Fundamentals of Data-Driven Simulation — Why Machine Learning? —

Practical Reservoir Engineering — Data-Driven Simulator Series, Session 1 (Detailed Edition)

Today's Agenda

■ Part A: Computational cost of conventional simulation and challenges in uncertainty assessment

■ Part B: The role and classification of proxy/surrogate models

■ Summary and preview of the next session / list of references and materials to purchase

Part A

Computational Cost of Conventional Simulation and Challenges in Uncertainty Assessment

- Aim of this part: understand why "machine learning" came to be needed, from the angle of computational cost
- Computational cost from increasing grid size and time-step refinement
- Iterative computational burden of history matching, uncertainty assessment, and optimization
- Trade-off with decision-making speed

A1: Growth in the Number of Grid Blocks and Computational Cost

■ With improving resolution, recent geological models have been reported to reach scales of millions to billions of grid cells

■ More grid cells directly means more unknowns in the system of equations, causing computation time and memory use to grow rapidly

Source: ResearchGate, 'Reservoir model with millions of grid blocks' related figure https://www.researchgate.net/figure/Reservoir-model-with-millions-of-grid-blocks_fig1_372290789

A2: Time-Step Refinement and Numerical Stability

■ Capturing rapid changes in multiphase flow and pressure/saturation requires finer time steps

■ Implicit methods offer superior stability but carry a high cost per step, while explicit methods have stability conditions that constrain the time-step size

■ Grid refinement and time-step refinement drive up computational cost synergistically

Source: OPM, 'The Open Porous Media Flow Reservoir Simulator' https://arxiv.org/pdf/1910.06059 (referenced as a general explanation of numerical methods)

Trade-off Between Grid Resolution and Computational Cost (Conceptual)

Coarse grid (few cells)

Computation time: short

Accuracy: hard to capture heterogeneity

Resolution
↑

Fine grid (many cells)

Computation time: long

Accuracy: reflects heterogeneity well

A3: Iterative Computational Cost of History Matching

■ History matching is an iterative computational process for fitting the model to actual data (production, pressure, etc.)

■ Manual history matching relies on engineers' trial and error and is time-consuming

■ Assisted (automated) history matching also requires many simulation runs, so its computational cost is high

Source: IntechOpen, 'Overview of History-Matching Approaches and Testing in Reservoir Simulation' https://www.intechopen.com/chapters/89743

A4: Automation via the Ensemble Kalman Filter (EnKF) and Its Computational Burden

■ Data-assimilation methods such as EnKF simulate many ensemble members (model realizations) in parallel and update them sequentially

■ A larger ensemble size improves accuracy but increases the number of simulation runs and the computational cost

■ Computational burden arises in both the forecast step and the model-update step (high cost is noted as an inherent challenge)

Source: ResearchGate, 'History Matching of Reservoir Models by Ensemble Kalman Filtering: The State of the Art and a Sensitivity Study' https://www.researchgate.net/publication/50297164_History_Matching_of_Reservoir_Models_by_Ensemble_Kalman_Filtering_The_State_of_the_Art_and_a_Sensitivity_Study

A5: Uncertainty Assessment via the Monte Carlo Method

■ Monte Carlo simulation (MCS) parameterizes geological and petrophysical uncertainty and obtains a probability distribution of results from many simulation runs

■ With few realizations the estimation error can reach several hundred percent; it has been reported that keeping the relative error below 10% may require realizations on the order of thousands

■ The spread of distributed parallel computing is making MCS itself feasible, but the cost of a single simulation run still dominates

Source: Stanford SUETRI-B, 'Monte Carlo simulation for uncertainty quantification in reservoir simulation: a convergence study' https://suetri-b.stanford.edu/publications/conference-papers/monte-carlo-simulation-uncertainty-quantification-reservoir

A6: Burden of Ensemble Simulation (Multiple Realizations)

■ To reflect geological-model uncertainty (facies distribution, faults, property distributions, etc.), it is common to prepare and evaluate multiple realizations

■ More realizations improve the reliability of the uncertainty assessment, but the number of simulation runs increases roughly proportionally

■ To stay within a realistic computation time usable for decision-making, one is often forced to limit the number of realizations

A7: Iterative Computational Burden of Well-Placement and Production Optimization

■ Well-placement and production-schedule optimization are iterative processes that run a simulation to evaluate each candidate design

■ Optimization methods such as genetic algorithms require many objective-function evaluations (i.e., simulation runs)

■ Using machine-learning-based proxy models to alleviate this iterative burden has been reported in recent years

Source: 'Recent Trends in Proxy Model Development for Well Placement Optimization Employing Machine Learning Techniques', Modelling 2024, https://doi.org/10.3390/modelling5040094

A8: The Role of Sensitivity Analysis and Design of Experiments (DOE)

■ Sensitivity analysis and Design of Experiments are often used together to narrow down, from many uncertain parameters, the factors with the greatest influence on results

■ DOE can reduce the number of required simulation runs, but tens to hundreds of runs are still often needed

■ Building a response surface (proxy model) from DOE results is a common workflow (detailed in Part B)

A9: Trade-off with Decision-Making Speed

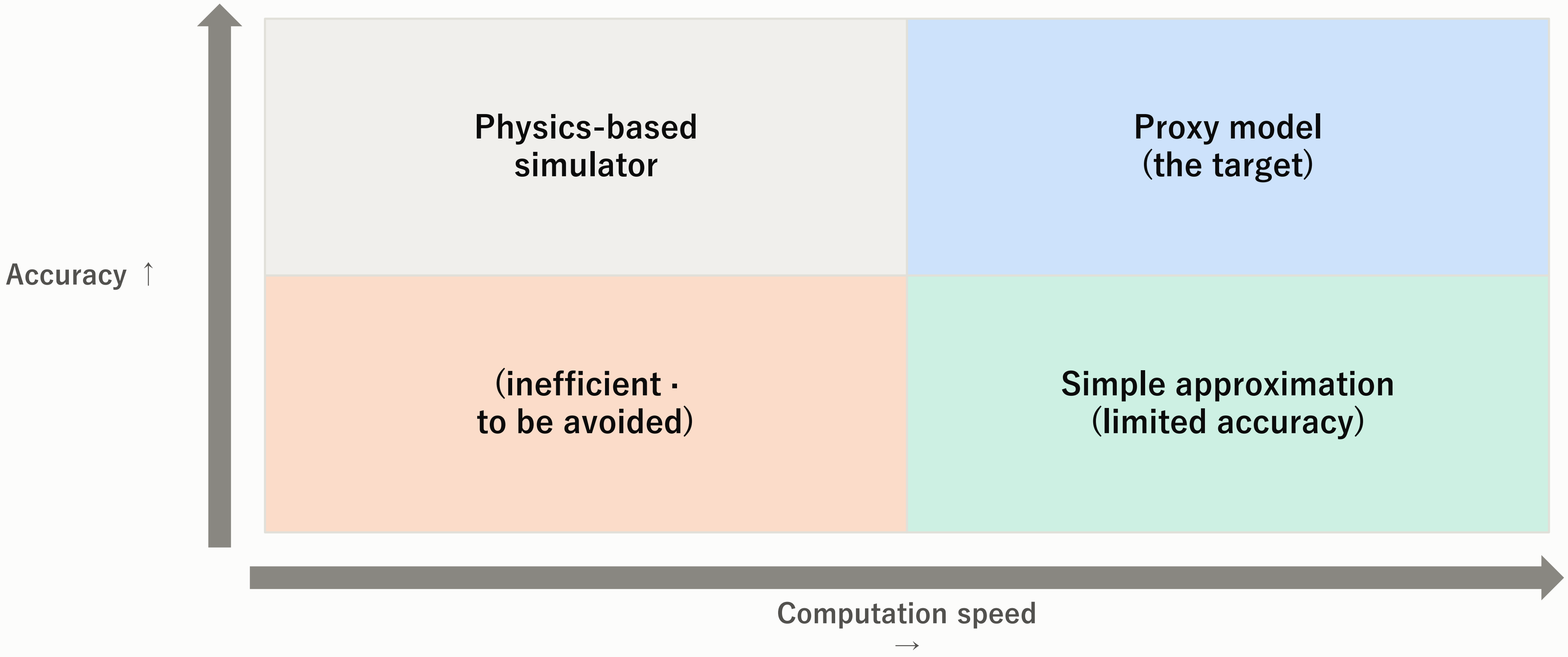
■ Investment decisions (drilling, EOR implementation, well-placement changes, etc.) must be made within limited time

■ Higher-accuracy simulation takes more computation time, conflicting with the need for rapid decisions

■ Proxy/surrogate models are drawing attention as a means of balancing "accuracy" and "speed"

Source: numberanalytics.com, 'Proxy Models: The Future of Reservoir Simulation' https://www.numberanalytics.com/blog/proxy-models-future-reservoir-simulation

Trade-off Between Accuracy and Speed (Conceptual Diagram)



A10: Part A Summary — Organizing the Computational-Cost Bottlenecks

■ For grid/time-step growth, history matching, uncertainty assessment (Monte Carlo, ensemble), and optimization alike, the essence of the bottleneck is "running the simulation many times"

■ The idea of replacing this iterative computation with a fast approximate model is the "proxy/surrogate model" (on to Part B)

■ The next part covers the classification and specific methods of these approximate models

Flow from the Challenges to Introducing a Proxy Model

Iterative computation needed (HM / uncertainty / optimization)

→

Computational cost becomes the bottleneck

→

Replace with a proxy model

→

Toward fast decision-making

Part B

The Role and Classification of Proxy/Surrogate Models

- Aim of this part: survey the definition, necessity, method classification, research examples, benefits and limits, and construction workflow of surrogate models
- Organized in the order: statistical methods → machine-learning methods → advanced methods (ROM, FNO, deep-learning surrogates)
- In Session 2 we actually build them hands-on (today's goal is to understand the classification)

B1: Definition of Proxy/Surrogate Models

- A low-cost substitute model that approximates the input–output relationship of a physics-based simulator
- "Proxy model" and "surrogate model" are often used almost synonymously; usage varies by the literature
- A trained model rapidly predicts the mapping from inputs (geology, properties, operating conditions, etc.) to outputs (production, pressure, etc.)

Source: Academia.edu, 'Development of proxy models for petroleum reservoir simulation: a systematic literature review and state-of-the-art'
https://www.academia.edu/44346638/Development_of_proxy_models_for_petroleum_reservoir_simulation_a_systematic_literature_review_and_state-of-the-art

B2: Why Needed (1): Reducing Computational Cost

■ As seen in Part A, history matching, uncertainty assessment, and optimization all require large numbers of simulation runs

■ A single run of a proxy model is far faster than a physics-based simulator (advanced methods have been reported to be up to three orders of magnitude faster than conventional solvers)

Source: Li, Z. et al., 'Fourier Neural Operator for Parametric Partial Differential Equations' (arXiv:2010.08895, 2020) https://arxiv.org/abs/2010.08895

B3: Why Needed (2): Application to Uncertainty Quantification and Optimization

- A fast proxy model makes Monte-Carlo-style many-run evaluations and iterative optimization feasible in realistic time
- Use cases in well-placement and production optimization have been reported

Source: 'Recent Trends in Proxy Model Development for Well Placement Optimization...' https://doi.org/10.3390/modelling5040094

B4: Overview of Proxy-Model Classification

■ (1) Statistical methods: response surface methodology (RSM), polynomial regression, kriging

■ (2) Machine-learning methods: random forest, support vector machine, neural network (MLP)

■ (3) Advanced methods: reduced-order models (ROM/POD), Fourier Neural Operator, deep-learning surrogates (CNN/LSTM, etc.)

■ (4) Hybrid: physics–data fusion (dual-driven), physics-informed neural networks (detailed in Session 4)

Source: ResearchGate, 'Advancements in Machine Learning–Driven Proxy Modelling Techniques for Reservoir Engineering: A Systematic Review'
https://www.researchgate.net/publication/399879336_Advancements_in_Machine_Learning-Driven_Proxy_Modelling_Techniques_for_Reservoir_Engineering_A_Systematic_Review

Classification Map of Proxy Models

Statistical methods	Machine-learning methods	Advanced methods	Hybrid
Response surface methodology (RSM)	Random forest	ROM/POD	dual-driven
Polynomial regression	SVM	Fourier Neural Operator	Physics-informed NN
Kriging	Neural network	Deep-learning surrogate	

B5: Statistical Methods (1): Response Surface Methodology (RSM)

■ Run simulations for a small number of input-parameter combinations and fit a function (e.g., a polynomial) to the results to build a response surface

■ A classic method widely used in practice for history matching and uncertainty assessment

Source: ResearchGate, 'Response Surface Methodology Approach for History Matching and Uncertainty Assessment of Reservoir Simulation Models'
https://www.researchgate.net/publication/254527939_Response_Surface_Methodology_Approach_for_History_Matching_and_Uncertainty_Assessment_of_Reservoir_Simulation_Models

B6: Statistical Methods (2): Regression and Kriging

■ In addition to polynomial regression, kriging based on spatial statistics is another representative surrogate-model method

■ A feature of kriging is that it yields not only the predicted value but also the prediction uncertainty (variance)

Source: Springer Nature, 'Response surface method for assessing energy production from geopressured geothermal reservoirs'
https://link.springer.com/article/10.1186/s40517-016-0057-5 (geothermal field, referenced for the method explanation)

B7: Machine-Learning Methods (1): Random Forest

■ An ensemble-learning method combining many decision trees; it handles strongly nonlinear input–output relationships well

■ Multiple applications to reservoir production forecasting and proxy-model construction have been reported

■ Because feature importance can be computed, it also helps interpret which input parameters affect the results

Source: SBC, 'Predicting oil field production using the Random Forest algorithm' https://sol.sbc.org.br/index.php/sibgrapi_estendido/article/download/23277/23104/

B8: Machine-Learning Methods (2): Neural Network (MLP)

■ A multilayer perceptron (MLP) is a basic neural network that learns the nonlinear mapping from inputs (properties, operating conditions) to outputs (production, pressure)

■ It is accurate given sufficient training data, but it depends heavily on data quantity and quality

Source: ResearchGate, 'Advancements in Machine Learning–Driven Proxy Modelling Techniques for Reservoir Engineering: A Systematic Review'
https://www.researchgate.net/publication/399879336_Advancements_in_Machine_Learning-Driven_Proxy_Modelling_Techniques_for_Reservoir_Engineering_A_Systematic_Review

B9: Advanced Methods (1): Reduced-Order Model (ROM) / POD

■ A method that uses Proper Orthogonal Decomposition (POD) and the like to compress per-grid state variables (pressure, saturation) into a few representative variables and compute in a low-dimensional space

■ Methods combined with Trajectory Piecewise Linearization (TPWL) have also been reported

■ Because it is based on physical equations, it is considered more robust to extrapolation than purely data-driven models

Source: Stanford SUETRI-B, 'Reduced-Order Modeling for Compositional Simulation by Use of Trajectory Piecewise Linearization' https://suetri-b.stanford.edu/publications/journal-articles/reduced-order-modeling-compositional-simulation-use-trajectory

B10: Advanced Methods (2): Fourier Neural Operator (FNO)

■ A type of "neural operator" that learns a mapping (operator) from function space to function space

■ By parameterizing the integral kernel in Fourier space, it achieves resolution-independent, efficient learning and inference

■ As an application to reservoir simulation, proxy models for cases with varying well conditions and permeability fields have been reported

Source: Li et al. (arXiv:2010.08895); ScienceDirect, 'Neural operator-based proxy for reservoir simulations considering varying well settings, locations, and permeability fields' https://www.sciencedirect.com/science/article/abs/pii/S0098300424003091

B11: Advanced Methods (3): Deep-Learning Surrogate (Handling Time-Varying Well Controls)

- Deep-learning surrogate models that handle time-varying well controls (production rate, BHP, etc.) have been reported
- Architectures capable of handling temporal and spatial structure, such as CNN and LSTM, are used

Source: ResearchGate, 'Deep-learning-based surrogate model for reservoir simulation with time-varying well controls'
https://www.researchgate.net/publication/340562536_Deep-learning-based_surrogate_model_for_reservoir_simulation_with_time-varying_well_controls

B12: Research Example (1): Physics–Data Fusion (Dual-Driven) Surrogate Model

■ Methods that combine physical laws (Darcy's law, conservation laws, etc.) with data-driven models to maintain high accuracy even with little data have been proposed

■ Compared with purely data-driven models, the potential to reduce dependence on data quantity has been reported

Source: Physics of Fluids (AIP), 'A data-physical dual-driven surrogate model for reservoir simulation' (2025)
https://pubs.aip.org/aip/pof/article/37/2/026632/3337562/A-data-physical-dual-driven-surrogate-model-for

B13: Research Example (2): History Matching × Deep-Learning Proxy

■ A history-matching method combining Bayesian inference with a proxy model based on an LSTM (long short-term memory) network has been reported

Source: SPE Journal (OnePetro), 'Robust Method for Reservoir Simulation History Matching Using Bayesian Inversion and Long-Short-Term Memory Network-Based Proxy' https://onepetro.org/SJ/article/28/03/983/514872/

B14: Benefits and Limitations

■ Benefit: major reduction in computational cost; good fit with optimization and uncertainty assessment that presuppose many evaluations

■ Limitation (1): accuracy tends to drop when extrapolating outside the training-data range (for purely data-driven models)

■ Limitation (2): heavy dependence on data quantity and quality (simulation output / real data)

■ Limitation (3): explainability of "why that prediction was made" (the model tends to become a black box)

Source: Academia.edu literature review (same as B1); Research/References/NewCourseProposal_DataDrivenSimulator.md (Chapter 1)

B15: Overview of the Surrogate-Model Construction Workflow

■ Step 1: Collect training data (simulation output / real data; sampling design via DOE)

■ Step 2: Feature design (selecting and normalizing input parameters)

■ Step 3: Model training (choose from statistical / machine-learning / advanced methods)

■ Step 4: Accuracy validation (cross-validation, testing on held-out data)

■ Step 5: Operation (embedding into optimization, uncertainty assessment, and decision support)

List of References (Sources) (1/5)

■ numberanalytics.com, 'Proxy Models: The Future of Reservoir Simulation' https://www.numberanalytics.com/blog/proxy-models-future-reservoir-simulation

■ AIP Publishing, Physics of Fluids, 'A data-physical dual-driven surrogate model for reservoir simulation'(2025) https://pubs.aip.org/aip/pof/article/37/2/026632/3337562/A-data-physical-dual-driven-surrogate-model-for

■ ResearchGate, 'Deep-learning-based surrogate model for reservoir simulation with time-varying well controls' https://www.researchgate.net/publication/340562536_Deep-learning-based_surrogate_model_for_reservoir_simulation_with_time-varying_well_controls

■ 'Recent Trends in Proxy Model Development for Well Placement Optimization Employing Machine Learning Techniques', Modelling(2024) https://doi.org/10.3390/modelling5040094

■ ScienceDirect, 'Neural operator-based proxy for reservoir simulations considering varying well settings, locations, and permeability fields' https://www.sciencedirect.com/science/article/abs/pii/S0098300424003091

■ SPE Energy Stream, 'Bridging The Gap Between Material Balance And Reservoir Simulation For History Matching And Probabilistic Forecasting Using Machine Learning' (training course)

List of References (Sources) (2/5)

■ JPT(SPE), 'Reservoir Simulation' https://jpt.spe.org/reservoir-simulation-2026

■ **SPE, 'Data-Driven Reservoir Modeling' (training course) https://www.spe.org/en/training/courses/data-driven-reservoir-modeling/**

■ **SPE, 'Python for Petroleum Data Analytics' (training course) https://www.spe.org/en/training/courses/python-for-petroleum-data-analytics/**

■ **Petroleum Engineers Association, 'Applied Machine Learning and Data Driven Modeling for Oil and Gas Using Python' (training course) https://www.peassociations.com/workshopdetails-NzQ=/**

■ Equinor, 'Volve field data set' https://www.equinor.com/energy/volve-data-sharing

■ energistics.org, 'Equinor's Volve Field Test Data' https://energistics.org/equinors-volve-field-test-data

■ **SINTEF, 'spe10: Access to the SPE10 benchmark case'(MRST documentation) https://www.sintef.no/contentassets/2551f5f85547478590ceca14bc13ad51/spe10.html**

■ OPM(Open Porous Media), 'SPE10 Comparative Solution Project Test Documentation'

List of References (Sources) (3/5)

■ Stanford SUETRI-B, 'Monte Carlo simulation for uncertainty quantification in reservoir simulation: a convergence study' https://suetri-b.stanford.edu/publications/conference-papers/monte-carlo-simulation-uncertainty-quantification-reservoir

■ SPE Journal(OnePetro), 'Robust Method for Reservoir Simulation History Matching Using Bayesian Inversion and Long-Short-Term Memory Network-Based Proxy' https://onepetro.org/SJ/article/28/03/983/514872/

■ IntechOpen, 'Overview of History-Matching Approaches and Testing in Reservoir Simulation' https://www.intechopen.com/chapters/89743

■ ResearchGate, 'Response Surface Methodology Approach for History Matching and Uncertainty Assessment of Reservoir Simulation Models' https://www.researchgate.net/publication/254527939_Response_Surface_Methodology_Approach_for_History_Matching_and_Uncertainty_Assessment_of_Reservoir_Simulation_Models

■ Academia.edu, 'Development of proxy models for petroleum reservoir simulation: a systematic literature review and state-of-the-art' https://www.academia.edu/44346638/Development_of_proxy_models_for_petroleum_reservoir_simulation_a_systematic_literature_review_and_state-of-the-art

List of References (Sources) (4/5)

- Li, Z. et al., 'Fourier Neural Operator for Parametric Partial Differential Equations'(arXiv:2010.08895, 2020) https://arxiv.org/abs/2010.08895
- Mohaghegh, S.D., 'Data-Driven Reservoir Modeling', SPE/PennWell(2017, ISBN 9781613995600) https://store.spe.org/Data-Driven-Reservoir-Modeling-P1054.aspx
- **Ertekin, T., Abou-Kassem, J.H., King, G.R., 'Basic Applied Reservoir Simulation' (SPE Textbook Series; book, not held in-house)**
- ResearchGate, 'History Matching of Reservoir Models by Ensemble Kalman Filtering: The State of the Art and a Sensitivity Study' https://www.researchgate.net/publication/50297164_History_Matching_of_Reservoir_Models_by_Ensemble_Kalman_Filtering_The_State_of_the_Art_and_a_Sensitivity_Study
- **Anaconda official documentation https://docs.anaconda.com/anaconda/install/**
- **Jupyter official documentation https://jupyter.org/install**
- **scikit-learn official documentation https://scikit-learn.org/stable/**
- **matplotlib official documentation https://matplotlib.org/**

List of References (Sources) (5/5)

■ Planning Department (internal document) 'NewCourseProposal_DataDrivenSimulator.md' (July 2026)

■ Planning Department (internal document) 'NewCourseProposal_Advanced.md' (July 2026)

List of References That Must Be Purchased (1/2)

■ Ertekin, T., Abou-Kassem, J.H., King, G.R., 'Basic Applied Reservoir Simulation' (SPE Textbook Series; book) — Needed for: primary-source support for the numerical formulation and computational cost in Part A 'A1: Growth in the Number of Grid Blocks and Computational Cost' and 'A2: Time-Step Refinement and Numerical Stability'

■ Mohaghegh, S.D., 'Data-Driven Reservoir Modeling' (SPE/PennWell, 2017, ISBN 9781613995600) — Needed for: systematic theoretical grounding of data-driven modeling for Part B 'B15: Overview of the Surrogate-Model Construction Workflow' and the course as a whole

■ SPE Journal (OnePetro) paywalled paper 'Robust Method for Reservoir Simulation History Matching Using Bayesian Inversion and Long-Short-Term Memory Network-Based Proxy' — Needed for: citing the method details and accuracy-validation results in Part B 'B13: Research Example (2): History Matching × Deep-Learning Proxy' (currently limited to an abstract-level introduction)

■ SPE members-only online training course 'Data-Driven Reservoir Modeling' (paid registration) — Needed for: reference material when concretizing Part E 'Ex.1 How to Proceed with the Exercise' and the design of materials from Session 2 onward against a practical training curriculum

List of References That Must Be Purchased (2/2)

■ ResearchGate/OnePetro paywalled paper 'Response Surface Methodology Approach for History Matching and Uncertainty Assessment of Reservoir Simulation Models' — Needed for: detailed citation of the equations and specific application examples in Part B 'B5: Statistical Methods (1): Response Surface Methodology (RSM)' (currently limited to an abstract-level citation)