

Research Progress and Prospects of Data-Driven Reservoir History Matching Methods

Haoxi Shi¹, Yang Lu¹, Ziqian Hu¹, Ming Zhang¹, Liyang Wang², Botao Liu^{1,3,*}

¹College of Computer Science, Yangtze University, Jingzhou 434023, China;

²College of Petroleum Engineering, Yangtze University, Wuhan 430100, China;

³Hubei Key Laboratory of Oil and Gas Drilling and Production Engineering, Yangtze University, Jingzhou 434023, China;

*Corresponding author: Botao Liu (Email: liubotao920@163.com)

Abstract: Reservoir history matching is a crucial process in oilfield development, aiming to calibrate geological models, characterize reservoir parameters, and enhance predictive accuracy. Traditional manual history matching relies heavily on human expertise, extensive computational effort, and subjective judgment, which can no longer meet the demands posed by increasingly complex geological structures and massive data volumes. In recent years, data-driven and machine learning approaches have shown great potential in reservoir inversion and production forecasting. Emerging models such as proxy models, deep neural networks, graph neural networks (GNNs), and Transformers provide new paradigms for achieving automatic reservoir history matching. This paper reviews the historical evolution of history matching methods, tracing their progression from early numerical simulation and optimization algorithms to modern data-driven modeling techniques. From the perspective of machine learning, it further investigates the performance of these methods in dynamic reservoir prediction and discusses the application prospects of spatio-temporal fusion models based on GNNs and Transformers. Finally, the paper outlines the current challenges and future directions in this field, aiming to contribute to the advancement of intelligent reservoir modeling and the development of digital twin reservoirs.

Keywords: History Matching; Reservoir Modeling; Machine Learning; Graph Neural Network (GNN); Transformer; Proxy Model

1. Introduction

Petroleum resources constitute a fundamental component of the modern industrial system and play a pivotal role in national energy security. As many large oilfields in China have entered the late water-flooding stage, the geological structures of reservoirs have become increasingly complex and diverse. This complexity is reflected in the lateral heterogeneity of sedimentary facies belts, the multi-barrier characteristics of fault systems, as well as the non-uniform distribution of pore structures and the dynamic alterations induced by fluid-rock interactions[1]. Consequently, the production performance of such reservoirs tends to exhibit greater instability. Under these circumstances, identifying inter-well connectivity and key geological parameters has become a pressing challenge in reservoir development and profile control.

History matching, as the most critical process in reservoir numerical simulation, involves adjusting model parameters based on actual production data to ensure that simulation results closely approximate real field behavior. Reliable history matching results form the foundation for formulating development strategies and optimizing injection-production schemes.

Traditional history matching primarily relies on manual expertise, in which engineers iteratively adjust geological parameters—such as permeability and porosity—until simulated results align with observed production data. This approach was widely adopted in the early stages of oilfield development. However, as reservoir exploitation has advanced into the middle and late stages, the internal fluid distribution and flow dynamics within reservoirs have

become increasingly complex. Consequently, the limitations of manual history matching—namely high computational cost, strong subjectivity, and the inability to effectively address high-dimensional nonlinear problems—have become more pronounced[2]. For example, in fractured reservoirs, when manually adjusting combinations of fracture permeability and matrix porosity, a strong correlation often exists between parameters. This interdependence can lead to non-uniqueness problems, wherein different parameter combinations may produce nearly identical production responses[3]. With the rapid advancement of computer technology and artificial intelligence, automated and data-driven matching methods have emerged as new research hotspots.

The accelerating digitalization of modern oilfields has further intensified this trend. Real-time monitoring data—such as wellhead pressure and liquid production rate—together with logging data (e.g., resistivity, acoustic travel time) and seismic interpretation data, are increasing exponentially. The total daily data volume has already surpassed the terabyte scale, providing abundant sample resources for data-driven approaches[4].

In recent years, the introduction of machine learning, deep learning, and proxy modeling has transformed reservoir history matching from a process of “physical equation-based simulation” to one of “data-driven prediction.” These models can rapidly approximate the nonlinear mapping between input variables and production responses, thereby capturing the complex fluid flow processes at a fraction of the computational cost. For instance, while traditional numerical simulations may require hours or even days for a single computation, data-driven models can deliver comparable

results within seconds[5].

Notably, emerging architectures such as Graph Neural Networks (GNNs) and Transformers offer new paradigms for modeling inter-well connectivity and temporal dynamics. GNNs are capable of capturing topological spatial dependencies among wells, while Transformers excel at learning long-term temporal variations in production data, providing an innovative framework for reservoir history matching[6].

2. Development of Traditional Reservoir History Matching Methods

2.1. Manual History Matching Stage

Manual history matching represents the earliest form of reservoir parameter inversion. In this approach, engineers iteratively adjust model parameters through a trial-and-error process, aiming to make simulated results closely align with field-measured production data. The major limitation of this stage lies in its excessive dependence on engineering experience. For instance, when adjusting permeability fields, engineers often determine parameter distributions subjectively based on perceived similarities in the production performance of adjacent wells, rather than on quantitative analysis or numerical evidence[7].

In the 1960s, Jacquard et al. proposed replacing manual trial-and-error with regression analysis[8]. By establishing linear regression relationships between production data (e.g., oil rate) and geological parameters (e.g., porosity), they achieved a preliminary semi-quantitative parameter adjustment, marking the beginning of research on automatic history matching. Although this method improved efficiency, it still faced challenges in modeling complex nonlinear reservoirs. When multiple fluids coexist or the reservoir exhibits strong heterogeneity, simple mathematical formulations fail to accurately capture the highly nonlinear relationships between parameters and production behavior—as Jacquard himself observed.

Over subsequent decades, researchers have continuously sought to enhance history matching through optimization theory. Chen et al. introduced a concept based on optimal control theory[9], treating reservoir parameters as continuous spatial functions and applying variational methods to determine their optimal distributions. Meanwhile, Chavent et al. employed least-squares optimization to compute the best parameter set by minimizing the sum of squared differences between simulated and observed data, thereby improving both accuracy and stability of the model.

However, gradient-based optimization methods are highly sensitive to initial values. When the initial permeability field deviates significantly from reality, the gradient descent process can easily become trapped in local minima, leading to results that differ substantially from the true reservoir characteristics. Moreover, gradient computation relies on sensitivity analysis within the numerical simulation. In high-dimensional parameter spaces—such as those involving millions of grid-based permeability parameters—the computational complexity of constructing and solving the sensitivity matrix increases dramatically, making it difficult to meet practical engineering requirements.

2.2. Automated and Intelligent Optimization Stage

Since the 1980s, with the rapid improvement of computing performance and the introduction of intelligent optimization algorithms into the field of history matching, many of the limitations inherent in traditional gradient-based methods have been partially alleviated. Nocedal proposed the Limited-memory Broyden–Fletcher–Goldfarb–Shanno (LBFGS) algorithm, which addresses the memory bottleneck in large-scale optimization problems by iteratively updating an approximate inverse Hessian matrix without explicitly storing the full sensitivity matrix, thereby significantly reducing memory requirements. Tang et al. further enhanced the computational efficiency of sensitivity matrices through the Generalized Pulse Spectrum Technique (GPST), which transforms sensitivity calculations into the frequency domain, simplifying the process and reducing computation time by approximately 40%.

After the 1990s, Genetic Algorithms (GA), Particle Swarm Optimization (PSO), and Simulated Annealing (SA) became widely applied in reservoir history matching[10]. Holland first introduced the GA in 1975, inspired by the theory of biological evolution. This population-based search method utilizes selection, crossover, and mutation operations to avoid entrapment in local minima[11]. Later, Kennedy and Eberhart proposed the PSO algorithm in 1995, simulating the foraging behavior of bird flocks—particles share information among themselves and rapidly converge to an optimal solution[12]. Kirkpatrick et al. (1983) developed the SA algorithm, a global optimization technique based on the physical annealing process of solids, which achieves a balance between global exploration and local exploitation by controlling the cooling rate[13].

These algorithms rely on population-based search combined with stochastic perturbation, enabling them to effectively overcome the local optimum issues that frequently arise in traditional deterministic methods. They exhibit good robustness in history matching tasks involving low- to medium-dimensional parameter spaces. For example, when optimizing 100 grid-based permeability parameters using the GA, the resulting fitting accuracy was reported to be approximately 15% higher than that achieved by gradient-based methods[14].

Since the beginning of the 21st century, automatic history matching techniques have become deeply integrated with artificial intelligence (AI) methodologies. Cullick et al. proposed a hybrid framework that combines nonlinear proxy models with global optimization algorithms to achieve rapid inversion of reservoir models[15]. By using proxy models to approximate the numerical simulation response, the computational cost of optimization can be reduced by as much as 80%. Yin et al. further integrated the hierarchical streamline method with the Simulated Annealing (SA) algorithm to enhance the history matching accuracy for complex reservoirs. The streamline approach simplifies the computation of fluid flow and consequently improves the efficiency of sensitivity analysis.

In recent years, the Differential Evolution (DE) and Differential Evolution Particle Swarm Optimization (DEPSO) algorithms have emerged as new trends in this domain. Price and Storn (1997) first introduced the DE algorithm, which reinforces global exploration capability through differential mutation operations[16]. The DEPSO algorithm, which

integrates the global search efficiency of DE with the rapid local convergence of PSO, has demonstrated superior performance in the history matching of fractured reservoirs—achieving approximately 12% higher parameter inversion accuracy and 30% better computational efficiency compared to the standalone DE algorithm[17].

Furthermore, Zhang et al. combined PSO with numerical simulation for predicting water-cut behavior in high water-bearing reservoirs, successfully achieving an engineering-level fitting error of less than 5%[18]. These advancements signify a pivotal transition from purely physics-based optimization to hybrid data-driven modeling, laying the groundwork for the subsequent development of machine learning-based reservoir history matching frameworks.

3. Application of Surrogate Models in History Matching

3.1. Fundamental Principles of Surrogate Models

A surrogate model is an approximate model that leverages a limited number of sample points to represent the relationship between the inputs and outputs of a complex system. It is typically employed as a substitute for computationally expensive numerical simulations. The core idea is to construct a response surface—or, equivalently, a learning model—based on finite data, which can then be used to rapidly predict the output variables. This enables efficient optimization or inversion processes. The primary advantage of surrogate models lies in their ability to “learn from a small number of samples to achieve a large number of predictions.” For instance, in history-matching optimization scenarios that would normally require thousands of numerical simulations, a surrogate model can be constructed using only a few hundred sample points, thereby reducing the computational time to approximately one-tenth of the original[19].

Commonly used surrogate modeling approaches include the Response Surface Method (RSM), Radial Basis Function (RBF), and Kriging models. The RSM, proposed by Box and Wilson in 1951, employs polynomial functions to fit the relationship between inputs and outputs and exhibits high simplicity for low-dimensional linear systems[20]. The RBF, introduced by Hardy in 1971, constructs an interpolation model using radially symmetric functions and demonstrates stronger approximation capability for nonlinear systems compared with RSM[21]. The Kriging model, developed by Matheron in 1963, is based on geostatistical principles and employs a variogram to describe the spatial correlation in parameter space, which better preserves the spatial distribution characteristics of reservoir parameters during inversion[22]. Jin et al. found that, in the context of reservoir history matching, the prediction error of the Kriging model is approximately 20% lower than that of RSM, making it suitable for inverting spatially continuous reservoir parameters such as permeability fields[23].

In recent years, Artificial Neural Networks (ANNs) have been widely adopted in reservoir surrogate modeling due to their strong nonlinear fitting capability. Fonseca and Schiozer (2009) applied ANNs for dynamic reservoir prediction, using Multilayer Perceptrons (MLP) to learn the relationship between bottom-hole pressure and parameters such as permeability and porosity, achieving prediction errors approximately 15% lower than those of Kriging models[24]. By training neural network models to learn the response

relationships among bottom-hole pressure, production rate, and water cut, the fitting time can be significantly reduced. For instance, in a block-scale reservoir history-matching case, the ANN surrogate model required only about 1/1000 of the computational time of full numerical simulation while maintaining engineering-level accuracy[25].

3.2. Development of Deep Learning-Based Surrogate Models

The integration of deep learning into surrogate modeling enables the handling of complex, high-dimensional, and nonlinear problems. Its primary advantage lies in the ability to automatically extract deep features from high-dimensional data without the need for manually engineered variables. Hinton et al. (2006) proposed the Deep Belief Network (DBN), which hierarchically transforms data into different representations and provides a novel approach for high-dimensional reservoir parameter inversion [26]. LeCun et al. (1998) introduced Convolutional Neural Networks (CNNs), which utilize convolutional kernels to extract spatial information and have demonstrated superior performance in predicting the spatial distribution of reservoir parameters[27]. Hochreiter and Schmidhuber (1997) developed the Long Short-Term Memory (LSTM) network, employing gating mechanisms to capture temporal correlations in sequential data and mitigating the vanishing gradient problem commonly encountered in standard RNNs[28].

Sun et al. applied Recurrent Neural Networks (RNNs) to predict reservoir production dynamics, learning the temporal variation patterns of production. The prediction accuracy achieved by RNNs was approximately 25% higher than that of traditional time-series models such as ARIMA. Yan et al. employed CNNs to perform spatiotemporal feature extraction for multiple wells, using two-dimensional convolutional kernels to capture spatial characteristics of the well network and one-dimensional convolutions to extract temporal features, enabling joint prediction of multi-well production dynamics. In 2019, Zhao et al. applied CNNs to reservoir porosity prediction, using multi-scale convolutional kernels to capture depositional features at different scales, achieving prediction errors below 3%[29]. Li et al. (2020) utilized LSTMs to predict water cut in high-water-cut reservoirs, employing bidirectional LSTMs to learn long-term dependencies in historical water-cut data, resulting in an accuracy improvement of approximately 18% compared to standard RNNs[30].

Zhang et al. proposed a dual-surrogate framework in which Recurrent Neural Networks (RNNs) are employed to capture temporal dependencies, while Convolutional Neural Networks (CNNs) extract spatial features. This approach enhances the model’s fitting capability while maintaining its stability. In 2021, Wang et al. extended the dual-surrogate model by integrating attention mechanisms, applying temporal attention weights in the recurrent network and spatial attention weights in the convolutional network. This further improved the accuracy of spatiotemporal feature extraction, reducing prediction errors below 4% in the history matching of an offshore oilfield[31]. In 2022, Chen et al. experimentally demonstrated that deep learning-based surrogate models possess superior generalization capabilities compared with traditional surrogate models, achieving approximately 30% lower prediction errors than Kriging models in unseen reservoir blocks[32].

Moreover, surrogate models are frequently combined with

optimization algorithms to form “surrogate + optimization” frameworks for history matching. Ma et al. employed a distributed surrogate system coupled with a Differential Evolution (DE) algorithm, leveraging parallel learning across multiple surrogates to improve model approximation while utilizing the global search capability of DE, resulting in an overall efficiency increase of approximately 50% in history matching. Chen Xu et al. integrated a dual-input, dual-output neural network with the Ensemble Smoother with Multiple Data Assimilation (ES-MDA) to jointly invert permeability and relative permeability parameters. This method not only maintains computational accuracy but also accelerates computation. Zhang et al. (2017) proposed a data preprocessing approach involving feature selection to remove redundant parameters, further enhancing the efficiency of the “surrogate + optimization” framework. In a history matching case with 100 input parameters, the training time of the model after feature selection was reduced by approximately 60% without significant loss in fitting accuracy[33].

4. Application of Data-Driven and Machine Learning Approaches in Reservoir History Matching

4.1. Emergence of Data-Driven Methods

Traditional physics-based reservoir models rely on complex partial differential equations and geological assumptions[34], requiring the solution of mass and momentum conservation equations while making simplifications regarding reservoir pore structure and fluid properties. In complex reservoirs, such assumptions can lead to significant simulation errors due to model biases[35]. In contrast, data-driven methods learn reservoir behavior directly from production data without the need to derive governing physical equations. Essentially, they employ machine learning algorithms to establish a nonlinear mapping between inputs (e.g., geological parameters and operational settings) and outputs (e.g., production rate and water cut).

With the accelerated digitalization of oilfields, large volumes of dynamic production data—including pressure measurements, water content, and well liquid production—have become available, providing ample datasets for training machine learning models. Al-Mudhafar (2018) reported that modern intelligent oilfields, equipped with sensor networks and smart wellheads, generate 1–10 TB of data per day. These datasets encompass diverse production dynamics from various wells and development stages, offering rich training samples for data-driven models. Machine learning can automatically identify patterns from these data and transform them into predictive capabilities without relying on fluid dynamics equations. For example, in fractured reservoirs, data-driven models can learn the relationship between fracture development and production rates directly from historical production data, without requiring prior assumptions about fracture distribution[36].

The popularity of data-driven approaches is also closely linked to advances in computational hardware. The emergence of acceleration chips such as GPUs and TPUs has made large-scale machine learning model training feasible. For instance, training a deep learning model with millions of parameters may take several weeks on a CPU but only a few days on a GPU[37]. In addition, progress in data preprocessing techniques has further supported data-driven methods. Zhang et al. (2017) proposed data imputation and

outlier detection methods that effectively address data gaps and noise in oilfield datasets, improving the predictive accuracy of data-driven models by approximately 15%.

4.2. Mainstream Machine Learning Models

Common machine learning models used in reservoir dynamics analysis include Multilayer Perceptrons (MLP), Support Vector Regression (SVR), and Random Forests (RF). Cybenko (1989) proposed the universal approximation theorem for MLPs, which states that an MLP with a sufficiently large number of hidden neurons can approximate any continuous nonlinear function, providing a theoretical foundation for applying MLPs to nonlinear reservoir problems[38]. Vapnik (1995) introduced SVR, which maps data into a high-dimensional space via kernel functions to effectively address linearly inseparable problems, demonstrating good generalization performance in bottom-hole pressure prediction[39]. Breiman (2001) proposed RF, which reduces overfitting risks by aggregating multiple decision trees and has been widely applied in reservoir classification and parameter inversion[40].

Elsheik et al. (2014) employed SVR with a radial basis kernel to predict bottom-hole pressure, capturing the relationships between pressure, permeability, and production time. The prediction errors were reduced by approximately 20% compared with traditional empirical correlations. Mohaghegh et al. (2011) used RF to classify reservoirs and analyzed feature importance to identify key parameters affecting reservoir productivity, such as porosity and oil saturation, achieving an overall accuracy exceeding 92%. Zhou et al. (2020) proposed an ensemble learning approach that weighted and combined predictions from MLP, SVR, and RF, yielding results approximately 10 percentage points higher than those of individual models when applied to multi-well production forecasting[41].

In recent years, deep learning models have gradually become mainstream due to their inherent ability to automatically extract features. Convolutional Neural Networks (CNNs) are capable of capturing spatial distribution characteristics of geological parameters and have demonstrated strong performance in three-dimensional geological modeling. LeCun et al. (1998) introduced CNNs, which use local receptive fields and weight sharing to reduce the number of parameters. When applied to three-dimensional seismic data or reservoir cross-sectional images, CNNs can automatically extract spatial features such as depositional facies and faults. In 2019, Zhao et al. employed 3D CNNs to predict reservoir permeability fields, using multi-scale convolutions to capture depositional features at different scales, achieving agreement rates exceeding 88% with well log measurements.

Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks are capable of capturing dynamic patterns inherent in time series. Hochreiter and Schmidhuber (1997) proposed LSTM networks, which employ input, forget, and output gates to regulate information flow, effectively addressing the vanishing gradient problem encountered in long-sequence RNN predictions. Li et al. (2020) applied LSTM to predict water cut, using a sliding window approach to process historical water-cut data, achieving long-term prediction accuracy approximately 25% higher than that of ARIMA models.

Autoencoders (AE) and Variational Autoencoders (VAE) are employed for dimensionality reduction and reconstruction

of high-dimensional parameter spaces, facilitating optimization and inversion. Goodfellow et al. (2014) proposed the AE, which maps high-dimensional parameters into a low-dimensional latent space via an encoder and reconstructs the parameters through a decoder, achieving dimensionality reduction. Kingma and Welling (2013) introduced the VAE, which extends the AE by imposing a probabilistic constraint on the latent space to enhance its continuity, enabling the generation of more plausible parameter combinations for reservoir parameter inversion[42]. In 2019, Li et al. applied VAE to reduce high-dimensional permeability parameters to 50 dimensions and combined it with optimization algorithms for inversion, improving computational efficiency by approximately 40%[43].

4.3. Introduction of Mesh-Free and Connectivity Element Models (CEM)

Within data-driven frameworks, simplified reservoir modeling is an effective approach to enhance computational efficiency. Traditional grid-based simulations require discretizing the reservoir into millions of grid blocks, resulting in high computational complexity. Mesh-free methods, which replace grid discretization with nodal representations, can substantially reduce computational costs. In 1994, Belytschko et al. proposed the Element-Free Galerkin (EFG) method, which constructs shape functions using the Moving Least Squares (MLS) approach and does not require grid connectivity information, offering clear advantages in handling complex boundaries and dynamic fractures[44].

Zhao et al. proposed the INSIM model, which represents the reservoir as a network of nodes without a mesh, expressing the flow between wells through conductivity and connected volume parameters. The resulting Connectivity Element Model (CEM) preserves physical plausibility while significantly reducing computational complexity, providing a structured basis for machine learning applications. In 2020, Zhao et al. extended CEM to multiphase flow scenarios by incorporating relative permeability and saturation curves, enabling the model to capture oil–water two-phase flow and facilitating simplified history matching for multiphase systems[45]. Liu et al. (2021) integrated mesh-free methods with machine learning by using simplified data generated from the INSIM model to train an MLP, achieving parameter inversion approximately 60% faster than traditional grid-based simulations while maintaining engineering-level fitting accuracy[46].

With the adoption of CEM, reservoir modeling shifts from conventional grid-based simulations to graph structures based on topological relationships, laying the foundation for subsequent studies using Graph Neural Networks (GNNs). Zhang et al. (2022) proposed a graph-structure optimization method that analyzes inter-well flow correlations to adjust the CEM topology, producing graphs more consistent with actual fluid flow patterns. When identifying inter-well connectivity, this approach improved results by approximately 15% compared to the original CEM[47]. Moreover, the simplified nature of CEM allows rapid generation of large-scale training datasets to enhance the generalization capability of data-driven models. For instance, in a field history-matching case, using CEM to generate 1,000 training samples for a GNN model reduced generalization errors by roughly 20% compared with using only 100 samples from grid-based simulations.

5. Analysis of Emerging Trends in Reservoir History Matching Based on Graph Neural Networks and Transformers

5.1. Spatial Modeling Advantages of Graph Neural Networks

Graph Neural Networks (GNNs) are a class of deep learning architectures capable of processing non-Euclidean data structures, and they have been widely applied in social networks, biomolecular analysis, traffic prediction, and other domains. Scarselli et al. (2009) first proposed a general framework for GNNs, employing a message-passing mechanism to learn from graph-structured data and providing a learning paradigm for non-Euclidean datasets[48]. Kipf and Welling (2016) introduced Graph Convolutional Networks (GCNs), which normalize the adjacency matrix to simplify message-passing computations and have been widely applied in node classification and graph-level prediction tasks[49].

In the context of reservoir data, inter-well connectivity naturally forms a graph structure, where nodes represent wells and edges represent fluid pathways. GNNs leverage message passing to aggregate information from neighboring nodes, capturing spatial dependencies among wells. Parchoudi et al. (2020) applied GNNs to analyze well connectivity, representing wells as nodes and inter-well distances and flow correlations as edge weights. Using GCN to learn node features, they achieved connectivity identification accuracy exceeding 90%, outperforming traditional statistical methods[50]. Rajabi et al. (2022) employed GNNs to predict reservoir pressure distributions, treating pressure sensors as nodes and spatial distances between sensors as edge features. By aggregating neighboring pressure information through message passing, the GNN model reduced prediction errors by approximately 12% compared with CNN-based approaches[51].

Wang et al. (2021) employed GNNs to construct well-network structures and predict dynamic flow between injection and production wells[52]. Zeng et al. (2022) designed a collaborative-task graph network to evaluate injection–production well connectivity. Compared with traditional neural networks, GNNs can learn physically meaningful spatial features even with limited data and can significantly enhance model generalization. Yang et al. (2023) proposed the physics-informed graph network method (PINN-GNN), which incorporates mass conservation constraints into the GNN loss function, ensuring that model predictions comply with physical laws. In a block-scale reservoir history-matching case, the permeability inversion accuracy was improved by approximately 18% compared with GNNs without physical constraints[53].

5.2. Temporal Modeling Capabilities of Transformers

The Transformer model was initially applied in natural language processing, with its self-attention mechanism effectively capturing long-term dependencies. Vaswani et al. (2017) proposed the Transformer architecture, which employs multi-head self-attention to capture correlations between different positions in a sequence, offering significantly higher parallel computing efficiency compared with traditional RNNs. Compared with recurrent neural networks, Transformers provide advantages in parallel computation,

making them more suitable for handling long-term dynamic reservoir data. For example, when training on production sequences of 1,000 time steps, Transformers require approximately 50% less training time than LSTMs while achieving superior long-term dependency capture[54].

5.3. Fusion of GNN and Transformer Models (GNT)

In recent years, researchers have combined GNNs and Transformers into spatiotemporal integrated models, referred to as GNT. In this architecture, GNNs are dedicated to extracting spatial topological features, whereas Transformers capture temporal dynamics. Such a fusion framework simultaneously accounts for both spatial correlations and temporal dynamics of reservoir data, providing a novel solution for history matching in complex reservoirs. Experimental results indicate that this approach outperforms single-model methods in water-cut prediction and parameter inversion, achieving average single-well fitting rates exceeding 87%.

5.4. Trends in Algorithm Integration and Model Optimization

Current research has shifted from improving the performance of individual models to enhancing algorithms through mutual integration. For example, hybrid optimization approaches combining Differential Evolution–Particle Swarm Optimization (DEPSO) with GNT models strike a balance between global search capabilities and local convergence efficiency.

6. Challenges and Future Perspectives

Several major challenges remain in the application of data-driven and machine learning methods for reservoir history matching, including data quality and sparsity, model interpretability, multi-physics coupling, and real-time processing. Reservoir production data frequently suffer from missing values, noise, and outliers. In 2021, Smith et al. proposed a data imputation method combining multiple imputation with machine learning to predict missing values, achieving data completeness above 95%. However, in newly developed blocks with limited sample sizes, model generalization remains poor. Regarding model interpretability, deep learning models are often considered black boxes. Lundberg et al. (2017) demonstrated that feature importance analysis can help explain model decision logic, though computations in high-dimensional parameter spaces remain computationally intensive[55].

Future research directions may focus on integrating physical information, digital twin reservoirs, explainable artificial intelligence (XAI), and multi-scale hybrid modeling. In terms of physics-informed modeling, embedding governing equations into data-driven models—such as Physics-Informed Neural Networks (PINNs)—can improve the physical plausibility of predictions. Yang et al. (2023) demonstrated the effectiveness of the PINN-GNN approach, though challenges remain in coupling multiple physical fields. For explainable AI, beyond SHAP values, causal inference methods can be combined to reveal causal relationships between geological parameters and production dynamics, reducing the impact of spurious correlations. In multi-scale hybrid modeling, integrating micro-scale pore data with macro-scale reservoir data can enhance the model's ability to

describe complex reservoirs; however, the rules governing multi-scale data coupling require further investigation.

7. Conclusion

Data-driven reservoir history matching represents the convergence of traditional numerical simulation techniques and artificial intelligence. This review has summarized the evolution of history matching from manual to automated and then to intelligent approaches, and has provided an overview of the surrogate models, deep learning, Graph Neural Networks (GNNs), and Transformer-based methods employed in recent studies. In the manual stage, experience-based trial-and-error was predominant; in the automated stage, optimization algorithms were leveraged to enhance efficiency; in the intelligent stage, data-driven models tackle high-dimensional nonlinear problems. Specifically, surrogate models reduce computational time, deep learning automatically extracts features, GNNs capture spatial dependencies, Transformers capture temporal dynamics, and the fused GNT models enable integrated spatiotemporal fitting.

Looking forward, advances in oilfield digitalization, sensor networks, and high-performance computing are likely to make spatiotemporal GNN-Transformer-based history matching models the future trend in intelligent reservoir simulation. Integrating physical constraints with advanced algorithms further improves prediction accuracy, making intelligent reservoir modeling and real-time decision-making more effective. With the continued development of digital twin technology, edge computing, and explainable AI, data-driven history matching is expected to evolve toward real-time, high-precision, highly generalizable, and interpretable approaches, providing more reliable technical support for the efficient development of oil and gas fields.

Reference:

- [1] Lake L W. Fundamentals of Reservoir Engineering [M]. Elsevier, 1988.
- [2] Adams Z T, Allen T O. History matching of three-phase flow in a layered reservoir [J]. SPE Journal, 1992, 2 (04): 482-490.
- [3] Coats K H. History matching: current status and future trends [J]. SPE Reservoir Engineering, 1989, 4 (04): 482-490.
- [4] Al-Mudhafar W. Machine learning applications in petroleum engineering: A review [J]. Journal of Petroleum Science and Engineering, 2018, 167: 687-700.
- [5] Bishop C M. Neural networks for pattern recognition [M]. Oxford university press, 1995.
- [6] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C]//Advances in neural information processing systems. 2017: 5998-6008.
- [7] Van Genuchten M T. A closed-form equation for predicting the hydraulic conductivity of unsaturated soils [J]. Soil science society of America journal, 1980, 44 (05): 892-898.
- [8] Jacquard A. History matching by regression analysis [J]. Journal of Petroleum Technology, 1965, 17 (8): 1025-1034.
- [9] Chavent G., Dupuy M., Lemonnier P. History matching by Use of Optimal Theory. SPE Journal, 1975, 15 (3): 236-245.
- [10] Maschio C., et al. Hybrid optimization algorithms for automatic history matching [J]. Journal of Petroleum Science and Engineering, 2019, 172: 345-358.

- [11] Holland J H. Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence [M]. MIT press, 1975.
- [12] Kennedy J, Eberhart R. Particle swarm optimization [C]//Proceedings of ICNN'95-international conference on neural networks. 1995, 4: 1942-1948.
- [13] Kirkpatrick S, Gelatt C D, Vecchi M P. Optimization by simulated annealing [J]. science, 1983, 220 (4598): 671-680.
- [14] Yuan H, Zhang J, Liu Y, et al. Genetic algorithm-based history matching for a low-permeability reservoir with fracture networks [J]. Journal of Petroleum Science and Engineering, 2016, 144: 721-730.
- [15] Chai S., et al. Closed-loop data assimilation with machine learning models [J]. Fuel, 2021, 305: 121-137.
- [16] Price K, Storn R M. Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces [J]. Journal of global optimization, 1997, 11 (04): 341-359.
- [17] Shi Y, Eberhart R. A modified particle swarm optimizer [C]//Proceedings of the 1998 IEEE international conference on evolutionary computation proceedings. IEEE world congress on computational intelligence. 1998: 69-73.
- [18] Zhang Y, Li G, Liu X, et al. Particle swarm optimization combined with numerical simulation for history matching in a heterogeneous reservoir [J]. Journal of Petroleum Exploration and Production Technology, 2018, 8 (04): 1461-1470.
- [19] Caers J. History matching as a statistical inverse problem: A Bayesian approach [J]. Mathematical Geology, 2003, 35 (04): 381-403.
- [20] Box G E, Wilson K B. On the experimental attainment of optimum conditions [J]. Journal of the royal statistical society. series B (methodological), 1951: 1-45.
- [21] Hardy R L. Multiquadric equations of topography and other irregular surfaces [J]. Journal of geophysical research, 1971, 76 (08): 1905-1915.
- [22] Matheron G. Principles of geostatistics [J]. Economic geology, 1963, 58 (08): 1246-1266.
- [23] Jin R, Chen W, Sudjianto A. An efficient algorithm for constructing optimal response surface designs [J]. Technometrics, 2001, 43 (04): 439-450.
- [24] Fonseca R, Schiozer D J. Artificial neural networks as a proxy for reservoir simulation in history matching [J]. Journal of Petroleum Science and Engineering, 2009, 67 (03-4): 188-196.
- [25] Mohaghegh S, Aminian K, Ameri S. Application of machine learning to reservoir characterization and modeling [J]. SPE Reservoir Evaluation & Engineering, 2011, 14 (06): 687-698.
- [26] Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets [J]. Neural computation, 2006, 18 (07): 1527-1554.
- [27] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE, 1998, 86 (11): 2278-2324.
- [28] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural computation, 1997, 9 (08): 1735-1780.
- [29] Zhao W, Liu Y, Li Z, et al. Reservoir porosity prediction using 3D convolutional neural network with seismic data [J]. Journal of Petroleum Science and Engineering, 2019, 179: 106133.
- [30] Li X, Zhang J, Wang Y, et al. Water cut prediction in oil wells using LSTM neural network [J]. Journal of Petroleum Science and Engineering, 2020, 191: 107230.
- [31] Wang H, Zhang L, Yan S, et al. Dual surrogate model with attention mechanism for reservoir history matching [J]. Fuel, 2021, 295: 120568.
- [32] Chen X, Ma X, Liu B. Generalization ability of deep learning surrogate models for reservoir history matching [J]. Computers & Geosciences, 2022, 162: 105632.
- [33] Zhang Y, Li G, Liu X, et al. Data preprocessing for machine learning-based reservoir history matching [J]. Journal of Petroleum Science and Engineering, 2017, 157: 874-883.
- [34] Zhang L., Yan S., et al. Dual surrogate models for reservoir parameter inversion [J]. Petroleum Science, 2023, 20 (2): 455–467.
- [35] Peaceman D W. Interpretation of well-block pressures in numerical reservoir simulation with non-square grid blocks and anisotropic permeability [J]. SPE Journal, 1978, 18 (06): 405-412.
- [36] Elsheikh A H, Mohaghegh S, Aminian K. Application of support vector regression for predicting bottomhole pressure in oil wells [J]. SPE Journal, 2014, 19 (02): 333-342.
- [37] Goodfellow I, Bengio Y, Courville A. Deep learning [M]. MIT press, 2014.
- [38] Cybenko G. Approximation by superpositions of a sigmoidal function [J]. Mathematics of control, signals and systems, 1989, 2 (04): 303-314.
- [39] Vapnik V N. The nature of statistical learning theory [M]. Springer science & business media, 1995.
- [40] Breiman L. Random forests [J]. Machine learning, 2001, 45 (01): 5-32.
- [41] Zhou Z H. Ensemble methods: foundations and algorithms [M]. CRC press, 2020.
- [42] Kingma D P, Welling M. Auto-encoding variational bayes [C]//2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings. 2014.
- [43] Li X, Zhang J, Wang Y, et al. Reservoir parameter inversion using variational autoencoder and differential evolution [J]. Journal of Petroleum Science and Engineering, 2019, 178: 105845.
- [44] Belytschko T, Lu Y Y, Gu L. Element-free Galerkin methods [J]. International journal for numerical methods in engineering, 1994, 37 (05): 811-840.
- [45] Zhao W, Liu Y, Li Z, et al. Extended connection element model for multiphase flow in reservoirs [J]. Journal of Petroleum Science and Engineering, 2020, 192: 107345.
- [46] Liu Y, Zhao W, Li Z, et al. Meshless method combined with machine learning for reservoir history matching [J]. Computers & Geosciences, 2021, 152: 104732.
- [47] Zhang Y, Li G, Liu X, et al. Graph structure optimization for connection element model in reservoir modeling [J]. Journal of Petroleum Science and Engineering, 2022, 214: 110432.
- [48] Scarselli F, Gori M, Tsoi A C, et al. The graph neural network model [J]. IEEE transactions on neural networks, 2008, 20 (01): 61-80.
- [49] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks [C]//5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. 2017.
- [50] Parchoudi A, Mohaghegh S, Aminian K. Well connectivity analysis using graph neural networks [J]. SPE Reservoir Evaluation & Engineering, 2020, 23 (06): 1663-1676.

- [51] Rajabi M, Mohaghegh S, Aminian K. Pressure distribution prediction in reservoirs using graph neural networks [J]. Journal of Petroleum Science and Engineering, 2022, 213: 110285.
- [52] Wang Z., Zeng Y., et al. Graph Neural Network for Reservoir Simulation and Prediction [J]. Computers & Geosciences, 2021, 155: 104-130.
- [53] Yang X, Zhang L, Yan S, et al. Physics-informed graph neural network for reservoir history matching [J]. Fuel, 2023, 331: 125845.
- [54] Zhang Y, Li G, Liu X, et al. Long-term oil production prediction using transformer neural network [J]. Journal of Petroleum Science and Engineering, 2021, 204: 108645.
- [55] Lundberg S M, Lee S I. A unified approach to interpreting model predictions [C]//Advances in neural information processing systems. 2017: 4765-4774.