Delft University of Technology

Physics-based data-driven model for production forecast

Blinovs, A.; Khait, M.; Voskov, D.

DOI
10.3997/2214-4609.202035240

Publication date
2020

Document Version
Final published version

Published in
ECMOR 2020 - 17th European Conference on the Mathematics of Oil Recovery

Citation (APA)
Blinovs, A., Khait, M., & Voskov, D. (2020). Physics-based data-driven model for production forecast. In *ECMOR 2020 - 17th European Conference on the Mathematics of Oil Recovery* (pp. 1-16). (ECMOR 2020 - 17th European Conference on the Mathematics of Oil Recovery). EAGE. https://doi.org/10.3997/2214-4609.202035240

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

Green Open Access added to TU Delft Institutional Repository

'You share, we take care!' - Taverne project

https://www.openaccess.nl/en/you-share-we-take-care

Otherwise as indicated in the copyright section: the publisher
is the copyright holder of this work and the author uses the
Dutch legislation to make this work public.

Physics-Based Data-Driven Model for Production Forecast

A. Blinovs[1], M. Khait[1], D. Voskov[1,2*]

[1] TU Delft; [2] Stanford University

Summary

A physics-based data-driven model is proposed in this study for the forecasting of secondary oil recovery. The model fully relies on production data and does not directly requires any in-depth knowledge of the reservoir geology or governing physics. In the proposed approach, we utilise Delft Advanced Reservoir Terra Simulator (DARTS) as a base for data-driven simulation. DARTS uses an Operator-Based Linearization technique which exploits an abstract interpretation of physics benefiting computational performance for a forward simulation. The proposed strategy was evaluated first on the two synthetic data ensembles and showed good prediction accuracy for a significantly reduced model size. Besides, the data-driven proxy methodology was compared with an advanced flow-based upscaling technique and demonstrated an improved accuracy for both ensembles. Besides, the proposed data-driven approach was examined on two realistic data sets. For the first case, the methodology demonstrates advanced predictive performance for training based on synthetic data generated from a high-fidelity simulation model with imposed random noise. To check the robustness of the proposed methodology, the control parameters for a forecast period were significantly changed in comparison to the training period. The data-driven model still manages to predict the forecast production quite close to the reference high-fidelity results. However, the training performed on another data set based on historical production from a real brownfield was not fully successful. We relate a bigger error in both training and forecast period for this model to poor data quality. The training procedure for this model led to a moderate accuracy in history matching for a long production period, where general production trends have resembled true data and water breakthrough time was restored in nearly all wells. However, there are still periods of poor accuracy, especially where shark peaks and falls are experienced.

Introduction

Computer technologies are progressing rapidly. Computational capacities that are available nowadays provide an opportunity for many subsurface applications to perform complex numerical simulations of high-resolution three-dimensional geo-cellular computer models. A prediction obtained from such models is an important factor governing efficient reservoir management and decision making. The models describe complex geological features through a set of grid blocks and associated rock and fluid properties. A high-resolution computer model can exceed a few million blocks and may take hours or even days to simulate. It is still not computationally feasible to perform history matching or reservoir development optimization at such resolution since it involves a large number of simulations runs.

Different methods have been developed to overcome the issue. Those methods fall into two categories: simplified full-field models, or data-driven approaches. Methods such as upscaling, multi-scale methods and streamline simulation fall into the first category. Upscaling is the process of numerical homogenization, where the high-resolution model is represented as a set of coarser grid blocks with assigned effective properties aiming to replicate high-fidelity model response (Durlofsky, 2005). Multi-scale methods are somewhat similar to upscaling, where the global flow is computed on a coarse grid, while the fine-scale heterogeneity is accounted for through basis functions (Jenny et al., 2003). The streamlined method (Batycky et al., 1997) is a Eulerian-Lagrangian approach with implicit pressure explicit saturation (IMPES) time approximation. In this approach, a full 3D transport solution is translated into a set of one-dimensional equations that are solved along streamlines.

All methods in the first category require an underlying geological characterization as a basis for construction. However, there are many cases when this information is not available or its reliability is questionable. Does it mean we cannot solve optimization or history matching problems efficiently? Methods from the second category resolve this issue. The data-driven method assumes the building of a proxy model with a sufficient amount of degrees of freedom to accurately mimic the high-fidelity model based on its calibration to the production data. With frequent, sustained, and accurate data being fed into a reliable regression framework, data-driven models can provide an accurate forecast for the given reservoir.

There are many data-driven approaches available in the industry including statistical data-driven model proposed by Jansen and Kelkar (1997); reduced-order models (Cardoso et al., 2009); a Capacitance Resistance Model (CRM) by Yousef et al. (2005) which improved an earlier model proposed by Albertoni and Lake (2003); flow-network model (Lerlertpakdee et al., 2014) where a complex 3D flow is represented as a set of 1D finite-difference reservoir models; Interwell Numerical Simulation Model (INSIM) (Zhao et al., 2015) and INSIM-ft (Guo et al., 2018) approaches which somewhat similar to both CRM and flow-network models; and many other alternative methods that rely on Artificial Intelligence (AI) (Mohaghegh, 2009) and data fitting (Zubarev, 2009). All of these approaches have specific advantages and limitations.

In this study, combining advantages of methods from both categories, we develop and validate a framework capable of performing an accurate forecast based on historical field data while respecting underlying physical processes at the same time. It is achieved through the utilization of the Operator-Based Linearization (OBL) technique (Voskov, 2017) and the highly efficient Delft Advanced Reservoir Terra Simulator (DARTS) framework (DARTS, 2019).

Methodology

In this section, we describe the main ingredients of the proposed data-driven physics-based simulation framework. It includes the generation of a connectivity graph, governing equations, nonlinear solution and training of the model.

Connectivity graph for proxy model

To connect spatial well locations with production data, we need to represent the domain of interest in a discrete form. The geometrical discretization of the reservoir is typically performed based on the control volume partitioning. For our proxy model, we use unstructured partitioning and finite-volume discretization suggested by Karimi-Fard et al. (2004). It results in a spatial connectivity graph which forms a discrete representation of the proxy reservoir model in terms of connections between control volumes and associated transmissibilities (Lim, 1995). In the proposed data-driven approach, we adopt this technique for the partitioning of the reservoir domain with a coarse resolution.

The discretized model is defined using boundaries which are gridded using the hierarchical approach: a volume (convex polyhedra) is bounded by a set of surfaces (convex polygons), a surface is bounded by a series of curves (segments), and a curve is bounded by two endpoints (nodes). The automatic open-source meshing software package Gmsh (Geuzaine and Remacle, 2009) was utilised in this work for the model gridding Exact grid parameters and rock properties (e.g., permeability, exact layer geometry) are considered to be unavailable, therefore corresponding initial guess for control parameters are computed using averaged values based on initial evaluations of the reservoir. When those values are not known, they can be estimated using an analogue field or just based on common sense. Typical unstructured mesh with corresponding well locations is shown in figure 1.

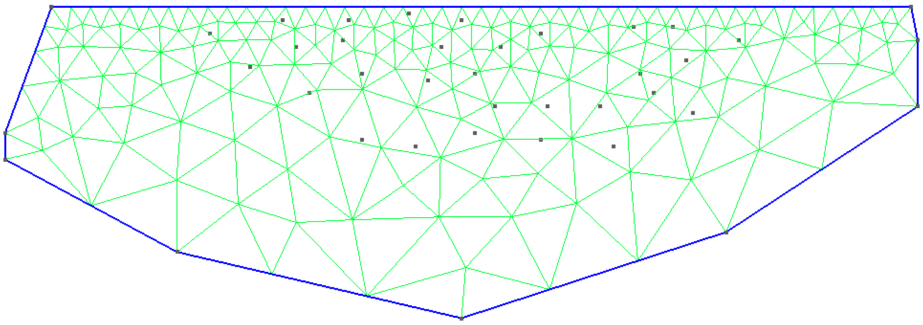

Figure 1 *Typical grid for connectivity graph generated for proxy model with corresponding border lines and well positions.*

Governing Equations

In this section, we describe the set of governing equations required for a general compositional numerical simulation. Transport equations for an isothermal system containing n_c components and n_p phases can be written as:

$$\frac{\partial}{\partial t}(\phi \sum_{j=1}^{n_p} x_{cj}\rho_j S_j) + \nabla \sum_{j=1}^{n_p} x_{cj}\rho_j \mathbf{v}_j + \sum_{j=1}^{n_p} x_{cj}\rho_j q_j^{\star} = 0, \quad c = 1,\ldots,n_c \tag{1}$$

where phase velocity is described by the Darcy's Law:

$$\mathbf{v}_j = -\left(\mathbf{K}\frac{k_{rj}}{\mu_j}(\nabla p_j - \gamma_j \nabla d)\right). \tag{2}$$

Here, ϕ - rock porosity, x_{cj} - mole fraction of component c in phase j, S_j - phase saturation, ρ_j - phase molar density, $\mathbf{v}_j$ - phase velocity, $q_j^{\star}$ - phase rate per unit volume, $\mathbf{K}$ - permeability tensor, k_{rj} - relative permeability, μ_j - phase viscosity, p_j - pressures of phase j, γ - gravity term, d is depth (positive downwards).

Eq. 1 can be written in a simplified (without gravity) discrete form by applying the finite-volume discretization in space and backward Euler approximation in time:

$$\mathbf{g} = \frac{V}{\Delta t}\left(\phi\sum_j \mathbf{x}_j\rho_j S_j + (\phi\sum_j \mathbf{x}_j\rho_j S_j)^n\right) - \sum_{l\varepsilon L}\left(\sum_j \mathbf{x}_j^l\rho_j^l T_j^l \Delta\psi^l\right) + \sum_j \rho_p \mathbf{x}_j q_j = 0, \tag{3}$$

where ψ^l is the pressure potential between two blocks. The fully implicit time approximation requires the flux term to be defined based on the nonlinear unknowns at the new timestep $(n+1)$, which introduces nonlinearity to the system of governing equations. We employ the overall molar formulation proposed in Collins et al. (1992). In the molar formulation, the primary nonlinear unknowns are pressure and overall composition, therefore the physical state ω is completely defined by these variables. The derivatives of all properties in eq. 3 with respect to nonlinear unknowns can be found by applying several closing assumptions.

Next, the Jacobian and the residual are constructed during the linearization stage. It is required by the Newton-Raphson method, where at each nonlinear iteration, the following linear system of equations is solved:

$$\frac{\partial \mathbf{g}(\omega^k)}{\partial \omega^k}(\omega^{k+1} - \omega^k) = -\mathbf{g}(\omega^k). \tag{4}$$

Here, $\mathbf{J}$ is the Jacobian matrix containing the derivatives with respect to primary unknowns, ω is the vector of nonlinear unknowns $\omega = \{p, z_c\}$, k is the nonlinear iteration step and $\mathbf{g}$ is the residual. The conventional nonlinear solution approach involves evaluation and storage of all properties and its derivatives with respect to the nonlinear unknowns, which is quite challenging. A new strategy for linearization was proposed in Voskov (2017) and will be briefly described next.

Operator-Based Linearization (OBL)

Eq. 3 can be written in a compact form as follows:

$$\mathbf{g}(\omega) = \frac{V(\xi)\phi_0(\xi)}{\Delta t}\left[\alpha_c(\omega) - \alpha_c(\omega_n)\right] + \sum_l \beta_c^l(\omega) T^{ab}(\xi)(p^a - p^b) + \gamma_c(\omega, \xi, \theta) = 0, \tag{5}$$

where ω defines physical state, while ξ represents spatial coordinates, ϕ_0 is initial porosity and T^{ab} is the geometric part of transmissibility. All involved operators are defined as:

$$\alpha_c(\omega) = (1 + c_r(p - p_{ref}))\sum_j x_{cj}\rho_j S_j \tag{6}$$

$$\beta_c(\omega) = \sum_p x_{cj}\frac{k_{rj}}{\mu_j}\rho_j \tag{7}$$

$$\gamma_c(\omega, \xi, \theta) = \sum_j \rho_j x_{cj} q_j(\omega, \xi, \theta) \tag{8}$$

Here, c_r is the rock compressibility, p_{ref} is initial reservoir pressure and θ is the vector of well controls.

In this form, the nonlinear system has a simplified description in terms of operators α_c and β_c, which depend only on the physical state and valid at any spatial location of a reservoir. When several regions for pressure, volume, temperature (PVT) properties or special core analysis laboratory (SCAL) properties are introduced in a reservoir, several sets of the operators can be accordingly utilized. The values of operators are uniquely determined in the parameter space which dimensionality is defined by the set of nonlinear unknowns p and z_c.

The OBL method suggests applying interpolation for evaluation of both operator values and their derivatives at any point in parameter space instead of continuous evaluation. Then, operator values are only computed at a limited set of supporting points. Moreover, operators are evaluated adaptively only at those supporting points in the discrete parameter space which are required to perform interpolation in the course of simulation (Khait and Voskov, 2018a). This approximated physical description allows for increased simulation performance, essential for gradient optimization problems, while approximation error remains under control. An extensive study on applications of OBL for various subsurface problems can be found in Khait and Voskov (2017, 2018b); Kala and Voskov (2019). We use the solution provided by OBL approach for solving both high-fidelity and proxy forward models.

Training of proxy model

The parameters of a model which are changed during the training stage are called *control variables*. A gradient-based optimization algorithm adjusts control variables to ensure that the data-driven proxy model response matches the "true" response based on either historical recorded data or reference high-resolution reservoir model response as close as possible.

Model training is done with the following objective function

$$J(\mathbf{u}) = \frac{1}{2}\left[\mathbf{q}(\mathbf{u}) - \mathbf{q}^{obs}\right]^T \mathbf{C}_D^{-1}\left[\mathbf{q}(\mathbf{u}) - \mathbf{q}^{obs}\right]. \tag{9}$$

Here, J is the objective function, $\mathbf{u}$ is the control variables vector, $\mathbf{q}(\mathbf{u})$ is the vector of production/injection rates from the proxy model (model response), $\mathbf{q}^{obs}$ is the vector of observed rates and $\mathbf{C}_D$ is the covariance matrix. The training is performed through the solution of the constrained optimization problem which can be formulated as

$$\min_{\mathbf{u}\in R^n} J(\mathbf{u}), \quad \mathbf{d}(\mathbf{u}) \leq 0, \tag{10}$$

where $\mathbf{d}(\mathbf{u})$ corresponds to constrains. To solve the constrained optimisation problem defined by Eq. 10, a gradient-based optimisation with the implementation of Sequential Least-Square Quadratic Programming (SLSQP) algorithm was used Kraft (1988). This approach requires the gradients of the objective function with respect to control variables. In the proposed approach, we use numerical gradients approximation which yields

$$\frac{\partial J}{\partial u_k} = \frac{J(\mathbf{u}+\delta_k\varepsilon) - J(\mathbf{u})}{\varepsilon} + O(\varepsilon), \tag{11}$$

where δ_k is Dirac's delta function. The disadvantage of the evaluation of the numerical gradient is that they lack the robustness (appropriate choice of ε) and computationally expensive (each derivative requires a forward run). However, this disadvantage was tackled to a certain extent with the parallel implementation of numerical gradient evaluation, which is an embarrassingly parallel procedure.

To ensure that the gradient optimisation process stays in the physical range, a large penalty term was imposed for non-physical regions. Moreover, the optimisation was penalised whenever the nonlinear convergence of the proxy model was not reached due to non-physical combination of parameters. Also, the scaling of the objective function and vectors of optimisation parameters was implemented in the training procedure to preserve reliable performance.

Governing relations for control variables

Since the training is performed via constrained optimization, every control variable is bounded by the minimum and maximum values. Using the bounds, every control variable is scaled to the interval $[0,1]$, as some of the regression algorithms are sensitive to the scale of a problem. Here, we briefly describe how the initial guess and constrained intervals for control variables can be calculated based on the available data and how they are included in the data-driven model.

Nonlinear control variables

We utilized a modified Brooks-Corey model (Brooks and Corey, 1964) to derive a nonlinear control variables for a multiphase flow representation in the data-driven proxy model. The modified Brooks-Corey model can be expressed as:

$$k_{ro} = k_{ro}^{e}\left(1 - S_w^{*}\right)^{n_o}, \tag{12}$$

$$k_{rw} = k^{e}_{rw} \left(S^{*}_{w}\right)^{n_w}, \tag{13}$$

$$S^{*}_{w} = \frac{S_w - S_{wc}}{1 - S_{wc} - S_{or}}. \tag{14}$$

where S^{*}_{w} is the normalised or effective water saturation, k_{rw} - water relative permeability, k^{e}_{rw} - endpoint water relative permeability, k^{e}_{ro} - endpoint oil relative permeability, n_w, n_o - exponents for water and oil, S_w - water saturation, S_{wc} - residual or connate water saturation and S_{or} is the residual oil saturation.

The relative permeability is included into the β_c-operator where it gets multiplied by $\frac{\rho_j}{\mu_j}$ term of the corresponding phase j, as can bee seen from (7). The vector of six nonlinear control variables was therefore defined as

$$v_n = \{S_{or}, S_{wc}, n_o, n_w, k^{e}_{rw}\rho_w/\mu_w, k^{e}_{ro}\rho_o/\mu_o\}\}, \tag{15}$$

subjected to the following constrains:

$$v_{n,min} = \{0.0, 0.0, 0, 0, 100, 10\}, \tag{16}$$

$$v_{n,max} = \{0.49, 0.49, 5, 5, 3000, 2000\}. \tag{17}$$

The first four parameters are unitless; the units of the last two are $[s * m^{-2}]$. These constraints were obtained based on the physical interpretation of control variables.

Linear control variables

In reservoir simulation, the general unstructured grid is usually characterized by a spatial connectivity graph represented as a connection graph Lim (1995). It involves the specification of the connections between grid blocks and associates transmissibility of those connections. The transmissibility T_{ij} between grid blocks i and j can be defined for a general unstructured grid Karimi-Fard et al. (2004) as:

$$\Gamma_{ij} = \frac{\alpha_i \alpha_j}{\sum_n \alpha_n}, \alpha_i = \frac{Ak_i}{D_i}. \tag{18}$$

Here, A is the interface area between two grid blocks, D_i is the distance from the pressure node to the interface along the line connecting two pressure nodes and k_i is the grid block permeability.

Transmissibility directly affects the flow dynamics in the reservoir since it is involved as a constant multiplier in the term T^{ab} of the convection operator in Eq. 5. In our approach, the linear control variables effectively change the transmissibilities in proxy model. As the initial guess, they can be defined based on two-point flux finite-volume discretization on unstructured grid using initial data about the reservoir properties (rock permeability) and approximate geometric parameters (thickness and net-to-gross ratio). Notice that transmissibility is a linear parameter with respect to the flow rate. However, it still remains highly nonlinear with respect to the objective function. In our approach, the amount of linear regression parameters equals to the number of connections. Through the regression course, linear control variables were constrained by $v_{l,min} = 1\ \mathrm{cP\,m^3/d/bar}$ and $v_{l,max} = 50{,}000\ \mathrm{cP\,m^3/d/bar}$.

Well control variables

A well index was proposed by Coats et al. (1974) in the steam-flood simulation to relate the grid block pressure/rate to wellbore flowing pressure/rate. The equation that relates a well and a reservoir grid block under the assumption of a single-phase flow can be written as:

$$q^{w}_{i} = \frac{WI_i}{\mu}(P_i - P^{w}_{i}) \tag{19}$$

where q^{w}_{i} is the well rate into (out of) the block i, WI_i is the well index of the grid block i intersected by a well, P_i is the grid block pressure and P^{w}_{i} is the well bottom hole pressure (BHP).

Once it is determined in single-phase assumptions, it is also applied to a multiphase flow. The equation to calculate WI was provided in Peaceman (1983) (Eq. 20) and can be used for evaluation of initial guess in proxy model following:

$$WI_i = \frac{2\pi \Delta z \sqrt{k_x k_y}}{\ln r_o / r_w + S}. \tag{20}$$

Similarly to transmissibility, a well index can be seen as a representation of the linear parameter between pressure difference and the flux between a wellbore and a reservoir block. The main difference of a well index is that it is involved in θ_c operator which directly relates the reservoir pressure with the wellbore pressure. Besides, the wellbore pressure can be used as either a control or a constraint in a simulation. In our study, the set of well control variables consisted of N_{well} well indexes. They were constrained by minimum value of $v_{w,min} = 1$ cP m^3/d/bar and $v_{w,max} = 1{,}000$ cP m^3/d/bar in the training stage.

Validation of proposed approach

We validate the proposed data-driven approach using two ensembles of stochastic fluvial models. For each high-fidelity model, two reduced-order models are constructed: the first – through conventional flow-based upscaling (Chen and Durlofsky, 2006); the second – using the proposed data-driven approach. Then, the accuracy of the two proxy models is compared for the entire ensemble under the assumption that reservoir physics is known and correct, and the same set of parameters was used for all three types of model tests. You can find these parameters in Appendix A.

Generation of ensemble of high-fidelity fluvial models

The stochastic ensemble of fluvial models was used to illustrate the accuracy of the proposed data-driven methodology. high-fidelity model ensembles were generated by two different modelling approaches:

- FLUMY Grappe et al. (2016): Process-based models using FLUMY software, see example in fig. 2(a),
- MSP Strebelle and Levy (2008): Multiple Point Statistics (MPS) models, see example in fig. 2(b).

This results in completely different model complexity between the ensembles. Each high-fidelity ensemble model has a size of 100 by 100 grid cells (cell dimensions are $10 \times 10 \times 10$ m). The models use a simple 5-spot vertical well set-up (one injector is located in the middle of the reservoir and surrounded by 4 producers located at reservoir edges). Injection wells are modelled by setting a rate control of 500 m^3 d^{-1} and production wells are modelled by setting a fixed BHP control of 100 bar. The simulation was limited by 2000 days with $\Delta t = 20$ days. The main difference between models generated by MPS and FLUMY is the main paleoflow orientation ranging from SW-NE to W-E. Besides, FLUMY model has a limited statistical variability in comparison to MPS model (de Hoop et al., 2018).

Typical porosity distribution of the high-fidelity realisations from MPS and FLUMY ensembles can be seen in figure 2. It is clear, that the model generated by MPS is more complex as the phase can flow only through distinct channels, which are usually smaller sized than the coarse grid block. In contrast, the model generated by FLUMY has many overlaying channels providing multiple possible flow paths, hence it is easier for capturing of the reservoir dynamics on a larger scale. More details about high-fidelity ensemble generation, upscaling and simulation properties can be found in de Hoop et al. (2018).

Generation of upscaled proxy models

Upscaled proxy models were generated using a global flow-based upscaling technique, which involves solving the fine-scale incompressible single-phase pressure equation and using it to obtain coarse-scale

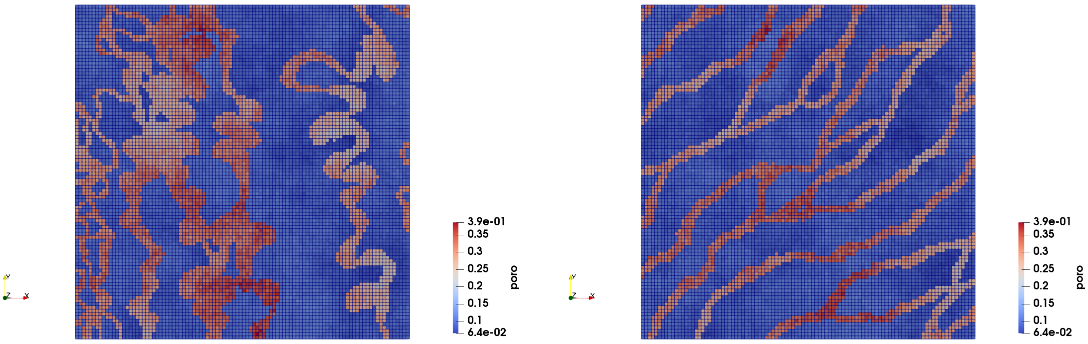

Figure 2 *Porosity distributions of a typical high-fidelity model realisation generated by process-based modeling approach with FLUMY software (a) and (b) stochastic modeling approach using Multiple Point Statistics (MPS) (b).*

transmissibility

$$-\nabla \cdot \left(\frac{K}{\mu} \nabla (P - \rho g) \right) = q_{well}. \tag{21}$$

Under the Two-Point Flux approximation, Eq. 21 can be written in the discrete form, in which the coarse properties can be evaluated

$$(q_x^c)_{i+1/2,j} = (T_x^c)_{i+1/2,j} \left(P_{i,j}^c - P_{i+1,j}^c \right). \tag{22}$$

Here, $(q_x^c)_{i+1/2,j}$ is the coarse flux across the interface $i+1/2, j$ simply defined as the integrated fine scale fluxes across the coarse interface, $(T_x^c)_{i+1/2,j}$ is the coarse transmissibility and ($P_{i,j}^c$ and $P_{i+1,j}^c$) are the coarse pressures obtained by arithmetic averaging the fine-scale pressures contained in each coarse-scale block respectively. A similar approach for a flow-based upscaled well-index for well α can be derived, given by the following equation

$$WI_\alpha = \frac{q_\alpha}{P_{i,j}^c - P_\alpha}. \tag{23}$$

The big advantage of the global flow-based upscaling technique is its computational efficiency and accuracy. However, in highly heterogeneous reservoirs it has one downfall: the resulting transmissibility values might be large or even negative (Chen and Durlofsky, 2006). This iterative upcsaling procedure is generally used to obtain a positive definite transmissibility matrix, which is typically reached within 5 iterations, see Holden and Nielsen (2000) for more details.

Each high-fidelity model was upscaled laterally by 100 times (10 × 10). The resulting upscaled transmissibility, porosity and well index was used to initialise the upscaled proxy model. Well controls and simulation time were kept identical to the high-fidelity model. Moreover, the same upscaled porosity was used in the data-driven proxy model to ensure pore volumes match between models.

Generation of data-driven models

Data-driven proxy models were generated with the same grid and well configuration as for upscaled proxy models. The initial guess for spatial connectivity and well indexes of the data-driven proxy model was chosen as a uniform distribution of 100 $\mathrm{cP\,m^3/d/bar}$ and 200 $\mathrm{cP\,m^3/d/bar}$ respectively.

Then, data-driven models are trained based on the linear control variables. The regression was limited by 100 iterations. However, most cases converged before reaching the imposed maximum. Thanks to computational efficiency of DARTS package, training of a single realization takes from 20 minutes to one hour on a single cluster node with four Intel Xeon CPU E5-2650 v3 processors. The performance of the regression framework can be improved further using adjoint gradients.

Comparison between upscaled and data-driven proxy models

For the correct comparison of the data-driven and upscaled model response, we have ensured consistency between model volume, physical properties and well controls. Those parameters are identical since the upscaling procedure is sensitive to the boundary conditions. We compare a stochastic response of the trained proxy results for two data-driven models against conventional flow-based upscaled models to validate the accuracy of the proxy modeling methodology.

Figure 3 illustrates the total water rate of all 100 realizations for high-fidelity, data-driven proxy and upscaled proxy cases. It can be seen that the stochastic response of the high-fidelity and data-driven models have a reasonably good agreement for both mean and individual realization water rates, whereas, the upscaled model rates matched worse, with a distinct delay in the water-breakthrough. The average error between data-driven and reference water rate throughout the simulation period is 3.4%, while error for the upscaled model is 14.9%. Error distribution graph for upscaled and trained data-driven models against the high-fidelity model can be found in the Appendix B.

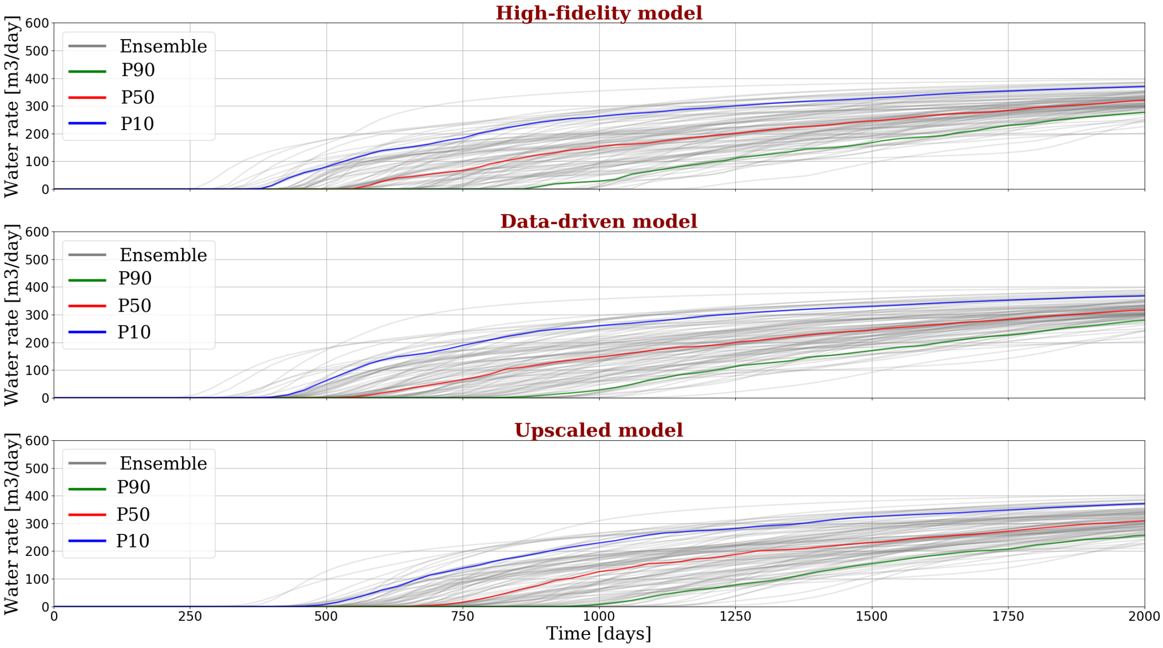

***Figure 3** The total water rate for the high-fidelity (reference) model with the size of the 100x100 grid block for the hundred FLUMY realizations, together with data-driven, and upscaled models (10x10) response. Grey lines indicate rates from a single model realization, whereas the red, blue and green lines indicate quantile response of the ensemble i.e the P10, P50 and P90.*

Then, the same test was performed for a more complicated model ensemble build with Multi-Point Statistics (MPS) modeling approach. Results are shown in figure 4. The mean error between total water rates for both data-driven and upscaled models increased to 7.4 % and 19.7 % respectively. Corresponding error distribution for the both types of proxy models against the high-fidelity model can be found in Appendix B.

It is an expected result as it is much more difficult to find a value for the effective property on a coarse scale that will accurately represent fine-scale features (e.g., small and poorly connected channels, which can be seen in figure 2,b). On the contrary, the channels in the FLUMY model overlap each other creating more distinct and rough flow paths, which are easier to capture on a coarse scale. The overall accuracy of the data-driven proxy model is still significantly higher than that for the upscaled proxy model. It confirms the applicability of the data-driven approach for uncertainty quantification analysis when a reliable and accurate high-fidelity model is not available.

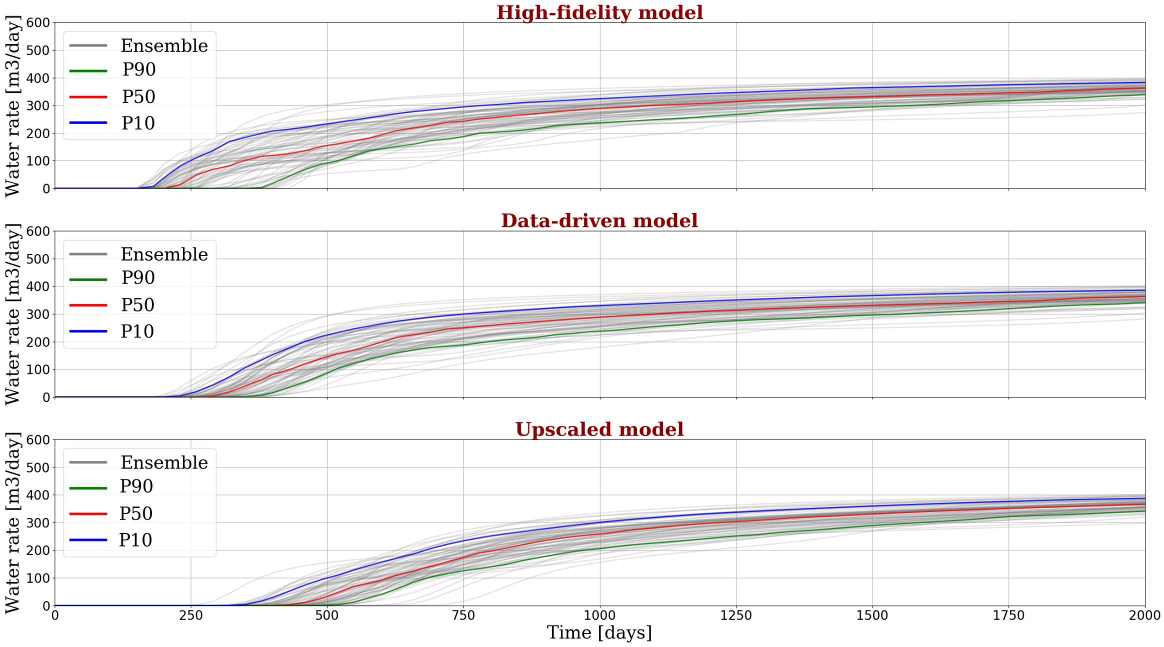

Figure 4 *the total water rate for the high-fidelity/reference model with the size of the 100x100 grid block for the hundred MPS realizations, together with data-driven, and upscaled models (10x10) response.*

Data-driven proxy model for realistic field

In this section, we give the description of a realistic field used in this study. Next, we describe a generation of the data-driven proxy-based ensembles and illustrate the performance of the proposed methodology.

High-fidelity and proxy model

We analyse the performance of the proposed framework using a reservoir model of the Brugge field (Peters et al., 2010). It is a synthetic model developed as a benchmark for optimisation of reservoir production. The structure of the Brugge field consists of an East-West elongated half-dome with a large boundary fault at the northern edge. The model has 30 wells (20 producers and 10 injectors, see 5(a)) located in the peripheral water drive. There are more than 100 realisations of this model. A single realisation encoded as FY-SS-KP-8-73 is used as the true high-fidelity model in this study and 40 other realisations with KP code are taken for the high-fidelity ensemble. The true data for history matching was obtained by 10 year-long simulation with monthly varying BHP well controls.

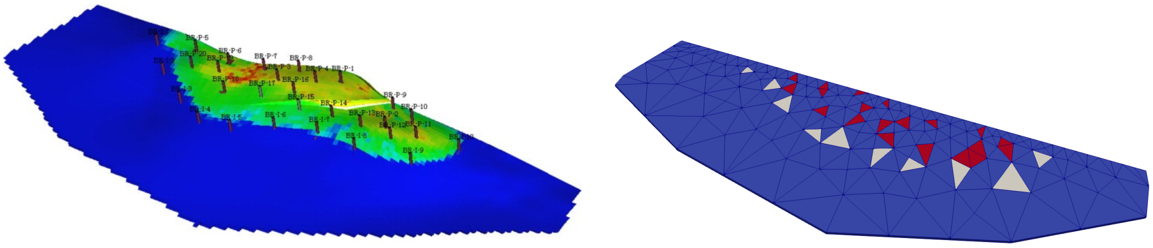

Figure 5 *High-fidelity Brugge model, used to generate the true model response (139x48x9 grid blocks) (a) Unstructured proxy model mesh with a 283 elements, red cells indicate the presence of a producing well, whereas white cells indicate injection well.*

Only the basic information about the field was used for the data-driven proxy model. For example, the reservoir boundaries were approximated based on several piece-wise linear segments as can be seen in figure 5(b). Next, the unstructured mesh was generated and extruded by an average value of reservoir thickness. The two-point stencil was utilized for the construction of unstructured grid to create the discrete connectivity graph for the proxy model. The mesh was constructed using two regions by Gmsh® software: an outer with coarser meshing and an inner with finer meshing, as seen in figure 5(b). It is done to preserve accuracy in the area where the main flow happens, whereas the outer cells were made larger to reduce the number of degrees of freedom along with the computational load. There are no significant flow dynamics in that area, hence the coarsening of that zone does not significantly affect simulation accuracy. The resulting performance of proxy vs. high-fidelity is approximately two orders of magnitude improvement (0.46 vs. 97 sec). This performance is farther enhanced utilizing shared-memory parallel implementation of numerical gradients calculations for training period which brings the proxy-model execution to around 20 runs per second.

Generation of proxy ensemble

Initial guess for control variables was calculated using a uniform permeability of $k = 1000$ mD and governing equations described in .The vector of nonlinear parameters $v_n = \{0.15, 0.25, 4, 3, 1800, 300\}$ and v_w was set to 200 cP m^3/d/bar for all wells. The total volume of the proxy model was adjusted with a model thickness, porosity and initial water saturations of individual cells to match fluid in-place volumes of the high-fidelity model. In a more general case, these parameters can be adjusted in the training period.

Obtained optimal linear and nonlinear parameters are used as the basis for proxy model ensemble generation, where we have applied a random perturbation of parameters using a normal distribution with 20% variance. The results of the training stage are shown in the figure 6. A good history matching was obtained based on the data-driven proxy model. However, some periods were not matched very well, which can be seen between 1000 and 2000 days for some wells. This is related to an insufficient number of degrees of freedom in the proxy model, which results in a limited variety of possible flow distribution.

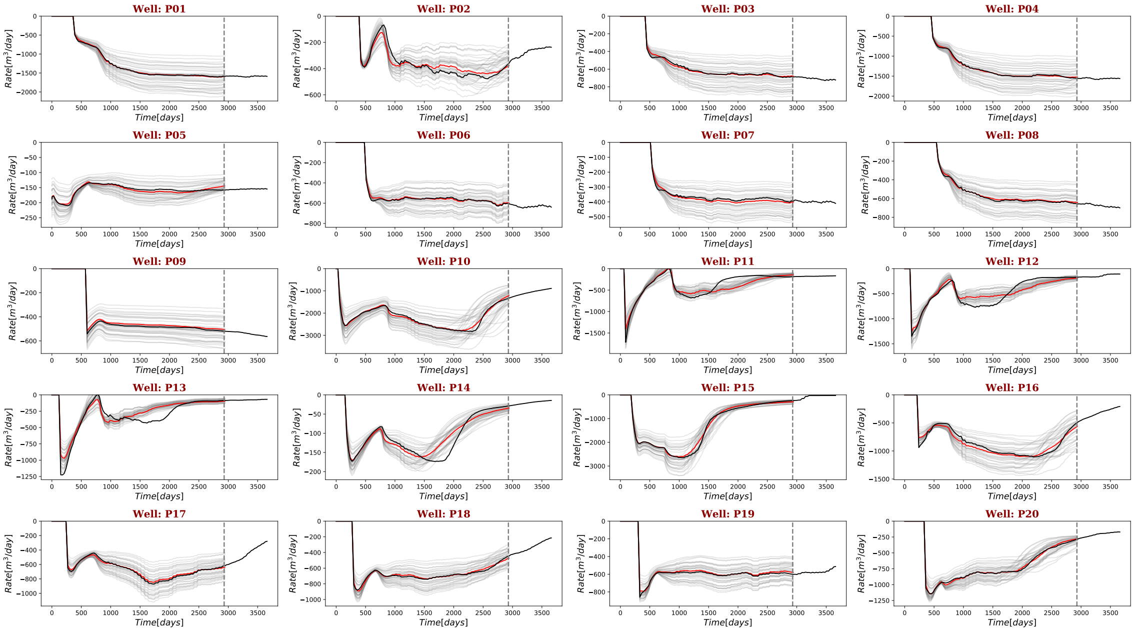

Figure 6 *The oil-production rates obtained with the history-matched data-driven proxy model. Grey line is oil-production rates from un-optimized proxy model, black line is a true response, and red line is history matched estimated data-driven proxy response.*

Training and forecast

Figure 7 illustrates a two-year production forecast of the Brugge model based on the data-driven prior ensemble training. For fair analysis of the model, we have generated a new set of well controls, different from the one used during training, with random perturbations within 20 percent of 100 bar for producers and 120 bar for injectors. The obtained results indicate a relatively accurate rate prediction for all 20 production wells. The mean error for training period is calculated using Eq. 24 and is equal to 1.83 %, whereas prediction period error reaches the value of 3.18 %

$$ME = \frac{\sum_{i=1}^{N} \left| \frac{Q^i_{o,opt} - Q^i_{o,truth}}{Q^i_{o,truth}} \right| * 100}{N} \tag{24}$$

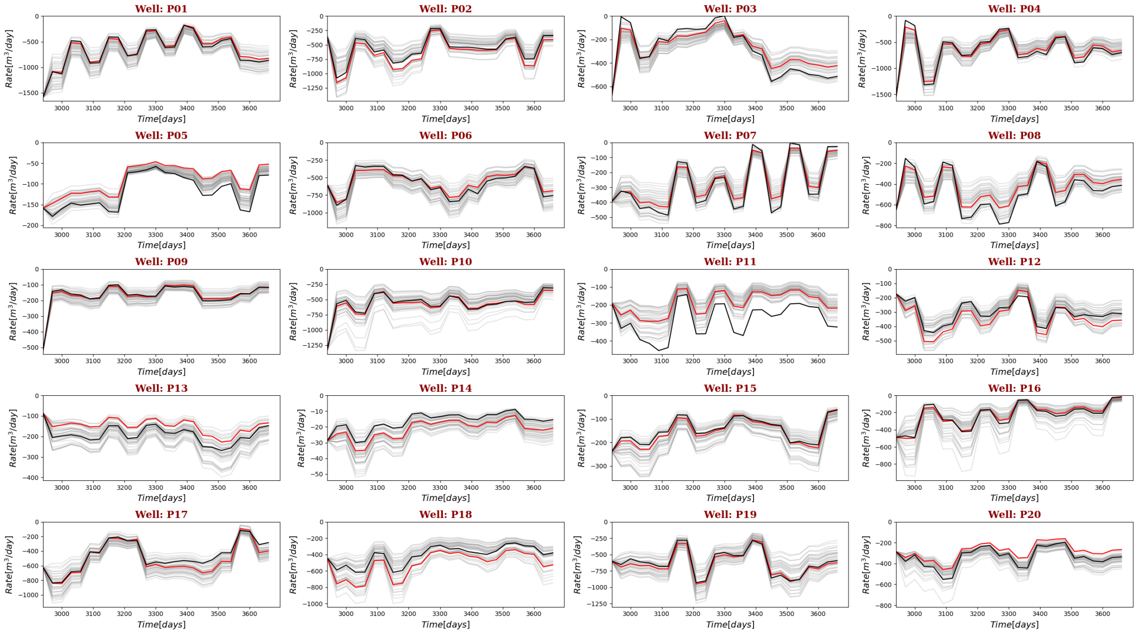

Figure 7 *The oil-production rates obtained with the history-matched data-driven proxy model. Grey lines are oil-production rates from multiple prior realisations of the high-fidelity model, black line is a true response, and red line is history matched estimated data-driven proxy response.*

For comparison, the data-driven model training can be performed using more accurate high-fidelity prior ensemble available for the Brugge model. It can be seen from the figure 8 that high-fidelity prior has less variation than proxy model-based prior. Consequently, this results in a smaller variation in the covariance matrix and is more constraining for the objective function. The mean error for the training period calculated to be equal to 1.67 % using eq. 24, whereas prediction period error came to 4.27 %. By comparison, there is no huge improvement in terms of model accuracy when trained on high-fidelity prior, whereas the time to generate its ensemble is more significant and usually is not feasible for situations when the geological model is not available.

It was observed that accuracy of the proxy model increases with the increase of data density for each well. In cases when well rate is small compared to the other wells, the output from those wells is less informative for the regression algorithm. Hence, it fails to adjust parameters in the reservoir for those wells with the same quality than for others. Moreover, the number of regression parameters should be chosen wisely as the model is prone to over-fitting. A superb history match can be achieved in this case, but the model will struggle to give an accurate prediction.

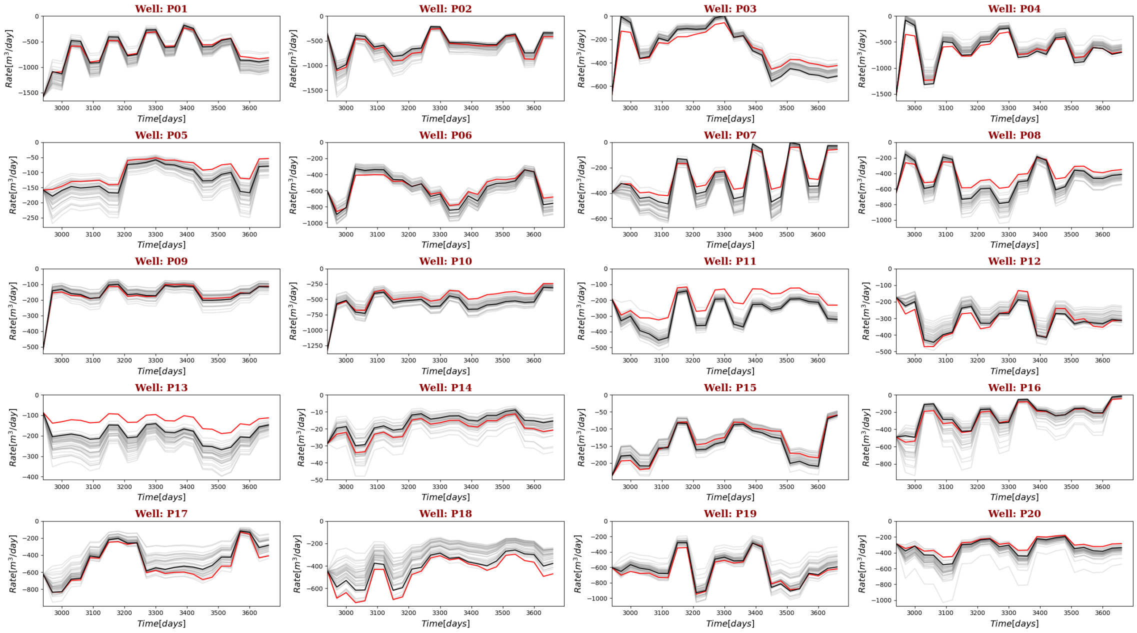

Figure 8 The oil-production rates obtained with the history-matched data-driven proxy model. Grey lines are oil-production rates from multiple prior realisations of the high-fidelity model, black line is a true response, and red line is history matched estimated data-driven proxy response. The vertical line separates the historical and future-prediction time periods.

Conclusion

In this work, a physics-based data-driven framework was developed based on the Delft Advanced Research Terra Simulation (DARTS) platform. The resulting strategy was evaluated on two synthetic data ensembles and showed good prediction accuracy for a significantly reduced model size. Both training and prediction accuracy is within a satisfactory level with good modelling of all well production rates. In addition, the data-driven proxy methodology was compared with an advanced flow-based upscaling technique and demonstrated an improved accuracy withing both stochastic ensembles.

The framework was examined on a more realistic Brugge field data set. The proposed data-driven methodology demonstrates advanced predictive performance for training based on synthetic data generated from high-resolution Brugge field. We compared the results of optimization with two types of covariance matrices based on high-fidelity prior ensemble and data-driven proxy prior ensemble. Both approaches behave equally accurate while proxy-based prior ensemble is more feasible in practical applications.

The proposed data-driven method offers a great opportunity to get a fast and reliable framework for solving many subsurface engineering problems, as was demonstrated in this work. The rising popularity of those techniques indicates their potential in a modern data-dependent world. There is still a wide range of methods that can be coupled with data-driven approaches to increase prediction capabilities and incorporate data-driven models into widely accepted engineering practice. The main advantage of the proposed approach is the potential for a natural extension of the proxy model for more complex production scenarios and physical processes.

Acknowledgements

We acknowledge the financial support from the Wintershall Dea GmbH. We would like to thank Vitaly Elichev for his valuable suggestions during the course of the project and Stephan de Hoop for the help with upscaling ensembles.

References

Albertoni, A. and Lake, L. [2003] Inferring interwell connectivity only from well-rate fluctuations in waterfloods. *SPE Reservoir Evaluation and Engineering*, **6**(1), 6–15.

Batycky, R., Blunt, M. and Thiele, M. [1997] A 3D Field-Scale Streamline-Based Reservoir Simulator. *SPE Reservoir Engineering (Society of Petroleum Engineers)*, **12**(4), 246–253.

Brooks, R. and Corey, A. [1964] Hydraulic properties of porous media, hydrology papers, no. 3, colorado state university, ft. *Collins, Colo*.

Cardoso, M., Durlofsky, L. and Sarma, P. [2009] Development and application of reduced-order modeling procedures for subsurface flow simulation. *International Journal for Numerical Methods in Engineering*, **77**(9), 1322–1350.

Chen, Y. and Durlofsky, L.J. [2006] Adaptive local–global upscaling for general flow scenarios in heterogeneous formations. *Transport in porous Media*, **62**(2), 157–185.

Coats, K., George, W., Marcum, B. and Chu, C. [1974] Three-dimensional simulation of steamflooding. *Soc Pet Eng AIME J*, **14**(6), 573–592.

Collins, D., Nghiem, L., Li, Y.K. and Grabenstetter, J. [1992] Efficient approach to adaptive-implicit compositional simulation with an equation of state. *SPE Reservoir Engineering (Society of Petroleum Engineers)*, **7**(2), 259–264.

DARTS [2019] Delft Advanced Research Terra Simulator.

Durlofsky, L. [2005] Upscaling and gridding of fine scale geological models for flow simulation. In: *8th International Forum on Reservoir Simulation Iles Borromees, Stresa, Italy*, 2024. 1–59.

Geuzaine, C. and Remacle, J.F. [2009] Gmsh: A 3-D finite element mesh generator with built-in pre- and post-processing facilities. *International Journal for Numerical Methods in Engineering*, **79**(11), 1309–1331.

Grappe, B., Cojan, I., Ors, F. and Rivoirard, J. [2016] Dynamic Modelling of Meandering Fluvial Systems at the Reservoir Scale, FLUMY Software. In: *Second Conference on Forward Modelling of Sedimentary Systems*. European Association of Geoscientists & Engineers, cp–483.

Guo, Z., Reynolds, A. and Zhao, H. [2018] A physics-based data-driven model for history matching, prediction, and characterization of waterflooding performance. *SPE Journal*, **23**(2), 367–395.

Holden, L. and Nielsen, B.F. [2000] Global upscaling of permeability in heterogeneous reservoirs; the output least squares (ols) method. *Transport in Porous Media*, **40**(2), 115–143.

de Hoop, S., Voskov, D., Vossepoel, F. and Jung, A. [2018] Quantification of coarsening effect on response uncertainty in reservoir simulation. *16th European Conference on the Mathematics of Oil Recovery, ECMOR 2018*.

Jansen, F. and Kelkar, M. [1997] Non-stationary estimation of reservoir properties using production data. *Proceedings - SPE Annual Technical Conference and Exhibition*, **Omega**, 131–138.

Jenny, P., Lee, S. and Tchelepi, H. [2003] Multi-scale finite-volume method for elliptic problems in subsurface flow simulation. *Journal of Computational Physics*, **187**(1), 47–67.

Kala, K. and Voskov, D. [2019] Element balance formulation in reactive compositional flow and transport with parameterization technique. *Computational Geosciences*.

Karimi-Fard, M., Durlofsky, L. and Aziz, K. [2004] An efficient discrete-fracture model applicable for general-purpose reservoir simulators. *SPE Journal*, **9**(2), 227–236.

Khait, M. and Voskov, D. [2018a] Adaptive Parameterization for Solving of Thermal/Compositional Nonlinear Flow and Transport With Buoyancy. *SPE Journal*, **33**(02), 522 – 534. SPE-182685-PA.

Khait, M. and Voskov, D. [2018b] Operator-based linearization for efficient modeling of geothermal processes. *Geothermics*, **74**, 7–18.

Khait, M. and Voskov, D.V. [2017] Operator-based linearization for general purpose reservoir simulation. *Journal of Petroleum Science and Engineering*, **157**, 990–998.

Kraft, D. [1988] A software package for sequential quadratic programming. *Forschungsbericht-Deutsche Forschungs- und Versuchsanstalt fur Luft- und Raumfahrt*.

Lerlertpakdee, P., Jafarpour, B. and Gildin, E. [2014] Efficient production optimization with flow-network models. *SPE Journal*, **19**(6), 1083–1095.

Lim, K.T. [1995] A new approach for residual and jacobian arrays construction in reservoir simulators. *SPE Computer Applications*, **7**(04), 93–96.

Mohaghegh, S.e.a. [2009] Artificial intelligence and data mining: enabling technology for smart fields.

The Way Ahead, **5**(03), 14–19.

Peaceman, D.W. [1983] Interpretation of well-block pressures in numerical reservoir simulation with nonsquare grid blocks and anisotropic permeability. *Society of Petroleum Engineers journal*, **23**(3), 531–543.

Peters, E., Arts, R., Brouwer, G., Geel, C., Cullick, S., Lorentzen, R., Chen, Y., Dunlop, K., Vossepoel, F., Xu, R., Sarma, P., Alhutali, A. and Reynolds, A. [2010] Results of the brugge benchmark study for flooding optimization and history matching. *SPE Reservoir Evaluation and Engineering*, **13**(3), 391–405.

Strebelle, S. and Levy, M. [2008] Using multiple-point statistics to build geologically realistic reservoir models: the MPS/FDM workflow. *Geological Society, London, Special Publications*, **309**(1), 67–74.

Voskov, D. [2017] Operator-based linearization approach for modeling of multiphase multi-component flow in porous media. *Journal of Computational Physics*, **337**, 275–288.

Yousef, A., Gentil, P., Jensery, J. and Lake, L. [2005] A capacitance model to infer interwell connectivity from production and injection rate fluctuations. *Proceedings - SPE Annual Technical Conference and Exhibition*, 323–341.

Zhao, H., Kang, Z., Zhang, X., Sun, H., Cao, L. and Reynolds, A. [2015] INSIM: A data-driven model for history matching and prediction for waterflooding monitoring and management with a field application. *Society of Petroleum Engineers - SPE Reservoir Simulation Symposium 2015*, **1**, 431–461.

Zubarev, D. [2009] Pros and cons of applying proxy-models as a substitute for full reservoir simulations. *Proceedings - SPE Annual Technical Conference and Exhibition*, **5**, 3234–3256.

Appendix A: Physical modeling properties

Table 1 *Hydrodynamic parameters*

Phase	Oil	Water
Fluid compressibility, 1/bar (c_j)	1.34 $\times 10^{-4}$	4.35$\times 10^{-5}$
Fluid densities, [kg/m^3] (ρ_j)	1002.8	897.0
Residual saturation (S_{jr})	0.15	0.225
End point relative permeability (K_{rje})	0.4	1.0
Saturation exponent (n_j)	3.0	3.0
Viscosity, cP (μ_j)	1.294	0.320

Appendix B: Errors between two proxy models and high-fidelity fluvial model

Below we present an ensemble errors of upscaling and data-driven approaches for the FLUMY-based ensemble by the end of the training stage.

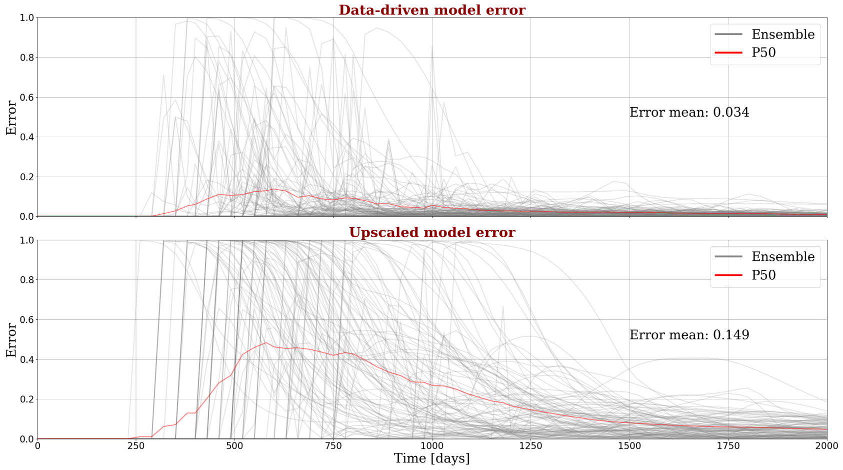

Figure 9 *The error between the response of reference models generated by FLUMY vs. data-driven and upscaled proxy models. Sharp peaks indicate water breakthrough timing mismatch.*

Similar graph is shown below for ensemble errors of upscaling and data-driven approaches for the MPS-based ensemble by the end of the training stage.

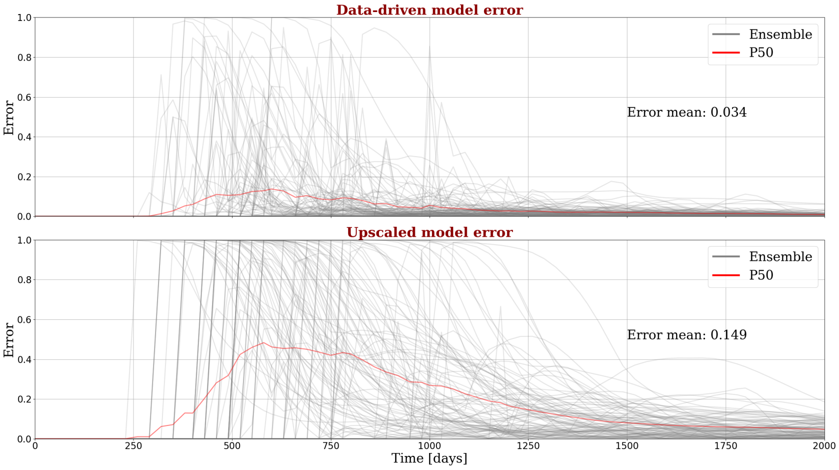

Figure 10 *The error between the response of reference models generated by MPS approach vs. data-driven and upscaled models. Sharp peaks indicate water breakthrough timing mismatch.*

Appendix C: Brugge model pressure controls

Lastly, we illustrate pressure controls for production wells in Brugge model for training and forecast stage.

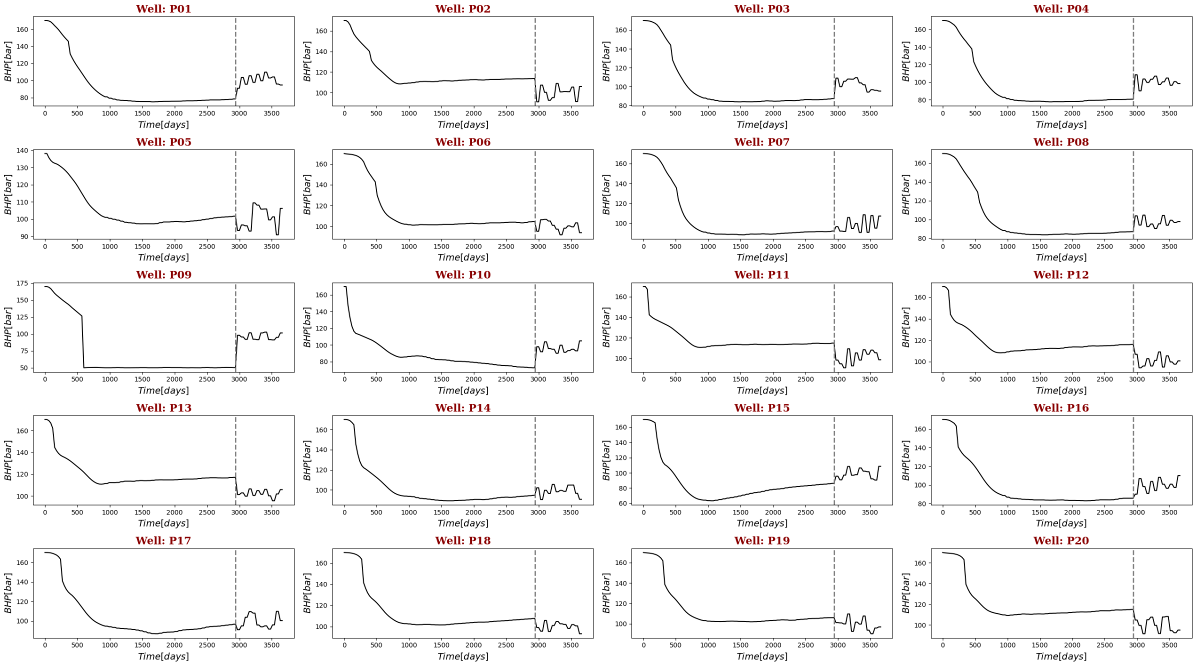

Figure 11 *Synthetically generated new BHP controls used for Brugge model production forecast steps. Dashed line indicate separation between model training and forecast periods*