Machine-Learning-Based Development of Regional PVT Correlations for Bubble-Point Pressure and Solution Gas-Oil Ratio for Western Kazakhstan Oilfields

Adel Sarsenova
Kazakh-British Technical University, Almaty, Kazakhstan
a.sarsenova@kbtu.kz (corresponding author)

Mariam Kumarkhan
Kazakh-British Technical University, Almaty, Kazakhstan
kumarhanmariam@gmail.com

Abdulakhat Ismailov
Kazakh-British Technical University, Almaty, Kazakhstan
a.ismailov@kbtu.kz

Received: 27 December 2025 | Revised: 2 February 2026 and 21 February 2026 | Accepted: 23 February 2026

 DOI: https://doi.org/10.48084/etasr.17205

ABSTRACT

Bubble-point pressure and solution gas-oil ratio are critical PVT properties that govern phase behavior, reservoir performance, and the accuracy of material balance, production forecasting, and numerical simulation studies. Existing empirical correlations have historically been calibrated using geographically limited datasets, resulting in significant prediction errors when applied to crude oils with different compositional and geological characteristics, such as those found in Kazakhstan. This study investigates the performance of widely used global correlations in an extensive set of Kazakhstan oilfield data and demonstrates their systematic limitations. To address these shortcomings, new regional correlations were developed for bubble-point pressure and solution gas-oil ratio using a machine-learning-assisted workflow. Symbolic regression was employed to derive an explicit analytical expression for the solution gas-oil ratio, while multivariate log-linear regression was used to formulate a physically interpretable correlation for bubble-point pressure. Both models were evaluated using industry-standard error metrics and compared with established global correlations. The newly developed regional correlations exhibit substantially lower prediction errors and markedly improved consistency across the full range of fluid properties. These findings highlight the importance of region-specific PVT model development and provide more reliable tools for reservoir engineering applications in settings where laboratory data are limited or incomplete.

Keywords-machine-learning-assisted regression; bubble-point pressure; solution gas-oil ratio; PVT correlations; Kazakhstan oilfields; log linear regression; reservoir engineering

I. INTRODUCTION

Bubble-point pressure and solution gas-oil ratio are fundamental PVT properties that influence phase behavior, drive mechanisms, oil formation volume factors, and the initialization of both material balance studies and numerical reservoir simulation. Accurate prediction of these properties is essential for forecasting performance, evaluating reserves, and designing production strategies. Although laboratory PVT experiments remain the most reliable means of determining these parameters, such measurements are often unavailable due to cost, logistical constraints, or incomplete sampling. In these circumstances, empirical correlations are widely employed to estimate bubble-point pressure and solution gas-oil ratio from commonly measured inputs such as oil gravity, gas specific gravity, reservoir temperature, and pressure.

Over the past decades, numerous empirical correlations have been proposed and evaluated in various regions [1-4]. These correlations have demonstrated acceptable accuracy within the compositional and geological domains for which they were originally developed. However, extensive evidence in the literature shows that their performance deteriorates significantly when applied to crude oils outside their calibration

range [5-7]. Each correlation inherently reflects the statistical behavior of the dataset used for its derivation, and when the fluid properties differ from these regional characteristics, large deviations between predicted and measured values typically occur. Several regional studies have concluded that no existing global correlation performs reliably across all reservoirs, reinforcing the need for locally calibrated relationships to achieve high prediction accuracy.

A significant gap exists regarding Kazakhstan's crude oils. Regional studies in Iraq, the UAE, Kuwait, and Nigeria demonstrate that global correlations often underperform when applied to crude oils with distinct compositional signatures [1, 3-6]. While limited recent studies have investigated specific aspects of Kazakhstani fluid characterization [8, 9], a comprehensive, scientifically validated regional PVT correlation applicable across the broader basin remains absent. Consequently, reservoir engineers rely on global correlations without prior regional validation, although the compositional and thermodynamic characteristics of Kazakhstan reservoirs differ markedly from those of the North Sea or Gulf of Mexico crudes. This discrepancy highlights the urgent need for region-specific PVT correlations calibrated to the unique behavior of Kazakhstan oil systems.

Recent methodological advances have integrated data-driven techniques with classical regression frameworks. While Artificial Neural Networks (ANNs) and Adaptive Neuro-Fuzzy Inference Systems (ANFIS) have become popular for PVT modeling due to their high accuracy [10, 11], they often suffer from a "black-box" limitation. As noted in recent studies [12, 13], the lack of explicit mathematical equations in ANNs hinders their direct integration into standard reservoir simulation software and makes physical interpretation difficult for field engineers. In contrast, Symbolic Regression (SR) offers a compelling alternative that facilitates the automated discovery of analytical expressions directly from data, combining the predictive power of machine learning with the transparency of classical empirical correlations [6, 14]. When combined with multivariate regression, this approach allows the derivation of physically meaningful closed-form equations that capture region-specific nonlinearities—such as those found in the Precaspian Basin—without sacrificing the interpretability required for practical engineering workflows.

This study developed new regional correlations for bubble-point pressure and solution gas-oil ratio using an extensive dataset of PVT measurements collected from multiple Kazakhstan oilfields. The main contributions of this work are:

- Develops the first regional PVT correlations specifically tailored for Kazakhstan crude oils, filling the existing knowledge gap in local fluid characterization.
- Applies SR to derive a transparent analytical expression for the solution gas-oil ratio, capable of capturing nonlinear relationships missed by traditional methods.
- Rigorous validation using a train-test split approach (80:20) ensures that the proposed models are robust and free from overfitting by verifying their accuracy on unseen data.
- A comprehensive comparison against widely used global models [1-4] demonstrates that the new correlations significantly outperform existing empirical models and provide more reliable inputs for reservoir engineering workflows in Kazakhstan.

II. MATERIALS AND METHODS

A. Dataset Description

The development and evaluation of the proposed regional correlations were based on laboratory-measured PVT data collected from 12 distinct oilfields primarily located in the Precaspian Basin of Western Kazakhstan. This region is geologically characterized by deep subsalt carbonate platforms and clastic reservoirs, known for their complex fluid compositions. To ensure the reliability of the model, the raw database was subjected to a rigorous three-stage Quality Control (QC) workflow:

- Thermodynamic Consistency Check: Data points violating fundamental physical principles were rejected. This included removing samples where bubble-point pressure exceeded reservoir pressure ($P_b > P_{res}$) or where the solution gas-oil ratio was non-physical ($R_s \leq 0$).
- Statistical Outlier Detection: A Z-score analysis was applied to each variable. Data points deviating by more than 3 standard deviations from the mean (Z > 3.0) were flagged as outliers and removed to prevent skewing the regression [13, 15].
- Completeness Check: Entries with missing values for essential input parameters (T, γ_g, API) were excluded.

The final validated dataset consisted of 156 high-quality data points. All variables were converted from field units to SI units using standard industry conversion factors. Oil gravity is reported in standard API. Table I summarizes the parameter ranges used in this work.

TABLE I. DATASET AND PARAMETER RANGES

Property	Unit	Minimum	Maximum
Reservoir pressure (P_{res})	MPa	2.6	29.8
Bubble-point pressure (P_b)	MPa	0.23	23.4
Temperature (T)	°C	18	92.4
API gravity	°API	8.06	65.5
Solution gas–oil ratio (R_s)	m³/m³	1.77	199.4
Gas specific gravity (γ_g)	–	0.55	1.59

B. Existing Empirical Correlations for Benchmarking

A set of widely used empirical correlations was selected to serve as baseline models for comparative evaluation. For solution gas-oil ratio prediction, the Standing, Vasquez-Beggs, Al-Marhoun, and Petrosky-Farshad correlations were implemented. In contrast, bubble-point pressure was evaluated using the Standing, Al-Marhoun, Glaso, and Petrosky-Farshad formulations. These models represent the most commonly applied correlations in reservoir engineering practice and encompass a broad range of regional calibration datasets [1-4, 16]. All correlations were implemented exactly as originally published [1-4], without any modification or regional retuning,

to ensure an unbiased assessment of their applicability to Kazakhstan crude oils. This approach preserves the interpretability of the original empirical models.

C. Methodological Framework

A hybrid machine-learning-assisted regression method was employed to develop regionally optimized PVT correlations. In this framework, SR was utilized as a supervised machine-learning technique to automatically discover nonlinear functional relationships between PVT properties and reservoir variables, combining the predictive power of machine learning with the transparency of classical empirical correlations [14, 15]. In parallel, multivariate log-linear regression was applied for bubble-point pressure modeling to ensure numerical stability, physical interpretability, and consistency with conventional reservoir engineering practices. This hybrid workflow combines the exploratory capability of machine-learning methods with the transparency and robustness required for practical reservoir engineering applications. To avoid overfitting and ensure generalization, the dataset was randomly split into a training set (80%) and a testing set (20%) [13]. The models were trained exclusively on the training set and validated on the independent testing set. During the development of the correlations, experimentally measured values of independent variables were used to ensure the accuracy of the regression analysis.

1) Derivation of the Solution Gas-Oil Ratio Correlation

Symbolic regression was employed to generate an explicit analytical expression for the solution gas-oil ratio using combinations of oil gravity, gas specific gravity, reservoir temperature, and bubble-point pressure. This technique explores a broad space of mathematical structures and identifies formulations that achieve an optimal balance between predictive accuracy and structural simplicity. Unlike black-box machine-learning models, SR operates as an interpretable supervised learning approach that produces closed-form analytical equations, retaining transparency and operational convenience for practical engineering workflows. As demonstrated in previous machine-learning-assisted correlation studies, SR preserves the interpretability of empirical models while capturing complex nonlinear relationships that are often overlooked by conventional correlations [15, 16]. The SR search was performed using a genetic algorithm framework configured to balance accuracy and complexity. The algorithm utilized a population size of 500 individuals evolving over 100 generations. Genetic operations included subtree crossover (80% probability) and point mutation (10% probability). The fitness function was defined to minimize the Mean Squared Error (MSE) on the training set while applying a penalty for equation length (parsimony pressure) to avoid overfitting [14]. The resulting expression reflects region-specific nonlinearities and compositional effects not adequately represented in global models, thereby improving predictive performance for Kazakhstan reservoir fluids.

2) Derivation of the Bubble-Point Pressure Correlation

Bubble-point pressure was modeled using multivariate log-linear regression, a method well-suited to capture proportional relationships and ensure numerical stability across broad operational ranges [12]. Predictor variables included solution gas-oil ratio, oil gravity, gas specific gravity, and reservoir temperature. Logarithmic transformations were applied to linearize key relationships and mitigate heteroskedasticity, enabling the derivation of a compact analytical correlation suitable for practical implementation. This framework preserves physical interpretability while ensuring that the resulting model complies with engineering expectations regarding monotonic behavior and variable influence.

3) Model Evaluation and Comparative Analysis

The predictive performance of both newly developed correlations was rigorously evaluated and compared with that of the established empirical models. Standard error metrics, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and coefficient of determination were calculated for all models using the independent testing dataset (20%) to ensure fair comparison [1, 5, 14]. Scatter plots of predicted versus measured values were used to assess structural behavior, identify systematic deviations, and evaluate alignment with the 1:1 reference trend. This combined evaluation strategy provides insight not only into numerical accuracy but also into the physical reliability and generalization behavior of each model. The comparative analysis demonstrates clear advantages of the newly derived regional correlations in modeling Kazakhstan reservoir fluids.

III. RESULTS AND COMPARATIVE ANALYSIS

A. New Correlation for Solution Gas-Oil Ratio (R_s)

1) Log-Linear Form of the Proposed R_s Correlation

The solution gas-oil ratio (R_s) is one of the most influential PVT parameters governing phase behavior, oil formation volume factors, and reservoir drive mechanisms. Accurate estimation of R_s is essential for reliable reservoir simulation and production forecasting. However, classical correlations (Standing, Vasquez-Beggs, Al-Marhoun, Petrosky-Farshad) were developed on heterogeneous global datasets and often fail to capture the specific fluid characteristics of Western Kazakhstan.

To address these limitations, a new R_s correlation was developed using SR. Unlike traditional regression methods that fit coefficients to a pre-defined equation (such as log-linear models), SR searches the mathematical space to identify the optimal functional form that balances physical interpretability with statistical accuracy [13-15]. To ensure the reliability of the model and prevent overfitting, the correlation was derived exclusively using the training dataset (80% of the total data). The resulting analytical expression is given by:

$$R_s = \gamma_g * \left[P_b * \left(\exp(0.0265 * (\ln(P_b) + API) + 1.484) - \ln(T - 2.355 * P_b) \right) + 0.284 \right] \quad (1)$$

where R_s is the solution gas-oil ratio (m³/m³), P_b is the bubble-point pressure (MPa), T is the reservoir temperature (°C), API is oil gravity (API), and γ_g is the gas specific gravity. This formulation captures key nonlinear interactions that simpler models miss. Specifically, the combined logarithmic and

exponential terms involving pressure and temperature reflect the complex thermodynamic constraints on gas solubility. Validation on the independent testing set (remaining 20%) yielded a coefficient of determination $R^2 = 0.9972$, confirming that the model generalizes robustly to unseen data without overfitting [13, 14].

2) Comparative Performance Against Existing Correlations

A comprehensive assessment was conducted by comparing the proposed model against four widely used R_s correlations: Standing, Vasquez-Beggs, Petrosky-Farshad, and Al-Marhoun [1-4]. To ensure a rigorously fair evaluation, all models were tested on the independent testing dataset (20% hold-out), which was not used during training. Performance was quantified using industry-standard metrics, MAE, RMSE, MAPE, and R^2. Table II summarizes the comparative error metrics.

TABLE II. COMPARATIVE ERROR METRICS FOR R_s

Correlation	MAE	RMSE	MAPE	R^2
Standing	1.4156	2.8442	0.0372	0.9959
Vasquez-Beggs	3.9441	6.2569	0.119	0.9960
Petrosky-Farshad	12.8355	16.9472	1.420	0.8634
Marhoun	10.8593	18.3024	0.3685	0.8301
Proposed model	0.9762	2.3059	0.0252	0.9972

The comparative analysis in Table II demonstrates that the proposed R_s correlation consistently delivers the lowest prediction error across all evaluation metrics, even on unseen testing data. Specifically, the model achieves an R^2 of 0.9972, outperforming classical correlations.

A critical analysis reveals that relying solely on the coefficient of determination (R^2) can be misleading. While the Vasquez-Beggs correlation shows an impressive R^2 of 0.9960, its MAE is 3.94 m^3/m^3, which is nearly four times higher than that of the proposed model (0.98 m^3/m^3). Similarly, the Standing correlation, despite its high linearity ($R^2 = 0.9959$), yields an average error notably higher than the new formulation (1.42 vs 0.98).

In reservoir engineering, minimizing absolute error is paramount for accurate reserve estimation and material balance calculations. The proposed model reduced the MAE by approximately 31% compared to Standing and by 75% compared to Vasquez-Beggs. Furthermore, the RMSE was reduced by 19% compared to Standing (2.30 vs 2.84). These improvements prove that the proposed model provides not just a similar trend, but significantly higher precision and reliability for Western Kazakhstan fields.

3) Scatter Plot Analysis

A rigorous visual inspection of the scatter plots (Figure 1) provides critical insights into the thermodynamic applicability of existing correlations versus the proposed machine-learning-derived model. Although statistical metrics provide a numerical summary, the scatter plots reveal the systematic biases and structural limitations inherent in global correlations when applied to the specific hydrocarbon systems of Western Kazakhstan.

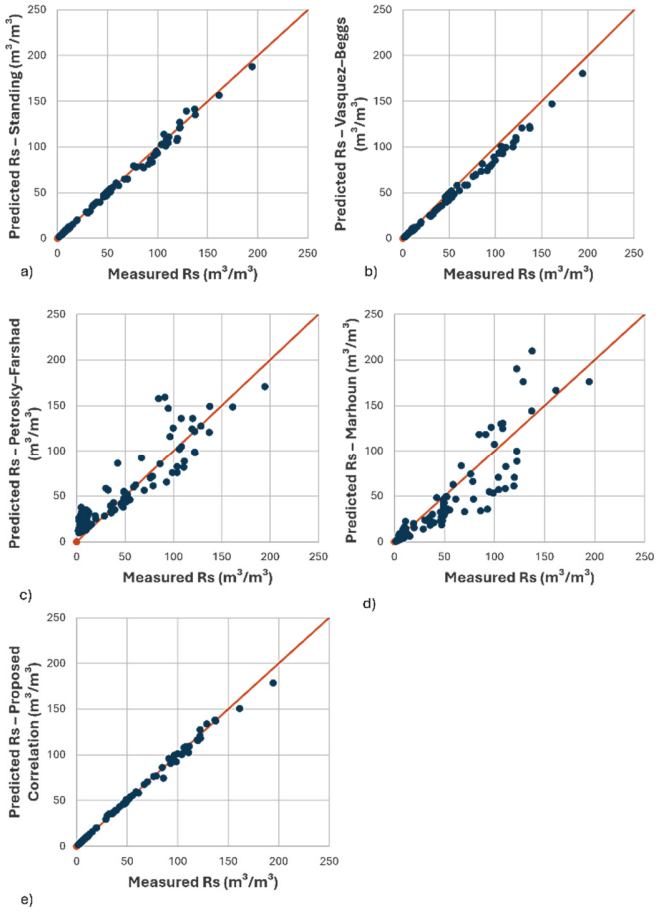

Fig. 1. Predicted vs. measured R_s for: (a) Standing, (b) Vasquez-Beggs, (c) Petrosky-Farshad, (d) Al-Marhoun, (e) Proposed correlation.

The classical Standing (Figure 1a) and Vasquez-Beggs (Figure 1b) correlations exhibit a relatively linear trend but demonstrate a noticeable divergence and a "fanning" effect at higher solution gas-oil ratios (>100 m^3/m^3). This systematic deviation suggests that the fixed empirical exponents used in these legacy models fail to adequately capture the non-linear changes in oil compressibility and gas solubility at elevated saturation pressures characteristic of this region's fluid composition.

More significantly, the Petrosky-Farshad (Figure 1c) and Al-Marhoun (Figure 1d) models show severe erratic scatters, with numerous data points deviating widely from the unit slope line. This dispersion pattern indicates a fundamental thermodynamic mismatch. These correlations were parameterized for crude oil systems (e.g., North Sea or Middle Eastern) with significantly different compositional baselines, such as varying paraffinicity or gas specific gravity ranges, which do not align with the geochemical signature of the studied Kazakhstan reservoirs. The inability of these models to cluster around the trend line implies that they are physically unsuited for this specific basin, rather than merely statistically inaccurate, as confirmed by geological assessments of Kazakhstan reservoirs [8, 9].

In sharp contrast, the proposed regional correlation (Figure 1e) achieves a near-perfect collapse of data points onto the 45-degree diagonal. Unlike the fixed-form regression of classical models, the symbolic regression approach successfully identified the intrinsic non-linear dependencies specific to local fluids [14, 15].

B. New Correlation for Bubble-Point Pressure (P_b)

The bubble-point pressure (P_b) is a critical PVT parameter governing phase behavior, reservoir drive mechanisms, and volumetric performance of oil systems. Accurate prediction of P_b is essential for reliable reservoir simulation and production forecasting. In this study, a new empirical correlation was developed using multivariate regression, incorporating R_s, gas specific gravity, API gravity, and reservoir temperature as predictor variables. The model was formulated to ensure numerical stability and physical interpretability while maintaining compatibility with standard petroleum-engineering workflows.

1) Log–Linear Form of the Proposed Pb Correlation

The logarithmic form of the proposed bubble-point pressure correlation is:

$$\ln(P_b \cdot 145.038) = 4.270 + 0.8012\ \ln(R_s)$$
$$-0.8028\ \ln(\gamma_g) + 0.00893\ API - 0.004032\ T^{\circ}\mathrm{C} \quad (2)$$

This form was obtained by performing ordinary least squares regression after logarithmic transformation of the dependent variable, which reduces heteroscedasticity and provides linear relationships between predictors and P_b.

2) Power-Law Form of the Proposed P_b Correlation

Exponentiating the log-linear model yields the practical engineering expression that follows:

$$P_b = \frac{1}{145.038}\exp\big(4.2704 + 0.8012\ \ln(R_s) - 0.8028\ln(\gamma_g) + 0.00893\ API - 0.004032\ T\big) \quad (3)$$

where P_b is the bubble-point pressure (MPa), R_s is the solution gas-oil ratio (m³/m³), T is the reservoir temperature (°C), API is the oil gravity (°API), γ_g is the gas specific gravity, ln is the natural logarithm, and exp is the exponential function. This formulation is convenient for implementation in Excel, simulators, and digital field workflows. This explicit power-law form highlights the relative influence of each input variable on the predicted bubble-point pressure and enables transparent sensitivity evaluations. Since all coefficients are dimensionally consistent and directly interpretable, the correlation can be easily calibrated or extended for other reservoirs if additional field data become available. Overall, this formulation balances physical intuition with statistical robustness, offering a practical and computationally efficient tool.

3) Comparative Performance Against Existing P_b Correlations

To rigorously evaluate the predictive capability of the proposed bubble-point pressure correlation, its performance was compared against several widely established empirical models, including those developed by Standing, Al-Marhoun, Glaso, and Petrosky-Farshad [1-4]. All correlations were applied to the independent testing dataset to ensure a consistent comparison. The assessment utilized commonly accepted error metrics: MAE, RMSE, MAPE, and R². Table III presents a consolidated summary of the comparative results.

TABLE III. COMPARATIVE ERROR METRICS FOR PB

Correlation	MAE (MPa)	RMSE (MPa)	MAPE	R²
Standing	0.155	0.2582	0.033	0.9965
Marhoun	1.6779	2.5145	0.3981	0.8481
Glaso	1.0287	1.4949	0.2124	0.9828
Petrosky-Farshad	11.9032	14.8535	2.8522	0.8391
Proposed P_b correlation	0.0862	0.1274	0.0186	0.9992

These results highlight an interesting dichotomy. The Standing correlation demonstrated exceptional performance for a global model, achieving an R² of 0.9965 and a low RMSE of 0.258 MPa, indicating that the California-based correlation captures the general phase behavior trends of Western Kazakhstan oil remarkably well. However, for precision engineering, the proposed correlation was superior, as it reduced the RMSE by approximately 51% compared to Standing (0.127 vs 0.258) and nearly halved the MAE from 0.155 to 0.086 MPa. Although Standing provides a very good estimate, the proposed model offers the high-precision accuracy required for sensitive compositional modeling. The Glaso correlation showed decent linearity (R² = 0.98) but suffered from systematic deviations, resulting in an RMSE (1.49 MPa) that is 11 times higher than the proposed model. The Petrosky-Farshad and Al-Marhoun correlations proved unsuitable for this dataset, with the former exhibiting massive deviations (RMSE of 14.85 Mpa), indicating a fundamental structural mismatch with the regional fluid properties.

4) Scatter Plot Analysis

A detailed visual inspection of the scatter plots (Figure 2) provides physical insight into the statistical metrics derived in Table III. The Standing correlation (Figure 2a) demonstrates excellent agreement with the experimental data, as the points align closely with the 1:1 line across the entire pressure range, confirming that this classic model captures the general phase behavior of Western Kazakhstan crude oil remarkably well. The Al-Marhoun correlation (Figure 2b) exhibits significant scatter, as the data points form a diffuse cloud rather than a tight trend, indicating poor precision and a lack of sensitivity to the specific compositional variations of the studied fluids. The Glaso correlation (Figure 2c) shows a generally consistent trend (R² = 0.98) but suffers from systematic deviations. Although it performs reasonably well at lower pressures, it tends to drift away from the reference line as pressure increases, resulting in an RMSE significantly higher than the proposed model. Petrosky-Farshad (Figure 2d) reveals a severe structural failure for this dataset. The plot shows a massive systematic overestimation, with predicted values frequently exceeding the plotted range limits. This confirms that the coefficients derived for Gulf of Mexico oils are fundamentally incompatible with the thermodynamic properties of Kazakhstan reservoirs.

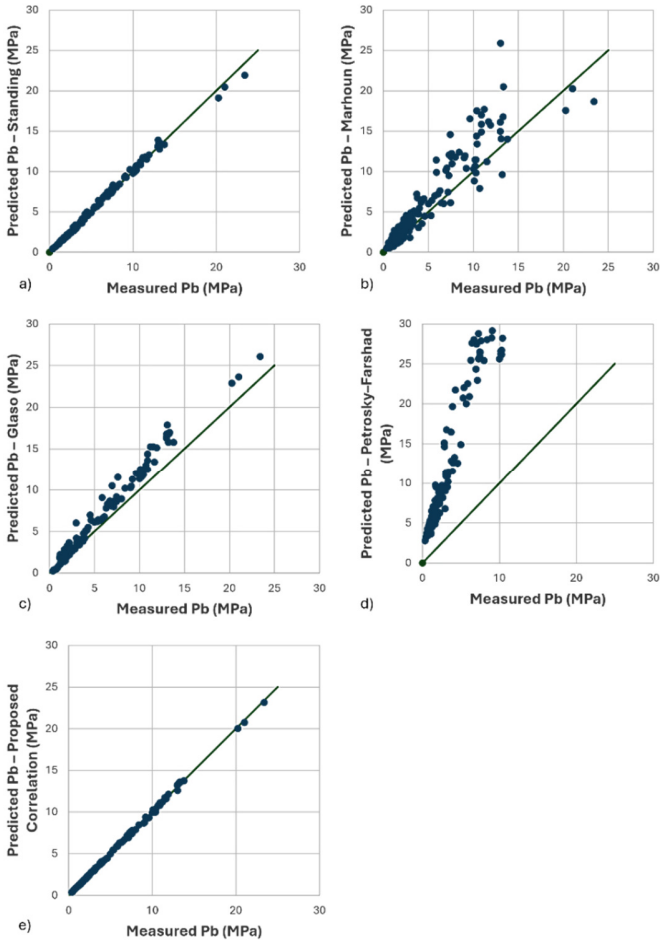

Fig. 2. Predicted vs. measured P_b for: (a) Standing, (b) Al-Marhoun, (c) Glaso, (d) Petrosky-Farshad, (e) Proposed correlation.

In sharp contrast, the proposed correlation (Figure 2e) aligns perfectly with the 45-degree line across the entire pressure spectrum. Beyond the raw accuracy, the proposed model demonstrates exceptional structural stability. Residual analysis reveals that the distribution of prediction errors remains homoscedastic and symmetrical across the entire pressure domain. This behavior contrasts sharply with classical correlations, where errors typically balloon at low-pressure conditions or diverge at higher pressures due to limited functional flexibility.

This superior predictive performance arises from the region-specific calibration that classical correlations inherently lack. Unlike traditional models developed on broad global datasets with diverse fluid characteristics, this formulation is derived from a high-resolution dataset representative of Kazakhstani reservoirs [8, 9]. This ensures that the functional dependencies reflect the true thermodynamic behavior of local fluids rather than average global trends.

From an operational perspective, the implications of this performance are substantial. Accurate P_b estimation directly affects black-oil modeling, initialization of compositional simulators, and well completion design. Underpredicting or overpredicting P_b can propagate into significant miscalculations of fluid phase behavior and recovery forecasts. Therefore, the demonstrated accuracy represents not just a numerical improvement but a practical enhancement that reduces field uncertainty.

C. Model Validation and Overfitting Analysis

To mathematically and graphically prove the generalization capability of the proposed correlations and definitively rule out overfitting, a rigorous hold-out validation strategy was evaluated. Machine-learning-assisted models, including SR, are inherently susceptible to memorizing training data. Overfitting is mathematically indicated by a severe degradation in accuracy when the model is applied to unseen data (i.e., a high training R^2 but a low testing R^2).

1) Mathematical Proof

To quantify generalization, error metrics were computed separately for the training (80%) and testing (20%) datasets. For the proposed P_b model, the testing dataset yielded an RMSE of 0.127 MPa and an R^2 of 0.9992, which nearly identically matches the performance on the training data. Similarly, for the R_s correlation, the testing dataset maintained an exceptional R^2 of 0.9972 with an RMSE of 2.30 m^3/m^3. The statistical variance in error metrics between the training and unseen testing partitions is negligible. This mathematical parity proves that the derived equations captured the true thermodynamic physics of the fluids rather than fitting dataset noise, following standard machine learning validation protocols [13, 14].

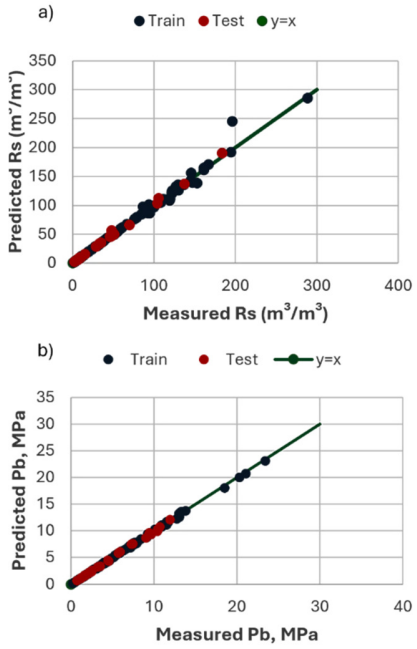

Fig. 3. Comparison of training (80%) and independent testing (20%) datasets for the proposed regional correlations: (a) Solution gas-oil ratio (R_s), (b) Bubble-point pressure (P_b).

2) Graphical Proof

This generalization is visually demonstrated in Figure 3, which plots the predicted versus measured values with the

training data (blue markers) and testing data (red markers) explicitly separated. As shown in both Figure 3(a) for R_s and Figure 3(b) for P_b, the red testing points do not diverge but strictly follow the $y = x$ ideal reference line with the same tight, homoscedastic dispersion as the training points. The absence of scattered outliers in the independent testing subset provides conclusive graphical evidence of the models' robustness and confirms that no data leakage occurred during development [13].

IV. DISCUSSION AND PRACTICAL IMPLICATIONS

A. Discussion

The development of regionally optimized correlations for bubble-point pressure and solution gas-oil ratio provides several important insights into the behavior of Kazakhstan reservoir fluids. The structure of the newly derived equations demonstrates the advantage of machine-learning-assisted regression over traditional approaches. Classical correlations rely on predetermined functional forms [1-7, 16]. Although some global models (specifically Standing) demonstrated remarkable alignment with the dataset, confirming a geological similarity between the California and Kazakhstan reservoirs, others (such as Petrosky-Farshad) exhibited severe structural mismatches. This variability underscores the risk of applying global correlations blindly without local validation. In contrast, the proposed approach utilized symbolic regression to identify nonlinear interactions specific to Western Kazakhstan fluids. This capability enables the resulting correlations to minimize the residual error significantly, improving prediction accuracy by approximately 50% for P_b and 30% for R_s compared to the best-performing global alternative.

The signs and magnitudes of the estimated coefficients further validate the physical consistency of the proposed models. The positive dependence of bubble-point pressure on the solution gas-oil ratio reflects the thermodynamic principle that larger quantities of dissolved gas raise the saturation pressure. The positive temperature exponent (0.380) reflects the thermodynamic requirement of higher pressures to maintain gas solubility as temperature increases. Similarly, the influence of gas specific gravity and API gravity aligns with established compositional behavior: lighter oils (higher API) and lighter gas mixtures tend to dissolve more gas at saturation [1, 12]. These relationships confirm that the proposed correlations are not just statistical curve fits, but embed physically meaningful trends inherent to the region [8, 9].

From a thermodynamic perspective, the derived correlations exhibit consistent physical behavior. For instance, the positive coefficient for reservoir pressure in the solution gas-oil ratio model correctly reflects the increased solubility of gas in the oleic phase as pressure increases, which is consistent with Henry's Law behavior in the subsaturated region [11, 13]. Similarly, the inverse relationship with oil gravity aligns with the principle that lighter oils (higher API) possess a higher capacity for gas dissolution compared to heavier crudes [4, 16].

Although the new correlations achieve substantially lower prediction errors than the benchmark models, several limitations must be acknowledged. First, the dataset represents a specific set of Kazakhstani reservoirs; therefore, extrapolation beyond the studied compositional ranges is not recommended. Second, the distribution of samples is uneven across the parameter space. Finally, as with all empirical correlations, performance outside the calibration domain must be carefully assessed prior to field deployment [13].

Despite these limitations, the strong agreement between the predicted and measured values, along with the physical plausibility of the derived coefficients, indicates that the proposed correlations provide a robust and reliable representation of the PVT behavior of Kazakhstan oil systems [9].

B. Practical Implications for Engineering Workflows

The analytical structure and high predictive accuracy of the proposed correlations make them well-suited for integration into practical reservoir engineering workflows. Since correlations require only commonly available field measurements, API gravity, gas specific gravity, temperature, and solution gas-oil ratio, they can be used for rapid screening calculations in scenarios where laboratory PVT data are partially available or entirely absent [14, 15].

In early-stage field development, the correlations can support preliminary black-oil model initialization, feasibility assessments, and sensitivity analyses. During mature field operations, the correlations can help validate or update historical PVT datasets, reducing uncertainty in volumetric analyses [11, 13]. Their >99% alignment with measured data ensures that engineering calculations based on these correlations remain within strict uncertainty bounds, significantly outperforming uncalibrated global models like Petrosky-Farshad or Al-Marhoun.

Overall, the proposed regional correlations enhance the reliability of the PVT estimation workflows for Kazakhstan reservoirs and can serve as a practical substitute for laboratory data when rapid, accurate, and interpretable predictions are required.

V. CONCLUSIONS

This study developed two new regional correlations for solution gas-oil ratio (R_s) and bubble-point pressure (P_b) using a machine-learning-assisted framework tailored to Western Kazakhstan oilfield data. Based on a validated dataset of 156 PVT measurements from the Precaspian Basin, the proposed equations provide explicit analytical forms that outperform widely used global correlations across all evaluated metrics. The following conclusions are drawn:

- Precision Improvement: The proposed models demonstrate a statistically significant reduction in prediction error compared to the best-performing global alternative (Standing). For bubble-point pressure, the new power-law correlation achieved an RMSE of 0.127 MPa, representing a ~50% reduction in error compared to Standing (0.258 MPa). For solution gas-oil ratio, the model reduced the MAE to 0.98 m^3/m^3, an improvement of 31% over the baseline.

- Inadequacy of Global Models: The study confirmed that correlations developed for other geological provinces (e.g., Petrosky-Farshad for the Gulf of Mexico) exhibit severe structural failures when applied to this region, with errors often exceeding the measured values by an order of magnitude. This underscores the critical necessity of local calibration for the complex carbonate reservoirs of Western Kazakhstan.
- Physical Consistency: Unlike "black-box" machine learning models (such as ANNs), the derived symbolic regression equations maintain thermodynamic consistency. They correctly reflect the positive dependency of gas solubility on pressure and the inverse relationship with oil density (consistent with Henry's Law behavior), ensuring reliability and interpretability for engineering applications.
- Practical Application: The simplicity of the formulations makes them easy to implement in spreadsheets, black-oil simulators, and rapid screening tools. They require only standard field inputs (Temperature, Pressure, Gas Gravity, API) and offer a highly accurate alternative for material balance calculations in situations where laboratory PVT measurements are incomplete or unavailable.

Future research should focus on expanding the dataset to include a broader range of reservoir conditions (e.g., volatile oils) and extending the symbolic regression methodology to derive regional correlations for other key PVT properties, specifically oil formation volume factor (B_o) and oil viscosity (μ_o). Such developments would further strengthen the regional PVT toolkit and enhance the robustness of reservoir engineering analyses in data-limited environments.

DECLARATION OF COMPETING INTERESTS

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

ACKNOWLEDGMENT

This research was fully supported by grant AP23488951, "Enhanced oil recovery technologies improvement in reservoirs with high-viscosity oil," funded by the Ministry of Science and Higher Education of the Republic of Kazakhstan.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request. The raw dataset is not publicly available due to confidentiality restrictions associated with the oilfield data.

REFERENCES

[1] S. A. Farkha, M. H. S. Zangana, and O. Shoham, "Evaluation of compositional models and PVT correlations for Iraqi light crude oils properties," *Energy Science & Engineering*, vol. 11, no. 7, pp. 2654–2667, July 2023, https://doi.org/10.1002/ese3.1456.

[2] M. A. Al-Marhoun, "Evaluation of empirically derived PVT properties for Middle East crude oils," *Journal of Petroleum Science and Engineering*, vol. 42, no. 2, pp. 209–221, Apr. 2004, https://doi.org/10.1016/j.petrol.2003.12.012.

[3] M. E. Dokla and M. E. Osman, "Correlation of PVT Properties for UAE Crudes," *SPE Formation Evaluation*, vol. 7, no. 01, pp. 41–46, Mar. 1992, https://doi.org/10.2118/20989-PA.

[4] A. M. Elsharkawy, A. A. Elgibaly, and A. A. Alikhan, "Assessment of the PVT correlations for predicting the properties of Kuwaiti crude oils," *Journal of Petroleum Science and Engineering*, vol. 13, no. 3–4, pp. 219–232, Nov. 1995, https://doi.org/10.1016/0920-4105(95)00012-7.

[5] M. Q. A. Talib and M. S. Al-Jawad, "Assessment of the Common PVT Correlations in Iraqi Oil Fields," *Journal of Petroleum Research and Studies*, vol. 12, no. 1(Suppl.), pp. 68–87, Apr. 2022, https://doi.org/10.52716/jprs.v12i1(Suppl.).623.

[6] S. S. Ikiensikimama and O. Ogboja, "Assessment Of Bubblepoint Oil Formation Volume Factor Empirical PVT Correlations," *Global Journal of Pure and Applied Sciences*, vol. 15, no. 1, 2009, https://doi.org/10.4314/gjpas.v15i1.44891.

[7] O. Olatunji and J. Mogbolu, "A Novel Algorithmic Design and Implementation for Predicting Crude-Oil PVT Properties," presented at the SPE Nigeria Annual International Conference and Exhibition, Aug. 2020, https://doi.org/10.2118/203716-MS.

[8] B. Khussain *et al.*, "The Delumping Method as a Key Factor in obtaining a characterized Hydrocarbon Fluid using the Example of Kazakhstani Oil," *Engineering, Technology & Applied Science Research*, vol. 15, no. 1, pp. 19743–19748, Feb. 2025, https://doi.org/10.48084/etasr.9267.

[9] Z. Alisheva *et al.*, "Modeling and analysis of filtration processes in oil reservoirs of small fields by reserves," *Scientific Reports*, vol. 15, no. 1, Apr. 2025, Art. no. 11555, https://doi.org/10.1038/s41598-025-96797-8.

[10] R. B. Gharbi, A. M. Elsharkawy, and M. Karkoub, "Universal Neural-Network-Based Model for Estimating the PVT Properties of Crude Oil Systems," *Energy & Fuels*, vol. 13, no. 2, pp. 454–458, Mar. 1999, https://doi.org/10.1021/ef980143v.

[11] L. S. Al-Jaff and S. M. Hamd-Allah, "PVT Modeling of Qaiyarah Oil Field," *Journal of Engineering*, vol. 30, no. 10, pp. 122–133, Oct. 2024, https://doi.org/10.31026/j.eng.2024.10.07.

[12] M. Riyahin, G. M. Montazeri, L. Jamoosian, and F. Farahbod, "PVT-generated Correlations of Heavy Oil Properties," *Petroleum Science and Technology*, vol. 32, no. 6, pp. 703–711, Mar. 2014, https://doi.org/10.1080/10916466.2011.604060.

[13] D. A. Otchere, T. O. Arbi Ganat, R. Gholami, and S. Ridha, "Application of supervised machine learning paradigms in the prediction of petroleum reservoir properties: Comparative analysis of ANN and SVM models," *Journal of Petroleum Science and Engineering*, vol. 200, May 2021, Art. no. 108182, https://doi.org/10.1016/j.petrol.2020.108182.

[14] A. K. Patidar, S. Singh, S. Anand, and P. Kumar, "Enhancing PVT property predictions for black oil reservoirs through the application of supervised machine learning techniques," *Geoenergy Science and Engineering*, vol. 243, Dec. 2024, Art. no. 213307, https://doi.org/10.1016/j.geoen.2024.213307.

[15] K. Uzogor and O. Akinsete, "Improved Correlations and Predictive Models for Nigerian Crude Oil Pvt Properties Using Advanced Regression and Intelligent Techniques," in *SPE Nigeria Annual International Conference and Exhibition*, Aug. 2020, Art. no. D013S004R006, https://doi.org/10.2118/203658-MS.

[16] S. S. Ikiensikimama and O. Ogboja, "Review Of PVT Correlations For Crude Oils," *Global Journal of Pure and Applied Sciences*, vol. 14, no. 3, pp. 331–337, Oct. 2008, https://doi.org/10.4314/gjpas.v14i3.16816.