Petroleum Science 23 (2026) 7249–7265

Contents lists available at ScienceDirect

Petroleum Science

journal homepage: www.keaipublishing.com/en/journals/petroleum-science

Original Paper

An improved multi-task learning framework for coupled prediction of pressure and temperature during well drilling

Xue-Rui Wang [a,b,*], Yuan-Ze Tian [a,b], Bao-Jiang Sun [c], Zhi-Yuan Wang [c], Yong-Hai Gao [c], Hao Li [c]

[a] College of Computer Science and Technology, China, University of Petroleum (East China), Qingdao, 266555, Shandong, China
[b] Shandong Key Laboratory of Intelligent Oil & Gas Industrial Software, China University of Petroleum (East China), Qingdao, 266555, Shandong, China
[c] School of Petroleum Engineering, China University of Petroleum (East China), Qingdao, 266555, Shandong, China

ARTICLE INFO

Article history:
Received 28 March 2025
Received in revised form
21 June 2026
Accepted 22 June 2026
Available online 25 June 2026

Edited by Jia-Jia Fei

Keywords:
Multi-task learning
Wellbore pressure-temperature prediction
Deep learning
TCN-BiLSTM
Customized gate control (CGC)
Coupling prediction

ABSTRACT

In recent years, petroleum exploration has increasingly shifted toward deepwater and ultra-deep formations, with the most critical technical challenges being the narrow safe drilling window. Accurate wellbore pressure prediction is essential to prevent drilling hazards. Current prediction methods face significant limitations: physical models rely on idealized assumptions and empirical formulas with limited applicability, while machine learning approaches predominantly focus on single field prediction, neglecting multi-field coupling effects. To address these limitations, this paper innovatively proposes a coupled prediction method for wellbore pressure and temperature. The main contribution lies in integrating a customized gate control (CGC) multi-task framework with an efficient channel attention (ECA) optimized TCN-BiLSTM architecture. This framework employs ECA-optimized temporal convolutional networks (TCN) as expert modules to dynamically extract static and dynamic features, intelligently emphasizing key high-dimensional time-series channels. Simultaneously, customized gating units achieve fine-grained feature sharing, preserving positive transfer between temperature and pressure tasks while isolating negative transfer. Task-specific bidirectional long short-term memory (BiLSTM) networks are then utilized for the final coupled prediction. Extensive experimental results quantitatively demonstrate the model's superiority. Compared to state-of-the-art single-task and physical models, the proposed method reduces the mean absolute percentage error (MAPE) by up to 94.98% for pressure and 95.28% for temperature, while improving the R^2 metric by 26.84% and 20.38%, respectively. Furthermore, compared to traditional multi-task learning models, it reduces pressure and temperature MAPE by 63.05% and 71.88%, respectively, maintaining exceptional accuracy even during abrupt changes and intense drilling fluctuations. Validated in field applications at an oilfield, this approach provides a novel technical solution for precise prediction and coordinated monitoring of wellbore pressure and temperature during drilling operations.
© 2026 The Authors. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

In recent years, the petroleum exploration has been increasingly confronted with complex geological conditions, particularly in deepwater and ultra-deep formations (Hu et al., 2020). These challenging strata not only impose higher requirements on drilling equipment and technologies (Gao et al., 2018), but also make wellbore pressure prediction and control one of the critical challenges. In deepwater and ultra-deep drilling environments, the narrow pressure safety window poses significant risks of well kicks and lost circulation. To address this challenge, managed pressure drilling (MPD) technology (Zhou et al., 2008) has emerged, preventing accidents through real-time adjustment of wellbore pressure. However, the implementation of MPD technology relies fundamentally on accurate wellbore pressure prediction, which serves as a crucial prerequisite for ensuring drilling safety and efficiency (Zhang et al., 2023).

* Corresponding author.
E-mail address: ruixue0713@126.com (X.-R. Wang).
Peer review under the responsibility of China University of Petroleum (Beijing).

https://doi.org/10.1016/j.petsci.2026.06.038
1995-8226/© 2026 The Authors. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

The complexity of wellbore pressure prediction lies in its coupling with multiple factors, including formation pressure, formation temperature, drilling fluid density, and flow characteristics (Caetano et al., 1992a, 1992b). Traditional wellbore pressure prediction methods depend on physical models based on fundamental theories of fluid mechanics, thermodynamics, and rock mechanics. These models predict wellbore pressure variations by describing fluid dynamic behaviors within the wellbore, combined with temperature, pressure, and drilling fluid properties. Zhang et al. (2022) proposed a novel method for constructing multi-gradient pressure profiles through the injection of hollow glass microspheres, employing Eulerian multiphase flow models to simulate the mixed flow of microspheres and drilling fluid, thereby achieving precise wellbore pressure control. Yang et al. (2024) developed a transient multiphase flow model suitable for oil-based drilling fluids by introducing interphase mass transfer kinetic equations, revealing the influence patterns of acidic gas solubility variations on pressure wave propagation. Yu et al. (2021) established a multiphase flow model considering drilling fluid rheological mutations in low-temperature environments (−50 °C) and permafrost-wellbore thermal coupling, successfully predicting critical pressure thresholds for well blowout accidents in Arctic drilling operations. Sun et al. (2020) addressed hydrate phase transition issues in deepwater drilling by establishing the first four-phase flow model incorporating supercritical, gas, liquid, and solid hydrate phases, significantly improving wellbore pressure prediction accuracy (40% error reduction compared to conventional models).

In recent years, machine learning-driven predictive approaches have gradually emerged as a research frontier. These methods can, by learning from extensive drilling datasets, uncover the underlying patterns of wellbore pressure variations. Representative studies are summarized in Table 1.

Given the complexities of the petroleum domain, especially regarding improvements through the integration of physical information, these methods exhibit significant development potential. Approaches integrating physical information can broadly be categorized into two branches: physics-informed neural networks (PINNs) and deep operator networks (DeepONet). Xu et al. (2023) integrated PINNs with partial differential equations (PDEs) governing two-phase wellbore hydraulics. By dynamically optimizing the loss function with adaptive weights, they addressed the accuracy issues of traditional models in pressure prediction under data-scarce scenarios. Liu et al. (2025) developed a PINN-driven method for solving the wellbore temperature field. This involved embedding the wellbore temperature field model as a loss function into a neural network and employing optimization methods with adaptive weights and adaptive learning rates to enhance training efficiency. They then analyzed the transient variations of wellbore temperature during drilling and gas well testing operations. Yang et al. (2026) established a surrogate model for reservoir numerical simulation based on Fourier neural operators (FNOs) to predict the pressure distribution, CO_2 mole fraction distribution, and oil saturation distribution in 2D heterogeneous CO_2 flooding reservoirs. However, these two types of physics-informed approaches are primarily designed to solve PDEs, and their predictive accuracy for real-world scenarios heavily relies on the precision of the embedded physical equations.

Comprehensive analysis reveals that traditional physical models face practical limitations due to high computational complexity and parameter sensitivity, making them susceptible to empirical errors and inadequate for high-precision prediction requirements. Existing machine learning methods neglect physical constraints and predominantly focus on single field prediction (e.g., pressure or temperature), failing to achieve multi-task collaborative modeling. Consequently, they cannot capture coupling relationships between different physical quantities, leading to weak adaptability in complex drilling scenarios. These limitations indicate that neither conventional methods nor modern machine learning techniques have fully resolved the technical bottlenecks in wellbore pressure prediction. Therefore, there is an urgent need for a novel prediction framework that integrates physical principles while fully exploiting data characteristics to achieve precise coupled prediction of wellbore pressure and temperature.

To address these challenges, this paper proposes a CGC multi-task learning to achieve coupled prediction method for wellbore pressure and temperature, aiming to resolve the coupled prediction problem of wellbore pressure and temperature. The core concept involves coupling temperature and pressure prediction tasks through a multi-task learning framework, thereby comprehensively considering their mutual influences. Compared with single-task prediction methods, multi-task learning can enhance prediction accuracy by leveraging inter-task correlations through shared feature extraction, while reducing feature extraction and training time compared to separate temperature and pressure predictions. This study employs efficient channel attention-optimized TCN networks as expert systems to extract spatiotemporal representations of dynamic and static features during drilling processes, combined with gating units to capture coupled temperature-pressure characteristics. Based on the extracted fused temperature-pressure features, task-specific layers utilize BiLSTM to construct sub-task models for coupled prediction. This design enables comprehensive characterization of both dynamic pressure variations within the wellbore and static formation parameters during drilling operations. The CGC framework enhances inter-task coupling modeling capabilities while preserving task independence through the deployment of task-specific experts, shared experts, and gating units for extracting distinct task features. Regarding loss function design, this paper simultaneously optimizes temperature and pressure prediction accuracy through weighted loss functions that balance their respective

Table 1
Summary of machine learning studies in petroleum engineering.

References	Main method	Research content and contribution
Zhou (2023)	FNN CNN	Developed deep learning models based on multiphase flow data to predict wellbore pressure
Zhu et al. (2022)	BP-LSTM CNN-LSTM	Considered temporal factors to predict bottomhole pressure
Akinsete and Adesiji (2019)	ANN	Directly predicted bottomhole pressure using wellhead and bottomhole data
Al-Fakih et al. (2025a, 2025b)	TSGAN SeqGAN	Generated synthetic logging data and completed missing data
Al-Fakih et al. (2025a, 2025b)	Pix2Geomodel (cGAN)	Developed Pix2Geomodel framework, to predict reservoir properties of the Groningen gas field's Rotliegend reservoir using a dataset from Aardolie Maatschappij
Al-Fakih et al. (2025a, 2025b)	Bayesian optimization ML models	Optimized geothermal temperature prediction accuracy and performance in western Yemen
Koeshidayatullah et al. (2024)	TimeGPT	Applied the TimeGPT architecture to analyze well logging data, predicting various key well logging responses from different locations and detecting anomalies in the logs with high precision

Note: FNN: Feedforward neural network; CNN: Convolutional neural network; ANN: Artificial neural network; BP: Back propagation; LSTM: Long short-term memory; TSGAN: Time-series generative adversarial network; SeqGAN: Sequence generative adversarial network; cGAN: Conditional generative adversarial network; ML: Machine learning.

contributions, ensuring the stability and efficiency of the multi-task learning framework. The framework's operational reliability has been systematically verified across diverse testing environments, including simulated data experiments and actual oilfield tests. Compared with traditional physical models, existing single-task machine learning methods, and current multi-task machine learning approaches, this method demonstrates significant improvements in both pressure and temperature prediction accuracy. Further experiments confirm that the method effectively captures coupling relationships between wellbore pressure and temperature, thereby substantially enhancing the model's physical interpretability.

2. Design of improved multi-task learning framework for coupled prediction of pressure and temperature

Single-task learning methods often fail to effectively handle the shared information between multiple related tasks, leading to poor performance in complex multi-task scenarios. Moreover, traditional multi-task learning methods (Caruana, 1997) often adopt simple sharing mechanisms that overlook the differences and coupling relationships between tasks, resulting in information conflicts or negative transfer, which can affect the overall performance of the model. To overcome these issues, this paper proposes a more flexible and efficient multi-task learning model as show in Fig. 1. The proposed CGC multi-task learning model consists of two parts: the multi-task learning layer and the specific task layers.

In the multi-task learning layer, to enhance model performance and capture multi-dimensional features, this paper adopts an efficient channel attention (ECA) optimized temporal convolutional network (TCN) as experts to effectively extract the complex spatiotemporal features during the drilling process. The experts are composed of three sub-networks: shared experts, temperature experts, and pressure experts. In the multi-task learning layer, each expert is responsible for extracting the spatiotemporal representations of dynamic and static features during the drilling process, while capturing the coupled features between temperature and pressure. To better integrate the features extracted by each expert network, the paper introduces gating units, which can flexibly adjust the contribution of each expert network based on the requirements of different tasks, thus achieving more effective information integration and task coordination. The shared expert focuses on extracting common features shared across different tasks, while the temperature and pressure experts focus on extracting task-specific features related to temperature and pressure.

In the specific task layer, for the temperature and pressure prediction tasks, bidirectional long short-term memory networks (BiLSTM) are used to predict temperature and pressure from the fused features. The BiLSTM model is capable of capturing both forward and backward temporal information of the input features, thereby improving the model's ability to handle sequential data. Through this layer, the predictions of temperature and pressure can better reflect their relationships with dynamic and static features.

Through this multi-task learning structure, the proposed model not only accurately predicts temperature and pressure but also effectively captures their coupled relationships, this enhancement elevates forecasting precision while strengthening cross-scenario adaptability of the model.

2.1. Overall network architecture based on CGC multi-task learning

Multi-task learning (MTL) is a machine learning method that contrasts with single-task learning. In traditional single-task learning, complex learning problems are often decomposed into several independent subproblems, each learned separately, and by combining the results of these subproblems, a complete

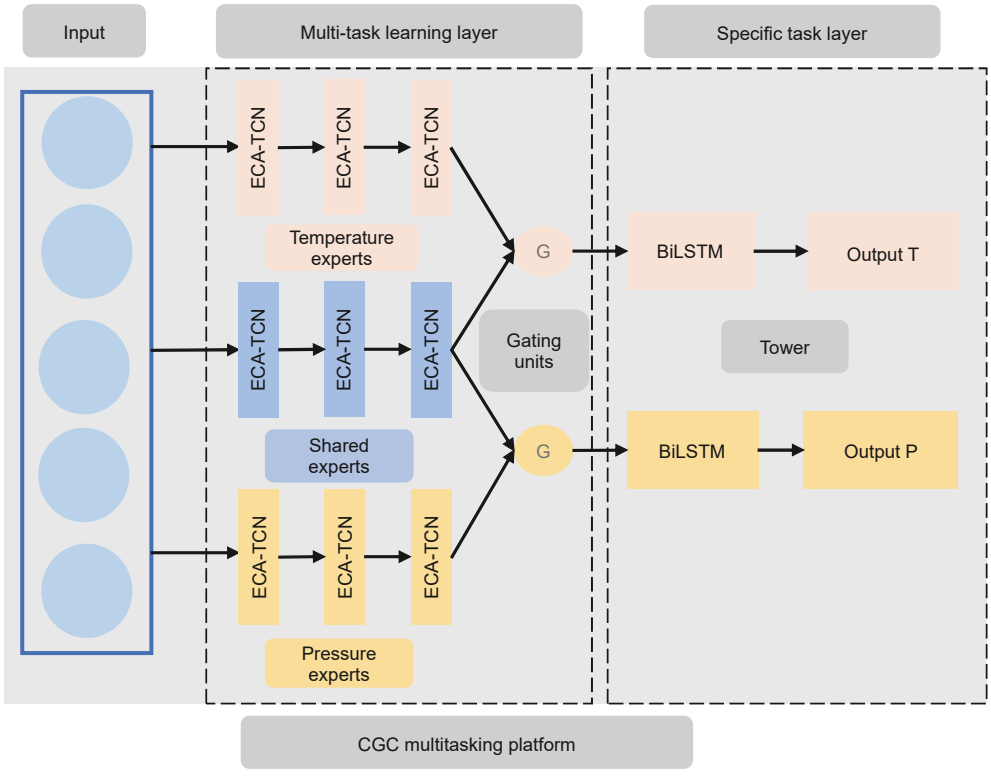


Fig. 1. Improved multi-task learning framework for coupled prediction of pressure and temperature.

mathematical model is constructed. This integrated learning paradigm enables parallel optimization of correlated tasks by leveraging task-shared feature spaces, where inter-task dependencies drive collaborative parameter refinement. The key advantage of multi-task learning lies in its shared layer design, which allows the model to learn and leverage auxiliary information provided by other related sub-tasks, thereby elevating forecasting precision while strengthening cross-scenario adaptability of the model.

In neural networks, multi-task learning commonly employs two parameter-sharing methods: hard sharing and soft sharing. Hard parameter sharing (Caruana, 1993), the most prevalent approach in neural network-based multi-task learning, involves sharing hidden layer parameters across all tasks. This mechanism effectively reduces overfitting risks and enhances the model's generalization capability. In contrast, soft parameter sharing assigns independent models and parameters to each task. While each task retains its own network and parameters, soft sharing ensures inter-task similarity by regularizing the distance between model parameters. Inspired by traditional multi-task regularization techniques, this method balances task independence and shared information utilization.

The CGC framework adopted in this paper is precisely based on a soft sharing mechanism. CGC is a gate-based model architecture typically applied in multi-task learning to achieve effective information sharing among different tasks and comprehensive extraction of task-specific features. The CGC network innovatively divides the expert structure into shared experts and task-specific experts, and equips each task with a customized gating network, enabling fine-grained management of feature sharing in multi-task learning. This design allows the model to dynamically adjust the combination ratio of shared and specific features based on the characteristics of the input sample. It retains valuable positive transfer effects between tasks while effectively isolating negative transfer interference from irrelevant features. This adaptively optimizes feature representation under varying task relevancies, achieving a 'share-on-demand, precise-sharing' effect, thereby significantly enhancing the overall performance and generalization capability of multi-task learning.

The network architecture is illustrated in Fig. 2. CGC establishes a three-layer feature sharing control system through its expert separation architecture and customized gating, enabling precise regulation of the sharing process: (1) Multi-task learning layer: This layer comprises expert networks and gating units. The expert networks decompose feature representations into two orthogonal spaces: a shared feature space and a task-specific feature space. Each expert module consists of multiple sub-networks (i.e., experts). The number of these sub-networks is an adjustable hyperparameter, with each sub-network specializing in extracting distinct feature representations. Specifically, the shared expert is responsible for learning common patterns across tasks, while the task-specific expert specializes in learning unique patterns pertinent to each individual task. Through physical isolation, CGC transforms the question of 'whether to share' into 'how much to share', fundamentally preventing the forced sharing of irrelevant features. The core of CGC is the task-customized gating unit, which dynamically determines how each task combines shared and specific features. The gating unit is responsible for dynamically regulating the feature sharing process among tasks. Through the gating mechanism, the model can allocate appropriate information between shared and task-specific experts, thereby achieving a balance across different tasks. Notably, gating operates at the

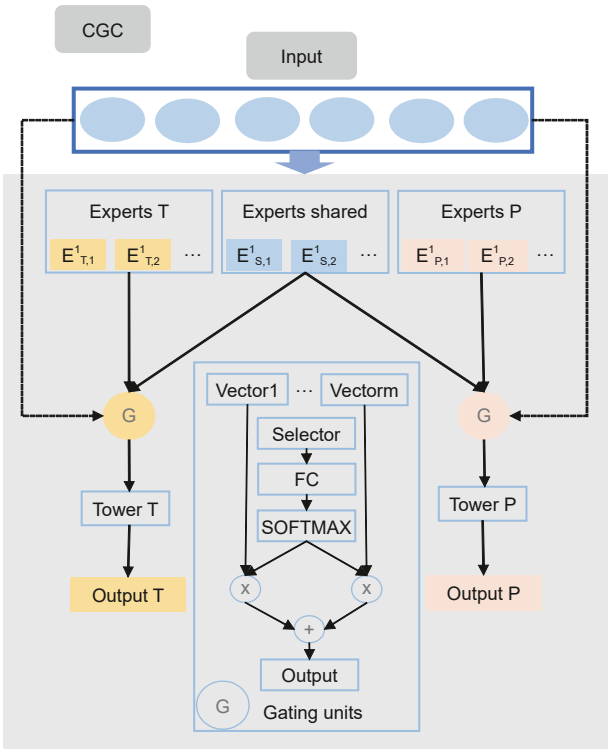


Fig. 2. CGC network architecture diagram.

expert granularity, where each expert represents a set of related features, rather than at the feature granularity. The gating unit employs a single-layer feedforward network structure with a softmax activation function. This gating mechanism allows the model to flexibly adjust the information flow and sharing patterns between tasks during the training process of each subtask, thereby preventing task conflicts and negative transfer. (2) Specific task layer: This layer consists of multiple task-specific towers, with each tower processing a distinct subtask. While the tower networks themselves do not directly manage feature sharing, they further optimize feature representations through task-specific non-linear transformations. Each tower learns different information from both shared experts and task-specific experts, and then performs task-specific predictions independently. The structure of these towers, including their width and depth, are adjustable hyperparameters.

Specifically, for k sub-tasks, the output of the gating unit and its internal calculations are given by (Tang et al., 2020):

$$g^k(x) = \omega^k(x)\mathbf{S}^k(x) \tag{1}$$

where x is the input vector representation; $\omega^k(x)$ represents the weights computed by the Softmax activation function, with the expression:

$$\omega^k(x) = \mathrm{Softmax}\left(\mathbf{W}_g^k x\right), \mathbf{W}_g^k \in R^{(mk+ms)*d} \tag{2}$$

Here, $\mathbf{W}_g^k$ is the weight matrix of the gating unit, mk and ms denote the number of task-specific experts and shared experts for task k respectively, d is the dimension of input representation. $\mathbf{S}^k(x)$ is a selection matrix representing the concatenation of the shared

experts and the task-specific experts for the k sub-task, $\mathbf{S}^k(x)$ explained as follows:

$$\mathbf{S}^k(x) = \left[\mathbf{E}^{\mathrm{T}}(k,1), \mathbf{E}^{\mathrm{T}}(k,2) \ , \ \ldots \ , \ \mathbf{E}^{\mathrm{T}}(k,mk), \mathbf{E}^{\mathrm{T}}(s,1), \right.$$
$$\left. \mathbf{E}^{\mathrm{T}}(s,2) \ , \ \ldots \ , \ \mathbf{E}^{\mathrm{T}}(s,ms)\right]^{\mathrm{T}} \tag{3}$$

where $\mathbf{E}^{\mathrm{T}}(k,i)$ denotes the transpose of the i expert for the k sub-task, and $\mathbf{E}^{\mathrm{T}}(s,i)$ denotes the transpose of the shared expert. Finally, the estimated output for the k sub-task, $y^k(x)$, is given by the upper tower t^k:

$$y^k(x) = t^k\left(g^k(x)\right) \tag{4}$$

where t^k represents the upper tower for the k sub-task, responsible for mapping the features processed by the gating unit to the final prediction.

2.2. Multi-task learning layer design

This study proposes an innovative approach by introducing the ECA mechanism into TCNs and further improving the traditional TCN residual connection structure. Traditional TCNs excel at capturing long-range temporal dependencies through their causal convolutions, dilated convolutions, and residual connections. However, their inherent design lacks adaptive weighting capabilities when processing feature channels, treating all channels indiscriminately. This limitation restricts fine-grained control over critical information flow and, consequently, impairs the in-depth extraction and representation of high-dimensional temporal features within the effective receptive field. Our work aims to address this fundamental limitation by strategically embedding the ECA module within the information flow of each TCN residual block, specifically after the dilated convolution operations. This pivotal modification allows the network to dynamically recalibrate channel-wise feature responses, ensuring that features propagated through the residual path are not only rich in temporal information but also optimally weighted in the channel dimension. This unique integration creates a powerful synergistic effect, enabling the TCN to better focus on critical local inter-channel interaction information arising from the dilated causal convolutions, thereby compensating for the limitations of TCN regarding their receptive field and significantly enhancing the representation of crucial high-dimensional temporal features. This fusion empowers the TCN to simultaneously capture complex temporal patterns by intelligently emphasizing key high-dimensional temporal feature channels, suppressing irrelevant information, and optimizing the path selection of information flow. This significantly enhances its discriminative capability of learned representations and its adaptability to complex dynamic changes, all achieved with minimal computational overhead, offering a more efficient, robust, and feature-expressive solution for time series analysis. The combined structure is illustrated in Fig. 3.

2.2.1. Temporal convolutional network

TCN is a convolutional neural network architecture that excels in processing ordered data, particularly time-series data. Its structure primarily comprises dilated causal convolutions and residual connections. In this study, TCN is applied to multi-task time series feature extraction. Leveraging its robust temporal modeling capabilities and scalability, TCN significantly enhances the model's ability to extract multi-task features and provides more accurate feature representations for subsequent task predictions.

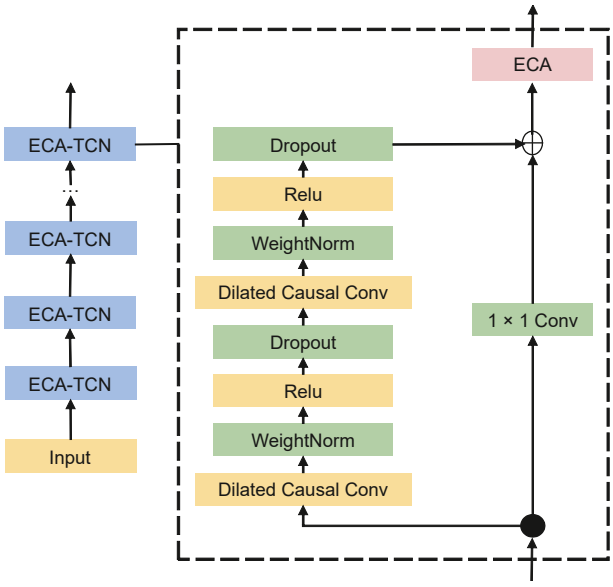


Fig. 3. ECA-TCN residual structure.

(1) Causal convolution and dilated convolution

To meet this causality requirement, TCN introduce causal convolution. This mechanism enforces a causal constraint on the convolution operation, ensuring that the receptive field of the convolutional kernel covers only current and past time steps.

Additionally, TCN employs dilated convolution. By skipping certain input data points during kernel application (via a dilation factor d), dilated convolution expands the kernel's receptive field. This technique effectively captures long-term dependencies or contextual information over extended time scales. The causal and dilated convolution architecture of the TCN is illustrated in Fig. 4.

Specifically, each TCN layer contains L convolutional layers. The computation for dilated convolution in TCN is formalized as (Bai et al., 2018):

$$F(s) = (x \times df)(s) = \sum_{i=0}^{k-1} f(i) \cdot x_{s-d \cdot i} \tag{5}$$

where $d \in [2^0, 2^1, \ldots, 2^{L-1}]$ (exponentially increasing dilation factors); k: Kernel size;

$f(i)$: i element of the convolution kernel; $x_{s-d \cdot i}$: Historical input at step $s - d \cdot i$.

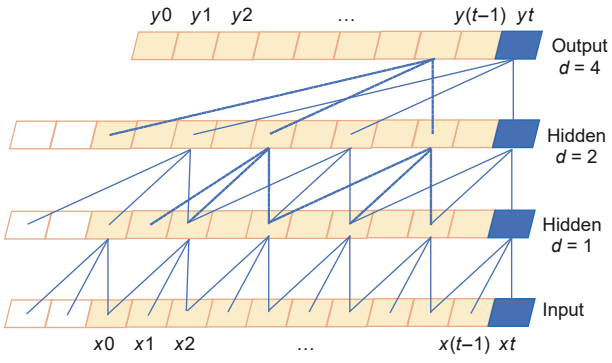


Fig. 4. TCN's causal and dilated convolutions.

(2) **Residual connections**

The core idea of residual connection is to create shortcuts in neural networks, allowing information to bypass certain layers during transmission. The residual connection architecture of the TCN is illustrated in Fig. 5.

Assume that the input to a given layer is x, and after a series of non-linear transformations, the output becomes $F(x)+x$. The residual connection then adds the input x directly to the output, resulting in the final output:

$$y = F(x) + x \tag{6}$$

2.2.2. Efficient channel attention network

ECA (Wang et al., 2020) is a module derived from the squeeze-and-excitation network (SE Net), designed to significantly enhance the performance of convolutional neural networks through an efficient channel attention mechanism. The core advantage of ECA net lies in its adoption of a local cross-channel interaction strategy. This approach achieves substantial performance improvements without dimensionality reduction and only a small increase in the number of parameters, making it particularly effective for tasks that require capturing complex feature relationships.

The ECA net module first applies dimension-preserving global average pooling to generate a $1\times1\times C$ vector. It then dynamically computes adaptive convolution kernel sizes to derive channel-specific weights. By leveraging weight sharing in its 1D convolution operations, ECA net significantly reduces the number of required weights from $C\times C$ to k. In this study, we integrate ECA net into the hidden layers of the multi-task learning TCN architecture. This enhancement modifies the baseline TCN residual connection structure to focus on critical local cross-channel interaction patterns during the two-stage dilated causal convolutions, addressing the limited receptive field of standard TCNs while strengthening the representation of high-dimensional temporal features. The ECA net architecture is illustrated in Fig. 6.

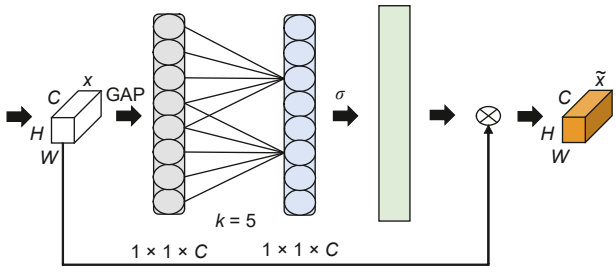


Fig. 6. ECA net structure.

2.3. Task-specific layer design

For the task-specific layers of the CGC framework, this paper selects the BiLSTM. The BiLSTM is an extension of the standard LSTM. Specifically, BiLSTM incorporates two LSTM layers: one processing the forward sequence and the other processing the reverse sequence. The structure of BiLSTM is shown in Fig. 7.

With BiLSTM, this paper achieves more accurate predictions of wellbore pressure and temperature changes, especially in the modeling of complex temporal data, thereby demonstrating the advantages of BiLSTM in multi-task learning.

Long short-term memory (LSTM) is a variant of recurrent neural networks (RNNs) designed specifically for sequence data, addressing the long-term dependency issues commonly encountered in time series. A key innovation of LSTM (Hochreiter and Schmidhuber, 1997) is its use of three gate structures—input gate, forget gate, and output gate—which regulate the flow of information and the updating of the cell state. These gating mechanisms allow the LSTM to determine, at each time step, which information to let in, retain, or output, thereby effectively capturing important long-term dependencies while filtering out irrelevant data. The LSTM structure is depicted in Fig. 8.

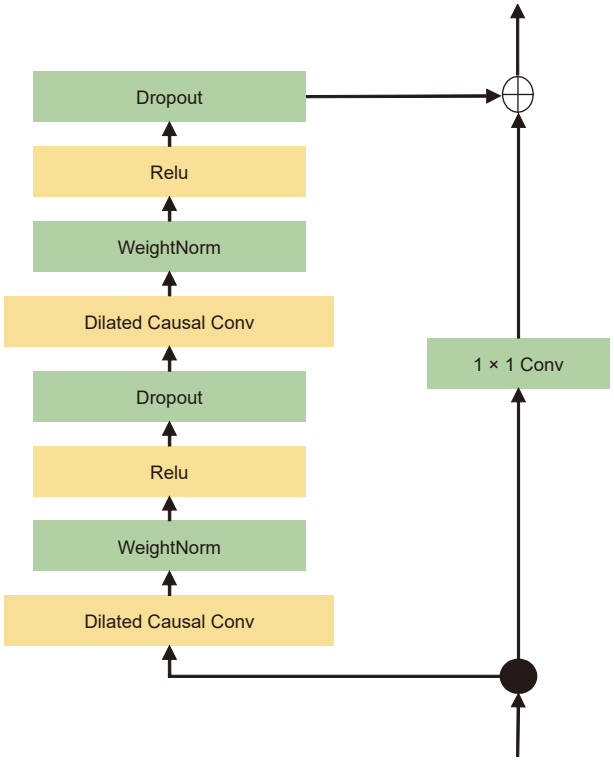


Fig. 5. TCN residual connection.

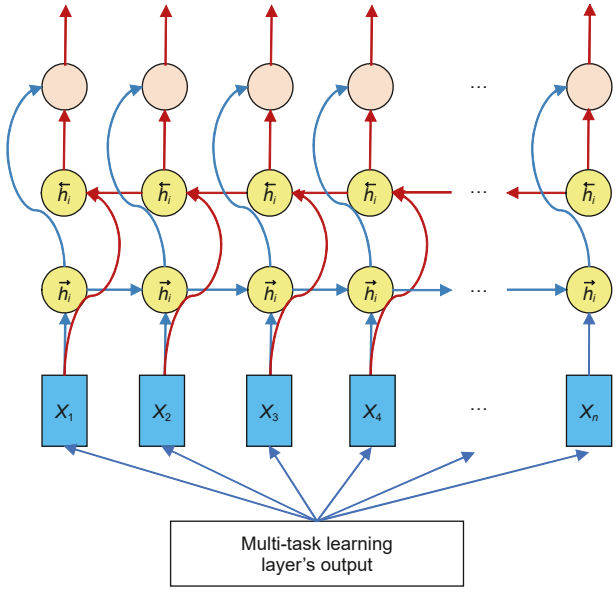


Fig. 7. BiLSTM network structure.

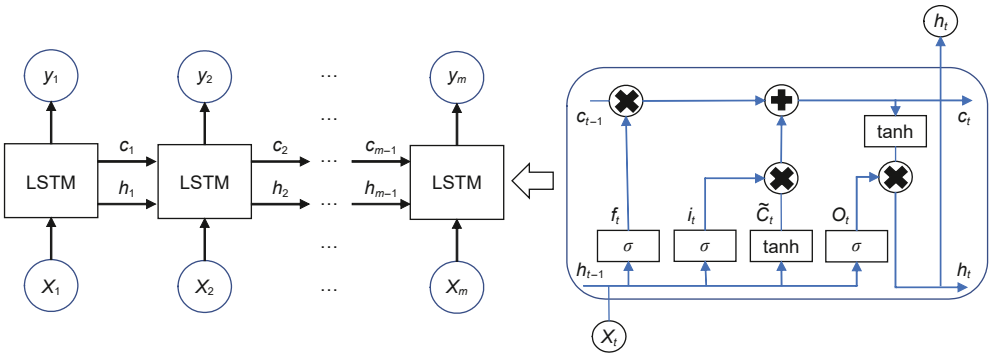


Fig. 8. LSTM structure.

3. Experiment design and implementation

3.1. Data preprocessing

The dataset used in this study originates from a deep collaboration between the research group and the China National Petroleum Corporation (CNPC) in the Xinjiang Oilfield and Tianjin Dagang Oilfield. It includes design reports and historical well logging data from multiple drilling sites. The dataset contains real drilling process records from 10 wells, with a total sample size of approximately 2 million data points. The selected data covers key technical indicators such as geological structure and drilling parameters, with rich temporal characteristics and multi-dimensional attributes. The data types include key parameters such as well depth, drilling pressure, rate of penetration, bit rotation speed, pump pressure, and bottom-hole pressure. The time span covers the entire drilling process, and the data is collected with a high frequency and fine temporal resolution, enabling the observation of dynamic changes during drilling.

The wells selected from the Xinjiang Oilfield are located in the Junggar Basin, one of China's most important petroleum and natural gas exploration and development regions. The geological structure of this area is complex, primarily consisting of pre-Jurassic basement, Jurassic-Cretaceous sedimentary cover, and Neogene-Quaternary sedimentary layers. Hydrocarbon reservoirs are mainly distributed in the Middle Triassic lower Karamay Formation, Upper Jurassic, and Lower Cretaceous strata. The geological features of the drilling sites include faults, folds, and significant variations in bed dip angles. Lithology is highly variable, the burial depth is considerable, and the region experiences high tectonic stress. These geological conditions directly affect the variations in key parameters such as pressure and temperature during the drilling process, which are crucial for data distribution and model construction.

The Dagang Oilfield is located in the Bohai Bay Basin on the North China Plain coastal region. It is one of China's first large to medium-sized onshore oilfields and is part of the Bohai Bay Basin. The region is dominated by Neogene and Quaternary sedimentary formations, with sedimentary environments primarily consisting of delta and lacustrine facies, resulting in relatively stable sedimentary layers. Hydrocarbons are mainly hosted in the Dongying, Guantao, and Shahejie Formations, with relatively shallow burial depths. The pressure system in the area is characterized by normal pressure and mild overpressure. The geological structure of the Dagang Oilfield is relatively simple, mainly consisting of fault-block structures with a dense fault network of moderate scale. The stress level in the region is relatively low, making it suitable for a variety of drilling and completion practices. Compared to the Junggar Basin, the geological conditions in the Dagang Oilfield are more stable, with more homogeneous lithology, providing a solid basis for validating methods in a low-stress, high-stability environment.

In selecting the well locations, this study prioritized representativeness and diversity in the data acquisition process. The selected wells cover various geological formations, burial depths, lithology combinations, and pressure systems, providing a comprehensive reflection of the main geological and engineering characteristics of different oil and gas blocks in China. Additionally, the wells vary significantly in equipment configuration, operational parameters, and geological conditions, offering a robust foundation for evaluating the method's performance across multiple conditions. Since the method used in this study relies on the temporal variation of drilling parameters and their interrelationships, such features are widely present in other sedimentary basins and oil-gas blocks. Therefore, the selected samples not only verify the applicability of the method in the Xinjiang and Dagang Oilfields but also lay the groundwork for broader regional application.

To ensure data quality, the original data were pre-processed, including removing outliers, filling in missing values, and normalizing the data, ensuring that the model performs well in real-world scenarios.

3.1.1. Data cleaning

Since the data is collected from well-logging sites, there are inevitable issues such as debugging of equipment and sensor failures, leading to a large amount of missing data. This paper uses interpolation to ensure the continuity of the data. For missing data, the average value of the previous and next time points is used as a substitute:

$$d_t' = \begin{cases} \dfrac{d_{t-1} + d_{t+1}}{2}, d_t = 0 \\ d_t, d_t \neq 0 \end{cases} \tag{7}$$

where d_{t-1} is the data from the previous time step, d_{t+1} is the data from the next time step, and d_t' and d_t represent the corrected and original data at time t.

In addition to missing data, the well logging data also contains outliers. Outliers may arise from equipment measurement errors, recording mistakes, or sudden environmental changes. This paper uses a statistical feature-based method combined with rule thresholds to detect and replace outliers. Specifically, the interquartile range (IQR) method is used to set threshold ranges:

$$\text{range} = [Q_1 - 1.5 \times \text{IQR}, Q_3 + 1.5 \times \text{IQR}] \tag{8}$$

where Q_1 and Q_3 are the first and third quartiles of the data, and IQR is the interquartile range. Data outside this range is considered an outlier. For detected outliers, this paper replaces them using interpolation from the previous and next time steps to ensure data continuity and validity.

Through these data cleaning steps, this paper effectively solves the issues of missing and outlier data, providing high-quality input data for subsequent model training and ensuring that the model can perform well in real-world scenarios.

3.1.2. Data normalization and partitioning strategy

The purpose of data normalization is to process data with different scales and units by rescaling them to a common interval or range, thereby reducing the impact of discrepancies in scale, feature, or distribution on the model. Since drilling is a sequential process that changes with depth, parameters such as well depth and drilling pressure do not follow a normal distribution. To eliminate dimensionality and reduce the impact of different magnitudes on model performance, this paper uses the max-min normalization method for well logging and drilling data. The formula is as follows:

$$\tilde{X}_i = \frac{X_i - \min(X_i)}{\max(X_i) - \min(X_i)} \tag{9}$$

where $\tilde{X}_i$ is the normalized result for the i-th variable, X_i is the original value of the i-th variable, min (X_i) is the minimum value of the i-th variable, and $\max(X_i)$ is the maximum value of the i-th variable.

The partitioning of the dataset directly impacts a model's generalization capability and evaluation effectiveness. To achieve a rational data split while fully capturing the temporal dynamics of pressure and temperature, the raw data were segmented into fixed time-step intervals. Specifically, the comprehensive dataset compiles real-world drilling data from all different wells. For dataset partitioning, we adopted a chronological splitting approach commonly applied to time-series data: the aggregated data from the entire oilfield were strictly partitioned in chronological order into a training set (60%), a validation set (20%), and a test set (20%). The training set is utilized for core parameter optimization, the validation set is used for hyperparameter tuning and evaluating generalization performance, and the test set is strictly reserved for the final objective assessment of the model. This sequential splitting strategy guarantees that the model is evaluated on entirely unseen data, thereby ensuring its robust performance in real-world scenarios.

3.2. Correlation analysis

Correlation analysis is a critical metric for quantifying the degree to which different factors vary in unison. Data often contain a significant amount of redundant or irrelevant features, and an excessive number of features can greatly increase model complexity and computation time. Data quality is a key factor influencing model performance.

In deep-water wells, the calculation of wellbore pressure and temperature is affected by various factors, such as wellbore structure, drilling fluid parameters, mechanical parameters, and formation pressure and temperature profiles. To improve model prediction performance, it is necessary to calculate the correlations among feature sequences and conduct rigorous feature selection and correlation analysis.

This paper adopts the distance correlation coefficient. Unlike the traditional Pearson correlation—which only captures linear relationships—the distance correlation can detect nonlinear associations. It does not require assumptions regarding the variables' distribution, offers high computational efficiency, and provides strong interpretability. The value of the distance correlation coefficient ranges from 0 to 1; values close to 1 indicate strong correlation, while values near 0 indicate weak correlation.

The formula for the distance correlation coefficient $R_n^2(X, Y)$ is given as follows (Székely et al., 2007):

$$R_n^2(X,Y) = \begin{cases} \dfrac{V_n^2(X,Y)}{\sqrt{V_n^2(X)V_n^2(Y)}}, V_n^2(X)V_n^2(Y) > 0 \\ 0, V_n^2(X)V_n^2(Y) = 0 \end{cases} \tag{10}$$

Here, $V_n^2(X, Y)$ denotes the distance covariance between X and Y; $V_n^2(X)$ is the distance variance of X; $V_n^2(Y)$ is the distance variance of Y; and $R_n^2(X, Y)$ is the distance correlation coefficient between X and Y.

Based on the factors influencing wellbore pressure and temperature, the dcor heatmap for the selected features is shown in Fig. 9. The target features—bottom-hole pressure (BHP) and bottom-hole temperature (BHT)—must be included as inputs. Static features are divided into drilling fluid-related parameters (e.g., drilling fluid density), wellbore geometric features (e.g., well depth, well diameter), and formation environmental characteristics (e.g., formation temperature). Dynamic parameters include those variable parameters that can be measured at the wellhead, such as outlet temperature, inlet temperature, drilling speed, rotary speed, riser pressure, and displacement.

3.3. Training process

To further ensure the reproducibility of the proposed method, the detailed software and hardware configurations are explicitly provided. All experiments and model training were conducted on a workstation equipped with an NVIDIA GeForce RTX 3090 GPU and 32 GB of RAM, running on the Windows 11 operating system. The deep learning models, including the proposed ideas and all baseline architectures, were implemented using the PyTorch 1.12.0 framework in Python 3.8.

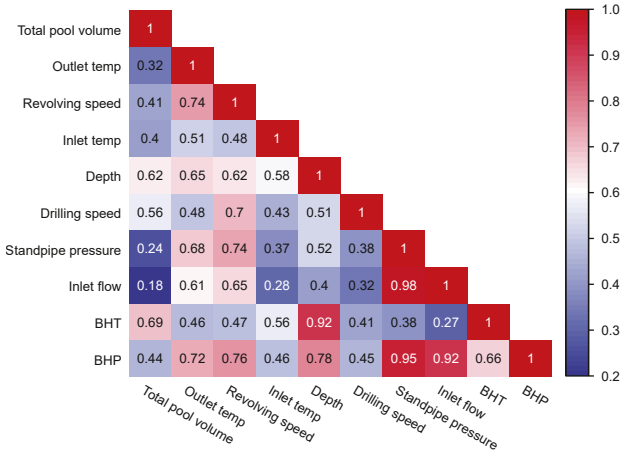


Fig. 9. Dcor correlation analysis heatmap.

Table 2
Hyperparameter values of each deep neural network model.

Predictive models	The value of the hyperparameter
TCN	The model employs a 5-layer TCN residual module with filter parameters of 128, 64, 32, 16, and 8 across successive layers, using a uniform convolution kernel size of 3 and dilation factors of 1, 2, 4, 8, and 16 respectively. Optimized via the Adam algorithm with a learning rate of 0.001, this configuration enables effective capture of both short-term and long-term dependencies in time-series data while enhancing feature extraction capabilities and prediction accuracy.
CNN-LSTM	The model architecture uses a convolution kernel size of 3. For temperature and pressure prediction tasks, the 2-layer LSTM networks are configured with [256, 128], [256, 128], and [128, 128] neurons respectively. The fully connected layers are structured as [128, 32, 1], [128, 64, 1], and [128, 32, 1] neurons. The Adam optimizer is employed with a learning rate of 0.001.
MTL-BiLSTM	The multi-task learning layer employs three fully connected networks, each with a neuron configuration of [256, 64, 1]. The BiLSTM parameters in the task-specific layers align with those of the LSTM in the CNN-LSTM model described earlier.
MMoE-BiLSTM	The multi-task learning layer utilizes three fully connected networks, each with a neuron configuration of [256, 64, 1]. The parameters of the BiLSTM in the task-specific layers align with those of the LSTM in the CNN-LSTM model.
This paper	The ECA-TCN residual module in the multi-task learning layer shares the same architecture as the TCN model, while the BiLSTM parameters in the task-specific layers align with those of the LSTM in the CNN-LSTM model.

Regarding computational cost, leveraging the parallel computing capabilities of the GPU, the training process is highly efficient. The computational efficiency is standardized by its convergence behavior: under the unified hyperparameter settings, the proposed multi-task coupled prediction model typically reaches stable convergence within 200 epochs. The final optimal network architecture and the specific hyperparameter search ranges are detailed comprehensively in the subsequent analysis and Table 2.

The prediction workflow, as illustrated in Fig. 10, comprises four main steps.

The first step involves data preprocessing, including merging dynamic and static data, imputing missing values, normalizing data sequences, and performing preliminary feature selection using the distance correlation coefficient (dcor) method.

The second step focuses on defining hyperparameter search ranges tailored to the prediction tasks. Bayesian optimization is employed to optimize hyperparameters and learning rates for both the multi-task learning layer and task-specific layers, with the Adam algorithm serving as the network optimizer.

The third step divides the dataset into training and testing sets. Multidimensional feature vectors are fed into the model, with temperature and pressure values as outputs. The model undergoes iterative training on the training set, while the test set validates its performance to determine the optimal network depth.

The fourth step evaluates the prediction accuracy of the proposed model using multiple evaluation metrics.

3.3.1. Model hyperparameter settings and optimization

In constructing the predictive model presented in this paper, the meticulous selection of hyperparameters was guided by a comprehensive consideration of the model architecture, data characteristics, computational efficiency, and empirical rules, aiming to balance the model's learning capacity with computational cost, and maximize its generalization performance.

In the multi-task learning layer, the ECA-TCN residual module employs a filter (channel) parameter range of [8, 16, 32, 64, 128]. This increasing channel count design adheres to a common pattern in deep learning, allowing the network to progressively extract increasingly abstract and complex temporal features from low to high levels as it deepens. Simultaneously, the ECA mechanism is utilized to efficiently and adaptively weight channel importance at different levels, thereby constructing a rich hierarchical feature structure. A convolutional kernel size of $k = 3$ is an efficient and parameter-economical choice for capturing local temporal dependencies in 1D convolutions, with multiple stacked layers effectively expanding the receptive field. The dilation coefficients d follows an exponential growth pattern [2^0, 2^1, …, 2^L], which is crucial for TCN's efficient capture of long-range temporal dependencies. This exponentially expands the model's receptive field without loss of resolution, ensuring comprehensive learning of multi-scale temporal patterns, and serving as an effective strategy to compensate for the limited receptive field of traditional sequential models.

In the specific task layer, the BiLSTM network neuron hyperparameter range of [32, 64, 128, 256, 512] is set to endow the model with sufficient capacity to further capture bidirectional global sequence dependencies and long-term memory beyond the local temporal features extracted by the TCN. This range is chosen to strike a balance between model expressiveness and avoiding overfitting. The fully connected layer neuron range of [8, 16, 32, 64, 128, 256] is responsible for mapping the features output by the BiLSTM to the dimensions required for the final task, optimizing the ultimate feature integration and prediction capabilities by adjusting the neuron count.

The choice of optimizer and training parameters is also critical. The Adam optimizer was selected due to its adaptive learning rate mechanism, which adjusts learning rates independently for each parameter, ensuring fast and stable convergence. An initial learning rate of 0.001 is an empirically proven optimal starting point for the Adam optimizer, balancing convergence speed and stability. An initial batch size of 256 aims to balance training efficiency (by utilizing GPU parallel computation) with gradient stability, while also promoting better model generalization. A dropout rate of 0.2 serves as an effective regularization technique, randomly "turning off" a portion of neurons to enhance model robustness and prevent overfitting, thereby forcing the network to learn more discriminative features. Finally, 1000 training iterations provide sufficient training epochs, allowing the model to fully learn and converge to optimal performance.

To find the optimal model configuration, Bayesian optimization is utilized to adjust the model's hyperparameters. Bayesian optimization effectively balances exploration and exploitation by dynamically selecting the best combination of hyperparameters, thereby reducing unnecessary search iterations. By controlling variables, the numbers of layers for the ECA-TCN, BiLSTM, and fully connected layers are set to 5, 2, and 3, respectively. The final hyperparameters for the proposed model and the other models are presented in Table 2.

3.3.2. Loss function design

In multi-task learning, the design of the loss function is a key factor that influences model performance. In this paper, we adopt a weighted loss function approach, where different weights are assigned to different subtasks. This weighted scheme combines the losses of each subtask to achieve an optimization balance among the tasks. Specifically, the overall loss function is defined as follows:

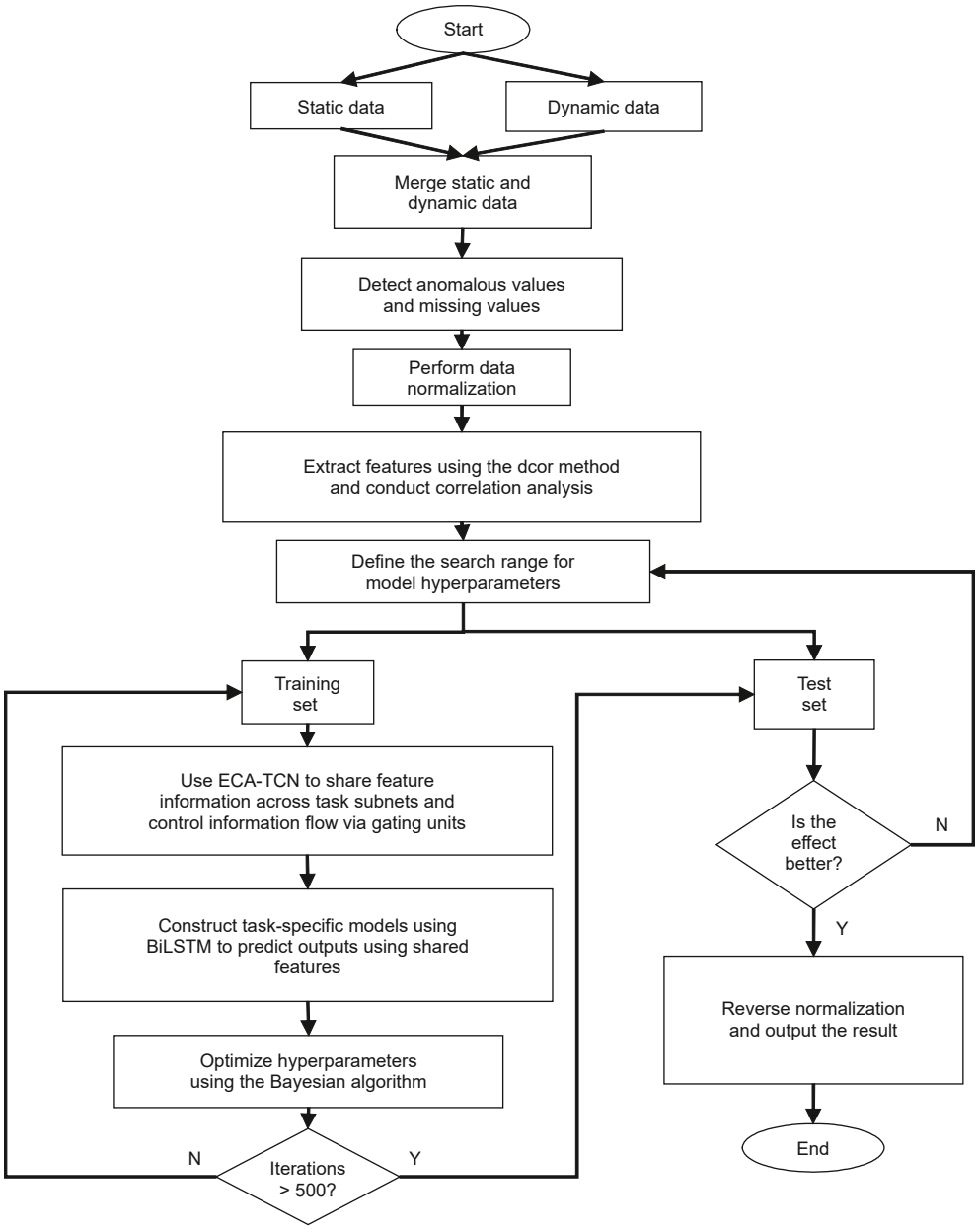


Fig. 10. Multi-task learning prediction process flowchart.

$$L(\theta_1, \ldots, \theta_K, \theta_s) = \sum_{k=1}^{k} \omega_k L_k(\theta_k, \theta_s) \tag{11}$$

Here, θ_s denotes the parameters shared by all subtasks; K is the total number of subtasks; L_k is the loss function for the kth subtask; ω_k is the weight factor for the kth subtask, which measures the importance of that task; and θ_K represents the task-specific parameters for the kth task.

3.3.3. Model evaluation

In this paper, we select the coefficient of determination R^2, mean absolute percentage error (MAPE), and mean absolute error (MAE) as the evaluation metrics for the model. The coefficient of determination R^2 measures the goodness-of-fit of the regression model, representing the percentage of the variance in the dependent variable that can be explained by the independent variables through the regression model. A higher R^2 indicates a better-fitting regression model. The MAE calculates the average absolute gaps between predicted results and true measurements, giving a direct view of how far off the predictions typically are. The MAPE employs absolute values to avoid the cancellation of positive and negative errors. A lower MAPE signifies a better fitting prediction model with higher accuracy. The calculation formulas are as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{12}$$

$$\mathrm{MAE} = \frac{1}{n}|y_i - \hat{y}_i| \tag{13}$$

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad (14)$$

Here, y_i represents the actual value, $\hat{y}_i$ represents the predicted value, $\bar{y}$ is the average of the actual values, and n denotes the number of samples.

3.4. Overall training results

3.4.1. Loss curve

In this study, we conducted rigorous training, validation, and testing of the model. Taking into account the importance of bottom-hole pressure and temperature during the drilling process, we assigned weight factors of 0.7 and 0.3 to pressure and temperature, respectively. The remaining training hyperparameters were configured according to the settings presented in Table 2. To visually assess the model's performance, we plotted loss function curves for both the training and validation datasets as shown in Fig. 11 and analyzed their trends.

The loss values on the training set gradually decreased as the training epochs increased, indicating that the model was effectively learning the data's features and fitting the training set. As the number of iterations grew, the loss values began to stabilize, suggesting that the model did not exhibit overfitting on the training set and that the training process was nearing convergence. The loss values stabilized after 200 training epochs.

The loss curve on the validation set typically displayed a trend of initial decline followed by stabilization. In the early stages, the validation loss decreased as the model was trained, demonstrating the model's ability to generalize well to new data. However, if the loss ceases to decline or even begins to rise after a certain number of epochs, this may indicate the onset of mild overfitting. To mitigate overfitting, we employed an early stopping strategy, halting the training when the validation loss reached its minimum value. Ultimately, the loss values for the training and validation sets at convergence were 0.001630302 and 0.004688, respectively.

3.4.2. Overall evaluation

In terms of the overall dataset results, this study conducted model evaluation and comparative analysis on wells from the Xinjiang Oilfield and the Dagang Oilfield. The corresponding results are presented in Table 3, while the prediction curves of representative case wells are shown in Fig. 12. Overall, the model achieved relatively stable prediction performance across datasets from different oilfields, indicating that the proposed method possesses a certain degree of generality and robustness under multi-regional and multi-operational conditions. It can be observed that, in the pressure and temperature prediction tasks, the R^2 values in the Dagang Oilfield increased by 0.0034 and 0.0082, respectively, compared with those in the Xinjiang Oilfield. This is primarily because the test wells in the Xinjiang Oilfield exceed 8000 m in depth and are characterized by complex geological structures and highly variable formation pressure systems, leading to relatively higher prediction errors. In contrast, the wells in the Dagang Oilfield are relatively shallower and located in more regular geological settings, which results in better predictive accuracy.

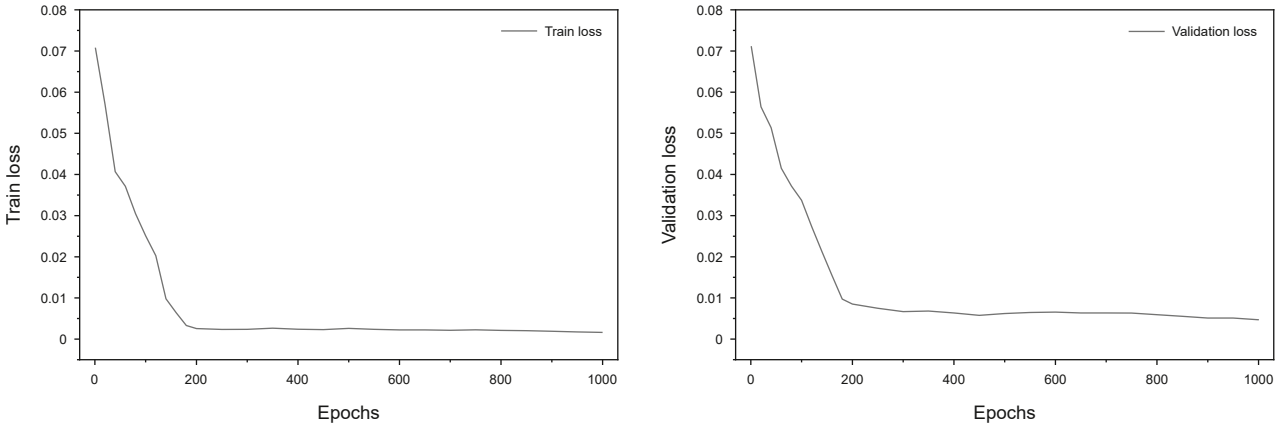
Table 3
Experimental evaluation results of the single-task model and the proposed model.

Evaluation criteria	Predictors	Xinjiang Oilfield	Dagang Oilfield
MAPE, %	Pressure	0.2037	0.1873
	Temperature	0.1791	0.1578
MAE	Pressure	0.0994	0.0754
	Temperature	0.2276	0.1925
R^2	Pressure	0.9878	0.9912
	Temperature	0.9852	0.9934

In summary, the performance differences of the model under various geological backgrounds and operational conditions are consistent with engineering practice. This further verifies that the data-driven approach can maintain good adaptability across diverse well environments, providing valuable insights for the subsequent optimization and broader application of the method.

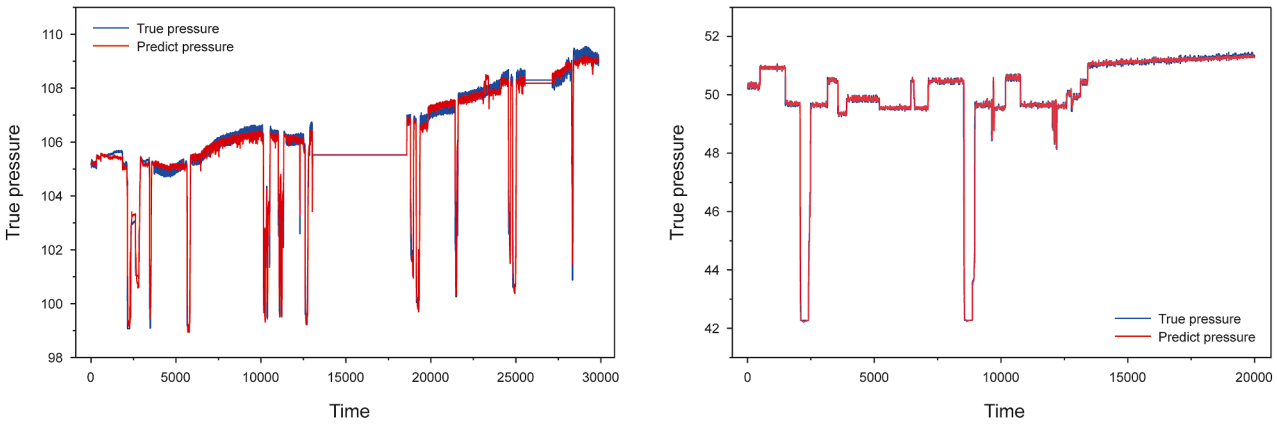
3.4.3. Ablation study

To comprehensively investigate and quantify the individual contributions of the core components within the proposed model to the final prediction performance, an ablation study was conducted under identical dataset splits and hyperparameter configurations. Four distinct model variants were configured for comparison. The performance metrics of these models on the pressure and temperature prediction tasks are presented in Table 4. Specifically, these variants involve: replacing the improved TCN proposed in this work with a standard TCN without the ECA module; substituting the CGC with a shared layer approach; removing the specific experts while retaining only the shared experts; and removing the BiLSTM module. The



(a) Loss curves for the training datasets **(b)** Loss curves for the validation datasets

Fig. 11. Loss curves for the training and validation datasets.



(a) Comparison of bottom-hole pressure prediction results for Well TW-1

(b) Comparison of bottom-hole pressure prediction results for Well BZ-7

Fig. 12. Comparison of representative wells from the two oilfields.

Table 4
Performance comparison of different module configurations in the ablation study.

Evaluation criteria	Predictors	This paper	w/o ECA	Shared-bottom	MMoE	w/o BiLSTM
MAPE, %	Pressure	0.1712	0.3212	0.4636	0.3636	0.2035
	Temperature	0.1423	0.3014	0.5061	0.2787	0.1879
MAE	Pressure	0.0642	0.1025	0.2657	0.1235	0.1020
	Temperature	0.1876	0.2586	0.4841	0.3783	0.1946
R^2	Pressure	0.9923	0.9695	0.9490	0.9521	0.9877
	Temperature	0.9856	0.9811	0.9362	0.9744	0.9839

experimental results indicate that the shared-bottom architecture exhibits the poorest prediction accuracy among all variants. Compared with the proposed full model, its MAPE for the pressure and temperature variables increased by 170.79% and 255.66%, respectively; its MAE increased by 313.86% and 158.05%, respectively; and its R^2 metric decreased by 4.36% and 5.01%, respectively. This demonstrates that the fundamental challenge in designing a multi-task learning architecture lies in balancing feature sharing and isolation across tasks. Forcing both tasks to share all bottom-level features significantly degrades the prediction accuracy. Consequently, it is evident that the gating units and task-specific expert networks are the most critical components influencing the final prediction performance of the proposed model.

4. Case analysis and discussion of results

The coupled pressure-temperature prediction model developed in this study was applied to analyze wellbore pressure and temperature variations during drilling operations in an example well, with its detailed parameters and wellbore architecture summarized in Tables 5 and 6.

Table 5
Basic data sheet of well.

Drilling fluid displacement	10 L/s	Total well depth	5035 m
Geothermal gradient	2.5 °C/100 m	Drilling fluid injection temperature	10 °C
Surface temperature	10 °C	Drilling fluid viscosity	50 s/qt
Well type	Straight wells	Drilling fluid-to-water ratio	82:18
Drilling fluid density	1.55 g/cm^3	Water loss of drilling fluids	1.6 mL/30 min

Table 6
Well depth structure table.

	Casing size, in	Casing outer diameter, mm	Casing inner diameter, mm	Top depth, m	Bottom depth, m
Casing	13-3/8	339.71	320.4	0	1930
	9-5/8	244.47	224.4	0	4271
	7	154.8	177.8	3663	4815.2

The well is a vertical borehole with a total depth of 5035 m. The surface casing is run to 1930 m, the intermediate casing to 4271 m, and the liner extends to 4815.2 m. Oil-based drilling fluid with a density of 1.55 g/cm^3 was utilized during drilling operations. To ensure a rigorous evaluation and prevent any risk of data leakage, the dataset compiles data from all different wells. Rather than random splitting, the data was partitioned sequentially based on the chronological drilling process: the first 60% of the continuous records was used for training, the subsequent 20% for validation, and the final 20% served strictly as the independent test set. By comparing the performance of different models on this identical and independent test set, the advantages of each architecture can be objectively identified, ultimately evaluating the best intelligent prediction model. The hyperparameter settings follow those outlined in Table 2.

4.1. CGC optimized weight results

The CGC network adopts a parameter soft-sharing approach, which allows different tasks to share underlying features while assigning independent feature weights and parameters to each task. The advantage of this mechanism is that it can flexibly adjust the allocation ratio of shared features based on the specific needs of each task. This approach maximizes the collaborative benefits of feature sharing while avoiding negative transfer among tasks. The gating unit essentially serves as a weight vector, indicating the contribution of each expert subnet's output as input to the subtask models, thereby enabling the subtask prediction model to select the most relevant shared features for training.

To analyze the working mechanism of CGC, the gating unit vector from the model was extracted, and a softmax function was applied to obtain the weight distribution for the subtasks, as

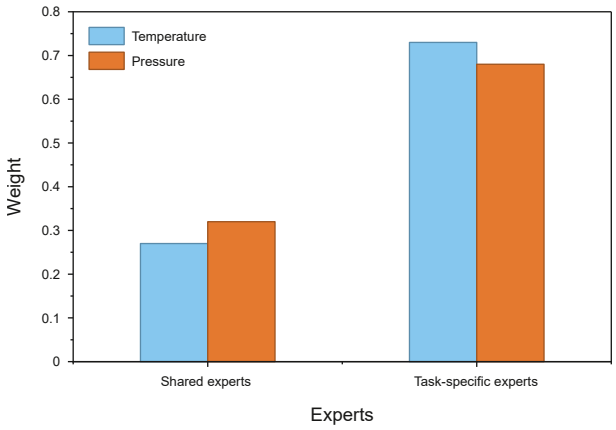


Fig. 13. Expert subnet weight optimization results.

shown in Fig. 13. Experimental results demonstrate that for both pressure and temperature prediction tasks, the CGC is able to capture the coupling relationship between the two while preserving their independent feature information. This, in turn, improves overall prediction accuracy and robustness.

Moreover, the gating mechanism of the CGC network is highly interpretable. By analyzing the output weights of the gating unit, one can directly observe the degree of dependency each task has on different features, which provides valuable insights for feature selection and model optimization. This interpretability not only enhances the model's credibility but also offers guidance for future task expansion and application.

Table 7
Performance comparisons between single-task methods and the new framework presented in this paper.

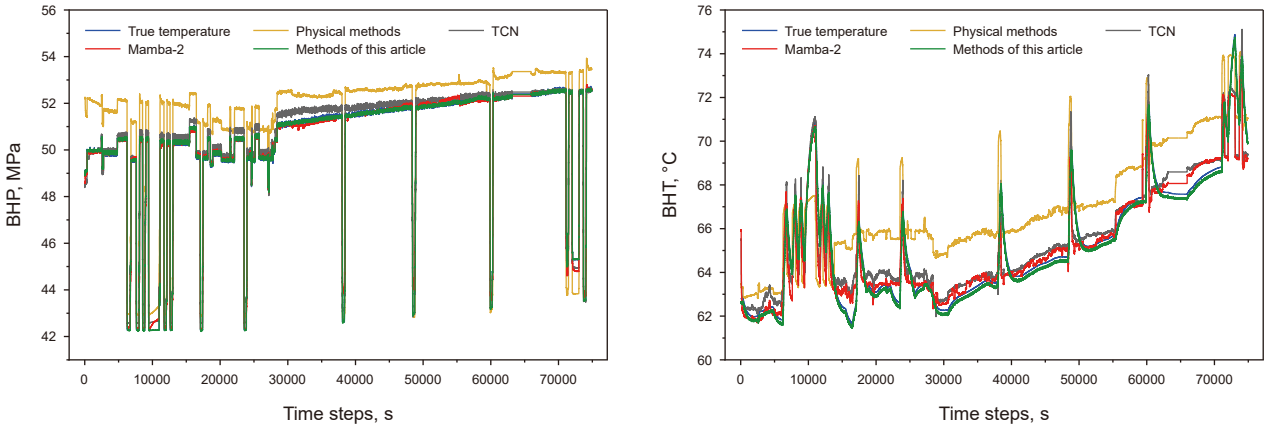
Evaluation criteria	Predictors	Physical methods	TCN	Mamba-2	This paper
MAPE, %	Pressure	3.7947	2.2353	1.0636	0.1923
	Temperature	3.3464	1.5965	1.1142	0.1578
MAE	Pressure	1.2455	0.7258	0.5432	0.0754
	Temperature	3.1246	1.8639	1.1234	0.1925
R^2	Pressure	0.7867	0.8993	0.9548	0.9912
	Temperature	0.8252	0.9123	0.9689	0.9934

4.2. Comparison between the proposed model and single-task prediction models

This study compares the proposed model with a physical model and single-task models (TCN and CNN-BiLSTM (Wang et al., 2016)) for temperature and pressure prediction. The results, summarized in Table 7 and Fig. 14(a) and (b), indicate that the physical method exhibits the largest error but captures general trends accurately. However, its error does not decrease with increased drilling data volume, as traditional physical models rely on polynomial fitting for estimation, sacrificing numerical precision despite adhering to physical laws. TCN and Mamba-2 (Gu and Dao, 2023) improve accuracy in stable phases by incorporating temporal factors but struggle at abrupt changes or during intense fluctuations. The proposed model demonstrates significant improvements: for pressure prediction, MAPE decreases by 94.98%, 91.38%, and 83.47% compared to other models, while temperature prediction MAPE decreases by 95.28%, 90.11%, and 86.71%. The MAE for pressure and temperature is reduced by up to 93.93% and 93.83%, respectively, with R^2 improvements of 26.84% (pressure) and 20.38% (temperature) over single-task models. These enhancements arise because single-task models ignore the coupling relationship between temperature and pressure, limiting performance even with advanced architectures. By integrating multi-task learning, the proposed model leverages shared feature information to enhance accuracy in both steady and dynamic drilling conditions.

4.3. Comparison between the proposed model and other multi-task prediction models

To verify that the proposed model can account for the different influences of various factors on temperature and pressure—thus reducing prediction error—this paper compares it with other multi-task learning models. The comparison involves the proposed model, a traditional MTL method, and a multi-gate mixture-of-experts network (MMoE). In the traditional multi-task learning method, the CGC multi-task learning layer is replaced with a parameter hard-sharing layer; the expert network is replaced by a fully connected network, and the gating unit network is removed. Then, a BiLSTM layer is used for load forecasting, and finally, a fully connected layer outputs the prediction results (Wu et al., 2022). The input features remain the same as those used in the proposed



(a) Comparison of pressure prediction effects

(b) Comparison of temperature prediction effects

Fig. 14. Prediction accuracy trends comparing standard temperature/pressure single-task models against the approach in this paper.

method. The customized gating network is essentially an upgrade from the multi-gate mixture-of-experts network. In MMoE, all parameters are shared among all tasks, with no explicit private parameters for individual tasks. When the relationships between tasks are weak, this may lead to a seesaw phenomenon where one task's improvement comes at the expense of another's performance (Wang et al., 2022).

The comparative results are presented in Table 8 and Fig. 15(a) and (b). Observing Fig. 15(a) and (b), both the proposed model and the traditional multi-task learning model achieve good prediction results for pressure and temperature, with the prediction curves closely matching the true curves. However, a closer look at the detailed regions of Fig. 15(a) and (b) reveals that the proposed model maintains a good fit even in parts with drastic fluctuations, with predictions closer to the true values. In contrast, the prediction errors of the MTL-BiLSTM and MMoE-BiLSTM models are higher, resulting in suboptimal performance. The proposed model reduces the MAPE by 63.05% and 52.91% for pressure, and 71.88% and 48.94% for temperature compared to other models. For the pressure and temperature prediction tasks, the MAE of the proposed method is reduced by up to 75.80% and 61.23%, respectively, while the R^2 metric improves by 4.52% and 5.22% compared to single-task prediction models. This performance enhancement stems from the model's emphasis on advanced feature extraction and effective sharing of feature information across tasks, which enables it to better capture task-specific and shared patterns compared to other multi-task approaches.

This is because the CGC network innovatively divides the expert structure into shared experts and task-specific experts, and equips each task with a customized gating network, achieving fine-grained management of feature sharing in multi-task learning. This design allows the model to dynamically adjust the combination ratio of shared and task-specific features based on the characteristics of the input samples. It preserves the valuable positive transfer effects between the pressure and temperature prediction tasks while effectively isolating the negative transfer interference caused by irrelevant features, thus adaptively optimizing feature representation under varying task correlations. Additionally, in each TCN network within the multi-task learning layer, especially after the dilated convolution operation, the ECA module is strategically embedded. This allows the TCN to better focus on key local cross-channel interaction information in the two dilated causal convolutions, compensating for the limitations of the receptive field in TCNs and enhancing the expression of critical high-dimensional time-series features. This unique fusion creates a powerful synergistic effect, enabling the TCN to simultaneously capture complex temporal patterns and significantly improve its adaptability to complex dynamic changes by intelligently emphasizing key high-dimensional time-series feature channels, suppressing irrelevant information, and optimizing the information flow path selection.

Table 8
Experimental evaluation results of the multi-task model and the model in this paper.

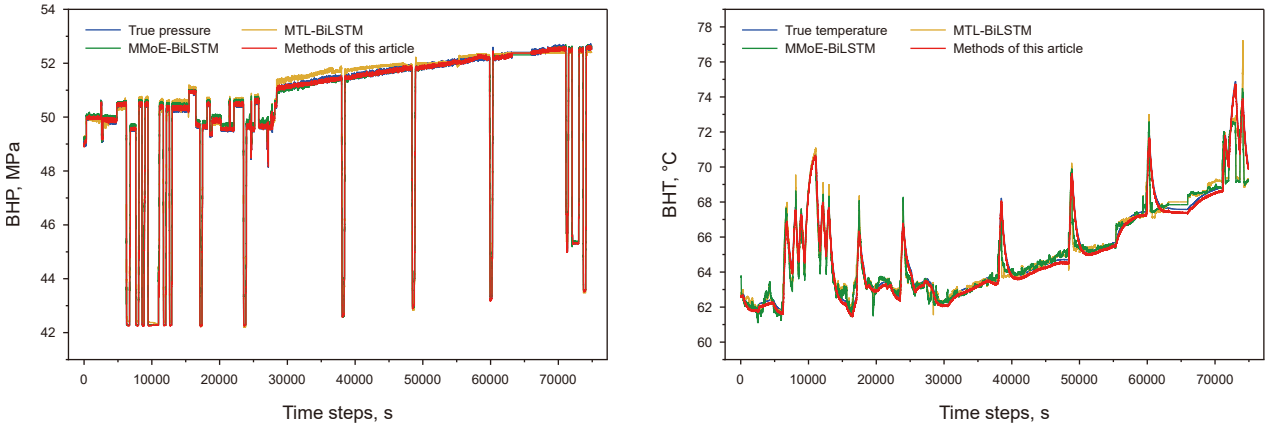
Evaluation criteria	Predictors	MTL-BiLSTM	MMoE-BiLSTM	This paper
MAPE, %	Pressure	0.4634	0.3636	0.1712
	Temperature	0.5062	0.2787	0.1423
MAE	Pressure	0.2658	0.1235	0.0642
	Temperature	0.4839	0.3783	0.1876
R^2	Pressure	0.9493	0.9521	0.9923
	Temperature	0.9367	0.9744	0.9856

4.4. Revealing physical laws through the model

Overall trend analysis systematically evaluates the strength of the proposed model and ensures its effectiveness in capturing physical phenomena. In this work, a synthetic dataset was created in which only one input parameter varies from its minimum to maximum value, while all other parameters remain constant at their average values. Using the dynamic variations of experimental conditions, the model's ability to capture the coupling effects between pressure and temperature under changes in the geothermal gradient and drilling fluid density was systematically evaluated. By adjusting key parameters, the coupling relationship between these two factors was revealed:

(1) **Variation of the geothermal gradient**

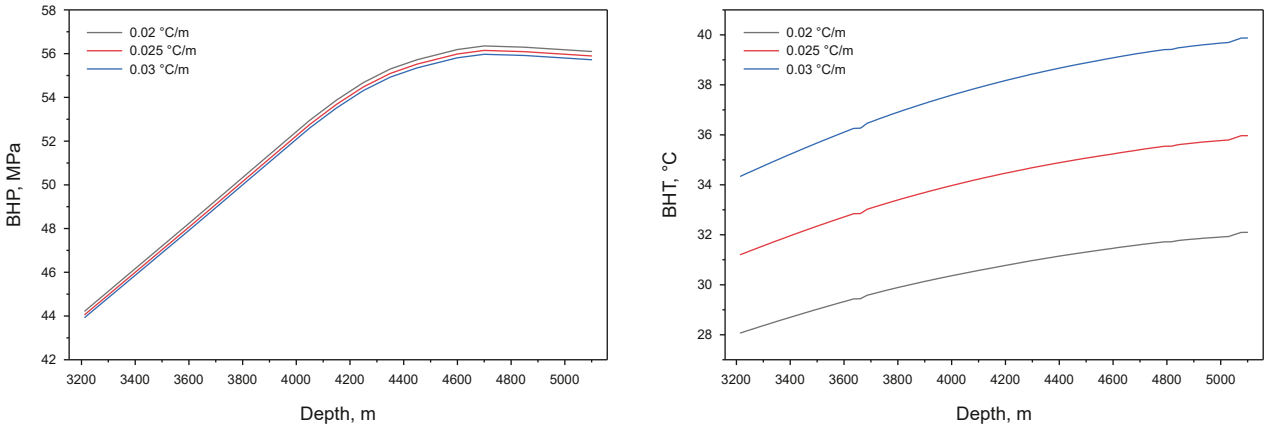
The results are presented in Fig. 16(a) and (b). Since the selected experimental well is a representative horizontal well, its longitudinal and lateral distribution characteristics provide valuable insights into the coupling relationship between bottom-hole pressure and bottom-hole temperature. From the lateral perspective, within the depth range of 4500–5000 m, the BHP remains nearly constant, while the BHT exhibits a significant increase. From the longitudinal trend, as the geothermal gradient



(a) Comparison of pressure prediction effects

(b) Comparison of temperature prediction effects

Fig. 15. Prediction accuracy trends comparing standard temperature/pressure multi-task models against the approach in this paper.



(a) Comparison of wellbore pressures under different geothermal gradients

(b) Wellbore temperatures are compared under different geothermal gradients

Fig. 16. Changes in wellbore pressure and temperature with varying geothermal gradient.

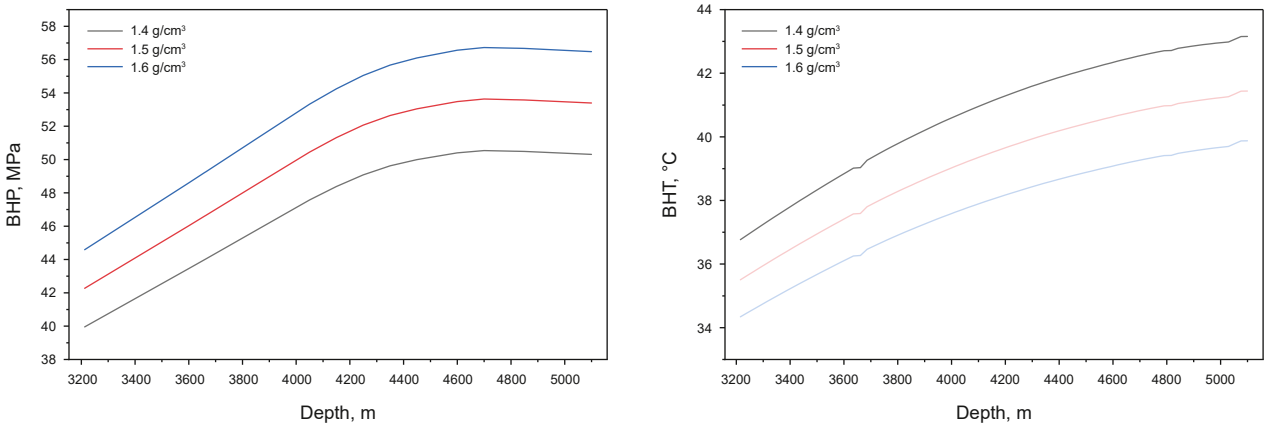
rises, the proposed method accurately captures the nonlinear variations: BHP decreases gradually, whereas BHT increases sharply. These results reveal the complex interaction mechanisms between temperature and pressure, indicating that the geothermal gradient not only directly affects the temperature distribution but also indirectly influences pressure variation by altering fluid properties and pressure transmission mechanisms within the formation. Consequently, a strong dynamic coupling relationship between pressure and temperature emerges.

It is noteworthy that conventional single-task neural network models typically treat pressure and temperature as independent variables, predicting them separately while neglecting their potential nonlinear dependencies and interactions. As a result, such models fail to realistically reflect the coupled effects of pressure and temperature in the wellbore environment, leading to insufficient characterization of dynamic behaviors under complex conditions. By contrast, the multi-task learning framework proposed in this study, through joint modeling and feature-sharing mechanisms, is able to effectively capture and characterize the interdependence and nonlinear coupling between pressure and temperature. This enables the model to achieve higher predictive accuracy and stronger generalization capability.

(2) **Variation of drilling fluid density**

The density of drilling fluid not only directly determines the magnitude of the hydrostatic pressure but also indirectly affects the distribution of BHP by altering the frictional resistance of fluid circulation and the associated dynamic pressure losses within the wellbore. The proper design of drilling fluid density is therefore a central issue in drilling engineering, as it enables effective balancing of formation pressures, thereby ensuring both operational safety and efficiency. In the simulation experiments, comparative analyses under different drilling fluid density conditions clearly reveal the coupling effects between fluid dynamic characteristics and the evolution of pressure and temperature.

The results are presented in Fig. 17(a) and (b). Fig. 17(a) illustrates the BHP profiles under drilling fluid densities of 1.4, 1.5, and 1.6 g/cm^3. It can be observed that, with increasing well depth, the BHP curves across different densities exhibit a consistent upward trend. However, at the same depth, higher drilling fluid density corresponds to a greater BHP. For instance, at a depth of 5100 m, an increase in drilling fluid density from 1.4 to 1.6 g/cm^3 results in a BHP rise from 50.3109 to 56.4738 MPa. This observation aligns well with the physical law that hydrostatic pressure increases with



(a) Comparison of wellbore pressures at different drilling fluid densities

(b) Comparison of wellbore temperature under different drilling fluid densities

Fig. 17. Changes in wellbore pressure and temperature with varying drilling fluid densities.

fluid density, thereby confirming that the model can accurately reproduce the underlying pressure mechanisms in the wellbore.

Fig. 17(b) shows the variation of temperature with drilling fluid density. The results clearly indicate that BHT is inversely correlated with drilling fluid density. At a depth of 5100 m, when the fluid density increases from 1.4 to 1.6 g/cm^3, the BHT decreases significantly from 543.16 to 539.87 °C. The physical explanation lies in the fact that, under constant flow rate, a higher fluid density leads to a slower circulation velocity, which raises the hydrostatic pressure and suppresses the volumetric expansion of bottom-hole fluids. In addition, denser drilling fluids enhance thermal conductivity and remove heat more efficiently from the bottom hole, thereby contributing to the reduction in BHT.

Overall, the trend analysis demonstrates that drilling fluid density exerts not only a direct impact on pressure distribution but also an indirect influence on bottom-hole temperature through fluid dynamics and heat transfer mechanisms. This coupled pressure–temperature relationship is clearly manifested in the experimental results. The proposed model successfully captures the underlying physical laws governing these processes and exhibits stable and consistent performance across different density conditions, thereby verifying its applicability and reliability under complex wellbore environments.

5. Conclusions

Current methods for predicting wellbore pressure and temperature face challenges such as high computational errors and poor adaptability in physical methods, as well as traditional machine learning approaches that often overlook physical laws and the coupled multi-field characteristics of the wellbore environment. This study addresses these issues by constructing a multi-task learning framework that fuses physical principles with data-driven features. Specifically, we propose a CGC multi-task channel attention optimized TCN-BiLSTM method for coupled prediction of wellbore pressure and temperature in managed-pressure drilling. The experimental results demonstrate that:

(1) The distance correlation analysis successfully selects the influential time and spatial features, confirming the high degree of coupling between wellbore pressure and temperature. Furthermore, trend analysis quantitatively verifies this physical coupling: for instance, at a depth of 5100 m, an increase in drilling fluid density from 1.4 to 1.6 g/cm^3 leads to a scientifically consistent BHP rise to 56.4738 MPa and a BHT decrease to 539.87 °C, proving the model's high physical interpretability.

(2) The TCN network within the CGC framework effectively mines high-dimensional features, and the supplementary ECA module further enhances the receptive field and the model's ability to capture key coupling information. Compared to advanced single-task models (e.g., TCN and Mamba-2), this architecture allows the proposed model to reduce the MAPE for pressure and temperature prediction by up to 94.98% and 95.28%, respectively, while improving the R^2 metric by 26.84% and 20.38%.

(3) Comparative analysis with various multi-task prediction models (such as MTL-BiLSTM and MMoE-BiLSTM) reveals that our proposed model better extracts coupled information and shared features. Specifically, it further reduces the MAPE by 63.05% for pressure and 71.88% for temperature, with MAE reductions of up to 75.80% and 61.23%, resulting in superior prediction efficiency and accuracy even under drastic drilling fluctuations.

The CGC multi-task channel attention optimized TCN-BiLSTM method offers an innovative solution for accurate wellbore pressure and temperature prediction. Future work could extend this method to other coupled prediction tasks, thereby providing a robust theoretical and technical foundation for the intelligent and efficient development of petroleum engineering.

Looking ahead, this research opens several promising avenues for future exploration. Firstly, the framework can be further extended to incorporate additional coupled prediction tasks, such as torque and drag, vibration, or hole cleaning, thus moving towards a more comprehensive wellbore digital twin. Secondly, integrating more explicit physics-informed constraints, such as PINNs and DeepONet or developing hybrid machine learning-physics models could further enhance prediction accuracy and model interpretability, especially in data-sparse scenarios or for novel drilling conditions. Thirdly, exploring uncertainty quantification in the predictions, perhaps through Bayesian neural networks or ensemble methods, would provide crucial confidence intervals for real-world operational decisions.

CRediT authorship contribution statement

Xue-Rui Wang: Writing – review & editing, Supervision, Resources, Conceptualization. **Yuan-Ze Tian:** Writing – original draft, Visualization, Software. **Bao-Jiang Sun:** Resources, Methodology, Conceptualization. **Zhi-Yuan Wang:** Writing – review & editing, Methodology. **Yong-Hai Gao:** Visualization, Data curation. **Hao Li:** Resources, Methodology.

Data availability statement

All data that support the findings of this study are included in this manuscript and its supplementary information files.

Conflict of interest

The authors declared that they have no conflicts of interest to this work.

We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

Acknowledgments

The authors acknowledge the support from the National Natural Science Foundation of China (52288101), National Science and Technology Major Project (025ZD1402206, 2025ZD1403201), Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (JYB2025XDXM301).

References

Akinsete, O., Adesiji, B.A., 2019. Bottom-hole pressure estimation from wellhead data using artificial neural network. In: SPE Nigeria Annual International Conference and Exhibition. SPE-198762-MS. https://doi.org/10.2118/198762-MS.

Al-Fakih, A., Koeshidayatullah, A., Mukerji, T., et al., 2025a. Well log data generation and imputation using sequence based generative adversarial networks. Sci. Rep. 15 (1), 11000. https://doi.org/10.1038/s41598-025-95709-0.

Al-Fakih, A., Al-khudafi, A., Koeshidayatullah, A., et al., 2025b. Forecasting geothermal temperature in Western Yemen with Bayesian-optimized machine learning regression models. Geothermal Energy 13 (1), 4. https://doi.org/10.1186/s40517-024-00324-3.

Bai, S., Kolter, J.Z., Koltun, V., 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv: 1803.01271. https://doi.org/10.48550/arXiv.1803.01271.

Caetano, E.F., Shoham, O., Brill, J.P., 1992a. Upward vertical two-phase flow through an annulus—Part I: Single-phase friction factor, Taylor bubble rise velocity, and

flow pattern prediction. J. Energy Resour. Technol. 114 (1), 1–13. https://doi.org/10.1115/1.2905917.
Caetano, E.F., Shoham, O., Brill, J.P., 1992b. Upward vertical two-phase flow through an annulus—Part II: Modeling bubble, slug, and annular flow. J. Energy Resour. Technol. 114 (1), 14–30. https://doi.org/10.1115/1.2905916.
Caruana, R., 1997. Multitask learning. Mach. Learn. 28 (1), 41–75. https://doi.org/10.1023/A:1007379606734.
Caruana, R.A., 1993. Multitask learning: A knowledge-based source of inductive bias. In: Machine Learning Proceedings 1993, pp. 41–48. https://doi.org/10.1016/b978-1-55860-307-3.50012-5.
Gao, Y., Liu, K., Zhao, X., et al., 2018. Prediction of wax precipitation region in wellbore during deep water oil well testing. Petrol. Explor. Dev. 45 (2), 351–357. https://doi.org/10.1016/S1876-3804(18)30039-9.
Gu, A., Dao, T., 2023. Mamba: Linear-time sequence modeling with selective state spaces. arXiv:2312.00752. https://doi.org/10.48550/arXiv.2312.00752.
Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.
Hu, S.Y., Li, J.Z., Wang, T.S., et al., 2020. CNPC oil and gas resource potential and exploration target selection. Petroleum Geology & Experiment 42 (5), 813–823. https://doi.org/10.11781/sysydz202005813 (in Chinese).
Koeshidayatullah, A., Al-Fakih, A., Kaka, S.I., 2024. Leveraging time-series foundation model for subsurface well logs prediction and anomaly detection. arXiv preprint arXiv: 2412.05681. https://doi.org/10.48550/arXiv.2412.05681.
Liu, X., Wang, Z., Wei, Q., et al., 2025. A method for solving wellbore temperature field driven by physical information neural network. Acta Pet. Sin. 46 (2), 413–425. https://doi.org/10.7623/syxb202502009 (in Chinese).
Sun, B.J., Wang, X.R., Sun, X.H., et al., 2020. Application and prospects of the wellbore four-phase flow theory in deepwater drilling and completion engineering and testing. Nat. Gas. Ind. 40 (12), 95–105. https://doi.org/10.3787/j.issn.1000-0976.2020.12.011 (in Chinese).
Székely, G.J., Rizzo, M.L., Bakirov, N.K., 2007. Measuring and testing dependence by correlation of distances. Ann. Stat. 35 (6), 2769–2794. https://doi.org/10.1214/009053607000000505.
Tang, H., Liu, J., Zhao, M., Gong, X., 2020. Progressive layered extraction (PLE): A novel multi-task learning (MTL) model for personalized recommendations. In: Proceedings of the Fourteenth ACM Conference on Recommender Systems, pp. 269–278. https://doi.org/10.1145/3383313.3412236.
Wang, D., Zhang, R., Zhao, L., 2022. A multivariate load prediction method of integrated energy systems based on MMoE-TCN. In: 2022 7th International Conference on Power and Renewable Energy (ICPRE). https://doi.org/10.1109/ICPRE55555.2022.9960446.
Wang, J., Yu, L.C., Lai, K.R., Zhang, X., 2016. Dimensional sentiment analysis using a regional CNN-LSTM model. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 225–230. https://doi.org/10.18653/v1/P16-2037.
Wang, Q., Wu, B., Zhu, P., et al., 2020. ECA-net: Efficient channel attention for deep convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11534–11542. https://doi.org/10.1109/CVPR42600.2020.01155.
Wu, C., Yao, J., Xue, G., et al., 2022. Load forecasting of integrated energy system based on MMoE multi-task learning and LSTM. Electric Power Automation Equipment 42 (7), 33–39. https://doi.org/10.16081/j.epae.202204083 (in Chinese).
Xu, B.C., Zhang, X.Z., Wang, Y.X., et al., 2023. Self-adaptive physical information neural network model for prediction of two-phase flow annulus pressure. Acta Pet. Sin. 44 (3), 545–555. https://doi.org/10.7623/syxb202303012 (in Chinese).
Yang, H.W., Li, J., Zhang, H., et al., 2024. Fully transient wellbore gas-liquid two-phase flow model considering interphase mass transfer. Acta Pet. Sin. 45 (11), 1680–1689. https://doi.org/10.7623/syxb202411009 (in Chinese).
Yang, Z., Liao, X., Dong, P., et al., 2026. Surrogate model for numerical simulation of CO_2 flooding based on Fourier neural operator. Pet. Geol. Oilfield Dev. Daqing 45 (1), 109–117. https://doi.org/10.19597/J.ISSN.1000-3754.202405039 (in Chinese).
Yu, Y., Wang, X., Ke, K., et al., 2021. Prediction model and distribution law study of temperature and pressure of the wellbore in drilling in arctic region. Petroleum Drilling Techniques 49 (3), 11–20. https://doi.org/10.11911/syztjs.2021047 (in Chinese).
Zhang, R., Li, J., Liu, G., 2022. Research on wellbore pressure prediction model during variable gradient drilling in deepwater. Pet. Sci. Bull. 7 (4), 564–575. https://doi.org/10.3969/j.issn.2096-1693.2022.04.048 (in Chinese).
Zhang, R., Li, J., Ye, D., et al., 2023. Indoor simulation experiment on the feasibility of multi-gradient managed pressure drilling. Petroleum Machinery 51 (5), 41–46. https://doi.org/10.16082/j.cnki.issn.1001-4578.2023.05.006 (in Chinese).
Zhou, D., 2023. Dynamic Prediction of Multiphase Flow in Wellbores Based on Deep Learning. Master's Thesis. Yangtze University. https://doi.org/10.26981/d.cnki.gjhsc.2023.000127.
Zhou, Y.C., Cui, M., Zha, Y.J., 2008. Discussion and prospect of managed pressure drilling technology. Petroleum Drilling Techniques 36 (4), 1–4. https://doi.org/10.3969/j.issn.1001-0890.2008.04.001 (in Chinese).
Zhu, Z., Song, X., Zhang, R., et al., 2022. A hybrid neural network model for predicting bottomhole pressure in managed pressure drilling. Appl. Sci. 12 (13), 6728. https://doi.org/10.3390/app12136728.