Energy Proceedings
Vol 64, 2026
ISSN 2004-2965

Study on Downhole CO_2 Injection Flow Measurement Model Based on Mechanism and Data Fusion#

Jiao Xinyu[1], Tan Chaodong[1,2,*], Shi Yipeng[2], He Jinsong[2], Chen Yanrun[2], Yang Aonan[2]

1 College of Petroleum Engineering, China University of Petroleum (Beijing)

2 College of Artificial Intelligence, China University of Petroleum (Beijing)
(Corresponding Author: Email: tanchaodong@cup.edu.cn)

ABSTRACT

The accurate measurement of downhole CO_2 injection rate is of great practical significance for optimizing enhanced oil recovery (EOR) effects, evaluating reservoir performance, and designing and adjusting development plans. To improve the accuracy of CO_2 injection well flow rate measurement, a downhole CO_2 injection flow rate measurement method integrating mechanism and data was proposed. Data were acquired and preprocessed using an indoor experimental platform for CO_2 flow rate measurement, and a mechanistic model, a data-driven model, and a hybrid-driven model were constructed. The accuracies of the three models were compared and analyzed. The research results show that: the mechanistic model can be used for downhole CO_2 flow rate measurement, with most errors ranging from 9% to 20%. Relying on its flexible application of multi-source data, relatively simple construction process, and strong adaptability, the data-driven model exhibits higher accuracy in CO_2 flow rate measurement than the mechanistic model, with an accuracy of 87.3%. The hybrid-driven model integrates the in-depth interpretability of the mechanistic model and the efficient learning ability of the data-driven model for massive data. Its comprehensive performance in accuracy, adaptability, and interpretability surpasses that of single models, with an accuracy of 94.7%. The application of this hybrid-driven model in an oil well in southern China shows that its accuracy is 92.3%, which can meet the actual on-site measurement requirements.

Keywords: downhole CO_2 Injection, CO_2 indoor experiment, mechanism-data fusion, hybrid-drive model, flow measurement

NONMENCLATURE

Abbreviations	
MSE	Mean Squared Error
MAE	Mean Absolute Error
R^2	Coefficient of Determination
LSTM	Long Short-Term Memory

1. INTRODUCTION

As an important method for tertiary oil recovery, CO_2 flooding is of great practical significance for expanding reservoir sweep efficiency, reducing residual oil saturation, and improving crude oil recovery. Meanwhile, CO_2 flooding provides an economical and effective approach for carbon capture and storage (CCS) [1]. However, blind CO_2 injection is prone to gas channeling, making it difficult to effectively utilize oil reservoirs. Additionally, the physical property parameters of CO_2 are highly affected by temperature and pressure; therefore, accurate measurement of CO_2 injection rate is particularly crucial.

There are three methods for CO_2 flow rate measurement: mechanistic models, data-driven models, and hybrid-driven models. Mechanism-simulation-based measurement models rely on prior knowledge of the internal structure of the process, and describe the correlation of system variables by deriving mass, energy, and momentum balance equations as well as empirical closure relationships. However, they require clear initial boundary conditions, and their solution algorithms are sensitive to boundary regions (minor adjustments to conditions lead to significant changes in calculation results), making them difficult to apply to real-time prediction [2]. Jiao Xinyu et al. [3] constructed a nozzle model for downhole CO_2 flow rate measurement, and obtained a flow rate measurement model by using Fluent to simulate the relationship between differential pressure and injection pressure under different temperatures and pressures. This study has certain practical significance for improving the test performance of nozzle flowmeters and the layered measurement and control of CO_2. Data-driven virtual measurement

This is a paper for the 17th International Conference on Applied Energy (ICAE2025), December 8-12, 2025, Bangkok, Thailand.

methods collect on-site, numerical simulation, and experimental test data, and fit mathematical models to model relevant processes. However, they have high requirements for data quantity and quality, and model reliability depends on the accuracy and completeness of sample data. Wang Lijuan et al. [4] combined Coriolis Mass Flowmeters (CMF) with differential pressure transmitters, and supplemented with machine learning algorithms to measure gas-liquid two-phase CO_2 in high-pressure pipelines. When the gas content is $\leqslant$ 80%, the uncertainties of horizontal and vertical pipelines reach ±1.5% and ±2.5%, respectively. Hotvedt Mathilde et al. [5] developed a mechanism-data integrated grey-box modeling method: taking a mechanical model as the baseline, they embedded data-driven elements to improve model flexibility, which not only reduces the error of mechanical virtual flow rate measurement but also enhances the scientific consistency of the data-driven model. Ishak M A B et al. [6] used a transient multiphase flow simulator and a data-driven multi-variate ensemble learning neural network in parallel, which can conduct real-time fault detection, verify the measurement results of physical phase flowmeters in well testing operations, and improve measurement reliability. Bikmukhmetov and Jäschke [7] combined the steady-state first principles of various components in the production system (throttle valves, oil pipes) with artificial neural networks (ANNs) and long short-term memory (LSTM) networks, and confirmed that this method has high accuracy and good interpretability in virtual measurement.

Aiming at the shortcomings of mechanistic models (poor flexibility) and pure data-driven models (weak interpretability), this study proposes a downhole CO_2 flow rate measurement model integrating mechanism and data to improve the accuracy and interpretability of the measurement model.

2. MATERIAL AND METHODS

Based on the ground experimental circulation system for CO_2 flowmeters, flow rate data under different temperature and pressure conditions were acquired. First, outliers were removed using boxplots to improve data quality; second, data normalization was performed to ensure consistent feature scales and accelerate model convergence; finally, correlation analysis was conducted to clarify the presence, strength, and direction of variable correlations. This process further enhances the accuracy and reliability of subsequent CO_2 flow rate measurement models.

2.1 Data sources

The main circulation system, as shown in Figure 1, includes a booster pump (pressurizing from 4–5 MPa to 45 MPa), a circulation pump (inlet/outlet diameter: 80/50 mm; pressure resistance: 45 MPa; temperature resistance: 150 °C; flow rate: 0–25 m^3/h), an electromagnetic heating tube, a Coriolis flowmeter (for calibration; pressure resistance: 45 MPa; accuracy: ±0.1%), and a Venturi flowmeter.

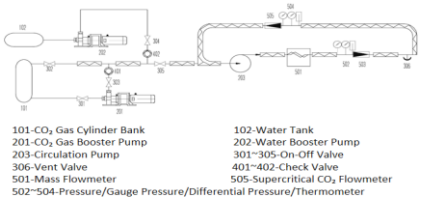


Fig. 1 CO_2 Flow Rate Measurement Experimental Platform

The gas boosting system, as shown in Figure 2, consists of a CO_2 gas cylinder, an air compressor, a gas booster pump, a high-pressure storage tank, and other components, and can realize pressure boosting, pressure stabilization, and flow rate regulation.

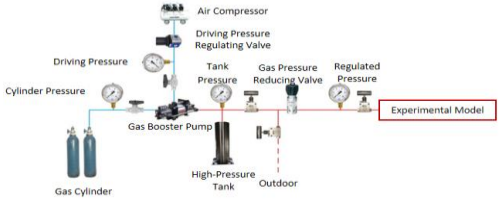


Fig. 2 Gas Boosting System

The gas boosting system was started: CO_2 was first decompressed, then pressurized by the air compressor, and the pressure was controlled to approximately 5 MPa; the heating tube was activated to reach the preset temperature. Next, the stepless speed-regulating magnetic circulation pump was started and adjusted to the preset displacement, and CO_2 flowed through the Coriolis flowmeter and Venturi flowmeter in sequence. After the readings of temperature, pressure, flow rate, and other parameters stabilized, the Venturi flowmeter recorded data for 3 minutes at a rate of 5 groups per second, which were transmitted to a computer via a serial port, and the calibration equipment parameters were recorded simultaneously. A total of 1915 groups of data were collected in the experiment, as shown in Table 1.

Table 1 Experimentally Collected Data Table

No.	T/℃	P/MPa	ρ/kg/m³	μ/mPa·s	ΔP/Pa	Q_g/kg/h
1	77.1	7.5	154.7	0.0215	3046.9	107.9
2	35.3	7.2	243.6	0.0223	2142.1	118.8
3	41.4	8.8	377.5	0.0453	23094.2	301.1
……	……	……	……	……		……

1915	83.3	9.9	217.7	0.0241	58423.6	365.3

2.2 Data Pre-processing

A box plot is a statistical graph used to display data distribution. As shown in Figure 3, it can concisely present key information such as the median, quartiles, and extreme values of data in a single graph, thereby intuitively reflecting the distribution characteristics of the data-including symmetry, dispersion degree, and the presence of outliers [8]. Points beyond the upper and lower edges of the box are recorded as outliers.

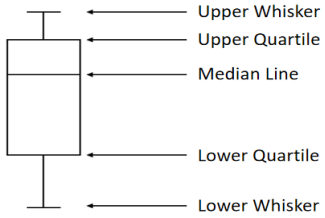


Fig. 3 Box Plot

A preliminary analysis was conducted on the experimentally collected data, which included a total of 1,915 data points with no missing values in any column. Subsequently, outliers were detected using box plots, and the box plot results for each parameter are presented in Figure 4. Points exceeding the upper or lower whiskers of the box plots were identified as outliers.

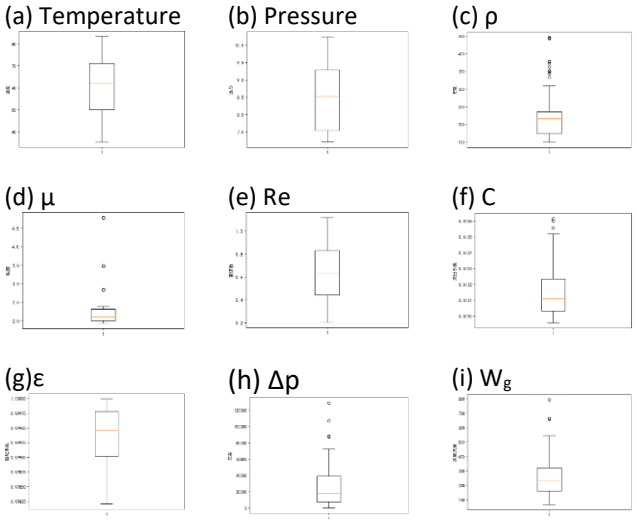


Fig. 4 Statistics of abnormal values of each parameter

As observed from the Figure 4, parameters such as density, viscosity, discharge coefficient, differential pressure, and mass flow rate all contained a certain number of outliers (to varying degrees). Given the small number of outliers, the entire rows containing these outliers were directly removed to ensure data quality for subsequent data analysis and model training.

To compare the accuracy of models trained on outlier-removed data versus raw data, both datasets were input into the Light Gradient Boosting Machine model for CO_2 mass flow rate prediction. The evaluation metrics of the models are presented in Table 2.

Table. 2 Model evaluation indexes of different data types

Data type	MSE	MAE	R^2
Raw data	0.00736	0.07791	0.806
Outlier-Removed Data	0.00510	0.06494	0.844

As shown in the table 2: the model built with outlier-removed data exhibited lower Mean Squared Error and Mean Absolute Error than the model built with raw data (without outlier removal), while its coefficient of determination was higher. This indicates that outlier removal has a positive effect on improving model accuracy. The min-max normalization method was used to normalize the dataset, ensuring all parameters were kept within the range of [0, 1] [9].

2.3 Pearson Correlation Analysis

Correlation analysis is a statistical method for studying the degree of association between variables. The Pearson correlation coefficient method is commonly used to measure the degree of linear correlation between two parameters, with a value range of [-1, 1][10]. The larger the absolute value, the stronger the correlation, and its value is calculated using Equation (1).

$$\gamma = \frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2}} \qquad (1)$$

Where: γ is the Pearson correlation coefficient; X_i is the i-th value of parameter X; $\overline{X}$ is the mean value of parameter X; Y_i is the i-th value of parameter Y; $\overline{Y}$ is the mean value of parameter Y; and n is the number of samples of the parameters.

Pearson correlation coefficient analysis was performed on the preprocessed parameters, and the results are shown in Figure 5. Warm colors represent positive correlation, while cool colors represent negative correlation. If the correlation between two parameters is extremely strong, it indicates that the two parameters can be converted into each other, and their variation and influence trends are consistent; thus, one of them can be eliminated without affecting the results.

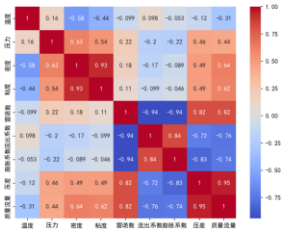


Fig. 5 Pearson Correlation Coefficient Results

As can be seen from Figure 5: the correlation coefficient between density and viscosity is 0.93, and the correlation coefficients between Reynolds number and discharge coefficient, and between Reynolds number and expansion coefficient are both -0.94, indicating an extremely strong pairwise correlation. Considering the data acquisition cost, difficulty level, and empirical importance, viscosity and Reynolds number were eliminated without affecting the CO_2 mass flow rate. Meanwhile, parameters with an absolute value of Pearson correlation coefficient greater than 0.4 with mass flow rate were retained. Finally, five parameters—pressure, density, discharge coefficient, expansion coefficient, and differential pressure—were selected for CO_2 flow rate prediction.

2.4 Model Establishment

To achieve accurate and reliable measurement of downhole CO_2 injection flow rate, a mechanistic model was first constructed to explore physical laws, and a data-driven model was built to mine hidden patterns in data. Subsequently, the advantages of both models were integrated to establish a mechanism-data-driven hybrid-driven model, which breaks through the bottlenecks of single models and provides a more suitable solution for downhole CO_2 injection flow rate measurement.

2.4.1 Mechanism model establishment

The measurement principle of the Venturi differential pressure flowmeter is based on the Bernoulli equation and the continuity equation. For the steady flow of downhole CO_2, the flow pressure drop relationship is obtained under the condition of neglecting potential energy change (assuming horizontal flow), as follows:

$$Q = \frac{\pi d_1^2}{4}\sqrt{\frac{2\Delta p}{\rho\left[\left(\frac{d_1}{d_2}\right)^4 - 1\right]}} \qquad (2)$$

Where: Q is the volumetric flow rate, m^3/h; d_1 is the inner diameter of the orifice, m; d_2 is the inner diameter of the pipeline, m; Δp is the differential pressure across the Venturi, Pa; and ρ is the fluid density, kg/m^3.

For compressible fluids, when the pressure change is significant, their density also changes remarkably, and the above formula requires further correction. In this case, the expansibility coefficient ε and discharge coefficient C need to be introduced to convert the volumetric flow rate to mass flow rate[11, 12], and the formula then becomes:

$$Q = C\cdot\varepsilon\cdot\frac{\pi d_1^2}{4}\sqrt{\frac{2\Delta p}{\rho\left[\left(\frac{d_1}{d_2}\right)^4 - 1\right]}} \qquad (3)$$

The flow rate data under different temperature and pressure conditions were calculated using Equation (3), i.e., the mechanistic model for downhole CO_2 flow rate measurement.

2.4.2 Data-driven model establishment

Based on the correlation analysis in Section 2.3, five mutually independent parameters—pressure, density, discharge coefficient, expansion coefficient, and differential pressure—were selected as model inputs, with mass flow rate as the output. The 1911 preprocessed samples were divided into a training set and a test set in an 8:2 ratio, and an LSTM-based downhole CO_2 flow rate measurement model was constructed, using MSE, MAE, and R^2 as evaluation metrics [13].

As a special recurrent neural network (RNN) designed to address the long-term dependency issue of long sequences in traditional RNNs, the LSTM network is centered on three core gating mechanisms: the forget gate, input gate, and output gate [14]. The optimal hyperparameters were determined via grid search [15]: the number of neural network layers is 2, with 64 and 32 neurons in each layer respectively; the activation function is ReLU; the optimization function is Adam; the number of training epochs is 500; and the batch size is 32.

During training, MSE was used as the evaluation metric for the loss function. Network convergence was determined when MSE was less than 0.01, and its loss function curve is shown in Figure 6.

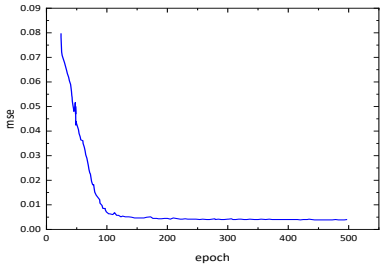


Fig. 6 Loss Function of the LSTM Model

As can be seen from Figure 6: when the number of training epochs reaches 200 (i.e., epoch = 200), the loss function begins to converge and its value is less than 0.01, which meets both the model accuracy requirements and the training convergence criteria.

2.4.3 Hybrid-drive model establishment

Although mechanistic models have strong interpretability, they are built on a series of assumptions. Pure data-driven models exhibit relatively accurate predictions but have high requirements for the quantity and quality of data, along with poor interpretability. To address the shortcomings of both models, this section reasonably combines them to construct a mechanism-LSTM hybrid model with more accurate downhole CO_2 flow rate prediction[16], as shown in Figure 7. This model embeds the data-driven model into the mechanistic model to supplement or optimize the functions of the mechanistic model.

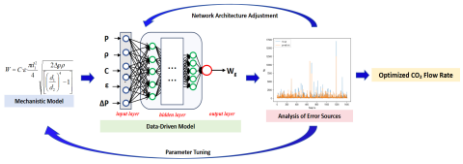


Fig. 7 Flow Chart of Downhole CO_2 Flow Rate Measurement for the Hybrid-Driven Model

3. RESULTS AND DISCUSSION

The calculation results of the mechanistic model were compared with the experimental test data, and the errors were plotted as a curve, as shown in Figure 8.

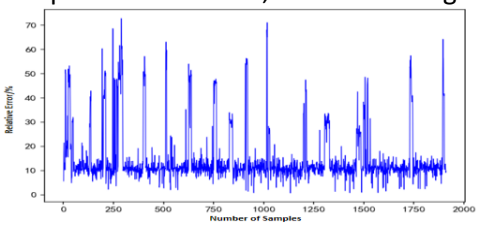
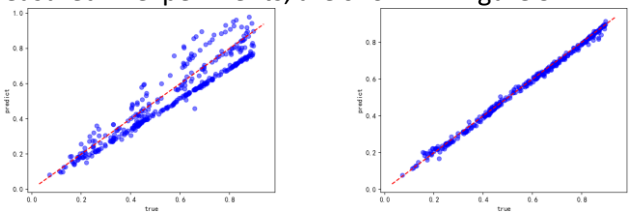


Fig. 8 Error Curve of the Mechanistic Model

As can be seen from Figure 8: the error range of the mechanistic model is [0.0071, 0.728], with most errors falling within [0.09, 0.20] and a maximum error of 72.8%. Although the model can simply predict downhole CO_2 flow rate, its error is relatively large. The core reason is that the curved downhole wellbore and existing valves in actual scenarios cause the fluid to exhibit complex transitional flow, leading to inconsistencies between pressure loss, flow velocity distribution, and the model assumptions. Additionally, downhole CO_2 flow rate measurement is a dynamic and nonlinear problem, and a pure mechanistic model cannot fully describe the working conditions or construct a complete and accurate measurement model.

The predicted values of the data-driven model and hybrid-driven model, along with the actual values measured in experiments, are shown in Figure 9.



(a) Data-Driven Model *(b) Hybrid-Driven Model*

Fig. 9 Predicted Values and Actual Values

As can be clearly seen from Figure 9, the data of the data-driven model are highly scattered on both sides of the reference line, indicating low model accuracy. To compare the accuracy of the two models more precisely, three evaluation metrics—MSE, MAE, and R^2—are used for comparative analysis, as shown in Table 3.

Table 3 Model evaluation index

Models	Evaluation Metrics		
	MSE	MAE	R^2
LSTM Model	0.00451	0.05418	0.873
Mechanistic-LSTM Model	0.00016	0.00974	0.947

As can be seen from Table 4: the hybrid-driven model exhibits lower MSE and MAE, with the highest R^2 reaching 94.7%. For the complex downhole CO_2 flow conditions, if the model relies solely on data, it may learn some "spurious patterns" that do not conform to physical reality. However, the mechanistic model in the hybrid-driven model can guide the data-driven model to align with physical processes, thereby improving accuracy.

To verify its on-site application effect, the model has been applied to a gas well in southern China, and the results are shown in Table 4.

Table 4 On-site application results

Evaluation Metrics	MSE	MAE	R^2
Value	0.00039	0.01162	0.923

As can be seen from Table 6, the model accuracy reaches 92.3%. In on-site operations, it can not only accurately reflect the real-time CO_2 injection flow rate and overcome interferences such as temperature fluctuations and pressure changes but also provide a solid theoretical basis for on-site CO_2 injection flow rate measurement.

4. CONCLUSIONS

1. The mechanistic model can simply measure downhole CO_2 flow rate, but it has relatively large

errors—most errors range from [0.09, 0.20], with a maximum error of 72.8%;

2. The accuracy of the pure data-driven model is higher than that of the mechanistic model, reaching 87.3%;

3. The hybrid-driven model achieves the highest accuracy at 94.7%, which is better than the two types of single models mentioned above;

4. The hybrid-driven model has been applied to an oil well in southern China, with an accuracy of 92.3%, and can provide theoretical guidance for on-site CO_2 injection flow rate measurement.

ACKNOWLEDGEMENT

This work was supported by the Major Science and Technology Project of PetroChina Company Limited, titled "Research and Demonstration of Key Technologies for the Whole Industry Chain of Large-Scale CO_2 Capture, EOR and Storage" (Project 3: "Fine Description of CCUS EOR Geological Bodies and Research on Key Reservoir Engineering Technologies", Grant No. 2021ZZ01-03); the 14th Five-Year Plan Major Scientific Research Project (CCUS Special Project) of China National Offshore Oil Corporation (CNOOC) (Grant No. KJGG-2022-12-CCUS); and the National Key R&D Program of China, titled "CO_2 EOR and Storage Safety Monitoring Technology" (Grant No. 2018YFB0605505).

REFERENCE

[1] Atena Sharbatian, Ali Abedini, ZhenBang Qi, et al.Full Characterization of CO2–Oil Properties On-Chip: Solubility, Diffusivity, Extraction Pressure, Miscibility, and Contact Angle[J]. Analytical Chemistry, 2018.

[2] BJARNE G, MATHILDE H, SANDNES A T, et al. Bayesian neural networks for virtual flow metering: An empirical study[J]. Applied Soft Computing Journal,2021.

[3] J. Xinyu et al., "Numerical simulation and parameter optimization of downhole stratified flow measurement with CO2 injection," 2024 36th Chinese Control and Decision Conference (CCDC), Xi'an, China, 2024, pp. 3613-3618, doi: 10.1109/CCDC62350.2024.10587829.

[4] Wang L, Liu J, Yan Y, et al. Gas-liquid two-phase flow measurement using coriolis flowmeters incorporating neural networks[C]// Instrumentation and Measurement Technology Conference Proceedings. IEEE, 2016:1-5.

[5] HOTVEDT M, GRIMSTAD B, LJUNGQUIST D, et al. On gray-box modeling for virtual flow metering[J]. Control Engineering Practice，2022，118.

[6] ISHAK M A B, ISMAIL I B, AL-QUTAMI T A H. Data Driven Versus Transient Multiphase Flow Simulator for Virtual Flow Meter Application[C]//8th International Conference on Intelligent and Advanced System (ICIAS), Electr Network,2021.

[7] BIKMUKHAMETOV T, JASCHKE J.Combining Machine Learning and Process Engineering Physics towards Enhanced Accuracy and Explainability of Data-Driven Models[J]. Computers & Chemical Engineering,2020.

[8] FU Wenlong, ZHAO Yibo, FU Yuchen, et al. Visc osity Prediction of Dissolved Gas in Transformer Oil Based on Improved Boxplot and ISMAHHO-KELM [J /OL]. High Voltage Apparatus,1-13[2025-01-13]. http: //kns.cnki.net/kcms/detail/61.1127.tm.20240604.1329. 002.html.

[9] Chunfeng Zheng, Ang Li, Song Wang, Cheng An, Tan, Chaodong, Guanghao Du, Huizhao Niu. Fault Diagnosis Model and Application of Water Injection Well Based on SPC Rules and Real-time Data Fusion[C]. Journal of Physics: Conference Series, 2021. DOI: http://dx.doi.org/10.1088/1742-6596/2095/1/012091

[10] TAN Chaodong, HUANG Xinchun, WANG Song, et al. Research progress on fault diagnosis and early warning theory of electric submersible pump lifting based on mechanism simulation and data driven fusion[J]. Oil Drilling & Production Technology, , 2021, 43(04): 483-488.

[11] OU Yuqiong, LIU Qilin, LAI Huanxin. Flow and Heat Transfer Characteristics of Compressible Impinging Jets at Low to Moderate Reynolds Numbers [J/OL]. Journal of East China University of Science and Technology, 1-9[2025-01-13]. https://doi.org/10.14135/j.cnki.1006-3080.20240911001.

[12] XIE Junliang, HAO Xiaolong, ZHANG Minggen, et al. Simulation Analysis of Supersonic Nozzle Discharge Coefficient, 2022,46(12):123-128.

[13] JIAO Xinyu, TAN Chaodong, ZHU Yongle, et al. Research status and prospect of virtual metering of multiphase flow in oil and gas wells [J]. Industrial Metrology,2024,34(04):90-97. DOI:10.13228/j.boyuan.issn1002-1183.2023.0206.

[14] Chaodong Tan, Song Wang, Hanwen Deng, et al. The Health Index Prediction Model and Application of PCP in CBM Wells Based on Deep Learning[J]. Geofluids, 2021. DOI: https://doi.org/10.1155/2021/6641395

[15] Chaodong Tan, Hanwen Deng, Ziming Feng, et al. Data-driven system efficiency prediction and pro duction parameter optimization for PW-LHM [J]. Jou rnal of Petroleum Science and Engineering, Volume 209,2022,109810, ISSN 0920-4105. https://doi.org/10.1016/j.petrol.2021.109810.

[16] CHEN Hongwei, MA Zherui, LV Chunwang. Parameter Calculation of Steam Pipeline Based on Hybrid Modeling [J]. Electric Power, 2020,53(09): 195-201.