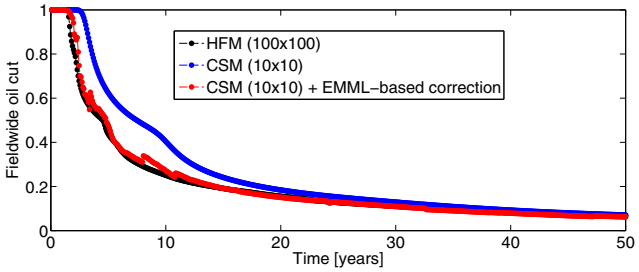


(b) Water production rate

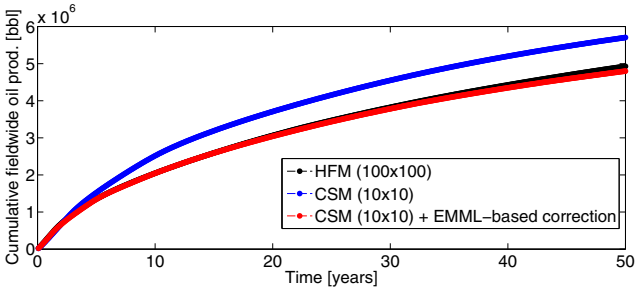


(c) Water injection rate

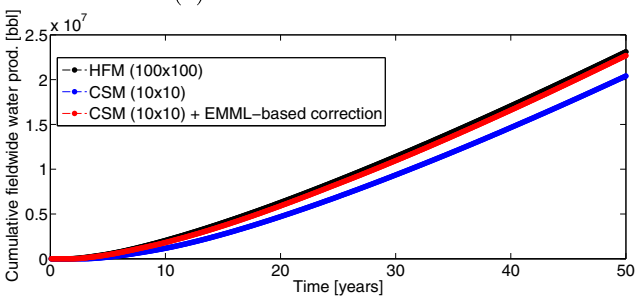
Fig. 7 EMML for QoI correction: Test Case 1. Production (P_1) and injection (I_1) rates for various models for two-dimensional bimodal channelized system, $N_{\text{train}} = 30$



(a) Fieldwide oil cut



(b) Fieldwide cumulative oil



(c) Fieldwide cumulative water

Fig. 8 EMML for QoI correction: Test Case 1. Fieldwide oil cut and cumulative production profiles for various models for two-dimensional bimodal channelized system, $N_{\text{train}} = 30$

degree of upscaling leads to a loss of resolution of the channels, which in turn results in late water breakthrough in the CSM. We observe, however, that the corrected solution provides significantly improved accuracy in all quantities, including breakthrough time (as is evident in Figs. 7b and 8a).

Note, however, that the corrected solutions for production and injection rates are not always smooth. This is because, when computing $\hat{\delta}_q^n$, we treat each time step as a distinct (independent) estimate due to the i.i.d. assumption. By contrast, results for the corrected cumulative production and injection, which entail time integration, are smooth, as is evident in Fig. 8b, c. It would of course be straightforward to apply a simple smoothing operation to the corrected results, which would act to reduce the fluctuations. It may also be possible to build more "memory" into the features, which could act to limit fluctuations in the EMML estimates.

3.1.3 QoI error modeling for additional models

We now present representative results for Test Cases 2–4. Time-integrated errors for the realizations in the set $\mathcal{T}_{\text{test}}$ will then be computed. Test Case 2 corresponds to a lower time-integrated error than Test Case 1, while Test Cases 3 and 4 correspond to similar or higher time-integrated errors than Test Case 1. Results for fieldwide oil cut for the three cases are shown in Figs. 9, 10, and 11. The corrected solutions display a high degree of accuracy relative to the HFM solutions for all cases (though fluctuations are again evident). Additional results for Test Cases 2–4 are presented in [51]. These results are generally comparable to those shown earlier for Test Case 1.

We now evaluate EMML performance for all 250 realizations. To enable this assessment, we define the following time-integrated error measures:

$$E_{o,w}^{\text{CSM}} = \frac{\sum_{d_p=1}^{n_{pw}} \sum_{n=1}^{N_t} \left| \left(\delta_{o,w}^n\right)_{d_p} \right| \Delta t^n}{\sum_{d_p=1}^{n_{pw}} \sum_{n=1}^{N_t} \left(q_{o,w}^n\right)_{d_p}^{\text{HFM}} \Delta t^n}, \tag{26a}$$

$$E_{o,w}^{\text{corr}} = \frac{\sum_{d_p=1}^{n_{pw}} \sum_{n=1}^{N_t} \left| \left(\hat{\delta}_{o,w}^n\right)_{d_p} - \left(\delta_{o,w}^n\right)_{d_p} \right| \Delta t^n}{\sum_{d_p=1}^{n_{pw}} \sum_{n=1}^{N_t} \left(q_{o,w}^n\right)_{d_p}^{\text{HFM}} \Delta t^n}, \tag{26b}$$

$$E_{inj}^{\text{CSM}} = \frac{\sum_{d_i=1}^{n_{iw}} \sum_{n=1}^{N_t} \left| \left(\delta_{inj}^n\right)_{d_i} \right| \Delta t^n}{\sum_{d_i=1}^{n_{iw}} \sum_{n=1}^{N_t} \left(q_{inj}^n\right)_{d_i}^{\text{HFM}} \Delta t^n}, \tag{26c}$$

$$E_{inj}^{\text{corr}} = \frac{\sum_{d_i=1}^{n_{iw}} \sum_{n=1}^{N_t} \left| \left(\hat{\delta}_{inj}^n\right)_{d_i} - \left(\delta_{inj}^n\right)_{d_i} \right| \Delta t^n}{\sum_{d_i=1}^{n_{iw}} \sum_{n=1}^{N_t} \left(q_{inj}^n\right)_{d_i}^{\text{HFM}} \Delta t^n}, \tag{26d}$$

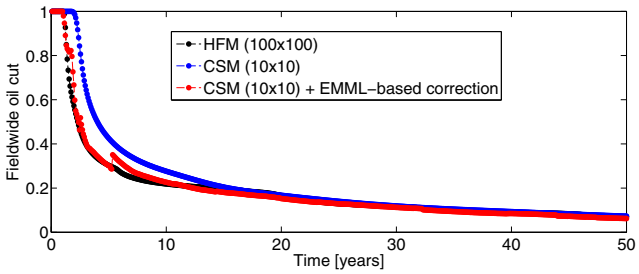


Fig. 9 EMML for QoI correction: Test Case 2. Fieldwide oil cut for various models for two-dimensional bimodal channelized system, $N_{\text{train}} = 30$

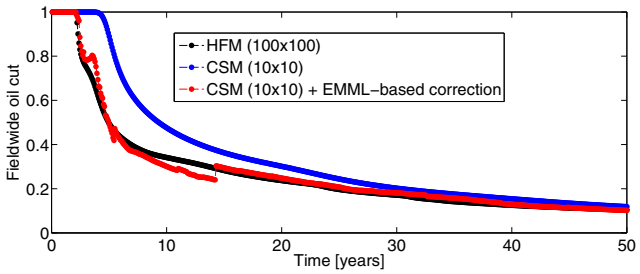


Fig. 10 EMML for QoI correction: Test Case 3. Fieldwide oil cut for various models for two-dimensional bimodal channelized system, $N_{\text{train}} = 30$

where $E^{\text{CSM}}_{o,w,inj}$ denotes the time-integrated error in the upscaled solutions, $E^{\text{corr}}_{o,w,inj}$ is the corresponding time-integrated error in the corrected solutions, o designates the oil production rate, w is water production rate and inj is the water injection rate, $\left(q^n_{o,w}\right)^{\text{HFM}}_{d_p}$ denotes the oil or water production rate for well d_p at time step n from the fine-scale simulation, $\left(q^n_{inj}\right)^{\text{HFM}}_{d_i}$ is the water injection rate for well d_i at time step n from the fine-scale simulation, and n_{p_w} and n_{i_w} are the number of production and injection wells.

The time-integrated errors (expressed as percentages) for all 250 realizations are shown in Fig. 12. For each realization in the full ensemble $\mathcal{T}_{\text{train}} \cup \mathcal{T}_{\text{test}}$, the time-integrated error in the upscaled solution is denoted by the blue points. The red points display the time-integrated error in the corrected solution for the 220 realizations in the EMML test dataset $\mathcal{T}_{\text{test}}$, and the green points show the error in the corrected solution for the 30 realizations in the EMML training set $\mathcal{T}_{\text{train}}$. The 250 realizations are sorted by increasing upscaling (CSM) error. Test Cases 1–4 (presented earlier) are identified in the figure. The median errors (%) corresponding to the EMML test dataset results in Fig. 12 are presented in Table 2.

From Fig. 12 and Table 2, we see that by correcting the CSM solution with the EMML error estimate, using $N_{\text{train}} = 30$, we reduce the three errors (E_o, E_w and E_{inj}) by about 61% on average. Based on the phase flow rate results

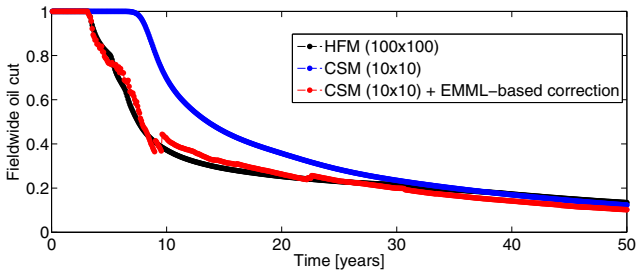
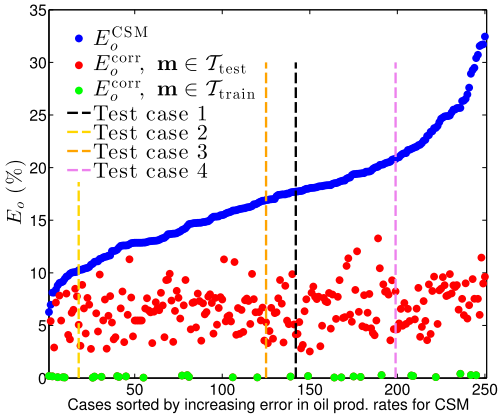
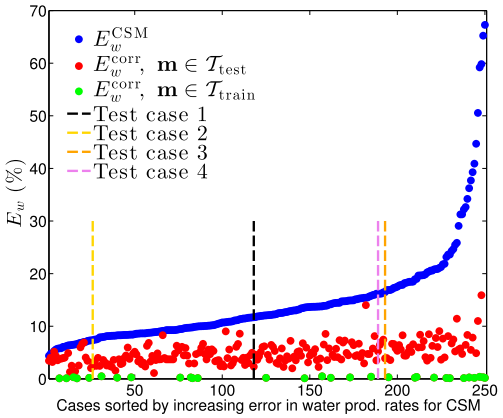


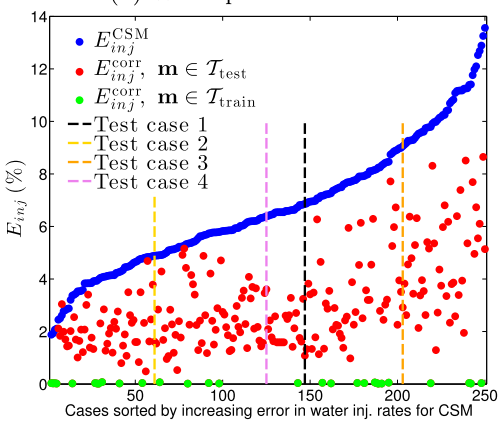
Fig. 11 EMML for QoI correction: Test Case 4. Fieldwide oil cut for various models for two-dimensional bimodal channelized system, $N_{\text{train}} = 30$



(a) Oil production rate



(b) Water production rate



(c) Water injection rate

Fig. 12 Time-integrated percent error in production and injection rates as defined by Eq. 26a–d (two-dimensional bimodal channelized system), $N_{\text{train}} = 30$

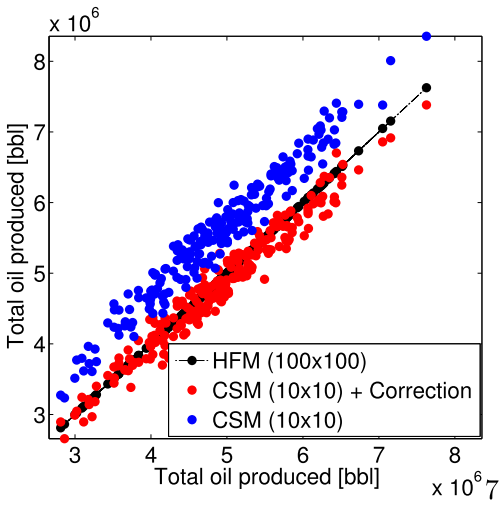
presented earlier (Figs. 7 and 11), this level of reduction in the time-integrated error may be less substantial than expected. The likely reason for this is that the EMML-based corrections reduce large errors, but they do not entirely eliminate more persistent small to medium errors. Similar behavior was observed in our previous study, where EMML was applied to correct error from a reduced-order

Table 2 EMML for QoI correction: Median value of the time-integrated errors with different EMML parameters for CSM and corrected solutions for cases in EMML test dataset (two-dimensional bimodal channelized system)

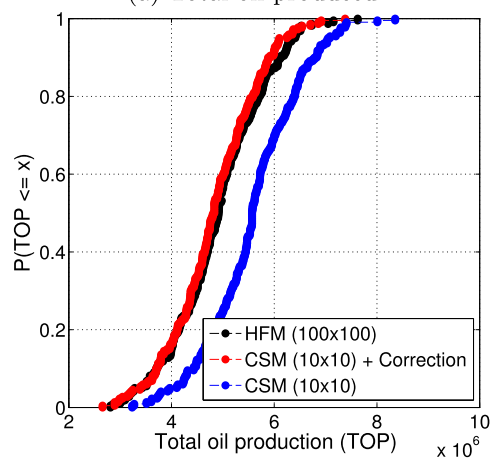
	E_o^i (%)	E_w^i (%)	E_{inj}^i (%)
i = CSM	17	12	6.4
i = corr, $N_{\text{train}} = 30$	6.6	4.5	2.6
i = corr, $N_{\text{train}} = 10$	7.0	7.9	2.8

model [52]. Note finally that the time-integrated errors in the corrected solutions corresponding to the realizations in $\mathcal{T}_{\text{train}}$ are very small but nonzero. This is because the error model $\hat{\hat{r}}_q$ does not overfit the EMML training dataset.

Realization-by-realization comparisons, and the cumulative distribution functions (CDFs) for the cumulative oil produced over the entire simulation ($T = 50$ years), are presented in Fig. 13. The CSM model predictions (blue points) are accompanied by a systematic bias in the total oil produced. The corrected solutions (red points) display improved accuracy in the realization-by-realization comparison, as well as in the CDF of the total oil produced. There is still some variance, however, evident from the scatter around the 45° line in Fig. 13a. Additional results for cumulative water production and total water injection are provided in [51]. For these quantities, CSM results appear to be more accurate than those in Fig. 13, though the corrected solutions still show improved accuracy. Note finally that, if we were only interested in predicting cumulative quantities over the entire simulation, we could potentially use a linear regression to fit the CSM and HFM results while correcting the bias. In this study, however, we are computing the integrated response directly from the corrected solution (i.e., we correct the rates at each time step and then compute the cumulative response).



(a) Total oil produced

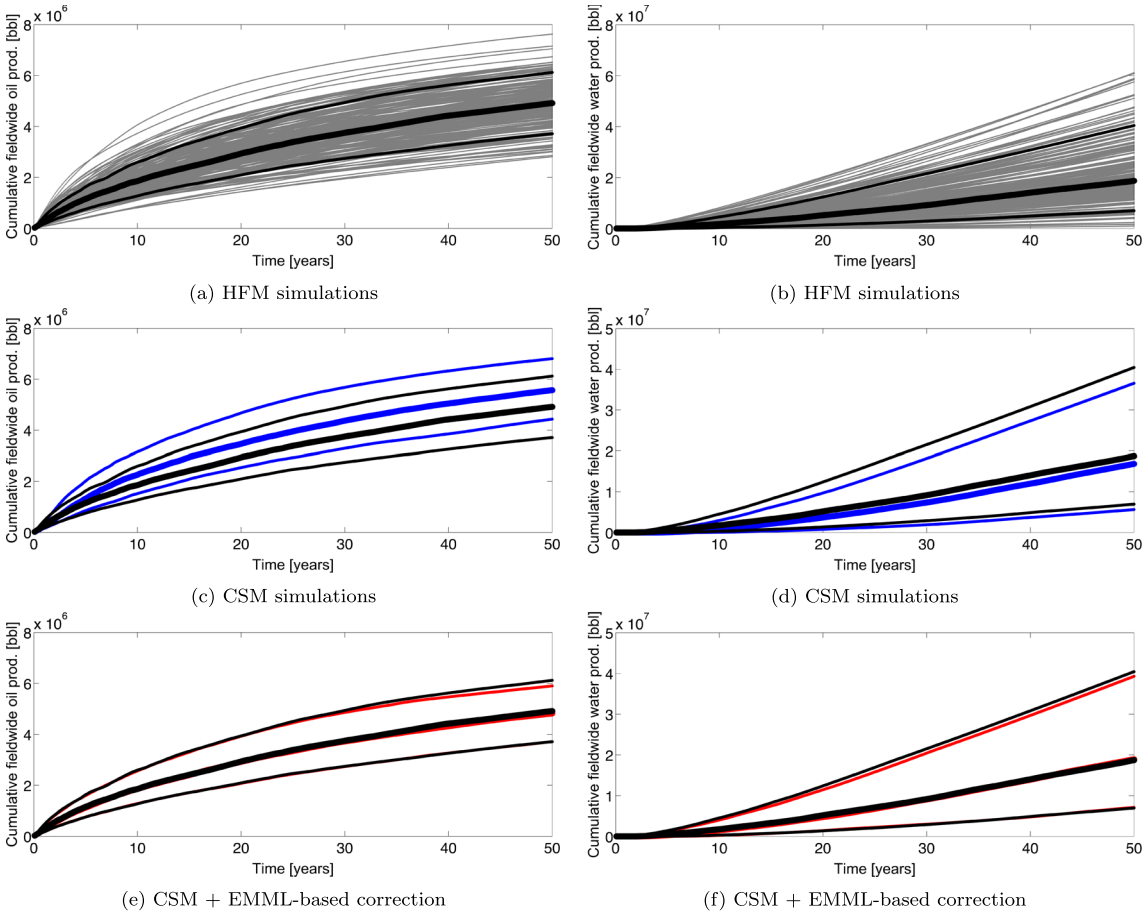


(b) CDF - Total oil produced

Fig. 13 Cross plots and CDFs for total oil production for realizations in $\mathcal{T}_{\text{test}}$ (two-dimensional bimodal channelized system), $N_{\text{train}} = 30$

We now assess the accuracy of the corrected solutions in terms of P10, P50, and P90 profiles for cumulative fieldwide oil and water production. These results, presented in Fig. 14, are obtained by computing the 10th, 50th, and 90th percentile of the cumulative fieldwide profile at each time step. The P50 profile is represented by the thick solid curves and the P10–P90 intervals by the thinner curves. The thin gray curves in Fig. 14a and b depict the fine-scale solutions for all realizations in the ensemble. The bias in the CSM solutions (also observed in Fig. 13) is evident in the overprediction of the P50 curve and the P10–P90 interval for the cumulative oil produced in Fig. 14c, and in the underprediction for the cumulative water produced in Fig. 14d. The corrected solutions demonstrate a high degree of accuracy, for the P10, P50, and P90 curves, in the cumulative fieldwide oil and water profiles shown in Fig. 14e, f. This is an important observation as it demonstrates that the EMML methodology is able to provide accurate results over a wide range of responses (i.e., not just for the P50 profiles), which is essential for UQ.

Because the error modeling procedure requires the offline simulation of 30 HFMs, it is important to demonstrate that the P10, P50, P90 results obtained from our procedure are more accurate than those computed directly from the 30 high-fidelity simulations used to construct the EMML training dataset. The P10, P50, and P90 results for these models, for cumulative fieldwide oil and water production, are shown in Fig. 15. It is clear that the EMML-generated results in Fig. 14e and f display higher overall accuracy than the results in Fig. 15. This is particularly evident for the P90 cumulative fieldwide water production result. It is possible that alternate selection/modeling procedures, such as those in [4, 44, 47], could perform better though this would have to be tested. In any event, the accuracy of the EMML-based P10, P50 and P90 results demonstrates the overall robustness of our framework.



(a) HFM simulations (b) HFM simulations

(c) CSM simulations (d) CSM simulations

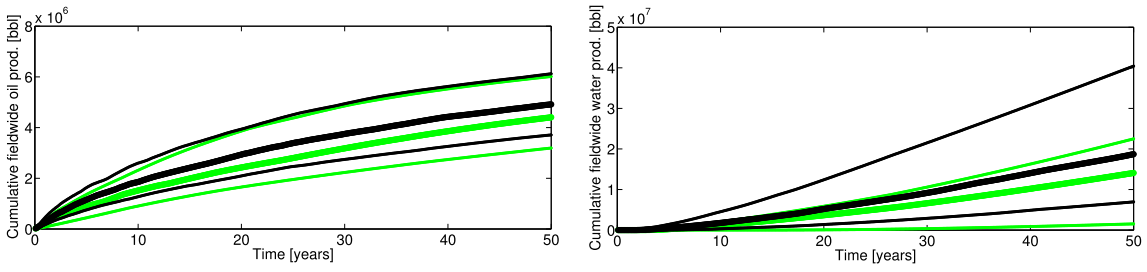
(e) CSM + EMML-based correction (f) CSM + EMML-based correction

Fig. 14 Statistical results for cumulative fieldwide oil (left) and water (right) production profiles for two-dimensional bimodal channelized system. In (**a**) and (**b**), the thin gray curves depict fine-scale solutions for individual realizations. In (**a**)–(**f**), P50 results shown as thick curves and P10–P90 interval as intermediate-weight curves. HFM solutions shown in black, CSM solutions in blue, and EMML-corrected solutions in red

We expect EMML performance to degrade if fewer HFMs are simulated in the training step. To quantify this decrease in accuracy, we constructed an error model $\hat{\hat{r}}_q$ using $N_{\text{train}} = 10$ (instead of $N_{\text{train}} = 30$ as presented above) HFM simulations. Detailed results are provided in [51], and median time-integrated errors are presented in Table 2. These errors are slightly higher than those for $N_{\text{train}} = 30$, though clear improvement in accuracy relative to the CSM results is still achieved. It is possible that the $N_{\text{train}} = 10$ results could be improved through use of an alternate training-realization-selection strategy.

3.1.4 Computational costs for 2D models

The timings for simulating the HFMs, the CSMs, and the offline (overhead) cost for constructing the error model are



(a) Oil (b) Water

Fig. 15 Comparison of P50 (thick curves) and P10–P90 interval (thinner curves) for cumulative fieldwide oil and water profiles between the HFM (black curves) and the EMML training simulations (green curves) for two-dimensional bimodal channelized system, $N_{\text{train}} = 30$

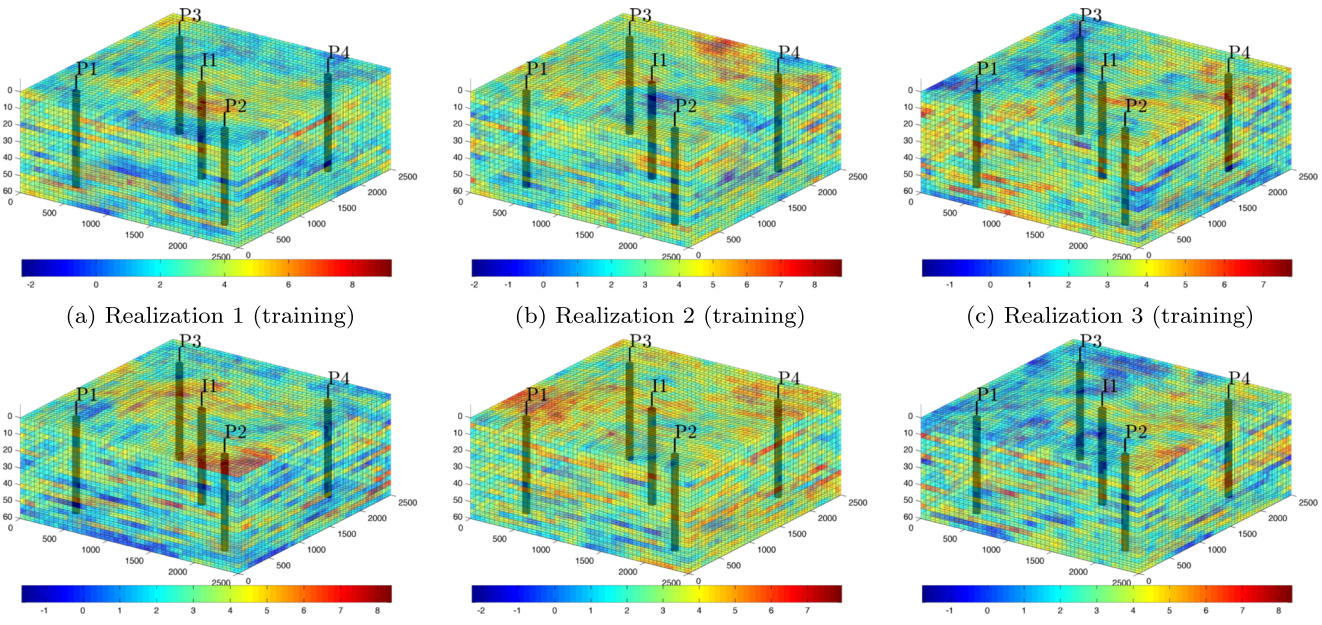
Table 3 Computational cost to simulate a HFM, CSM and to construct error model $\hat{\hat{r}}_q$ for 2D models

Description	Time (seconds)
Average time to simulate a 100×100 HFM	601
Average time to simulate a 10×10 CSM	22
Average time to solve Eq. 17 to generate features, for one model (required for realizations in both $\mathcal{T}_{\text{train}}$ and $\mathcal{T}_{\text{test}}$)	9
Average time to construct the classification model for each production well	804
Average time to construct $\hat{\hat{r}}_q$, $q = o, w$ for each production well	776
Average time to construct $\hat{\hat{r}}_q$, $q = inj$ for injection well	554
Time to simulate all realizations (on fine and coarse scales) in EMML training set $\mathcal{T}_{\text{train}}$ in a parallel environment using 10 cores	1803

summarized in Table 3. These timings are for $N_{\text{train}} = 30$. We use MRST [28, 32], which is a Matlab code, for all fine and coarse-scale simulations. The combined cost of simulating a CSM and solving Eq. 17 (needed to generate features) provides a speedup of about 20 relative to HFM simulation. More substantial speedups (of at least a factor of N_c/N_c^*) would be expected for realistic examples run with a commercial-type simulator.

The offline cost for the EMML framework involves (1) the construction of the EMML training dataset $\mathcal{T}_{\text{train}}$, (2) the construction of the classification model for each production well $d \in \mathcal{D}_P$, (3) the construction of $\hat{\hat{r}}_q$, $q = o, w$, for each production well $d \in \mathcal{D}_P$, and (4) the construction of $\hat{\hat{r}}_q$, $q = inj$, for the injection well $d \in \mathcal{D}_I$. The average wall-clock time to construct the classification model for a production well is 804 s, the average time to construct $\hat{\hat{r}}_q$ $(q = o, w)$ for a production well is 776 s, and the average time to construct $\hat{\hat{r}}_q$ $(q = inj)$ for an injection well is 554 s. Assuming serial processing, the total offline cost corresponding to the construction of all of the statistical models is 9978 s, which is comparable to the cost of about 17 HFM simulations. We emphasize that the offline cost, though significant, is only incurred once, and these computations can be easily accelerated through use of parallel computation.

The total serial time for simulating 250 HFMs is 41.7 h. By contrast, the total serial time to execute the full EMML-based procedure, including simulating the fine and coarse-scale models in $\mathcal{T}_{\text{train}}$ (with $N_{\text{train}} = 30$), constructing the statistical models for all wells, and simulating the additional 220 models in $\mathcal{T}_{\text{test}}$, is 9.9 h. Thus the method provides a relatively modest speedup of a factor of 4.2 for this case. Note however that these timings are for a basic Matlab implementation—we expect the construction



(a) Realization 1 (training) (b) Realization 2 (training) (c) Realization 3 (training)

(d) Test Case 1 (e) Test Case 2 (f) Test Case 3

Fig. 16 Realizations of the permeability field ($\log k$) from the ensemble of three-dimensional Gaussian models. Well locations are also shown

of the statistical models to be faster in an optimized code. In addition, although the coarse model contains 1% of the grid blocks of the fine model, it runs only 27 times faster, which indicates that the coarse model timings are strongly impacted by overhead. The cost of constructing the error model is independent of the HFM physics, since it depends only on the number of samples and the number of features in the EMML training dataset. We therefore expect considerably better speedup for larger and more complex problems. Offline cost will, however, increase with the number of wells, since an error model must be constructed for each well. The online cost of computing $\hat{\delta}_q^n$ is negligible—only about 0.0004 s in this example.

3.2 Error modeling for 3D Gaussian models

The three-dimensional Gaussian models, defined on a grid of dimensions $50 \times 50 \times 20$, contain four production wells (P_1–P_4) and one injection well (I_1). The wells are fully penetrating and are arranged in a five-spot pattern. The upscaled models correspond to a grid of dimensions $10 \times 10 \times 2$. The fine-grid cells are 50 ft $\times$ 50 ft $\times$ 3 ft, while the coarse-grid cells are 250 ft $\times$ 250 ft $\times$ 30 ft. The initial reservoir pressure is 4200 psi, and the initial water saturation is 0.001.

We consider an ensemble of 300 unconditioned realizations. Figure 16 shows six realizations from the ensemble: three realizations belong to $\mathcal{T}_{\text{train}}$, and three belong to $\mathcal{T}_{\text{test}}$. These realizations are generated using sequential Gaussian simulation [41], with the mean of log-permeability set to 3 and the standard deviation of log-permeability set to 1.5. The dimensionless permeability correlation lengths (normalized by the system length in the corresponding direction) are 0.5 in the x- and y-direction and 0.05 in the z-direction. All of the wells are BHP controlled, with the producer BHPs set to 2700 psi and the injector BHP set to 7000 psi.

Gravitational effects are included in this example. Thus, Darcy's law is now as follows:

$$\mathbf{v}_j = -\lambda_j \mathbf{k} \cdot \left(\nabla p - \rho_j g \nabla D\right), \tag{27}$$

where D is depth, g is gravitational acceleration, and all other variables are as previously defined. The computation of the upscaled quantities (ϕ^*, T^* and WI^*), the single-phase velocity $\mathbf{v}_{1\text{p}}$, and the volume averaging procedure (described in Section 2.3) are generalized to three dimensions and to include gravity. For each coarse block the well intersects, we compute the features described in Section 2.4, and then concatenate them to construct the feature vector $\boldsymbol{f}^n, n = 1, \ldots, N_t$.

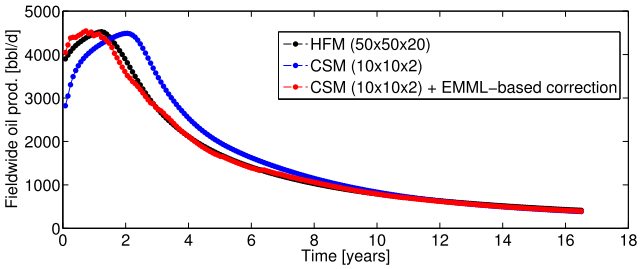
We next evaluate the quality of the EMML error estimates for different realizations in the EMML test dataset. As in the 2D case, the error model $\hat{\hat{r}}_q$ is constructed by performing classification prior to regression. The EMML training is accomplished by simulating $N_{\text{train}} = 30$ HFM realizations. The misclassification percentage, based on all of the samples in $\mathcal{T}_{\text{test}}$, is 8.5%.

3.2.1 Upscaling error modeling for 3D models

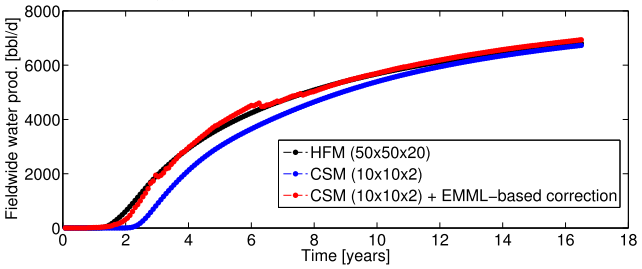
Because the results for the 3D models are very consistent with those for the 2D models considered in Section 3.1, our presentation here is quite concise. For more extensive 3D results (including results for individual wells) and additional discussion, please refer to [51].

Fieldwide production and injection rates for Test Case 1 ($\log k$ field shown in Fig. 16d), are displayed in Fig. 17. All curves are as previously defined. A noticeable error in the CSM solution is evident in all three fieldwide quantities. By correcting the CSM solution using the EMML-driven error estimates $\hat{\delta}_q^n, n = 1, \ldots N_t, d \in \mathcal{D}, q = o, w, inj$, we achieve clear improvement in accuracy for each of the QoI.

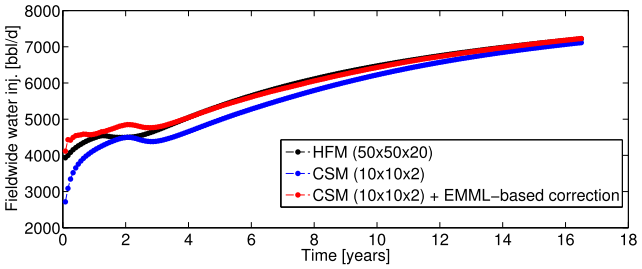
Fieldwide oil cut results for Test Cases 2 and 3 ($\log k$ fields shown in Fig. 16e, f) are presented in Figs. 18 and 19. Test Case 2 corresponds to smaller time-integrated



(a) Fieldwide oil production rate



(b) Fieldwide water production rate



(c) Fieldwide water injection rate

Fig. 17 EMML for QoI correction: Test Case 1. Fieldwide production and injection rates for various models for three-dimensional Gaussian system, $N_{\text{train}} = 30$

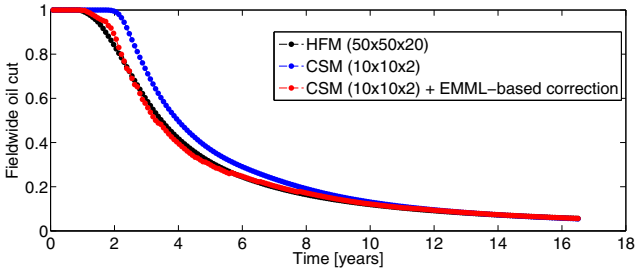


Fig. 18 EMML for QoI correction: Test Case 2. Fieldwide oil cut for various models for three-dimensional Gaussian system, $N_{\text{train}} = 30$

errors relative to Test Case 1, while Test Case 3 corresponds to larger errors. The corrected oil cut results display clear improvement over the CSM solutions, though some discrepancy is still evident at early time in Fig. 19.

Figure 20 displays results for the time-integrated errors (E_o^{corr}, E_w^{corr}, E_{inj}^{corr}) for the CSMs and corrected solutions. All points are as previously defined, and Test Cases 1–3 are identified. The test cases are again sorted by increasing upscaling error. We see that the corrected solutions (red points) demonstrate significantly improved accuracy relative to the CSM solutions (blue points). The time-integrated errors in the corrected solutions corresponding to the realizations in $\mathcal{T}_{\text{train}}$ (green points) are again small but nonzero. The median errors for these results are as follows: $E_o^{\text{CSM}} = 13.2\%$, $E_o^{\text{corr}} = 2.5\%$; $E_w^{\text{CSM}} = 6.3\%$, $E_w^{\text{corr}} = 1.5\%$; and $E_{inj}^{\text{CSM}} = 3.7\%$, $E_{inj}^{\text{corr}} = 0.6\%$. The corrected solutions are clearly quite accurate for this three-dimensional system.

Realization-by-realization comparisons and CDFs for total oil production, obtained using the HFMs, CSMs and corrected solutions, are shown in Fig. 21. As was observed earlier in Fig. 13, the bias exhibited by the CSM solutions is largely eliminated through application of the error model, though some variance (scatter) still exists. Note that the overall variation between realizations is much less in this case than in the two-dimensional bimodal channelized system (as is evident from the scales in Figs. 21 and 13).

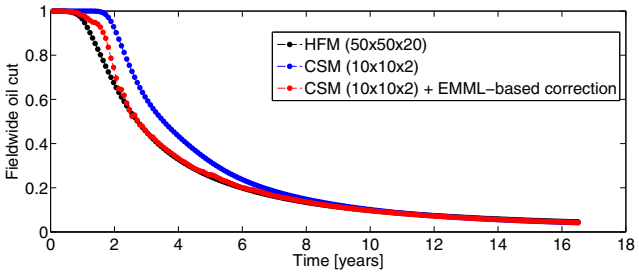
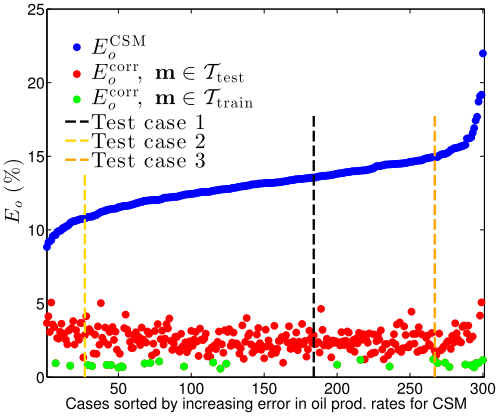
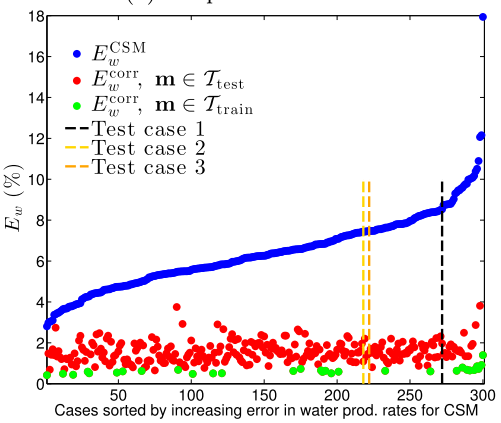


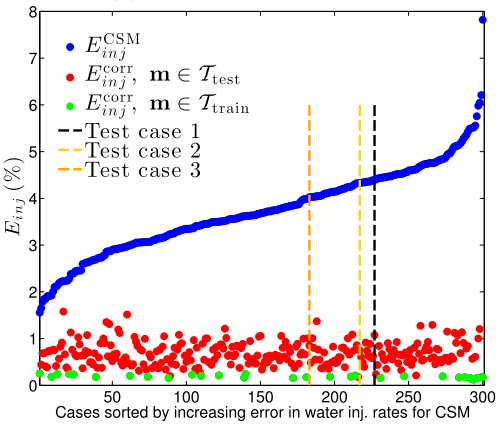
Fig. 19 EMML for QoI correction: Test Case 3. Fieldwide oil cut for various models for three-dimensional Gaussian system, $N_{\text{train}} = 30$



(a) Oil production rate



(b) Water production rate

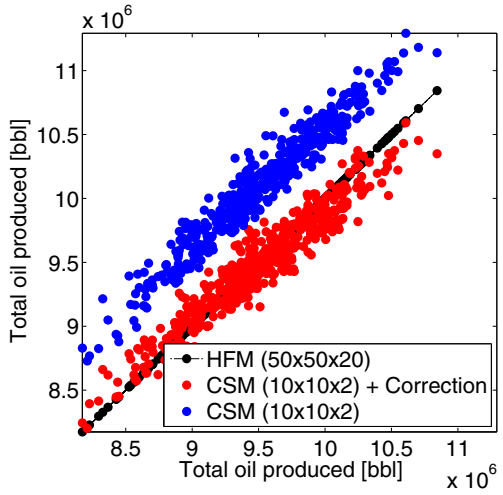


(c) Water injection rate

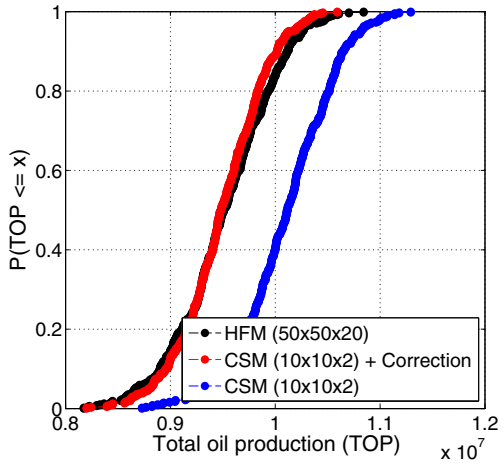
Fig. 20 Time-integrated percent error in production and injection rates as defined by Eq. 26a–d (three-dimensional Gaussian system), $N_{\text{train}} = 30$

3.2.2 Computational costs for 3D models

Timings for this example are provided in Table 4. As before, all simulations are run using MRST [28, 32]. For this system, performing the coarse-scale simulation and solving Eq. 17 corresponds to a speedup of about 50 relative to



(a) Total oil produced



(b) CDF - Total oil produced

Fig. 21 Cross plots and CDFs for total oil production for realizations in $\mathcal{T}_{\text{test}}$ (three-dimensional Gaussian system), $N_{\text{train}} = 30$

a fine-scale simulation. Assuming serial processing, the construction cost of the statistical models (described in Section 3.1.4) is 157.5 minutes, which is about the same as the time required for a fine-scale simulation. Elapsed time could again be reduced through parallel processing. The EMML offline cost, relative to the fine-scale simulation cost, is much lower in this case (here this cost ratio is about 1, while for the two-dimensional case it was about 17). This is due to the large computational effort needed to simulate the fine-scale three-dimensional models in Matlab. Note that the elapsed time required to construct $\hat{\hat{r}}_q$ ($q = o, w, inj$) for a well is similar to that for the two-dimensional model. The online cost of computing $\hat{\delta}_q^n$ is again only about 0.0004 s.

The total serial time required to simulate 250 HFMs in this case is 625 h, while the serial time to execute the full EMML-based procedure is about 90 h. Thus, the method results in a speedup factor of about 7 for this three-dimensional example. Again, these timings would be modified by more efficient implementations, use of commercial simulation software, and parallel processing.

4 Concluding remarks

In this work, we applied machine learning to estimate upscaling error over multiple geological realizations in oil–water flow simulations. Our treatments and results are directly relevant to uncertainty quantification, where the goal is to evaluate reservoir performance over an ensemble of geomodels. The upscaled models considered here were generated using global single-phase transmissibility upscaling, which is (essentially) the most accurate such technique available. Subgrid transport effects, however, are not modeled in this procedure, and this leads to error in coarse-scale simulations. Estimates for these subgrid effects, generated by solving an approximate fine-scale transport equation (with a constant-in-time velocity field), were used to provide the features required by the machine learning framework. The set of features also included global properties such as well-pair volumes and well-allocation factors. The error model was constructed by regressing the true upscaling error in the two-phase simulation outputs (known for all of the training samples) on the features. This involves first constructing the EMML training dataset $\mathcal{T}_{\text{train}}$ by simulating both the fine- and coarse-scale models corresponding to some (representative) realizations in the

Table 4 Computational cost to simulate the HFM and CSM (using MRST) and to construct $\hat{\hat{r}}_q$ for 3D models

Description	Time (minutes)
Average time to simulate a $50 \times 50 \times 20$ HFM	150
Average time to simulate a $10 \times 10 \times 2$ CSM	0.85
Average time to solve Eq. 17 to generate features, for one model (required for realizations in both $\mathcal{T}_{\text{train}}$ and $\mathcal{T}_{\text{test}}$)	2.1
Average time to construct the classification model for each production well	13.8
Average time to construct $\hat{\hat{r}}_q$, $q = o, w$ for each production well	9.7
Average time to construct $\hat{\hat{r}}_q$, $q = inj$ for injection well	24.7
Time to simulate all realizations (on fine and coarse scales) in EMML training set $\mathcal{T}_{\text{train}}$ in a parallel environment using 10 cores (offline)	450

ensemble. For the remaining realizations (i.e., the EMML test dataset $\mathcal{T}_{\text{test}}$), the coarse-scale two-phase simulation outputs were corrected using the error model. Classification of the samples into categories was applied prior to regression. Both the classification and regression models were constructed using random forest.

We successfully applied the EMML framework to two example cases involving two-dimensional channelized models and three-dimensional Gaussian models. In both cases, improved accuracy was achieved for each well for the quantities of interest (phase injection and production rates) over a large number of test cases. The corrected solutions displayed a generally high level of accuracy both on a realization-by-realization basis and in statistical quantities such as CDFs and P10, P50, P90 predictions for cumulative oil and water production. The overhead (offline) cost associated with the EMML procedure is substantial, as $O(10-30)$ fine-scale simulations must be performed, and regression models must then be constructed on a well-by-well basis (classification models are also used for production wells). High degrees of speedup are achieved in the online computations, however, as all simulations are then on the coarse scale and the cost of the error computation is negligible.

Future work should be directed toward improving the current EMML implementation and to extending the methodology to other problems. In terms of algorithmic enhancements, it will be useful to explore approaches that introduce smoothness into the computation of $\mathring{\delta}_q^n$, which could lead to the elimination of the fluctuations observed in rate quantities at some time steps. It will also be of interest to evaluate the performance of other realization-selection methods in the EMML training step, as this may enable us to perform fewer fine-scale training simulations. Other application areas include error modeling in more complex upscaling problems, in history matching, and in well placement and well control optimization. The complex upscaling problems could include, for example, error assessment in compositional upscaling [30, 31, 42] or in upscaling discrete fracture models. Errors deriving from the use of coarse-scale models in history matching, or through the use of parameterized geological models (as in [54]), should also be assessed.

The use of our EMML-based methodology could be particularly useful in the context of well placement optimization under uncertainty. In this problem, the goal might be to maximize, e.g., the net present value of the project averaged over a large number of geological realizations. The use of error modeling in conjunction with a multilevel optimization method, such as that presented in [1], could greatly accelerate these computations. The framework presented here would require some extension for this application. In particular, the error model would now need to account for the particular well configuration under evaluation in the optimization. This would require the inclusion of additional features, such as well-to-well distance and near-well geology quantities, into the random forest regression. The use of error modeling in combination with well control optimization under uncertainty, where time-varying well rates or BHPs are varied to maximize an objective function, could also be considered. In this case, well rate or BHP information would need to be incorporated into the feature set, as was done in [52]. Finally, error models for the joint optimization of well location and well control, which could build on the findings for the two individual problems, could also be devised.

Acknowledgments We thank the sponsors of the Stanford Smart Fields Consortium for financial support. We are grateful to Hai X. Vo for providing the channelized geological models used in this study, and to Olav Møyner for his support with MRST. We also thank Stanford's Center for Computational Earth & Environmental Science for providing computational resources.

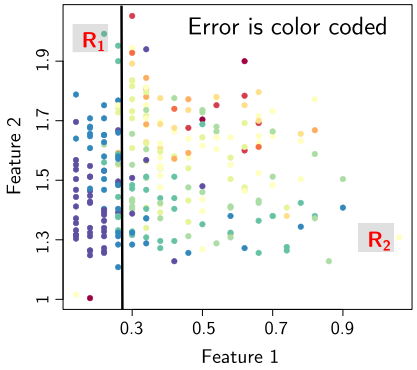
Appendix: Random-forest regression

Random forest is a decision-tree-based supervised statistical technique used here to construct both the classification and regression models from the EMML training dataset. We provide a succinct description here – see [5] for more details. A single decision tree can be used to recursively segment the domain of the EMML training dataset in the feature space, as shown in Fig. 22. Segmentation is achieved by minimizing the following expression in a top-down greedy approach:

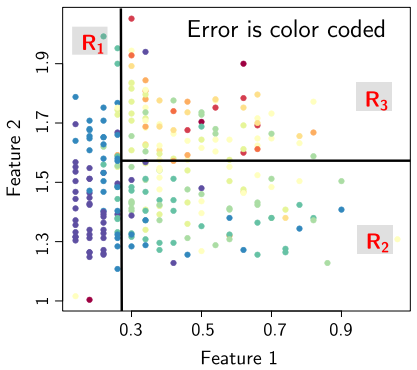
$$\sum_{n,k:\boldsymbol{f}^n\in R_1(j,s)} (\delta^n - \hat{\delta}_{R_1})^2 + \sum_{n,k:\boldsymbol{f}^n\in R_2(j,s)} (\delta^n - \hat{\delta}_{R_2})^2. \quad (28)$$

Here, $R_1(j,s) = \{\boldsymbol{f} \mid f_j < s\}$ and $R_2(j,s) = \{\boldsymbol{f} \mid f_j \geq s\}$ denote the feature-space regions as shown in Fig. 22, where $j \in \{1,\ldots,N_f\}$ is the feature index and $s \in \mathbb{R}$ is the cut-point that segments the domain, and $\hat{\delta}_{R_k} \in \mathbb{R}$, $k = 1,2$ denotes the mean value of the error for all of the training samples in R_k. The recursive segmentation enables the nonlinear interactions between the features to be captured. Methods of this type are known as tree-based because the recursive segmentation can be interpreted as a decision tree.

Single decision trees, however, tend to overfit the data. As a result the classification or regression model may suffer from low bias and high variance. To avoid this problem, in the random forest procedure we construct an ensemble of decision trees to improve prediction accuracy. This is accomplished by first selecting a random subsample of the EMML training dataset from the overall set. The subsamples are drawn with replacement (bootstrapping)



(a) Segmenting domain into regions R_1 and R_2 with $j = 1$, $s = 0.27$.



(b) Segmenting region R_2 into R_2 and R_3 with $j = 2$, $s = 1.6$.

Fig. 22 Schematic of segmentation of the feature space using a decision tree. Figure modified from [24]

such that the size of the new set is the same as that of the original set. Then, when constructing each decision tree, at any node of the tree, random forest selects a random subset of features from all possible features, and then segments the data based on this subset. This entails solving Eq. 28 using the subset of features. As a result of these treatments, random forest leads to reduction in the variance without a corresponding increase in the bias. The random forest implementation used in this work is provided in [6].

References

1. Aliyev, E., Durlofsky, L.J.: Multilevel field development optimization under uncertainty using a sequence of upscaled models. Math. Geosci. **49**(3), 307–339 (2017). https://doi.org/10.1007/s11004-016-9643-0
2. Arnold, D., Demyanov, V., Christie, M., Bakay, A., Gopa, K.: Optimisation of decision making under uncertainty throughout field lifetime: A fractured reservoir example. Comput. Geosci. **95**, 123–139 (2016)
3. Bakay, A., Demyanov, V., Arnold, D.: Uncertainty quantification in fractured reservoirs based on outcrop modelling from northeast Brazil. In: 7th EAGE international conference and exhibition (2016)
4. Bardy, G., Biver, P.: Sorting reservoir models according to flow criteria: A methodology, using fast marching methods and multi-dimensional scaling. In: Mathematics of Planet Earth: Proceedings of the 15th Annual Conference of the International Association for Math. Geosci., pp. 643–646. Springer. https://doi.org/10.1007/978-3-642-32408-6_140 (2014)
5. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
6. Breiman, L., Cutler, A., Liaw, A., Wiener, M.: Package random forest version 4.6–12. https://cran.r-project.org/web/packages/randomForest/randomFo rest.pdf (2015)
7. Cardoso, M., Durlofsky, L.J.: Linearized reduced-order models for subsurface flow simulation. J. Comput. Phys. **229**(3), 681–700 (2010)
8. Chen, Y., Durlofsky, L.J.: Ensemble-level upscaling for efficient estimation of fine-scale production statistics. SPE J. **13**(4), 400–411 (2008)
9. Chen, Y., Durlofsky, L.J., Gerritsen, M., Wen, X.H.: A coupled local–global upscaling approach for simulating flow in highly heterogeneous formations. Adv. Water Resour. **26**(10), 1041–1060 (2003)
10. Chen, Y., Mallison, B.T., Durlofsky, L.J.: Nonlinear two-point flux approximation for modeling full-tensor effects in subsurface flow simulations. Comput. Geosci. **12**(3), 317–335 (2008)
11. Chen, Y., Park, K., Durlofsky, L.J.: Statistical assignment of upscaled flow functions for an ensemble of geological models. Comput. Geosci. **15**(1), 35–51 (2011)
12. Drohmann, M., Carlberg, K.: The ROMES method for statistical modeling of reduced-order-model error. SIAM/ASA J. Uncertain. Quantif. **3**(1), 116–145 (2015)
13. Durlofsky, L.J.: Coarse scale models of two phase flow in heterogeneous reservoirs: Volume averaged equations and their relationship to existing upscaling techniques. Comput. Geosci. **2**(2), 73–92 (1998)
14. Durlofsky, L.J.: Upscaling and gridding of fine scale geological models for flow simulation. In: 8th International Forum on Reservoir Simulation (2005)
15. Durlofsky, L.J., Chen, Y.: Uncertainty quantification for subsurface flow problems using coarse-scale models. In: Numerical Analysis of Multiscale Problems, pp. 163–202. Springer (2012)
16. Efendiev, Y., Datta-Gupta, A., Ma, X., Mallick, B.: Efficient sampling techniques for uncertainty quantification in history matching using nonlinear error models and ensemble level upscaling techniques. Water Resources Research 45(11) (2009)
17. Efendiev, Y.R., Durlofsky, L.J.: A generalized convection-diffusion model for subgrid transport in porous media. Multiscale Model. Simul. **1**(3), 504–526 (2003)
18. Efendiev, Y.R., Durlofsky, L.J.: Accurate subgrid models for two-phase flow in heterogeneous reservoirs. SPE J. **9**(2), 219–226 (2004)
19. Floris, F., Bush, M., Cuypers, M., Roggero, F., Syversveen, A.R.: Methods for quantifying the uncertainty of production forecasts: A comparative study. Pet. Geosci. **7**(S), S87—S96 (2001)
20. Glimm, J., Hou, S., Lee, Y., Sharp, D., Ye, K.: Solution error models for uncertainty quantification. Contemp. Math. **327**, 115–140 (2003)
21. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. **3**, 1157–1182 (2003)
22. Hastie, T., Tibshirani, R., Friedman, J., Franklin, J.: The elements of statistical learning: Data mining, Inference, and prediction, vol. 27. Springer-Verlag, New York (2005)
23. He, J., Durlofsky, L.J.: Constraint reduction procedures for reduced-order subsurface flow models based on POD-TPWL. Int. J. Numer. Methods Eng. **103**(1), 1–30 (2015)
24. James, G., Witten, D., Hastie, T., Tibshirani, R.: An introduction to statistical learning, vol. 6. Springer, New York (2013)
25. Josset, L., Ginsbourger, D., Lunati, I.: Functional error modeling for uncertainty quantification in hydrogeology. Water Resour. Res. **51**(2), 1050–1068 (2015)

26. Khodabakhshi, M., Jafarpour, B., King, M.J.: Field applications of a multi-scale multi-physics history matching approach. In: SPE Reservoir Simulation Symposium, SPE 173239-MS (2015)
27. Kovscek, A., Wang, Y.: Geologic storage of carbon dioxide and enhanced oil recovery. I. Uncertainty quantification employing a streamline based proxy for reservoir flow simulation. Energy Convers. Manag. **46**(11), 1920–1940 (2005)
28. Krogstad, S., Lie, K.A., Møyner, O., Nilsen, H.M., Raynaud, X., Skaflestad, B.: MRST-AD–An open-source framework for rapid prototyping and evaluation of reservoir simulation problems. In: SPE Reservoir Simulation Symposium, SPE 173317-MS (2015)
29. Krogstad, S., Raynaud, X., Nilsen, H.M.: Reservoir management optimization using well-specific upscaling and control switching. Comput. Geosci. **20**(3), 695–706 (2016)
30. Li, H., Durlofsky, L.J.: Ensemble level upscaling for compositional flow simulation. Comput. Geosci. **20**(3), 525–540 (2016)
31. Li, H., Durlofsky, L.J.: Local–global upscaling for compositional subsurface flow simulation. Transp. Porous Media **111**(3), 701–730 (2016)
32. Lie, K.A., Krogstad, S., Ligaarden, I.S., Natvig, J.R., Nilsen, H.M., Skaflestad, B.: Open-source MATLAB implementation of consistent discretisations on complex grids. Comput. Geosci. **16**(2), 297–322 (2012)
33. Lødøen, O.P.: Bayesian calibration of reservoir models using a coarse-scale reservoir simulator in the prior specification. In: EAGE Conference on Petroleum Geostatistics (2007)
34. Lødøen, O.P., Omre, H.: Scale-corrected ensemble Kalman filtering applied to production-history conditioning in reservoir evaluation. SPE J. **13**(2), 177–194 (2008)
35. Lødøen, O.P., Omre, H., Durlofsky, L.J., Chen, Y.: Assessment of uncertainty in reservoir production forecasts using upscaled flow models. In: Geostatistics Banff, pp. 713–722. Springer (2005)
36. Ma, X., Al-Harbi, M., Datta-Gupta, A., Efendiev, Y.: An efficient two-stage sampling method for uncertainty quantification in history matching geological models. SPE J. **13**(1), 77–87 (2008)
37. Møyner, O., Krogstad, S., Lie, K.A.: The application of flow diagnostics for reservoir management. SPE J. **20**(2), 306–323 (2015)
38. Ng, L.W.T., Eldred, M.: Multifidelity uncertainty quantification using non-intrusive polynomial chaos and stochastic collocation. In: 14th AIAA Non-Deterministic Approaches Conference, vol. 43 (2012)
39. Omre, H., Lødøen, O.P.: Improved production forecasts and history matching using approximate fluid-flow simulators. SPE J. **9**(3), 339–351 (2004)
40. Peaceman, D.W.: Interpretation of well-block pressures in numerical reservoir simulation with nonsquare grid blocks and anisotropic permeability. SPE J. **23**(3), 531–543 (1983)
41. Remy, N., Boucher, A., Wu, J.: Applied geostatistics with SGeMS: A user's guide. Cambridge University Press, New York (2009)
42. Salehi, A., Voskov, D., Tchelepi, H.: Thermodynamically consistent transport coefficients for upscaling of compositional processes. In: SPE Reservoir Simulation Symposium, SPE 163576-MS (2013)
43. Scheidt, C., Caers, J.: Representing spatial uncertainty using distances and kernels. Math. Geosci. **41**(4), 397–419 (2009)
44. Scheidt, C., Caers, J.: Uncertainty quantification in reservoir performance using distances and kernel methods–application to a west Africa deepwater turbidite reservoir. SPE J. **14**(4), 680–692 (2009)
45. Scheidt, C., Caers, J., Chen, Y., Durlofsky, L.J.: A multi-resolution workflow to generate high-resolution models constrained to dynamic data. Comput. Geosci. **15**(3), 545–563 (2011)
46. Shahvali, M., Mallison, B., Wei, K., Gross, H.: An alternative to streamlines for flow diagnostics on structured and unstructured grids. SPE J. **17**(3), 768–778 (2012)
47. Shirangi, M.G., Durlofsky, L.J.: A general method to select representative models for decision making and optimization under uncertainty. Comput. Geosci. **96**, 109–123 (2016)
48. Shook, G.M., Mitchell, K.M.: A robust measure of heterogeneity for ranking earth models: The F-PHI curve and dynamic Lorenz coefficient. In: SPE Annual Technical Conference and Exhibition, SPE 124625-MS (2009)
49. Strebelle, S.: Conditional simulation of complex geological structures using multiple-point statistics. Math. Geol. **34**(1), 1–21 (2002)
50. Suzuki, S., Caers, J.K.: History matching with an uncertain geological scenario. In: SPE Annual Technical Conference and Exhibition, SPE 102154-MS (2006)
51. Trehan, S.: Surrogate modeling for subsurface flow: A new reduced-order model and error estimation procedures. Ph.D. thesis, Stanford University (2016)
52. Trehan, S., Carlberg, K.T., Durlofsky, L.J.: Error modeling for surrogates of dynamical systems using machine learning. Int. J. Numer. Methods Eng. **112**(12), 1801–1827 (2017). https://doi.org/10.1002/nme.5583
53. Trehan, S., Durlofsky, L.J.: Trajectory piecewise quadratic reduced-order model for subsurface flow, with application to PDE-constrained optimization. J. Comput. Phys. **326**, 446–473 (2016)
54. Vo, H.X., Durlofsky, L.J.: Data assimilation and uncertainty assessment for complex geological models using a new PCA-based parameterization. Comput. Geosci. **19**(4), 747–767 (2015)
55. Zhang, P., Pickup, G.E., Christie, M.A.: A new practical method for upscaling in highly heterogeneous reservoir models. SPE J. **13**(1), 68–76 (2008)