Energy Proceedings
Vol 55, 2025
ISSN 2004-2965

Dynamic Production Forecasting Method Based on the EMD-iTransformer Model

Lizhe Li[1], Fujian Zhou[1*]

1 National Key Laboratory of Petroleum Resources and Engineering, China University of Petroleum-Beijing, Beijing 102249, China

(Fujian Zhou: zhoufj@cup.edu.cn)

ABSTRACT

In the development of unconventional oil and gas resources, dynamic production forecasting plays a crucial role in guiding production optimization and resource allocation. However, production variations are influenced by a combination of multi-source heterogeneous factors, including geological conditions, well control parameters, and historical production data. Consequently, forecasting models must handle high-dimensional, nonlinear, and temporally complex data while maintaining strong generalization ability and adaptability.

To address these challenges, this study proposes a dynamic production forecasting method tailored for multi-source heterogeneous data. The approach applies Empirical Mode Decomposition (EMD) to perform multi-scale decomposition of the original time series, enabling the extraction of key features across different frequency levels. These features are then integrated into an iTransformer-based predictive model, enhancing the model's capability to process non-stationary and multivariate dynamic data. This chapter first establishes and preprocesses the multi-source dataset, covering key dimensions such as well control parameters, geological attributes, and production data, followed by data cleaning and normalization. Subsequently, the structure and optimization process of the dynamic forecasting model are presented in detail, with a focus on the EMD decomposition strategy and enhancements to the iTransformer architecture. Experiments are conducted under both single-well and multi-well scenarios. The results demonstrate that the proposed method achieves high prediction accuracy and strong robustness in dynamic production forecasting tasks. Notably, it exhibits excellent trend-capturing capability in forecasting newly drilled wells. This study provides a novel approach and technical foundation for multi-source data-driven production prediction, offering promising potential for digital modeling and intelligent forecasting of complex reservoir production processes.

Keywords: AI-Driven Production Forecasting, Multi-Source Data Fusion, Nonlinear Time Series Modeling, Energy Intelligence for Sustainable Development

1. INTRODUCTION

Oil and gas production curves are typical time series, characterized by continuous changes over time and considered a form of dynamic data in petroleum engineering. The task of forecasting production curves using dynamic data is referred to as the dynamic production forecasting problem. Research in this area typically involves analyzing the autocorrelation, trend, and periodic patterns of production data. Some researchers have combined neural networks with traditional decline curve analysis (DCA) methods to improve forecasting accuracy. DCA remains a widely used approach in the oil and gas industry for modeling production decline behavior. For example, Ockree et al. (2018) and Pan et al. (2019) integrated neural networks with DCA to forecast production in hydraulically fractured wells, demonstrating high prediction accuracy on test samples through surrogate modeling. However, the DCA method itself suffers from significant limitations, including error propagation and oversimplification, which can lead to inaccurate production forecasts, especially under complex reservoir dynamics.

Several researchers have conducted production forecasting studies by treating oil and gas production curves as time series. Sagheer and Kotb (2019) compared the performance of Simple RNN and LSTM in forecasting oil and gas productivity time series, and found that LSTM significantly outperformed Simple RNN. Their work highlights the importance of evaluating different recurrent neural networks (RNNs) to identify the most suitable architecture for a given dataset.

Fargalla et al. (2024) embedded Time2Vec encoding and attention mechanisms into a CNN–BiGRU architecture for modeling shale and sandstone gas

This is a paper for the 11th Applied Energy Symposium: Low Carbon Cities & Urban Energy Systems (CUE2025), July 18-22, 2024, Kitakyushu, Japan.

reservoir dynamics. Their model accurately captured production evolution patterns under varying geological conditions. Xue et al. (2024) developed a CNN-based global sensitivity analysis (GSA) method to identify key influencing factors in shale gas production. They proposed three GSA scenarios: (1) full production phase analysis, (2) fixed-interval analysis, and (3) decline rate analysis. Wei et al. (2024) focused on identifying key control factors in coalbed methane (CBM) development. They applied smoothing and denoising techniques to time series data and proposed a deep learning–based CBM productivity prediction model, which effectively captured production characteristics.

Xu and Leung (2024) employed a spatio-temporal graph convolutional network (ST-GCN) for shale gas production forecasting, explicitly incorporating inter-well interference effects into the model. This approach resulted in significantly improved prediction accuracy by capturing both spatial and temporal dependencies among wells. Similarly, Liu et al. (2024) also considered inter-well interference in their forecasting framework for shale gas production, further highlighting the importance of modeling spatial interactions in multi-well systems.

Many researchers have introduced transfer learning and ensemble strategies to enable cross-scenario production forecasting. Mohd Razak et al. (2022) developed a deep recurrent neural network (RNN) model for long-term production forecasting in unconventional reservoirs. By integrating early production data from a target well with long-term dynamics learned from other wells, the model achieved knowledge transfer, thereby enhancing generalization in data-scarce scenarios. Dong et al. (2024) proposed a multi-source model fine-tuning approach, where a pre-trained source-domain model is adapted using a limited amount of target-domain data. This method ensures fast convergence and strong predictive performance in the target domain. In the context of coalbed methane (CBM) production forecasting, H.-C. Wang et al. (2024) introduced a learn-to-learn strategy to better utilize historical data from different types of reservoirs. They proposed a few-shot learning framework based on orthogonally constrained Riemannian meta-learning, which captures common patterns across reservoir categories while reducing interference from heterogeneous samples. This significantly enhances transferability between dissimilar source and target reservoirs.

Compared to traditional reservoir engineering and numerical simulation methods, machine learning (ML) approaches offer several notable advantages: 1.Data-driven modeling: ML eliminates the need for complex physical modeling, enabling more streamlined workflows with broader applicability. 2.Superior nonlinear mapping: ML models are capable of capturing intricate, nonlinear relationships among production variables with higher fidelity. 3.Reduced expert dependency: Unlike traditional methods that rely heavily on domain expertise, ML models can be trained and applied across similar reservoirs with minimal human intervention. 4.Self-adaptive learning: ML models can automatically update parameters through gradient-based or gradient-free optimization, making them suitable for analyzing multi-stage production data under evolving field conditions.

In recent years, Transformer-based architectures have been increasingly introduced into oil and gas production forecasting due to their powerful ability to capture long-range dependencies. Originally proposed by Vaswani et al. (2017), the Transformer model leverages the attention mechanism, which enables efficient parallel computation and superior modeling capacity. Its applications have spanned a wide range of fields, including speech recognition, image processing, time series forecasting, and recommender systems. Several studies have demonstrated the potential of Transformer variants in petroleum engineering tasks. For example, Huang et al. (2024) proposed a time-block dynamic attention Transformer, which segments production data into temporal blocks and adaptively weighs their contributions. This dynamic attention mechanism improves the robustness of forecasting under fluctuating field conditions.

However, there are still several limitations when applying Transformer models to time series forecasting tasks: First, the Transformer architecture was originally designed for Natural Language Processing (NLP), where both inputs and outputs are discrete word embeddings. When transferred to continuous time series data, the model fails to effectively capture intrinsic properties such as continuity, trend, and seasonality, which are essential for accurate forecasting. Second, the self-attention mechanism in the standard Transformer incurs high computational overhead. For a sequence of length L, the complexity of the vanilla self-attention operation is O($L2$), making the model inefficient and resource-intensive for long sequence forecasting tasks. Third, due to the high model complexity and large number of hyperparameters, the Transformer is prone to overfitting when applied to small-scale datasets, leading to poor generalization performance. Fourth, the original Transformer employs an auto-regressive prediction

strategy, where it forecasts one future time step at a time and recursively feeds the predicted value as input for the next step. This process leads to prediction error accumulation, as the model gradually shifts from using ground-truth values to relying on its own predictions, causing accuracy to degrade over extended forecasting horizons. Finally, extensive empirical evidence shows that the vanilla Transformer is not well-suited for multivariate forecasting. It primarily focuses on the temporal dimension and lacks mechanisms to effectively capture inter-variable dependencies. In real-world applications such as oil production forecasting, time series are influenced not only by historical trends of the target variable (e.g., production rate), but also by multiple structured physical parameters such as water cut, reservoir pressure, and choke size. The original Transformer struggles to model these multi-source interactions, limiting its predictive accuracy in complex physical systems.

The existing models face several critical limitations: Lack of explicit coupling between structured physical variables (e.g., pressure, water cut, and flow rate) and the target production sequence. Ignoring their dynamic dependencies can significantly impair forecasting accuracy. Architectural bottlenecks, especially in long-sequence modeling, where traditional attention-based frameworks suffer from high computational complexity and poor scalability, limiting their ability to capture global trends. Insufficient noise robustness, as field production data often exhibit strong multi-scale noise and mixed-frequency characteristics. Conventional single-scale models struggle to disentangle trend and disturbance components, leading to poor prediction stability.

To address the limitations of traditional models in long-sequence modeling, inadequate integration of structured features, and low inference efficiency, this study proposes a comprehensive framework improvement along the following dimensions: 1. Incorporation of EMD: EMD is employed as a preprocessing strategy to enhance the model's capability in handling non-stationary time series by decomposing complex signals into intrinsic mode functions of different frequency bands. 2. Enhanced long-range dependency modeling: The architecture is optimized to improve the extraction of global trends across extended time horizons, thereby mitigating the degradation of forecasting accuracy in long-term predictions. 3. Support for multi-dimensional time series forecasting: The proposed method is designed to accommodate high-dimensional inputs, capturing complex interactions among various operational variables under real-world production conditions. 4. Coupling mechanism for structured physical parameters: A fusion mechanism is introduced to explicitly model the interactions between physical attributes (e.g., pressure, water cut) and the primary production time series, thereby improving the physical interpretability of the prediction results. By integrating these enhancements, the proposed approach aims to significantly improve both the accuracy and robustness of production forecasting models, offering a more reliable tool for dynamic reservoir management and decision-making.

2. DATA CONSTRUCTION

2.1 Well control parameters

Well control parameters are critical factors influencing hydraulic fracturing performance and well productivity. These include fracturing fluid volume, pumping rate, proppant ratio, fracture length, fracture conductivity, and completion type. These parameters determine fracture morphology and its connectivity with the reservoir, serving as key indicators for optimizing fracturing design. In this study, well control data were extracted from field fracturing operation reports and organized into a structured database for subsequent analysis and modeling. The collected data fall into the following three categories:

(1) Well trajectory information: including well name, kickoff point depth (m), surface elevation (m), measured total vertical depth (TVD, m), measured depth (MD, m), build-up point depth (m), and depths of key trajectory points (e.g., points A and B) in both TVD and MD. These parameters describe the spatial geometry of the wellbore and its relationship with the target reservoir; (2) Cluster and stage design parameters: such as the number of perforation clusters per stage and the location of perforation centers, which characterize the detailed layout of fracturing stages; (3) Operational pressure data: including pumping volume, rate, proppant concentration, breakdown pressure, and shut-in pressure for each stage. These parameters provide a basis for analyzing the mechanical response during fracturing and the underlying fracture initiation mechanisms；

2.2 Geological parameters

Geological parameters have a direct impact on the seepage behavior in reservoirs, including reservoir permeability, porosity, lithology, fracture conductivity, and reservoir thickness, etc. Accurate measurement and

characterization of these parameters are essential for developing reliable and physically consistent productivity prediction models. In this study, reservoir petrophysical properties were systematically characterized based on logging interpretation data, including the length of the interpretation section (m), the thickness of the oil layer (m), sweet spot encounter rate (%), the evaluation of various geological desserts (Class I, Class II, Class III geological dessert thickness, m), porosity (%), permeability (mD), and oil saturation (%) and other key indicators.

2.3 Production Data

The production data capture the dynamic performance of oil and gas wells during field development and serve as a crucial basis for evaluating the effectiveness of reservoir stimulation and validating the reliability of productivity prediction models. The main production indicators include daily oil production, daily gas production, water content, cumulative production, bottom hole or formation pressure variations, and choke size. A systematic time-series analysis of these data enables assessment of the conductivity of hydraulic fractures and their connectivity with the reservoir. This, in turn, allows for iterative calibration of the prediction model to improve its adaptability and accuracy. In this study, production data from a representative well were processed and analyzed. As shown in Fig.1, the production trend illustrates the temporal evolution of reservoir deliverability.

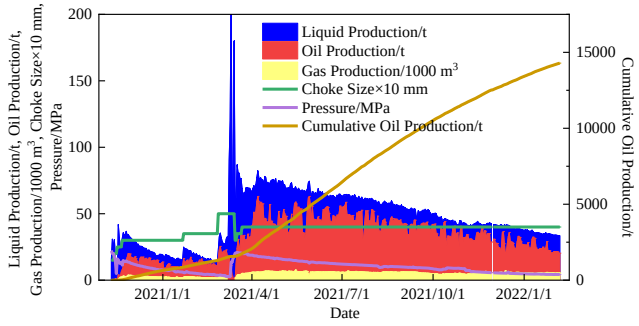

Fig. 1 Trends in production

3. MATERIAL AND METHODS

3.1 Data cleaning and standardization

Due to variations in data sources, issues such as missing values, outliers, and inconsistent units are commonly encountered. Therefore, data cleaning is a necessary step, involving outlier detection, imputation of missing values, and unit conversion. To eliminate dimensional discrepancies and improve model convergence, normalization or standardization techniques are employed during preprocessing.

In the field production data, instances of incomplete or missing entries are observed, while well logging data often contain missing values or placeholder entries for invalid measurements. Consequently, a thorough outlier detection and missing value check is required to ensure data quality. Appropriate imputation methods are selected based on data characteristics to ensure completeness and rationality.

For production data gaps caused by shut-in or operational adjustments, constant-value imputation is used to maintain consistency. In contrast, for missing measurements in logging and reservoir datasets, different strategies are adopted depending on the context. When the missing data occur outside the target stratigraphic intervals or involve large gaps, forward or backward filling is applied to leverage existing sequential information—suitable for datasets with continuity in stratigraphy. Within target intervals, interpolation methods are employed to better reconstruct measurement trends and enhance the continuity and reliability of the data.

These imputation strategies help minimize the impact of missing data on subsequent analyses and provide more accurate and consistent data support for modeling and interpretation.

3.2 Design of dynamic production forecasting algorithm

This study proposes a hybrid model architecture that integrates data-driven techniques with physics-informed constraints. The overall modeling framework features three key components:

1. Data Decomposition: The original production time series is decomposed into multiple components across different scales, including trend, fluctuation, and noise terms. This enhances the model's ability to capture complex temporal variations and multiscale dynamics.

2. Physical Constraints: Domain knowledge and physical principles-such as hydraulic fracturing mechanisms and flow governing equations are embedded into the modeling process to constrain and guide the prediction outputs. This integration improves the physical consistency and interpretability of the results.

3. iTransformer Forecasting: An improved Transformer architecture based on self-attention mechanisms is employed to effectively model long-range temporal dependencies in time-series data. The model

incorporates structured and unstructured inputs—such as well control parameters, geological attributes, and production records-enabling cross-domain learning and coupled analysis across multiple information dimensions.

3.3 Data Decomposition Design

The empirical mode decomposition (EMD) data decomposition method is introduced and compared with the singular spectrum analysis (SSA) and variational mode decomposition (VMD) methods. EMD effectively extracts characteristic patterns embedded in production time-series data, enhancing analytical accuracy and robustness.

EMD, originally proposed by Huang et al. (1998), is a data-driven signal processing method well-suited for analyzing nonlinear and non-stationary time series. It is designed to extract intrinsic oscillatory modes from complex signals without requiring a predefined basis. The EMD algorithm decomposes the original signal into a finite set of Intrinsic Mode Functions (IMFs) and a residual component. Each IMF represents local oscillations at a specific temporal scale, reflecting physically meaningful signal features.

To qualify as an IMF, a function must satisfy the following two criteria:

1.The number of extrema and the number of zero-crossings must either be equal or differ at most by one throughout the dataset.

2.At any point in time, the mean value of the upper and lower envelopes (defined by the local maxima and minima, respectively) must be zero.

Through an iterative sifting process, EMD decomposes the original signal $S(t)$ into a series of IMFs and residuals $R(t)$ through an iterative screening process, such that:

$$S(t) = \sum_{i=1}^{n} IMF_i(t) + R_n(t) \tag{1}$$

Where，$S(t)$ represents the original signal or raw time series; *n* denotes the number of decomposition iterations; $IMF_i(t)$ refers to the *i*-th Intrinsic Mode Function obtained during the decomposition, $R(t)$ is the residual component remaining after *n* rounds of sifting.

Specifically, the EMD decomposition procedure involves the following steps:

First, identify all local maxima and minima of the original time series $S(t)$. Use cubic spline interpolation to construct the upper envelope and lower envelope of the signal.

Secondly, compute the mean $l(t)$ envelope and subtract the mean from the original signal $S(t)$ to obtain the first trial signal $h_1(t)$:

$$h_1(t) = S(t) - l(t) \tag{2}$$

Where, $h_1(t)$ denotes the iterative signal, and $l(t)$ denotes the mean of the upper and lower envelopes. The signal $h_1(t)$ is then checked to determine whether it satisfies the two conditions of IMF. If the conditions are satisfied, $IMF_1 = h_1(t)$; If not, $h_1(t)$ is treated as the new input signal and the above steps are repeated until the IMF conditions are met. Once an IMF is extracted, subtract it from the current signal to obtain the residual. This iterative process continues until the residual becomes a monotonic function or contains only a low-frequency trend component.

3.4 Improved iTransformer model

As previously discussed, hydraulic fracturing production forecasting is a complex multivariate time-series prediction task. Accurate prediction requires modeling not only the temporal evolution of production rates but also the multidimensional interactions among various external factors. These external factors include well control parameters, geological attributes, and production-related data, as outlined in Section 2. Each variable constitutes an independent time series, corresponding to different physical entities (e.g., reservoir, fractured well, fracturing stage) and representing distinct meanings (e.g., pressure, flow rate, water cut), often with heterogeneous units and statistical distributions.

Conventional Transformer models do not explicitly account for inter-variable differences in multivariate time series. To address this limitation and improve performance in sequence modeling and generation tasks, this study introduces an enhanced Transformer architecture: the iTransformer.

The original Transformer architecture consists of two main components: an encoder that maps input sequences to latent representations, and a decoder that translates these representations into output sequences. Compared with traditional models such as RNN, LSTM, and GRU, Transformer leverages a fully connected self-attention mechanism, enabling global parallel computation and removing the constraint of sequential dependencies. This significantly enhances its capability to model long-range temporal dependencies.

In conventional Transformer-based time-series applications, variables at a given time step are aggregated into a single Temporal Token. However, this approach introduces two key issues: 1. It fuses

semantically distinct variables into a single high-dimensional token, making it difficult for the model to differentiate and learn meaningful inter-variable dependencies; 2. The semantic representation of each token is weak, and the limited receptive field prevents the model from effectively capturing asynchronous or variably sampled inputs.

To overcome these limitations, iTransformer introduces a transposed structure that replaces the traditional temporal-token formulation with Variate Tokens. Instead of treating each time slice as a token, iTransformer maps the entire temporal evolution of each variable into a distinct token. This design allows the model to preserve variable-specific semantics, avoid inter-variable entanglement, and better exploit temporal patterns at the variable level. Additionally, a normalization module is applied to eliminate differences in unit scales and value ranges across variables, ensuring comparability and model stability. In summary, unlike conventional Transformer models that emphasize temporal sampling, the iTransformer emphasizes global temporal patterns across variables, making it especially suitable for high-dimensional, strongly correlated multivariate time series prediction scenarios.

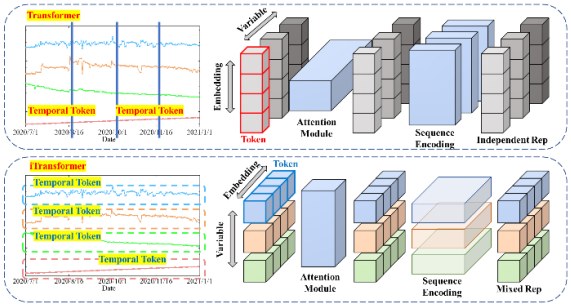

Fig. 2 Token Design for iTransformer

The architecture of the iTransformer model consists of three main modules: an input module, a feature extraction module, and an output module. The overall model framework is illustrated in Fig. 2.

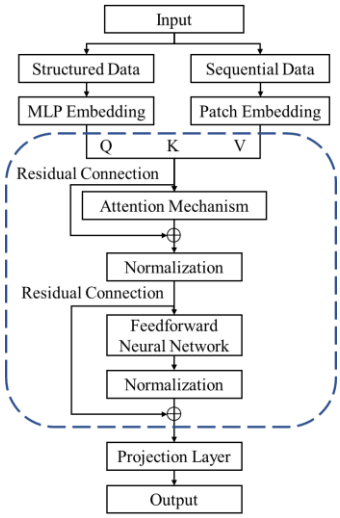

Fig. 3 iTransformer Model Architecture

The input module is primarily responsible for encoding the input data. In this study, the input consists of both structured data and time-series data, which are processed through different embedding layers to generate low-dimensional representations suitable for model learning. Specifically, the input data are composed of structured data $\mathbf{X}_s \in \mathbb{R}^{n \times d_s}$ and time-series data $\mathbf{X}_t \in \mathbb{R}^{T \times d_t}$.

Structured data and time-series data are embedded through independent layers:

$$\mathbf{E}_s = \text{Embedding}_s(\mathbf{X}_s) \in \mathbb{R}^{n \times d} \quad (3)$$

$$\mathbf{E}_t = \text{Embedding}_t(\mathbf{X}_t) \in \mathbb{R}^{T \times d} \quad (4)$$

The two types of data are concatenated or fused as the input sequence of the iTransformer:

$$\mathbf{E} = \text{Concat}(\mathbf{E}_s, \mathbf{E}_t) \in \mathbb{R}^{(n+T) \times d} \quad (5)$$

Where, *n* is the number of structured features, *T* is the length of time-series, d_s and d_t are dimensions.

The feature extraction module of the iTransformer follows the standard Transformer architecture, using only the encoder component, which consists of multiple layers of self-attention and feedforward neural networks. Specifically, to reduce computational complexity and improve the efficiency of modeling long time series, the iTransformer incorporates local attention and sparse attention mechanisms. Each encoder layer includes multi-head attention and a feedforward network. The local attention mechanism restricts the attention computation within a predefined window, effectively reducing the overall computational burden. Meanwhile, global attention is selectively applied to ensure that the model can still capture long-range dependencies across the entire sequence.

The attention mechanism applies a linear transformation to the input sequence *E*, producing three

matrices: the Query matrix *Q*, the Key matrix *K*, and the Value matrix *V*. The attention is then computed using the following formula:

$$\text{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\mathbf{V} \quad (6)$$

Where, $\mathbf{Q}\mathbf{K}^{\mathrm{T}}$ refers to the weight matrix, and calculates the similarity between the query and the key; $\sqrt{d_k}$ is the scaling factor, to prevent the inner product value is too large to cause the gradient to disappear; $\text{softmax}(\cdot)$ normalize each query location to form attention weights.

Standard attention can only learn dependencies from a single representation subspace, whereas multi-head attention splits the input into multiple subspaces and learns attention information independently within each. This allows the model to capture richer semantic features.

$$\text{MultiHead}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{Concat}(\text{head}_1,\cdots,\text{head}_h)\mathbf{W}^O \quad (7)$$

Each attention head is computed as follows:

$$\text{head}_i = \text{Attention}(\mathbf{Q}_i,\mathbf{K}_i,\mathbf{V}_i) \quad (8)$$

Where, $Q_i = EW_i^Q$, $K_i = EW_i^K$, $V_i = EW_i^V$ is the subspace expression of each attention head; W_i^Q, W_i^K, W_i^V is the mapping matrix of each head; W^O is the projection matrix used to linearly transform the concatenated heads back to the original embedding dimension.

In addition, iTransformer incorporates a local attention mechanism, where attention is computed only within a fixed window around each query position. This strategy significantly reduces the overall computational complexity.

$$\text{LocalAttention}(\mathbf{Q}_i,\mathbf{K}_{i-w:i+w},\mathbf{V}_{i-w:i+w}) = \sum_{j=i-w}^{i+w}\alpha_{ij}\mathbf{V}_j \quad (9)$$

Where, the attention weight $\alpha_{i,j}$ is:

$$\alpha_{ij} = \frac{\exp\left(\frac{\mathbf{Q}_i\cdot\mathbf{K}_j^{\mathrm{T}}}{\sqrt{d_k}}\right)}{\sum_{k=i-w}^{i+w}\exp\left(\frac{\mathbf{Q}_i\cdot\mathbf{K}_k^{\mathrm{T}}}{\sqrt{d_k}}\right)} \quad (10)$$

Where, w is the window size, meaning that *w* time steps are considered on both sides of each query position; d_k denotes the dimensionality of the attention vectors. The computation range of local attention is restricted to a neighborhood window of size 2*w*+1, resulting in a reduced complexity of O(*Lw*).

Sparse attention is a technique that sparsifies the attention weight matrix by computing attention only at a subset of predefined positions. This approach effectively controls both memory usage and computational cost.

$$\text{SparseAttention}(\mathbf{Q}_i,\mathbf{K},\mathbf{V}) = \sum_{j\in\mathcal{N}(i)}\alpha_{ij}\mathbf{V}_j \quad (11)$$

The attention weights in sparse attention are computed as:

$$\alpha_{ij} = \frac{\exp\left(\frac{\mathbf{Q}_i\cdot\mathbf{K}_j^{\mathrm{T}}}{\sqrt{d_k}}\right)}{\sum_{k\in\mathcal{N}(i)}\exp\left(\frac{\mathbf{Q}_i\cdot\mathbf{K}_k^{\mathrm{T}}}{\sqrt{d_k}}\right)} \quad (12)$$

Where, $\mathcal{N}(i)$ is the set of predefined positions defined by the sparse connectivity structure:

$$\mathcal{N}(i) = \{i-w,\ldots,i+w\}\cup \{i+ks \mid k\in\mathbb{Z}, i+ks\in[0,L)\} \quad (13)$$

For each position i, the attention computation range is defined as a combination of local and jump positions. In the formula, the first term represents the local window, while the second term corresponds to jump connections, where dependencies are retained every s time steps. Under this sparse attention scheme, the computational complexity is reduced to $O\left(L\sqrt{L}\right)$.

Overall, the iTransformer adopts a hybrid attention strategy that integrates local, sparse, and global attention mechanisms, namely:

$$\begin{aligned} Attention = {} & LocalAttention + SparseAttention \\ & + \lambda\cdot GlobalTokenAttention \end{aligned} \quad (14)$$

Further, after normalization and feedforward neural network module, a set of perfect feature extraction module is formed. The overall formula is expressed as follows:

$$E^{m+1} = \text{TrmBlock}(E^m), m = 0, 1, \cdots, M-1 \quad (15)$$

Finally, the output module mainly consists of the projection layer. Through the projection layer, the output prediction result is obtained, namely:

$$Y = \text{projection}(E^M) \quad (16)$$

Among them, Y represents the predicted value, that is, the predicted future time point value.

Overall, the model architecture designed in this chapter offers several key advantages. First, it enables structural fusion, effectively integrating structured data and time-series data into a unified input representation. Second, the incorporation of local and sparse attention mechanisms significantly reduces the computational complexity associated with long sequences. Third, by retaining the global attention mechanism, the model maintains a strong capability for capturing long-range dependencies, which is essential for long-horizon time-

series forecasting tasks. the iTransformer emphasizes global temporal patterns across variables, making it especially suitable for high-dimensional, strongly correlated multivariate time series prediction scenarios.

4. EXPERIMENTS AND RESULTS

4.1 Experiments and evaluations of EMD methods

Experiments and evaluations of three decomposition methods were conducted based on six-dimensional data, starting with qualitative analysis:

SSA (Singular Spectrum Analysis) is more suitable for processing stationary or slowly varying processes, and it performs well in extracting trends and separating periodic components. However, its performance is highly sensitive to the window size parameter, and it lacks robustness in handling non-stationary signals with abrupt fluctuations.

VMD (Variational Mode Decomposition) offers both frequency separation capability and optimization stability. It allows for precise control over the number of modes, and the resulting components are mutually exclusive, smooth, and physically interpretable. Nevertheless, its effectiveness strongly depends on the proper selection of the parameter K, which requires the aid of empirical criteria or validation mechanisms.

EMD excels in adaptability and requires no prior assumptions, making it well-suited for non-stationary and nonlinear signals. It is particularly useful for preliminary exploration of intrinsic structures in data. However, EMD suffers from stability issues, including sensitivity to noise, mode mixing, and boundary instability, which may affect the reliability of the decomposition results.

Next, multiple decomposition methods were applied to the six-dimensional data, the reconstructed data was analyzed, and a quantitative evaluation of the methods was conducted. The six-dimensional features are designated as follows: parameter 1 represents daily fluid production (t), parameter 2 represents daily water production (t), parameter 3 represents daily gas production (1000 m^3), parameter 4 represents daily oil production (t), parameter 5 represents cumulative oil production (t), and parameter 6 represents flowback rate (%).

Fig. 4 displays a comparison between the original time series data corresponding to parameter 1 (represented by a black dashed line) and the results of various data reconstruction methods, including MSSA, MSSA-Corrected, VMD-K1, VMD-K2, VMD-set, and EMD, represented by blue solid line, light blue solid line, green solid line, orange solid line, purple solid line, and red solid line, respectively. The comparison reveals differences in how various data reconstruction strategies capture the main trends, local fluctuations, and anomalous peak characteristics of the original sequence. First, both MSSA and MSSA-Corrected methods effectively preserve the fluctuation characteristics of the original time series, demonstrating high fidelity in capturing high-frequency oscillations and peak variations, particularly in reproducing the main trends and periodic fluctuations of the curve. Specifically, the MSSA method exhibits a slight forward shift in the predicted curve, indicating a time alignment deviation possibly due to errors in extracting lagged features from the sequence. In contrast, the MSSA-Corrected method incorporates phase correction, resulting in higher temporal consistency with the original sequence, especially in accurately aligning peaks and valleys. Second, the VMD methods (including K1, K2, and set configurations) tend to smooth the data. Under the K1 and K2 configurations, the reconstructed curves lose significant local fluctuation information from the original data, particularly in the anomalous peak region around the 100th time point, where these features are almost entirely attenuated or eliminated. Although VMD-set improves the reproduction of fluctuation details to some extent, it still fails to fully capture the drastic changes in the original data, with limited responsiveness to anomalous peaks, making it more suitable for processing smooth signals. Finally, the EMD method performs most prominently overall, accurately capturing the global trends of the curve while demonstrating strong responsiveness to local details, particularly in regions with abrupt peak changes. Its reconstructed curve closely matches the original sequence in shape, balancing global stability and local accuracy, thus exhibiting superior reconstruction capabilities.

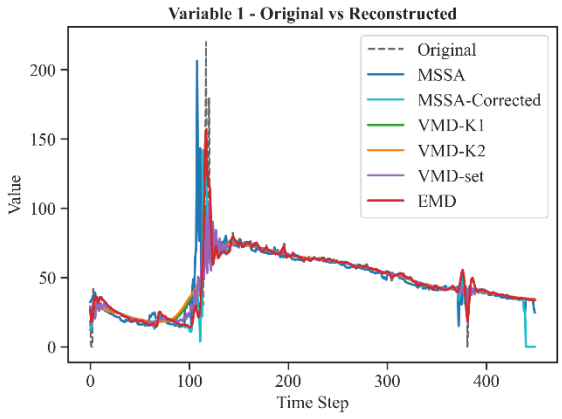

Fig. 4 The comparison between the original time series data and the results of various data reconstruction methods corresponding to parameter 1

Fig. 5 and Fig. 6 illustrate the comparison between the original time series data for parameter 2 and parameter 3 and the results of various data reconstruction methods. Overall, the EMD method demonstrates the most balanced performance in this comparison, effectively reproducing global trends while exhibiting high responsiveness near abrupt change points. However, during periods of rapid oscillations or abrupt zero-value transitions, the MSSA, MSSA-Corrected, and EMD methods erroneously amplify fluctuations, resulting in unreasonable negative values in the predicted curves, leading to some physically implausible outcomes. Subsequent data processing to address negative values or the incorporation of physical constraints is necessary to mitigate these issues.

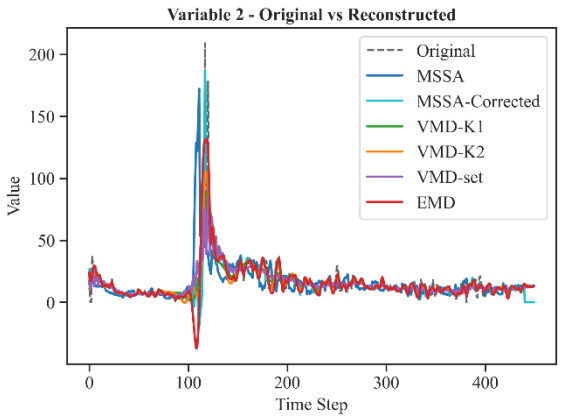

Fig. 5 The comparison between the original time series data and the results of various data reconstruction methods corresponding to parameter 2

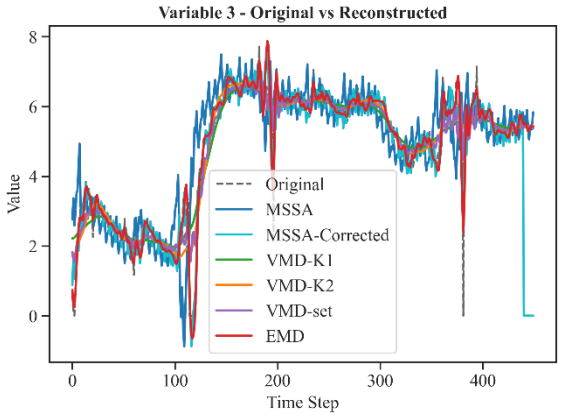

Fig. 6 The comparison between the original time series data and the results of various data reconstruction methods corresponding to parameter 3

Fig. 7 presents a comparison between the original time series data for parameter 4 and the results of various data reconstruction methods. Overall, all methods effectively capture the rising-plateau-declining structure of the time series. The EMD method's reconstructed curve demonstrates high sensitivity in responding to abrupt changes at multiple key points, while maintaining consistent fluctuations and physical plausibility throughout the plateau period. Compared to other methods, EMD exhibits superior performance in both global fitting accuracy and local responsiveness.

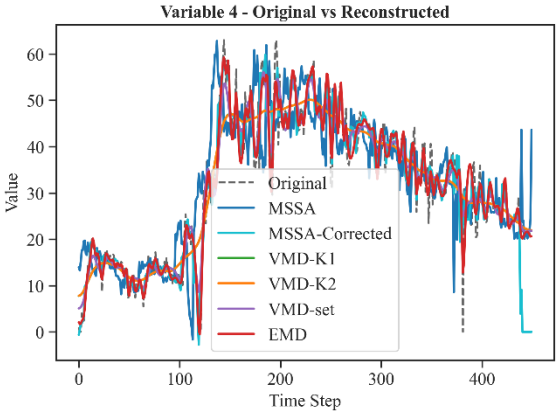

Fig. 7 The comparison between the original time series data and the results of various data reconstruction methods corresponding to parameter 4

Fig. 8 and Fig. 9 respectively illustrate the comparison between the original time series data for parameter 5 and parameter 6 and the results of various data reconstruction methods. Both sequences exhibit a smooth upward trend with minimal fluctuations and relatively simple change structures. In terms of overall trend reproduction, all methods accurately capture the smooth growth characteristics of the parameters over time, demonstrating high consistency and indicating robust adaptability for handling single-trend signals. However, the MSSA method slightly anticipates the trend as the time step approaches the sequence's end. The MSSA-Corrected method's reconstructed curve shows a sharp, cliff-like drop at the tail due to missing data, revealing deficiencies in boundary handling. VMD-K2 slightly deviates at the tail, marginally underestimating the true growth magnitude. Overall, the EMD method performs relatively well, effectively reproducing the global trend while appropriately preserving structural changes near inflection points, demonstrating strong local adaptability.

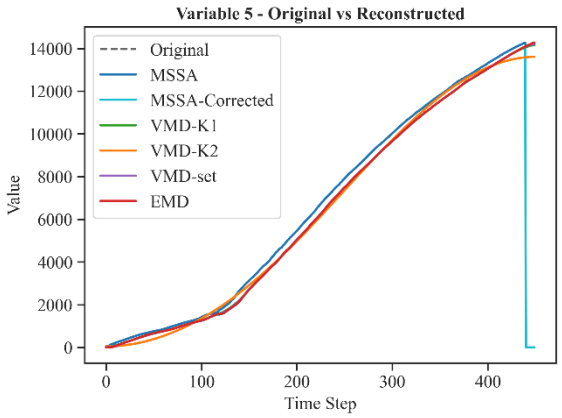

Fig. 8 The comparison between the original time series data and the results of various data reconstruction methods corresponding to parameter 5

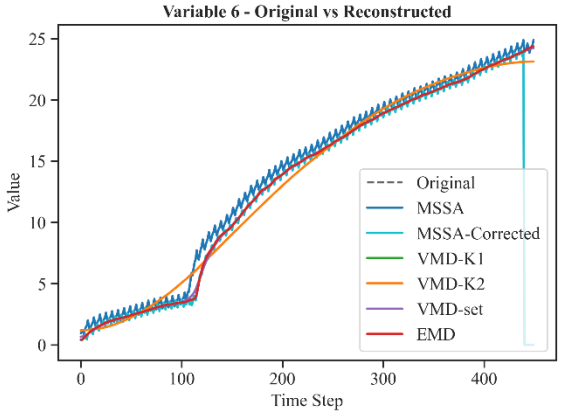

Fig. 9 The comparison between the original time series data and the results of various data reconstruction methods corresponding to parameter 6

Fig. 10 presents the comparison curves of various models across multiple evaluation metrics (MSE, RMSE, MAE, PCC), with the horizontal axis representing different predicted variables and the vertical axis indicating the corresponding metric values. The overall trend shows that the EMD method consistently outperforms others across all evaluation metrics, with its curves positioned at the lowest values for error-based metrics (MSE, RMSE, MAE) and the highest for PCC, indicating superior predictive accuracy and trend consistency. Specifically, for the three error-based metrics (MSE, RMSE, MAE), EMD achieves the lowest values, demonstrating minimal prediction errors across most variables. Regarding the correlation metric (PCC), EMD consistently maintains values above 0.9, approaching 1, reflecting a strong ability to fit the original data's change trends. Additionally, VMD-set and VMD-K2 follow closely behind EMD in most metrics, outperforming MSSA and its corrected version. MSSA exhibits poorer overall performance, with higher errors and greater PCC variability, indicating limited adaptability.

Therefore, combining the qualitative analysis from previous sections with the quantitative evaluation of error metrics, the EMD method demonstrates the best performance in terms of predictive accuracy, stability, and trend consistency, significantly surpassing other compared methods. Consequently, this chapter selects EMD as the core method for subsequent data processing.

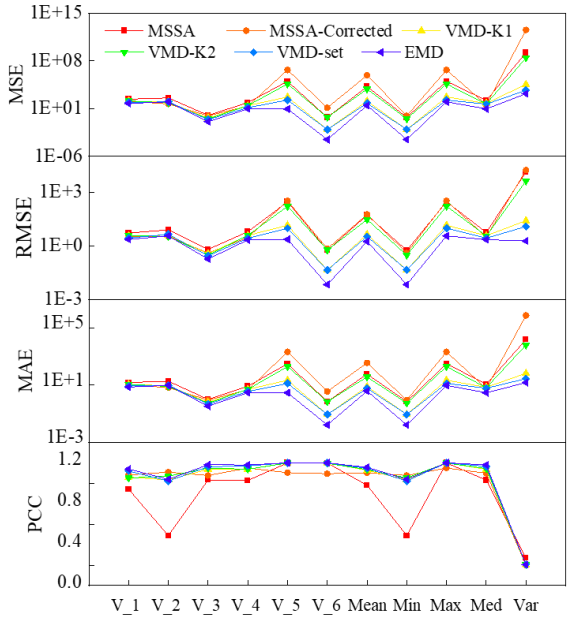

Fig. 10 Comparison curves of models across multiple evaluation metrics

4.2 EMD-ITransformer model evaluation experimental

To further validate the applicability and predictive capability of the EMD-ITransformer model in real-world production scenarios, this study designed and conducted two sets of experiments: one for single-well production forecasting, predicting future production based on historical results; the other for multi-well production forecasting, predicting new well production based on historical data from older wells. The following sections elaborate on the experimental design and comparison with existing methods.

(1) Experiment 1

Using historical six-dimensional feature time series data from a single well, the goal is to predict the well's future production (i.e., parameter 4: daily oil production (t)). The six-dimensional data were split into a training set and a test set at a 7:3 ratio: the training set was used for model parameter learning, and the test set was used

post-training to evaluate the model's generalization ability. Empirical Mode Decomposition (EMD) was applied to decompose each input variable, extracting effective intrinsic mode functions (IMFs) as new feature inputs to achieve multi-scale information enhancement. This strategy significantly improved the model's ability to identify abrupt changes and periodic variations. Subsequently, the iTransformer's attention mechanism was employed to model the reconstructed sequences, capturing long-term trends and short-term fluctuation characteristics. The hyperparameter settings for the ITransformer model are as follows: Learning Rate sets to 0.001, controlling the step size of model weight updates; too high a value leads to model instability, while too low a value results in slow convergence. Batch Size sets to 32, determining the number of data samples processed during each gradient update. Training was conducted for 50 epochs, ensuring the model fully learns data features without overfitting. The Adam optimizer was used, offering faster convergence compared to traditional methods like SGD. Mean Squared Error (MSE) was used to measure the difference between predicted and actual values.

(2) Experiment 2

To further validate the performance of the EMD-ITransformer model in multi-well production forecasting, this study predicts the future production of a new well (primarily parameter 4: daily oil production (t)) based on historical six-dimensional feature time series data from older wells (parameter 1: daily fluid production (t), parameter 2: daily water production (t), parameter 3: daily gas production (1000 m^3), parameter 4: daily oil production (t), parameter 5: cumulative oil production (t), parameter 6: flowback rate (%)). To investigate the impact of different data processing methods on model performance, four datasets (D1, D2, D3, D4) were designed, and comparative experiments were conducted around two key hyperparameters: window size and prediction length.

Dataset Features:

D1: Raw oil production curves, containing only the time-varying daily oil production (parameter 4) of each well.

D2: Oil production curves + auxiliary parameters, incorporating static auxiliary parameters (e.g., formation thickness, completion method, porosity, permeability) into D1 to enhance modeling of inter-well differences.

D3: EMD-processed oil production curves, applying Empirical Mode Decomposition (EMD) to D1's raw oil production curves to extract primary intrinsic mode functions (IMFs) and reconstruct denoised oil production curves to reduce noise interference.

D4: EMD-processed oil production curves + auxiliary parameters, combining D3's denoised curves with D2's auxiliary parameters to construct enhanced input features integrating multi-scale temporal information and well-control attributes.

Data Preprocessing: All datasets underwent standardization to eliminate magnitude differences between wells. Auxiliary parameters were processed using normalization or Z-score standardization to ensure data consistency.

Hyperparameter Settings:

Window Size (Historical Sequence Length): {5, 10, 20, 40, 60}, representing the number of historical data points used for each prediction, corresponding to short- to long-term temporal dependencies.

Prediction Length (Future Prediction Steps): {1, 2, 5, 10}, representing the future time steps predicted, corresponding to daily, weekly, and monthly production forecasting needs.

Other Hyperparameters: Consistent with the single-well experiment settings: learning rate of 0.001, batch size of 32, 50 epochs, Adam optimizer, and Mean Squared Error (MSE) loss function. The training set and test set were divided at a 7:3 ratio.

4.3 EMD-iTransformer model evaluation results

(1) Experiment 1

Fig. 11 shows the Comparison of the effect of each model on the prediction of daily oil production of oil wells. From Fig. 11, it is evident that all models closely approximate the overall trend of the actual daily oil production curve, indicating a certain degree of trend-capturing capability. The iTransformer (green) and Transformer (brown) respond most promptly at inflection points, and their predicted fluctuations align closely with the actual values. For instance, during the sharp decline and subsequent oscillations in early December, both models precisely tracked the trend changes, with their prediction curves nearly overlapping the ground truth.

This observation is consistent with the quantitative error metrics: iTransformer achieved the best performance with MAE of 0.745, RMSE of 1.192, and R2 of 0.951, followed closely by Transformer, further confirming their advantage in modeling complex, non-stationary time series data. In contrast, CNN (light blue) and GRU (orange) maintain the general trend reasonably well but show deviations in the face of abrupt changes.

For example, during the trough in early December 2021, neither model accurately captured the minimum point. Their corresponding RMSE values were 5.03 and 5.48, with R^2 scores below 0.20, indicating limited sensitivity to local variations.

LSTM (red) and RNN (purple) exhibit significant lag and smoothing across multiple time periods, especially near local troughs and peaks, failing to reflect abrupt changes in production accurately. As shown in the figure, their predictions at inflection points often deviate from the true trend, revealing a weakness in modeling short-term high-frequency disturbances. This is corroborated by their poor error metrics: RMSEs of 4.97 and 4.25 for LSTM and RNN respectively, with R^2 values below 0.30, indicating relatively weak fitting performance.

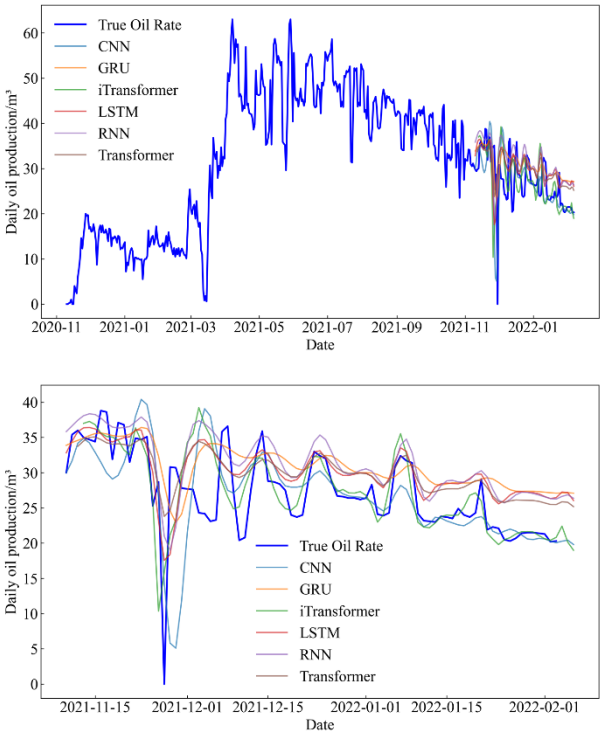

Fig. 11 Comparison of the effect of each model on the prediction of daily oil production of oil wells

In summary, iTransformer demonstrates the strongest predictive capability, accurately capturing overall trends while responding promptly and stably to fluctuations. The Transformer model also performs well, with good trend synchronization and moderate responsiveness to volatility. GRU and CNN exhibit stable performance but limited responsiveness to extreme variations. RNN and LSTM, constrained by their architectures, suffer from noticeable delays and amplitude attenuation when handling non-stationary changes. Therefore, considering both visual and error-based evaluations, the iTransformer model, based on a multi-scale modeling strategy, shows significant advantages in real-world production forecasting and is the preferred structure for this task.

Table 1 Comparison of prediction errors of daily oil production from wells by models

Model	MAE	MSE	RMSE	R2
GRU	4.4206	30.0617	5.4828	0.00093
CNN	2.9358	25.3447	5.0343	0.15769
RNN	4.1398	23.0519	4.8012	0.23389
Transformer	3.2903	16.7269	4.0898	0.43835
LSTM	3.1792	14.8972	3.8596	0.50490
iTransformer	0.7452	1.42005	1.1916	0.95131

(1) Experiment 2

Without introducing any external auxiliary information or data augmentation techniques, the modeling performance of iTransformer on the original oil production dataset (D1) exhibits certain fluctuations. As shown in Table 2, the model's performance is highly sensitive to the input window length. When the window length is 5, the model achieves the highest R^2 value (0.5259) and the lowest MAE (3.6111). However, as the window length increases to 20, performance deteriorates significantly. This suggests that, in the absence of auxiliary information, excessively long historical input sequences may introduce irrelevant or redundant features, thereby impairing model effectiveness.

Table 2 Prediction performance of D2 dataset with different window lengths (prediction step of 1)

	R2	MSE	RMSE	MAE	MAPE
5	0.5259	30.2405	5.4991	3.6111	163.4801
10	0.4638	34.1538	5.8441	3.9054	202.4594
20	0.4453	35.3362	5.9444	4.0455	189.6078
40	0.4875	32.5928	5.7090	3.9053	183.9960
60	0.5028	30.7958	5.5494	3.8809	165.0965

With the input window length fixed at 40, the prediction accuracy of the model declines consistently as the forecasting horizon increases from 1 to 10 steps. As shown in Table 3, the R^2 value decreases from 0.4875 to 0.1441, while MAPE rises markedly from 184% to 434.65%. This indicates that the noise components and non-stationarity in the raw data have a significant impact on long-horizon forecasting. Relying solely on the original

time series for modeling is insufficient to support high-quality long-term prediction tasks.

Table 3 Prediction performance of D3 dataset with different prediction steps (window length of 40)

	R^2	MSE	RMSE	MAE	MAPE
1	0.49	32.5928	5.70901	3.905309	183.9961
2	0.36	40.4218	6.35782	4.395757	274.1812
5	0.23	48.8472	6.98908	4.971432	357.0521
10	0.49	32.5928	5.70901	3.905309	183.9961

Table 4 Prediction performance of D2 dataset with different window lengths (prediction step of 1)

	R^2	MSE	RMSE	MAE	MAPE
5	0.4403	35.7003	5.9749	3.7858	213.5361
10	0.4982	31.9649	5.6537	3.8448	187.5501
20	0.5371	29.4858	5.4300	3.7201	202.4966
40	0.4986	31.8834	5.6465	3.8864	215.4361
60	0.5370	28.6757	5.3549	3.7404	211.3277

After incorporating static auxiliary parameters (e.g., completion type, permeability), the overall performance of the model on the D2 dataset shows improvement. In particular, when the window lengths are set to 20 and 60, the model achieves R^2 values of 0.5372 and 0.5371, respectively—both significantly higher than the corresponding results on D1. This suggests that static information can effectively enhance the model's generalization ability and predictive accuracy.

Table 5 Prediction performance of D3 dataset with different prediction steps (window length of 40)

	R^2	MSE	RMSE	MAE	MAPE
1	0.4986	31.8834	5.6465	3.8864	215.4361
2	0.3863	39.0379	6.2480	4.4473	262.6912
5	0.2407	48.3113	6.9506	4.9990	396.3425
10	0.1452	53.2459	7.2969	5.3650	423.2382

Under increasing prediction horizons, the D2 dataset exhibits greater robustness compared to D1. Although the overall performance still shows a declining trend, the decreases in R^2 and MAE are relatively moderate. Notably, at a prediction step of 2, the R^2 remains at 0.3863—slightly higher than the 0.3646 observed for D1—indicating that the inclusion of auxiliary parameters helps mitigate information loss in long-term forecasting to some extent.

Table 6 Prediction performance of D3 dataset with different window lengths (prediction step of 1)

	R^2	MSE	RMSE	MAE	MAPE
5	0.9485	2.8259	1.6810	1.1108	56.6603
10	0.9486	2.8162	1.6781	1.1468	67.5743
20	0.9416	3.1973	1.7881	1.2458	80.5241
40	0.9483	2.8283	1.6817	1.1716	65.1958
60	0.9436	3.0285	1.7402	1.2454	96.0169

Although D3 performs exceptionally well at a prediction step of 1, its performance deteriorates significantly as the prediction horizon increases. The R^2 drops from 0.948 to 0.0564, and MAPE rises to over 500%. This suggests that EMD is more suitable for extracting short-term patterns, while its capability in modeling long-term trends remains limited.

Table 7 Prediction performance of D3 dataset with different prediction steps (window length of 40)

	R^2	MSE	RMSE	MAE	MAPE
1	0.9482	2.8282	1.6817	1.17160	65.19581
2	0.6634	18.4133	4.2910	3.00754	195.2987
5	0.2372	41.7608	6.4622	4.68866	464.7270
10	0.0563	50.9263	7.1362	5.2945	500.5442

The R^2 remains consistently between 0.93 and 0.94, with both MAE and MAPE reaching their lowest levels. The model achieves its best performance at a window length of 5, with R^2 = 0.9432, MAE = 1.169, and MAPE = 58.65%. This indicates that, following high-quality preprocessing, a short input window is sufficient to capture the primary dynamic characteristics of the well.

Table 8 Prediction performance of D4 dataset with different window lengths (prediction step of 1)

	R^2	MSE	RMSE	MAE	MAPE
5	0.94324	3.11479	1.76487	1.168633	58.65327
10	0.94254	3.14921	1.77460	1.196673	61.85385
20	0.94269	3.13866	1.77162	1.201258	61.39678
40	0.94150	3.19958	1.78873	1.232639	72.27941
60	0.93393	3.54720	1.88340	1.283815	71.41304

At a prediction step of 2, D4 maintains an R^2 of 0.6738, significantly higher than that of the other datasets. This indicates that the enhanced feature representations from EMD, combined with auxiliary parameters, offer complementary advantages in modeling long-term trends.

Table 9 Prediction performance of D4 dataset with different prediction steps (window length of 40)

	R^2	MSE	RMSE	MAE	MAPE
1	0.94150	3.19958	1.7887	1.2326	72.27941
2	0.67380	17.8480	4.2246	2.9983	213.2648
5	0.18759	44.4801	6.6693	4.8199	460.0655
10	0.06176	50.6343	7.1157	5.2306	515.1403

The results demonstrate that applying EMD alone (D3) can significantly improve prediction accuracy. When further combined with auxiliary parameters (D4), the model not only maintains high accuracy but also achieves greater stability in long-term forecasting.

As the prediction horizon increases, the performance of all datasets declines; however, D4 consistently maintains relatively higher accuracy, confirming its robustness under complex input conditions.

In summary, both window length and prediction horizon have a significant impact on model performance. Short input windows yield better results after denoising; EMD processing proves to be a key technique for enhancing short-term forecasting accuracy. The inclusion of auxiliary parameters effectively improves the model's generalization ability and stability in long-term prediction. Among all datasets, D4 (EMD + auxiliary parameters) achieves the best performance across all evaluation metrics, providing a strong basis for selecting input data configurations in future modeling and deployment.

Firstly, we examine the effect of historical window length on short-term forecasting performance (prediction step P = 1). Experimental results show that the model achieves optimal accuracy when the window length is set to 5-time steps. As the window increases to 10 and 20 steps, performance remains relatively stable. However, further extending the window to 40 and 60 steps leads to increased prediction error. This suggests that, after effective denoising through EMD, the temporal autocorrelation range is significantly compressed, allowing the model to capture key dynamic features with a shorter history. In contrast, overly long input windows introduce redundant information and dilute sample effectiveness.

Next, with the window length fixed at 40 steps, we evaluate the model's generalization ability under varying prediction horizons. The results indicate that the model maintains high accuracy when the prediction step increases from 1 to 2, outperforming the baseline datasets (without denoising or auxiliary parameters) by over 30 percentage points. However, when the prediction step is further extended to 5 and 10, model performance deteriorates sharply.

In conclusion, short input windows (5 to 20 steps) combined with EMD-based denoising can significantly enhance short-term forecasting accuracy while maintaining low computational cost. Meanwhile, conventional models exhibit an effective forecasting horizon of approximately two steps; beyond this limit, a single end-to-end deep network struggles to capture long-term evolutionary patterns accurately. This highlights the need for multi-scale recursive architectures or hybrid modeling strategies incorporating physical constraints. Finally, the EMD-iTransformer model not only improves prediction accuracy but also delays the onset of performance degradation, effectively extending the length of the "predictable window."

5. CONCLUSIONS

This study addresses the problem of dynamic production forecasting driven by multi-source heterogeneous data and proposes a novel forecasting approach based on the EMD-iTransformer architecture. The method reconstructs the original time series using EMD, and leverages the iTransformer to better capture relationships among heterogeneous parameters. This enhances the model's ability to handle non-stationary, multivariate data and improves overall prediction accuracy.

(1) A systematic framework for the construction and preprocessing of multi-source data was established, encompassing key dimensions such as well control parameters, geological attributes, and production data, with subsequent data cleaning and normalization.

(2) The EMD-based decomposition strategy demonstrated superior performance in prediction accuracy, stability, and trend alignment, significantly outperforming baseline methods.

(3) The temporal modeling capability of iTransformer effectively captures inter-parameter dependencies, making it particularly well-suited for high-dimensional, multivariate coupling scenarios encountered in field applications.

(4) Short input windows (5 to 20 steps), when combined with EMD-based denoising, significantly enhance prediction accuracy while keeping computational costs low.

(5) The integration of EMD with multi-dimensional static features not only improves prediction accuracy but also delays the onset of performance degradation, thereby extending the "predictable window" and

providing more lead time for dynamic reservoir management.

ACKNOWLEDGEMENT

The financial support from the National Natural Science Foundation of China (Nos. U23B2084 and 52174045)

REFERENCE

[1] Cao, Q., Banerjee, R., Gupta, S., & Al, E. (2016). Data Driven Production Forecasting Using Machine Learning. Day 2 Thu, June 02, 2016, D021S006R001.

[2] Dong, Y., Jin, J., Zhang, J., Wu, K., Liu, S., & Bai, J. (2024). Multisource modeling method for petroleum production prediction models: A case study of CO2-flooding performance. SPE Journal, 29(11), 5761–5778.

[3] Fargalla, M. A. M., Yan, W., Deng, J., Wu, T., Kiyingi, W., Li, G., & Zhang, W. (2024). TimeNet: Time2Vec attention-based CNN-BiGRU neural network for predicting production in shale and sandstone gas reservoirs. Energy, 290, 130184.

[4] Huang, T., Qian, H., Huang, Z., Xu, N., Huang, X., Yin, D., & Wang, B. (2024). A time patch dynamic attention transformer for enhanced well production forecasting in complex oilfield operations. Energy, 309, 133186.

[5] Liu, W.-C., Qiao, C.-C., Wang, P., Huang, W.-S., Kong, X.-W., Sun, Y.-P., Sun, H.-D., & Jia, Y.-P. (2024). Study of inter-well interference in shale gas reservoirs by a robust production data analysis method based on deconvolution. Petroleum Science, 21(4), 2502–2519.

[6] Matthew Ockree, Kenneth G. Brown, Joseph Frantz, Michael Deasy, & Ramey John. (2018). Integrating Big Data Analytics Into Development Planning Optimization. Day 4 Wed, October 10, 2018, 2018-Octob, D043S005R001.

[7] Mohd Razak, S., Cornelio, J., Cho, Y., Liu, H.-H., Vaidya, R., & Jafarpour, B. (2022). Transfer Learning with Recurrent Neural Networks for Long-Term Production Forecasting in Unconventional Reservoirs. SPE Journal, 27(04), 2425–2442.

[8] Pan*, Y., Bi, R., Zhou, P., Deng, L., & Lee, J. (2019). An effective physics-based deep learning model for enhancing production surveillance and analysis in unconventional reservoirs. In Unconventional Resources Technology Conference, Denver, Colorado, 22?24 July 2019 (pp. 2579–2601). Unconventional Resources Technology Conference (URTeC); Society of Exploration Geophysicists.

[9] Sagheer, A., & Kotb, M. (2019). Time series forecasting of petroleum production using deep LSTM recurrent networks. Neurocomputing, 323, 203–213.

[10] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). Attention is all you need (No. arXiv:1706.03762). arXiv. https://doi.org/10.48550/arXiv.1706.03762

[11] Wang, H.-C., Zhang, K., Chen, N., Zhou, W.-S., Liu, C., Wang, J.-F., Zhang, L.-M., Yu, Z.-G., Cui, S.-T., & Yang, M.-C. (2024). Better use of experience from other reservoirs for accurate production forecasting by learn-to-learn method. Petroleum Science, 21(1), 716–728.

[12] Wei, X., Huang, W., Liu, L., Wang, J., Cui, Z., & Xue, L. (2024). Low-rank coalbed methane production capacity prediction method based on time-series deep learning. Energy, 311, 133247.

[13] Xu, Z., & Leung, J. Y. (2024). Shale gas production forecasting with well interference based on spatial-temporal graph convolutional network. SPE Journal, 29(10), 5120–5131.

[14] Xue, L., Xu, S., Nie, J., Qin, J., Han, J.-X., Liu, Y.-T., & Liao, Q.-Z. (2024). An efficient data-driven global sensitivity analysis method of shale gas production through convolutional neural network. Petroleum Science, 21(4), 2475–2484.