

Supplementary file

Quantum machine learning-driven surrogate modeling for efficient multi-objective optimization of CO₂ storage and geothermal energy extraction

Babak Mohammadi¹, Mingjie Chen^{2*}, Mohammad Reza Nikoo³, Ali Al-Maktoumi²

¹ *Hydrology Research Unit, Swedish Meteorological and Hydrological Institute, Norrköping 601 76, Sweden*

² *Water Research Center, Sultan Qaboos University, Muscat 123, Oman*

³ *Department of Civil and Architectural Engineering, Sultan Qaboos University, Muscat 123, Oman*

E-mail address: babak.mohammadi@smhi.se (B. Mohammadi); mingjie@squ.edu.om (M. Chen);

m.reza@squ.edu.om (M. R. Nikoo); ali4530@squ.edu.om (A. Al-Maktoumi)

*Corresponding author: (ORCID: 0000-0002-9955-637X (Mingjie Chen))

Mohammadi, B., Chen, M., Nikoo, M. R., Al-Maktoumi, A. Quantum machine learning-driven surrogate modeling for efficient multi-objective optimization of CO₂ storage and geothermal energy extraction. Advances in Geo-Energy Research, 2025, 18(2): 137-152.

The link to this file is: <https://doi.org/10.46690/ager.2025.11.04>

Supplementary A

A 1. NUFT model framework and spatial discretization

Non-isothermal Unsaturated-saturated Flow and Transport (NUFT) solves a system of coupled nonlinear partial differential equations governing mass conservation for CO₂ and brine phases, energy conservation, and thermodynamic equilibrium relationships using an integrated finite difference method with fully implicit time discretization (Nitao, 1998). The governing equations account for advective and dispersive transport, capillary pressure effects described by van Genuchten relationships ($\alpha = 5.1 \times 10^{-5} \text{ Pa}^{-1}$, $n = 0.46$) (Zhou et al., 2008; Buscheck et al., 2012; Chen et al., 2023), relative permeability functions following Brooks-Corey models, and temperature-dependent fluid properties including CO₂ phase transitions between gaseous, liquid, and supercritical states.

The three-dimensional computational domain represents a fault-bounded reservoir block in North Oman (Chen et al., 2023), measuring $2,000 \text{ m} \times 3,000 \text{ m} \times 20 \text{ m}$ ($X \times Y \times Z$), inclined at a 5° dip angle around the Y-axis with the anchor point (0,0) positioned at 1,500 m depth (Table A 1). The domain is discretized into a structured grid with uniform spacing of $\Delta x = 50 \text{ m}$, $\Delta y = 50 \text{ m}$, and $\Delta z = 1 \text{ m}$, resulting in $40 \times 60 \times 20 = 48,000$ computational cells that provide sufficient resolution to capture thermal fronts and CO₂ plume migration while maintaining computational tractability. A horizontal well doublet configuration is implemented with 1-km long wells placed parallel to the Y-axis at the middle Y-range ($Y = 1,500 \text{ m}$), where the injection well is fixed at the top-left corner ($X = 0$, $Z = 0$) and the production well location varies between 400 and 1,000 m distance for optimization purposes. The reservoir boundaries are treated as no-flow (closed) conditions for all faces except the right boundary ($X = 2,000 \text{ m}$), with its cell volume magnified by 10^6 to represent far-field conditions, while the bottom boundary receives a constant geothermal heat flux of 0.05 W/m^2 , representative of continental crust heat flow (Moridis and Pruess, 1992; Babaei et al., 2016; Chen et al., 2023). The initial conditions and simulation scenarios are described in detail in Appendix A of the supplementary information.

Table. A 1. Reservoir model geometry and discretization parameters.

Parameter	Value	Unit	Description
Domain size ($X \times Y \times Z$)	2,000×3,000×20	m	Fault-bounded reservoir block
Grid spacing ($\Delta x \times \Delta y \times \Delta z$)	50×50×1	m	Uniform structured grid
Number of cells	48,000	-	40×60×20 grid cells Inclination around Y-axis
Reservoir dip	5	degrees	
Anchor depth	1,500	m	Top-left corner depth
Well length	1,000	m	Horizontal well extent
Well placement	$Y = 1,500$	m	Middle of Y-domain

Supplementary B

B 1. Initial conditions and simulation scenarios

The reservoir pressure is initialized with hydrostatic pressure distributions calculated based on depth, with the anchor point (0.0) at 1500 depth is specified as 17 MPa. The initial reservoir temperature is specified at a uniform 90 °C (Zhao et al., 2012). The reservoir formation is initially saturated by 70 vol% CO₂ and 30 vol% brine with 180 ppt of salinity (Table B 1). Cold CO₂ (30 °C) is injected and hot CO₂ is extracted via a production well to a surface power plant. CO₂ circulation between the injection-production well doublet is driven by specified injection overpressure and production pressure fixed at its initial value for up to 100 years or until produced fluid temperature drops below 80 °C. The heterogeneous porosity-permeability fields are generated using Sequential Gaussian Simulation with correlation lengths varying from 100 to 1000 m in horizontal directions and 1-20 m vertically, with permeability derived from porosity using the Kozeny-Carman relationship: $k = 192\phi^3/(1-\phi)^2$, where mean porosity is 0.25 corresponding to mean permeability of 5.3 mD (Rajabi et al., 2021).

Table. B 1. Reservoir properties and operational parameters (Chen et al., 2023).

Parameter	Value/Range	Unit	Description
Initial temperature	90	°C	Whole domain
Heat flux (bottom)	0.05	W/m ²	Basal heat source
Brine salinity	180	ppt	Formation water
Rock density	2,650	kg/m ³	Carbonate formation
Rock specific heat	1,000	J/kg·°C	Thermal capacity
Thermal conductivity	2.1	W/m·°C	Rock matrix
Mean porosity	0.25	-	Geostatistical mean
Mean permeability	5.3	mD	Derived from porosity
Injection temperature	30	°C	Cold CO ₂ injection
Injection overpressure	5-15	MPa	Optimization variable
Well spacing	400-1,000	m	Optimization variable
CPG cutoff temperature	80	°C	Economic threshold

Supplementary C

C 1. Feature selection algorithms

Three feature selection techniques, namely Boruta, Chi-Square, and Pearson correlation, were employed to ensure the identification of the most relevant predictors as inputs for the models. The Boruta method represents a wrapper-based feature selection technique that leverages Random Forest classifier as its core computational engine for determining feature relevance (Kursa & Rudnicki, 2010). This algorithm operates by generating shadow attributes (randomized duplicates of original features) and conducting comparative analysis to distinguish truly important variables from those that may appear significant due to random chance. The selection criterion employed by Boruta is the z-score metric, and features achieving Z-scores above a predefined threshold are classified as statistically significant and retained as important variables. The Chi-square test was utilized to assess the independence between categorical features and the target variable (Plackett, 1983). Since the original continuous variables required discretization, features were categorized into discrete bins before analysis. The Chi-square statistic was calculated as (Pandis, 2016):

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \quad (1)$$

where χ^2 is the Chi-square statistic, O_i is the observed frequency in category i , E_i is the expected frequency in category i under the null hypothesis of independence, and k is the number of categories.

Higher Chi-square values indicate stronger dependence between the feature and target variable, suggesting greater discriminative power. Features with statistically significant Chi-square values (higher scores) were considered more important for classification tasks. Pearson correlation coefficient was employed to measure the linear relationship between each feature and the target variable. The correlation coefficient quantifies the strength and direction of linear association, calculated using the following equation (Pearson, 1895; Benesty et al., 2009):

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}} \quad (2)$$

where r is the correlation coefficient, X_i and Y_i are individual data points, $\bar{X}$ and $\bar{Y}$ are the means of variables X and Y , respectively, and n is the sample size. The correlation coefficient ranges from -1 to +1, where values closer to ± 1 indicate stronger linear relationships. To enable comparability among methods, the importance weights of all features were normalized to a $[0, 1]$ scale, and features with standardized weights ≥ 0.75 were selected as significant predictors.

C 2. Quantum machine learning models and classical baseline development

The QML framework employs a sophisticated data encoding pipeline that transforms classical features into quantum-compatible representations through angle encoding and quantum state preparation (Schuld et al., 2015). To prepare quantum-compatible input features, the original dataset underwent power transformation using the Yeo-Johnson method (Yeo & Johnson, 2000), followed by principal component analysis (PCA) to reduce dimensionality while preserving variance (Jolliffe & Cadima, 2016). The resulting features were then standardized and transformed using the arctan function with $\pi/2$ offset to map all values into the range of $[0, \pi]$, ensuring optimal performance for quantum rotation gate operations (Tudisco et al., 2025). The dataset was split using an 80/20 strategy, where 80% of the samples were used for training and the remaining 20% for testing. In this study, two QML-based models were applied as surrogate models, including a QNN model (Beer et al., 2020) and a HQER model.

The QNN represents a hybrid quantum-classical approach that mimics the structure of classical neural networks using quantum circuits (Liang et al., 2021; Arthur & Date, 2022). This model utilizes the principles of quantum superposition and entanglement to process input data in ways that may offer advantages over classical neural networks for certain types of problems. The QNN architecture consists of a data encoding layer that maps input features to quantum states using rotation gates (RY), followed by multiple variational quantum layers. Each variational layer contains parameterized rotation gates (RX , RY , RZ) applied to each qubit, representing the quantum equivalent of weight transformations in classical neurons, followed by CNOT entanglement gates that create quantum correlations between qubits, analogous to connections between neurons in classical networks. The QNN uses 4 qubits (or fewer if the input has fewer features) and 4 quantum layers, resulting in 48 quantum parameters (3 rotations per qubit per layer). The classical linear layer adds 5 additional parameters (4 weights + 1 bias). The choice of 4 qubits was guided by consideration of classical simulation tractability and computational efficiency for the iterative training process required for surrogate model development, and also by empirical validation through preliminary experiments showing that 4 qubits combined with PCA dimensionality reduction and proper feature encoding can achieve acceptable prediction performance for the CPG optimization task. The model is trained using a gradient-free optimization approach over 500 epochs, with random perturbations (Gaussian noise with standard deviation 0.1) added to parameters. Quantum weights are initialized from a Gaussian distribution (mean=0, std=0.1), while classical weights are uniformly initialized between -1 and 1. The quantum circuit is implemented using PennyLane with the default.qubit simulator, and target values are normalized during training and denormalized for evaluation.

The HQER model adapts the Quantum Approximate Optimization Algorithm (QAOA) framework (Farhi et al., 2014) for continuous regression tasks by treating feature extraction as an optimization problem. This model uses a parameterized quantum circuit to extract quantum features from input data, which are then combined with classical features in an ensemble of classical regression models

to make predictions. The HQER architecture begins by initializing qubits in superposition using Hadamard gates, followed by multiple QAOA layers with alternating problem Hamiltonians (cost layers) that encode input data and mixer Hamiltonians (mixing layers) that enhance exploration (Bengtsson et al., 2020; Majumder et al., 2024). The quantum circuit extracts features through Pauli-Z measurements on each qubit, which are then concatenated with the original (scaled) classical features. These combined features are fed into an ensemble of classical regression models (Ridge, ElasticNet, Random Forest, and Gradient Boosting), with final predictions obtained through a weighted average of the ensemble outputs. The HQER model uses 4 qubits (or fewer if the input has fewer features) and 3 QAOA layers, with 3 alpha parameters (range $[0, 2\pi]$) for the problem Hamiltonian and 3 beta parameters (range $[0, \pi]$) for the mixer Hamiltonian. The quantum circuit parameters are optimized over 500 iterations using a gradient-free approach with Gaussian noise ($\text{std}=0.2$). The ensemble consists of Ridge regression ($\alpha=0.5$, $\text{weight}=0.3$), ElasticNet ($\alpha=0.1$, $\text{ll_ratio}=0.5$, $\text{weight}=0.25$), Random Forest ($n_estimators=200$, $\text{max_depth}=8$, $\text{weight}=0.25$), and Gradient Boosting ($n_estimators=200$, $\text{learning_rate}=0.1$, $\text{max_depth}=5$, $\text{weight}=0.2$). Original features are standardized using StandardScaler before combining with quantum features, and the quantum circuit is implemented using PennyLane with the default qubit simulator.

The adoption of quantum feature encoding in this study is motivated by several theoretical advantages that are particularly relevant for capturing the complex, nonlinear physics governing CPG systems. First, quantum feature encoding through rotation gates and entanglement operations enables the creation of nonlinear feature transformations in quantum state space, where quantum superposition and parameterized unitary operations can generate complex feature representations that complement classical methods. While 4-qubit circuits operate at a modest scale that does not demonstrate exponential quantum advantage, the quantum operations can still create nonlinear feature transformations and capture higher-order feature interactions that may be valuable for capturing the complex physics of CPG systems. Classical polynomial feature expansion could theoretically achieve similar interaction terms, but the number of features grows combinatorially, leading to the curse of dimensionality, whereas quantum entanglement implicitly encodes these interactions in the quantum state structure without explicit feature enumeration. Third, the variational nature of quantum circuits, where parameterized gates are optimized during training, provides a form of learnable nonlinear basis function that adapts to the specific structure of CPG simulation data, analogous to kernel methods but with kernels defined by quantum operations rather than fixed classical functions. This adaptive quantum kernel approach may be particularly effective for problems where the optimal feature representation is unknown a priori and differs substantially from standard polynomial or radial basis function kernels commonly employed in classical ML. Fourth, the hybrid quantum-classical architecture (especially in HQER) combines quantum feature extraction with classical ensemble methods, leveraging the complementary strengths of both paradigms: quantum circuits explore nonlinear feature spaces that may be difficult to access classically, while classical regressors provide stable, interpretable predictions and uncertainty quantification. This synergy is theoretically expected to outperform either purely quantum or purely classical approaches, particularly for scientific computing problems where both complex feature interactions and reliable predictions are essential. To benchmark the performance of the developed quantum-based models, two widely-used classical ML algorithms were implemented as baseline comparisons: (1) Generalized Boosted Regression (GBR), a powerful regression approach that iteratively builds an ensemble of weak learners to capture complex nonlinear patterns, and (2) Gradient Boosted Trees (GBT), also known as gradient boosting machine which is a gradient-based ensemble method recognized as a state-of-the-art algorithm with strong predictive performance across diverse regression tasks. The GBT model was implemented using the `h2o` package in R (Fryda et al., 2024), while the GBR model was implemented using the `gbm` package in R (Ridgeway, et al., 2024). Both GBR and GBT models were trained using 500 trees, with an interaction depth of 3 for the GBR model and a maximum depth of 5 for the GBT model.

Supplementary D

D 1. Model accuracy assessment

The performance of the developed QML models was assessed using six complementary statistical metrics that capture different aspects of prediction accuracy and model reliability. These include mean absolute error (MAE) and root mean square error (RMSE), relative error expressed as the RMSE–standard deviation ratio (RSR), and Refined Index of Agreement (dr) (Willmott et al., 2012) and the coefficient of determination (R^2). The mathematical formulations of these metrics are provided in Eqs (3) to (7):

$$MAE = \frac{\sum_{i=1}^n |Y_i - \hat{Y}_i|}{n} \quad (3)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n}} \quad (4)$$

$$RSR = \frac{RMSE}{SD_{obs}} \quad (5)$$

$$dr = \begin{cases} 1 - \frac{\sum_{i=1}^n |\hat{Y}_i - Y_i|}{c \cdot \sum_{i=1}^n |Y_i - \bar{Y}|}, & \text{if } \sum_{i=1}^n |\hat{Y}_i - Y_i| \leq c \cdot \sum_{i=1}^n |Y_i - \bar{Y}| \\ \frac{c \cdot \sum_{i=1}^n |Y_i - \bar{Y}|}{\sum_{i=1}^n |\hat{Y}_i - Y_i|} - 1, & \text{if } \sum_{i=1}^n |\hat{Y}_i - Y_i| > c \cdot \sum_{i=1}^n |Y_i - \bar{Y}| \end{cases} \quad (6)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \quad (7)$$

where n is the number of data points in the dataset, Y_i and $\hat{Y}_i$ indicate the observed and predicted values for the i -th data point, respectively. $\bar{Y}$ represents the mean of observed values, SD_{obs} is the standard deviation of observed values, and c is a constant value (set as 2).

Supplementary E

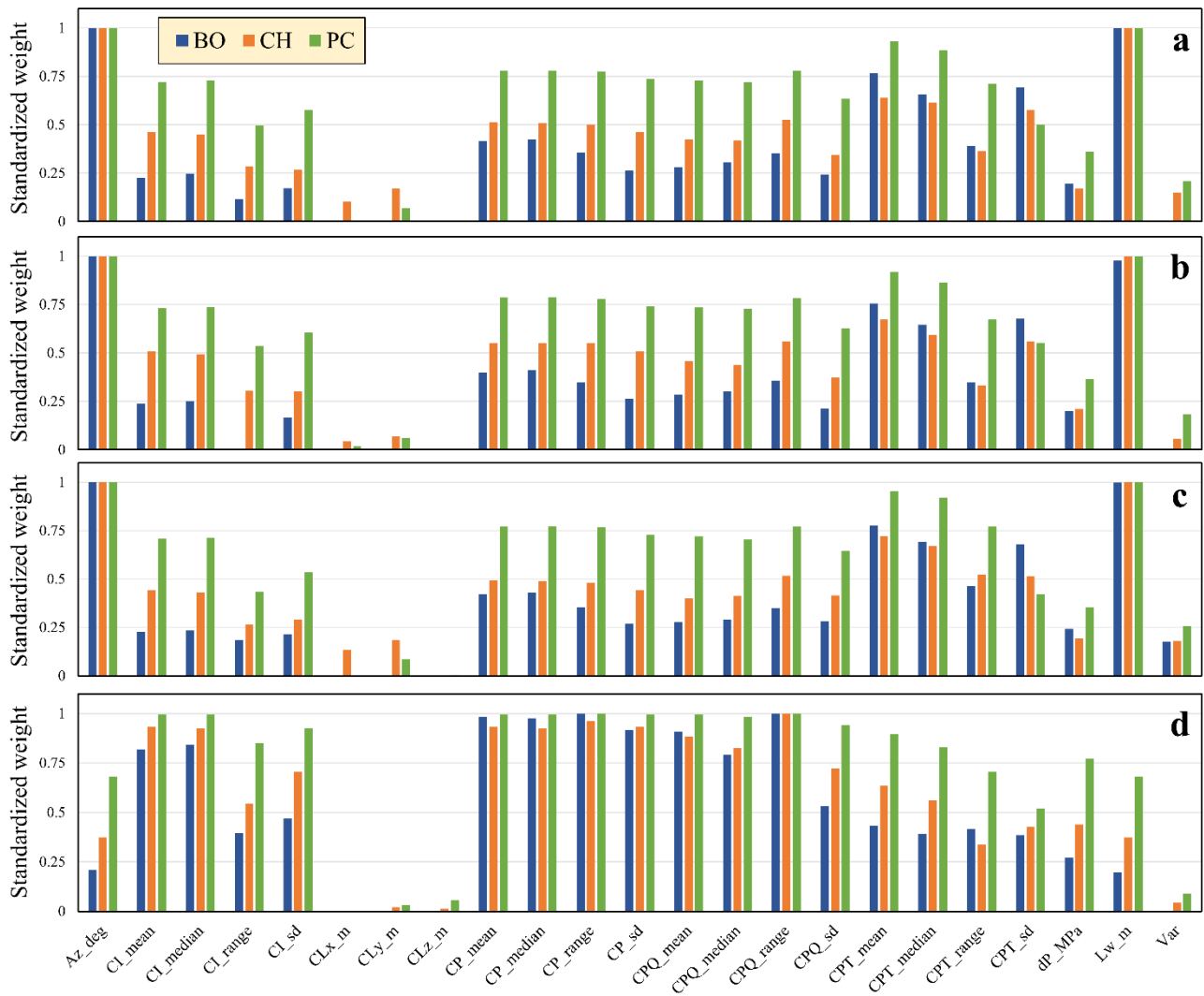


Fig. E 1. Standardized feature importance weights (0-1 scale) for iCO₂ (a), eCO₂ (b), sCO₂ (c), and extQh (d) across three feature selection methods.

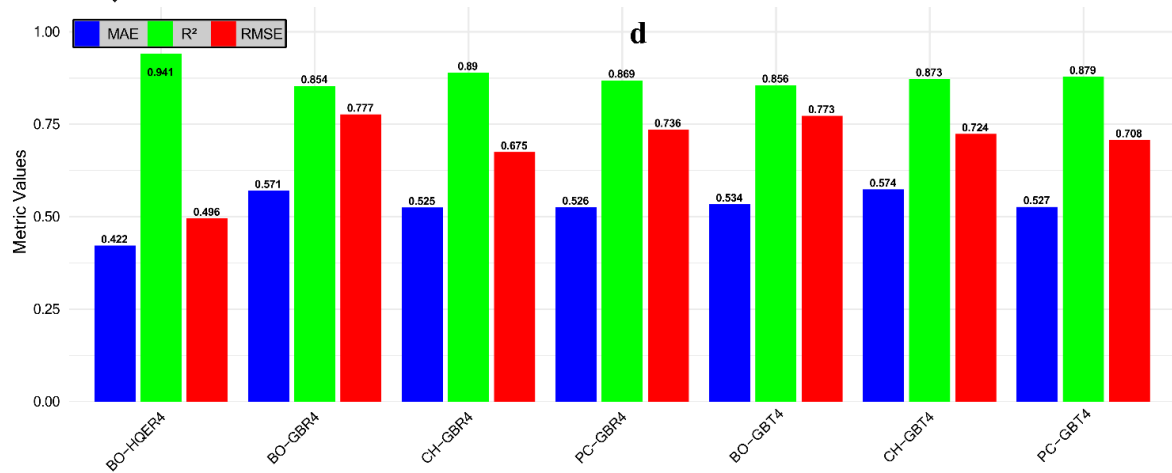
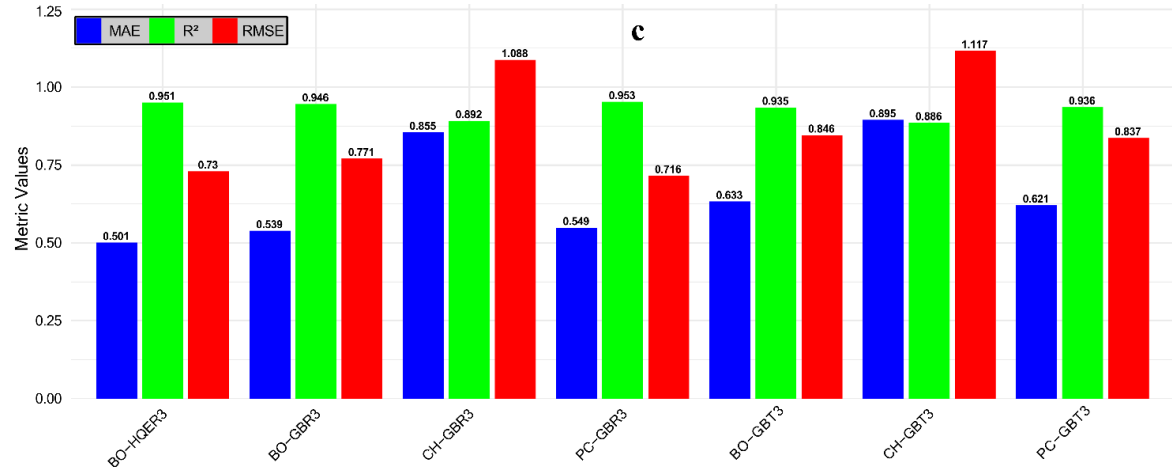
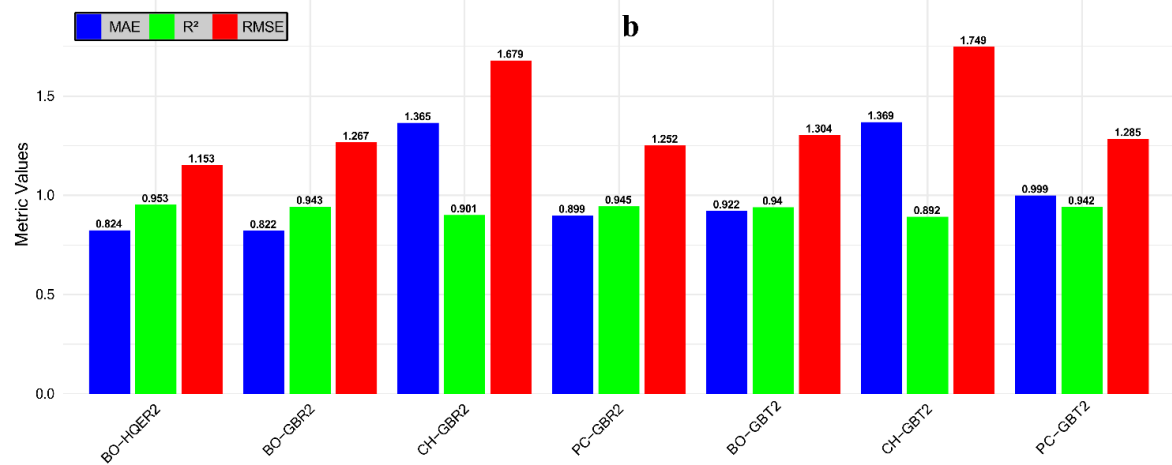
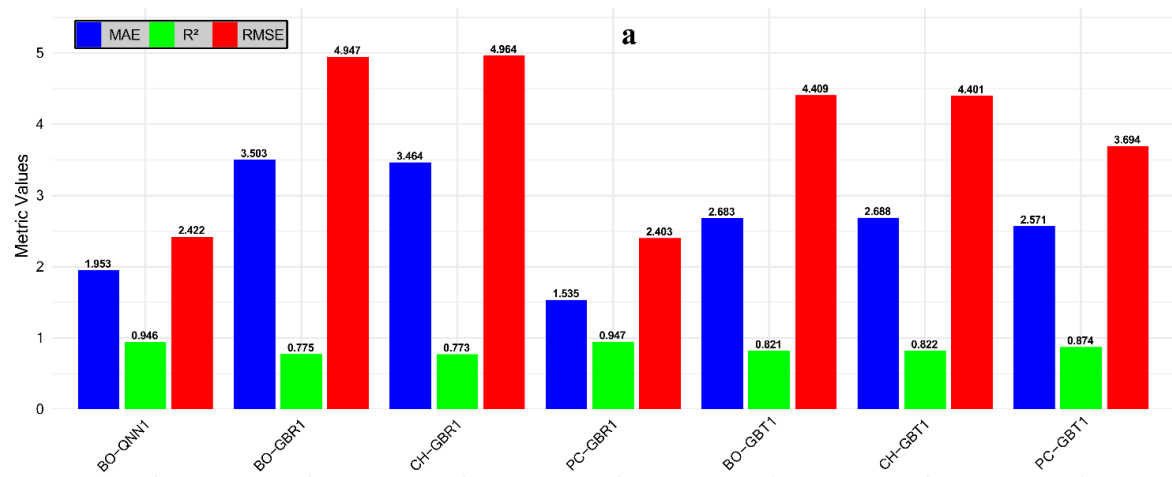
Supplementary F

F 1. Comparative performance of developed quantum-based and classical baseline models

The comparative analysis between the best-performing quantum-based models and classical baseline models (GBR and GBT) across all six target variables reveals consistent superiority of the quantum-inspired architectures, while also demonstrating that classical models achieve competitive performance with appropriate feature selection (Fig. F 1). The comparison evaluates the test set performance using three metrics (MAE, R², and RMSE) to provide comprehensive assessment of prediction accuracy, with quantum models represented by either HQER or QNN architectures depending on which achieved the best test performance for each target variable. Comparative analysis of quantum-based versus classical approaches revealed that quantum-enhanced models consistently achieve superior or competitive performance using substantially fewer input features than classical models require, demonstrating quantum advantage in data efficiency and feature economy.

For tLife prediction, BO-QNN1 served as the quantum benchmark, achieving test MAE of 1.953 years, R^2 of 0.946, and RMSE of 2.422 years, demonstrating the strongest overall performance among all tested models. Among classical baselines, PC-GBR1 achieved the best performance with MAE of 1.535 years, R^2 of 0.947, and RMSE of 2.403 years. This near-parity suggests that for tLife prediction, the comprehensive 15-feature set selected by Pearson correlation combined with gradient boosting regression effectively captures the temperature-dominated dynamics governing operational lifetime. However, BO-GBR1 and CH-GBR1 showed substantially degraded performance (R^2 of 0.775 and 0.773, respectively) with RMSE exceeding 4.9 years, indicating that compact feature sets (2-3 features) are insufficient for classical models to accurately predict system lifetime, whereas the quantum model achieved strong performance with only 2 features (CPT_mean, CPT_median). The iCO₂ prediction comparison demonstrates clear quantum advantage, with BO-HQER2 achieving test MAE of 0.824 Mt, R^2 of 0.953, and RMSE of 1.153 Mt, substantially outperforming all classical baselines (in terms of R^2 and RMSE). PC-GBR2 emerged as the strongest classical model with MAE of 0.899 Mt, R^2 of 0.945, and RMSE of 1.252 Mt. The Boruta and Chi-squared feature selections (using only 2-3 features) resulted in significantly degraded classical model performance, with BO-GBR2 and CH-GBR2 achieving R^2 of only 0.943 and 0.901, respectively, accompanied by substantially higher errors (MAE of 0.822-1.365 Mt, RMSE of 1.267-1.679 Mt). This pattern indicates that minimal feature sets fail to provide adequate information for classical tree-based methods to accurately predict CO₂ injection dynamics. The GBT models showed consistent but inferior performance compared to GBR counterparts across all feature selection methods (R^2 ranging from 0.892-0.942 versus 0.901-0.945 for GBR).

For eCO₂ prediction, BO-HQER3 achieved exceptional test performance with MAE of 0.501 Mt, R^2 of 0.951, and RMSE of 0.73 Mt, establishing a strong benchmark against which classical models were compared. PC-GBR3 delivered the best classical performance with MAE of 0.549 Mt, R^2 of 0.953, and RMSE of 0.716 Mt. However, this classical performance required the comprehensive 8-feature Pearson correlation set, whereas the quantum model achieved nearly equivalent accuracy using only 3 features (Az_deg, Lw_m, CPT_mean) selected by Boruta. The sCO₂ prediction results reveal the strongest quantum advantage among all CO₂-related variables, with BO-HQER4 achieving test MAE of 0.422 Mt, R^2 of 0.941, and RMSE of 0.496 Mt, substantially outperforming all classical models. CH-GBR4 emerged as the best classical baseline with MAE of 0.525 Mt, R^2 of 0.89, and RMSE of 0.675 Mt. For extEn prediction, BO-HQER5 established the performance benchmark with test MAE of 0.212 PJ, R^2 of 0.95, and RMSE of 0.307 PJ, demonstrating exceptional accuracy for this critical performance metric. PC-GBR5 achieved the strongest classical performance with MAE of 0.223 PJ, R^2 of 0.954, and RMSE of 0.295 PJ. However, this classical model required the comprehensive 8-feature Pearson correlation set, whereas the quantum model achieved near-equivalent performance using only 3 Boruta-selected features (Az_deg, Lw_m, CPT_mean). The extQh prediction comparison reveals the most pronounced quantum advantage observed across all target variables, with BO-HQER6 achieving test MAE of 0.419 MW, R^2 of 0.95, and RMSE of 0.993 MW, outperforming all classical baselines. PC-GBR6 emerged as the strongest classical model with MAE of 0.797 MW, R^2 of 0.909, and RMSE of 1.337 MW. The BO-GBR6 and CH-GBR6 models with compact feature sets (9 features selected by Boruta and Chi-squared) showed severely degraded performance with identical R^2 values of 0.893, MAE of 0.71 MW, and RMSE of 1.447 MW, indicating that even relatively feature-rich selections (9 features) prove insufficient for classical methods to accurately predict instantaneous heat flux dynamics.



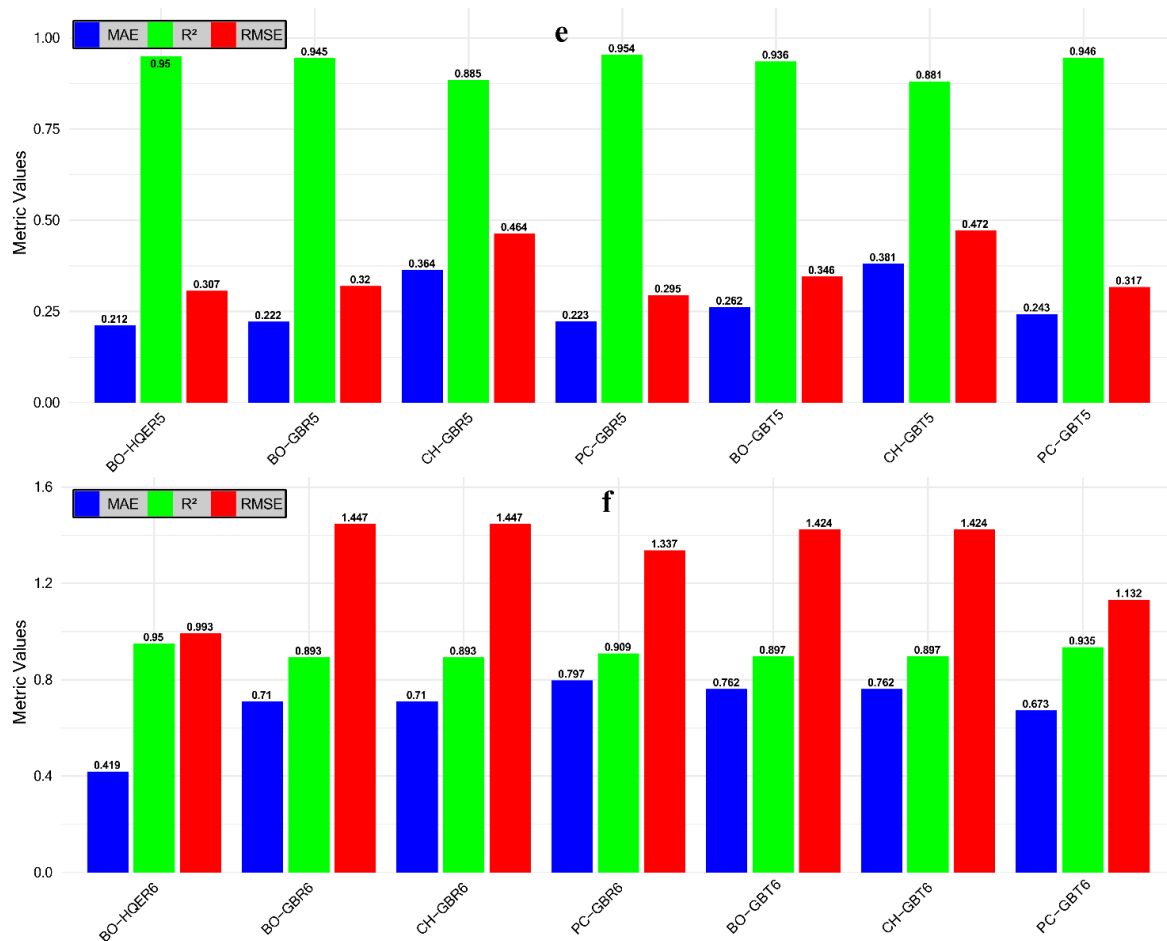


Fig. F 1. Comparative performance analysis of quantum-based models versus classical baseline models (GBR and GBT) across tLife (a), iCO₂ (b), eCO₂ (c), sCO₂ (d), extEn (e), and extQh (f) variables for the test dataset. Each panel displays three performance metrics for the best-performing quantum model (the model achieving the highest performance during test phase) and classical baselines across three applied feature selection methods.

Supplementary G

Fig. G 1 shows the Taylor diagrams of developed storage QML models across all six target variables for the entire data set (train and test together). The remarkable clustering of models near the observed reference point demonstrates exceptional accuracy, with all models achieving high correlations, indicating that developed models successfully capture both the magnitude and variability of six target variables. The tLife predictions show slightly more dispersion but maintain high performance, with HQER models consistently positioned closer to the reference point than their QNN. The most notable pattern emerges in the extQh diagram, where models display greater spread with QNN variants showing larger RMSE (positioned farther from the reference), while HQER models maintain tighter clustering near the observed point. Across all diagrams, the Boruta-based models (red and cyan markers) consistently achieve optimal or near-optimal positioning. The overall tight clustering and high correlations across diverse target variables demonstrate that the developed QML models' capabilities to capture the complex physics of CPG systems.

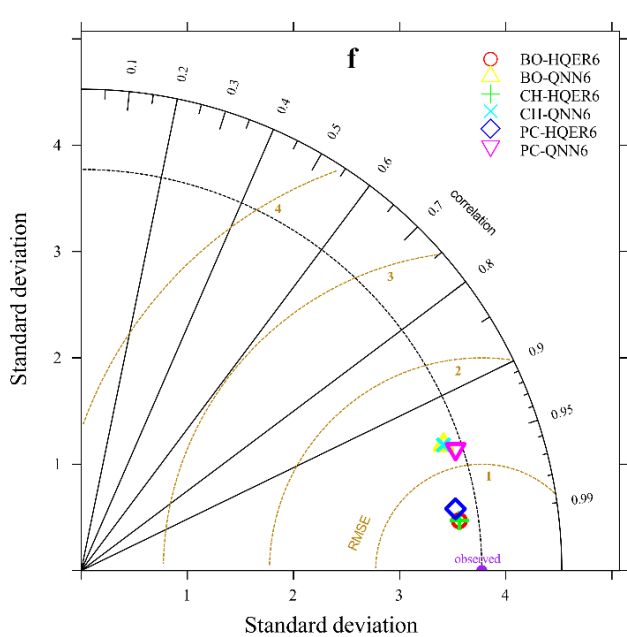
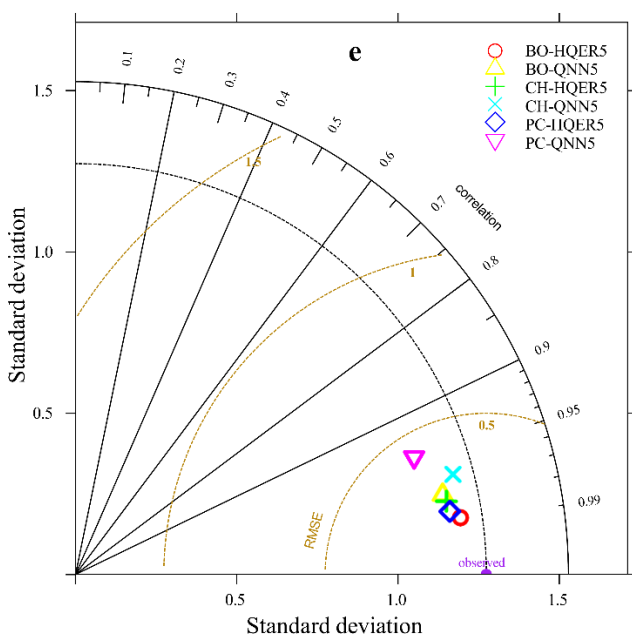
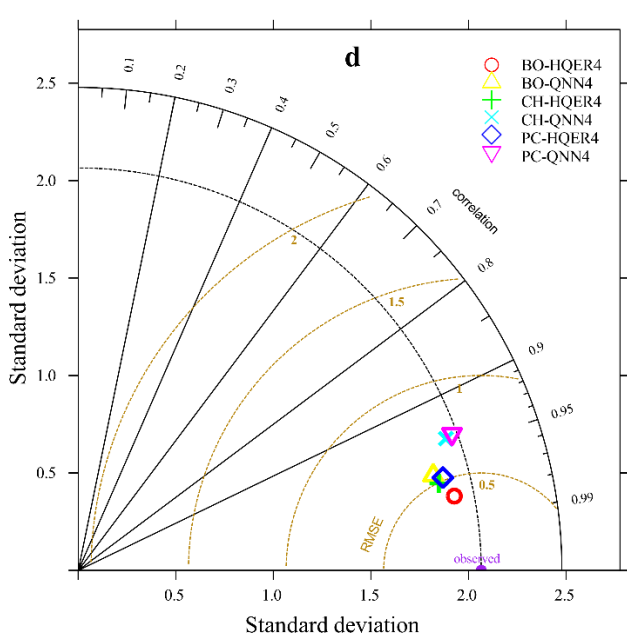
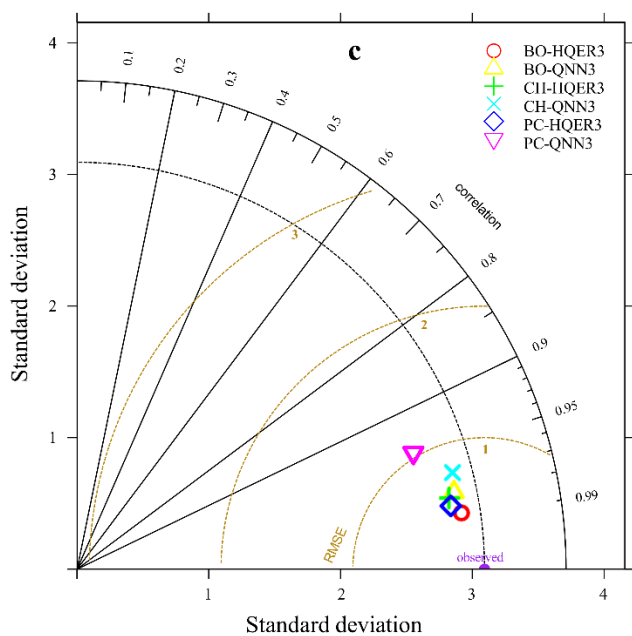
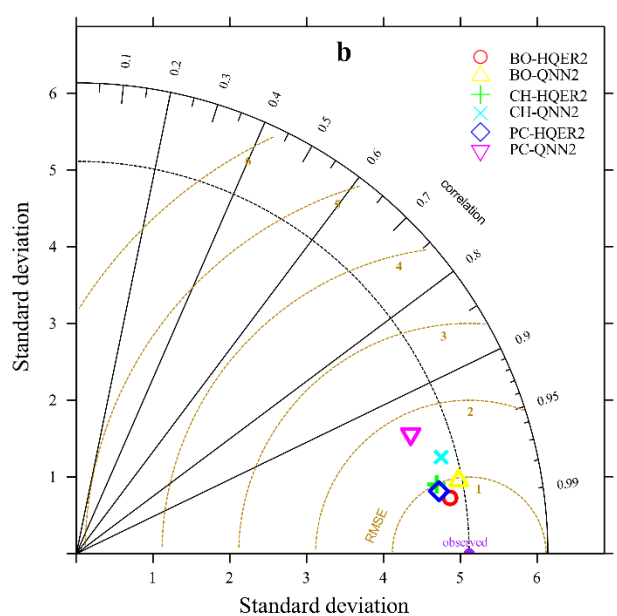
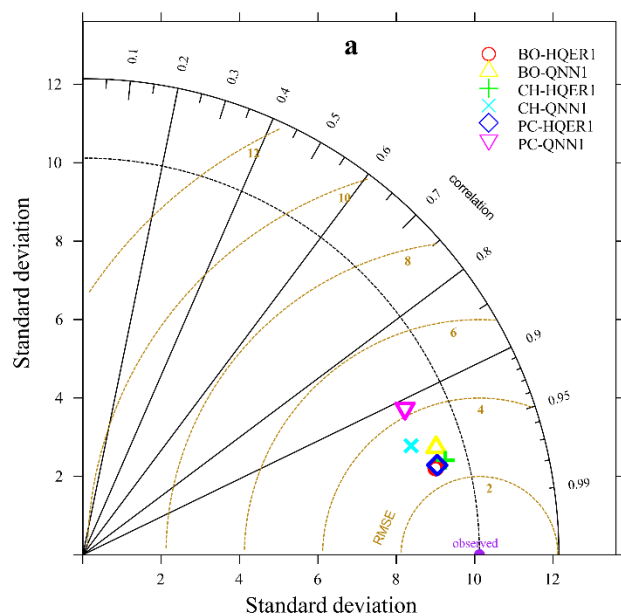


Fig. G 1. Taylor diagrams comparing the performance of six developed QML models for tLife (a), iCO₂ (b), eCO₂ (c), sCO₂ (d), extEn (e), and extQh (f) variables using the entire dataset (all 100 samples, including both training and testing data).

Supplementary H

H 1. Multi-objective optimization

Table H 1 and Fig. H 1 present the optimization results based on the best developed surrogate models for each variable. Results reveal distinct patterns in optimal decision variables and objective trade-offs across six scenarios of increasing complexity, with the solution selection strategy prioritizing maximum extEn from the Pareto fronts. Single-objective optimization using MFO (Scenario 1) achieved the baseline maximum extEn of 5.787 PJ with dP_MPa of 5.305 (MPa) and well spacing (Lw_m) of 991.293 (m). The bi-objective optimization for energy metrics (Scenario 2) required significantly higher overpressure at the injection well (12.158 MPa) to balance extEn (5.517 PJ) with heat extraction rate (9.249 MW). When system lifetime was added as the third objective (Scenario 3), the optimizer converged to lower pressure (5.094 MPa) and slightly reduced well spacing (969.316 m), achieving superior extEn (5.862 PJ) while maintaining acceptable heat flux (4.126 MW) and extending operational lifetime to 39.502 years. The storage-focused scenarios (4-6) consistently converged to nearly identical solutions (dP $\approx$ 5.1 to 5.3 MPa, Lw $\approx$ 991 to 969 m), suggesting a robust optimal configuration for maximizing both energy extraction and CO₂ sequestration, with stored CO₂ around 8 Mt and injected CO₂ approximately 22.4 Mt for scenarios involving injection minimization. The convergence of multiple multi-objective scenarios to similar optimal solutions (particularly Scenarios 3, 5, and 6) validates the robustness of the identified configuration and suggests that lower injection pressures combined with well spacing around 970 m represents a CPG system design that naturally balances energy extraction, storage effectiveness, and operational longevity.

Table. H 1. Optimal solutions and corresponding objective values for six optimization scenarios via MFO (for single-objective optimization) and NSGA-II (for multi-objective optimization). * Marks indicate objectives not included in the respective optimization scenario.

Scenario	Number of Objectives	Solution for dP_MPa	Solution for Lw_m	Best objective (extEn)	Best objective (extQh)	Best objective (tLife)	Best objective (sCO ₂)	Best objective (iCO ₂)	Best objective (extQh)	Best objective (tLife)
Scenario 1 (MFO)	1	5.305	991.293	5.787	*	*	*	*	*	*
Scenario 2 (NSGA-II)	2	12.158	994.889	5.517	9.249	*	*	*	*	*
Scenario 3 (NSGA-II)	3	5.094	969.316	5.862	4.126	39.502	*	*	*	*
Scenario 4 (NSGA-II)	2	5.305	991.291	5.786	*	*	7.816	*	*	*
Scenario 5 (NSGA-II)	3	5.094	969.319	5.863	*	*	8.044	22.444	*	*
Scenario 6 (NSGA-II)	5	5.094	969.316	5.862	*	*	8.041	22.437	4.125	39.507

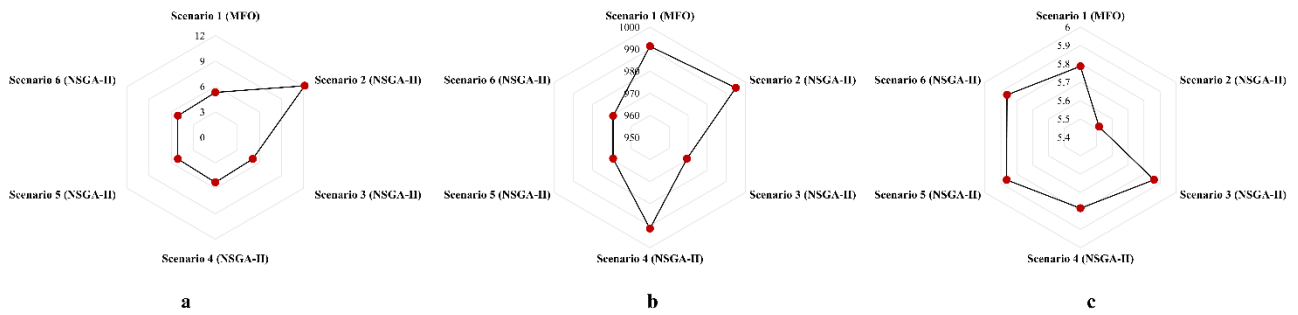


Fig. H 1. Spider diagrams comparing recommended solutions for dp_MPa (a), Lw_m (b), and $extEn$ (c) via various scenarios showing trade-offs between different CPG performance metrics.

The convergence of Scenario 3 to the 5.094 MPa - 969.316 m configuration and its recurrence in storage-focused scenarios (5 and 6) with nearly identical parameter values warrants detailed examination to understand why this solution occupies a dominant position on the Pareto frontier. This configuration achieves superior multi-objective performance by balancing three competing mechanisms governing CPG system behavior: thermal sweep efficiency, CO₂ plume containment, and pressure-induced fracturing risk. The moderate injection overpressure of ~5.1 MPa provides sufficient driving force to maintain continuous CO₂ circulation and heat extraction ($extEn = 5.862$ PJ, $extQh = 4.126$ MW) while remaining well below the estimated fracture pressure threshold (~15 MPa), thereby minimizing induced seismicity risk and formation damage that could compromise long-term system integrity. The well spacing of ~970 m represents an optimal trade-off between thermal breakthrough time and plume connectivity: closer spacing (<600 m) would cause premature thermal breakthrough as cold injected CO₂ reaches the production well before sufficient heat extraction, reducing both $extEn$ and $tLife$, while wider spacing (>1200 m) would require excessive injection pressures to establish hydraulic communication, increasing iCO_2 without proportional gains in sCO_2 or $extEn$.

The optimization results reveal fundamental physical mechanisms that govern trade-offs among competing CPG performance objectives, which can be understood through the coupled thermodynamics and fluid mechanics of CO₂-brine-heat transport in porous media. The stark contrast between Scenario 2's high-pressure solution (12.158 MPa, 994.889 m) and Scenario 3's moderate-pressure configuration (5.094 MPa, 969.316 m) reflects competing physical processes: high injection overpressures increase CO₂ mass flux through enhanced pressure gradients, accelerating both injection and production rates which elevate instantaneous heat extraction ($extQh = 9.249$ MW), but simultaneously cause rapid thermal sweep that depletes the reservoir's thermal energy inventory more quickly, reducing system lifetime and cumulative energy recovery. This aggressive extraction regime creates steep thermal fronts that propagate rapidly from the injection well toward the production well, with convective heat transport dominating over conductive spreading, leading to premature thermal breakthrough when cold injected CO₂ reaches the production well before sufficient heat exchange with the formation. Conversely, moderate injection pressures in Scenario 3 allow more uniform thermal sweep with greater conductive heat diffusion from surrounding rock into the CO₂ plume, maximizing the effective thermal contact area and enabling sustained heat extraction over extended operational periods ($tLife = 39.502$ years). The well spacing optimization reflects a balance between two opposing physical constraints: the percolation threshold for hydraulic connectivity, which requires sufficient pressure gradients to establish flow pathways through heterogeneous permeability fields (requiring closer spacing or higher pressures), and the thermal interference distance, beyond which separate circulation cells can operate independently without premature thermal breakthrough (favoring wider spacing). The convergence to ~970 m spacing across multiple scenarios suggests this distance optimally balances these competing requirements for the North Oman reservoir properties (mean permeability 5.3 mD, porosity 0.25, thermal conductivity typical of sandstone formations). The trade-off between iCO_2 and sCO_2 in Scenarios 5-6 reflects the fundamental distinction between mobile and immobile CO₂ fractions: high injection pressures mobilize larger CO₂ volumes but also enhance production efficiency through increased pressure gradient between injection and production

well, reducing the fraction of injected CO₂ that remains trapped through residual saturation, capillary trapping, or dissolution in formation brine. The optimal solutions consistently minimize iCO₂ while maximizing sCO₂, indicating that effective storage occurs primarily through mechanisms favored by moderate pressures and long residence times, capillary trapping in pore throats, gravity-driven structural trapping beneath low-permeability layers, and diffusive dissolution into brine, rather than through brute-force injection of massive CO₂ volumes. The robust recurrence of the 5.1 MPa - 970 m configuration across storage-focused scenarios (4-6) demonstrates that this operational regime naturally aligns multiple physical objectives: it maintains subcritical stress states that avoid geomechanical risks (injection pressure remains well below minimum principal stress plus tensile strength), optimizes sweep efficiency through balanced advective-diffusive transport, and maximizes residence time for enhanced dissolution trapping.

Supplementary I

I 1. Limitations and future work

I 1.1. Computational implementation and quantum hardware constraints

In this study all quantum computations were performed using classical simulators (PennyLane's default.qubit simulator) and while the simulated quantum circuits successfully demonstrate the conceptual framework and achieve exceptional prediction accuracy, the practical feasibility of implementing these QML models on current near-term quantum devices remains uncertain. Contemporary quantum hardware faces some challenges including limited qubit coherence times, high gate error rates, restricted circuit depth due to decoherence, and scalability constraints that severely limit the number of qubits available for practical applications. The quantum circuits developed in this study, while modest in qubit requirements (4 qubits), would still face implementation challenges on real quantum hardware where noise accumulation over multiple quantum layers could compromise prediction accuracy.

I 1.2. Technical implementation details

Several technical details require more comprehensive elaboration to enable reproducibility and deeper understanding of the quantum-enhanced framework. The quantum scaling feature encoding process, while described as transforming standardized features through arctan functions with $\pi/2$ offset to map values into the $[0, \pi]$ range for quantum rotation gates, lacks detailed explanation of several implementation aspects. Specifically, the current study does not fully address how this particular encoding scheme was selected over alternative quantum embedding strategies (such as amplitude encoding, basis encoding, or more sophisticated variational encoding circuits). Future work should include detailed sensitivity analyses examining how variations in quantum circuit architecture (number of layers, types of entangling gates, measurement strategies) affect the quality and interpretability of extracted quantum features, and should provide quantitative metrics for assessing the added value of quantum feature extraction beyond what classical nonlinear transformations (such as kernel methods, neural network embeddings, or polynomial features) could achieve with comparable computational resources.

I 1.3. Physical interpretability and conservation law consistency

Another limitation of the developed QML surrogate models is the reduced physical interpretability compared to the underlying NUFT simulations, which are grounded in rigorous conservation laws for mass, energy, and momentum. While the quantum and hybrid quantum-classical models achieve exceptional prediction accuracy across all six target variables ($R^2 > 0.95$ for most cases), the black-box nature of these ML approaches raises important questions about whether the models truly capture the fundamental physics governing CPG systems or merely perform sophisticated pattern matching

that could fail under conditions outside the training distribution. The NUFT simulator explicitly enforces conservation of CO₂ and brine mass through coupled nonlinear partial differential equations, ensures energy conservation through enthalpy balance equations accounting for advective and conductive heat transport, and maintains thermodynamic consistency by computing CO₂ phase behavior (gaseous, liquid, supercritical transitions) according to established equations of state. In contrast, the QML surrogate models learn statistical relationships between input features and output variables without explicit constraints enforcing these physical principles, potentially allowing predictions that violate fundamental conservation laws under certain parameter combinations. For instance, the models could theoretically predict eCO₂, iCO₂, or thermal energy extraction rates inconsistent with the available enthalpy of the produced fluid, although such violations were not observed in the validation datasets. Future work should incorporate physics-informed constraints into the QML training process, either through penalization terms in the loss function that discourage conservation law violations, through post-processing corrections that project predictions onto physically feasible manifolds, or through hybrid architectures that combine data-driven quantum feature extraction with physically-constrained classical output layers. Additionally, systematic testing of the surrogate models under extreme parameter combinations not well-represented in the training data would help assess whether the models maintain physical plausibility or produce spurious predictions that could mislead design optimization efforts.

I 1.4. Feature selection methodology and sensitivity analysis

The comparative analysis of three feature selection methods (Boruta, Chi-squared, Pearson correlation) revealed substantial differences in selected feature sets, yet the practical implications of these differences for model interpretability, robustness, and generalization require deeper investigation. Boruta consistently identified compact feature sets emphasizing the most statistically significant predictors (e.g., CPT_mean, CPT_median for tLife; Lw_m, Az_deg for CO₂-related outputs), achieving strong performance particularly when combined with HQER architectures, suggesting that minimal feature sets can capture the essential physics when quantum-enhanced feature processing is employed. Chi-squared feature selection produced similarly parsimonious selections, often converging to identical feature subsets as Boruta for several target variables, indicating robust agreement on the most discriminative predictors. In contrast, Pearson correlation systematically selected larger feature sets (up to 15 features for tLife and extQh), incorporating broader coverage of injection, production, temperature, and heat flux statistics. While the expanded feature sets achieved strong performance in HQER models (test $R^2 > 0.94$ for most variables), they showed mixed results in QNN architectures, with PC-QNN variants sometimes suffering larger train-test degradation than their Boruta or Chi-squared counterparts. This suggests that compact feature sets may provide better regularization for simpler quantum architectures, while ensemble methods can more effectively leverage diverse feature information without overfitting.

A critical limitation is that the feature selection threshold of 0.75 for standardized importance weights was chosen based on practical considerations rather than through systematic sensitivity analysis. The current study does not provide empirical justification for this specific threshold value, nor does it examine how variations in the threshold would affect model performance, feature set composition, and computational efficiency. The threshold of 0.75 represents a relatively stringent criterion, potentially excluding features with moderate but still meaningful predictive value (weights between 0.5-0.75), while a more lenient threshold might introduce noise from weakly relevant features. Future work should conduct comprehensive sensitivity analyses systematically varying the feature selection threshold across a range (e.g., 0.5, 0.6, 0.7, 0.75, 0.8, 0.85, 0.9) and evaluating the resulting models across multiple metrics, generalization performance (train-test gap), computational efficiency (number of features, training time), and interpretability. Such analyses would help establish data-driven guidelines for threshold selection tailored to different target variables and model architectures. Additionally, more sophisticated feature selection approaches could be explored, including recursive feature elimination that iteratively removes the least important features while monitoring model performance, ensemble feature selection that combines rankings from multiple methods through

voting or averaging, and adaptive threshold selection that optimizes the threshold value through cross-validation to maximize generalization performance for each specific prediction task.

I 1.5. Practical implications for decision-making and multi-objective trade-offs

While the multi-objective optimization framework successfully identified optimal CPG system configurations across six scenarios of increasing complexity, the practical implications of the inherent trade-offs between competing objectives require more explicit discussion to effectively guide engineering decision-making and policy development. The optimization results reveal several fundamental tensions in CPG system design that operators and stakeholders must navigate based on their specific priorities and constraints. The most prominent trade-off emerges between maximizing extQh versus tLife , as evidenced by the stark contrast between Scenario 2 (requiring 12.158 MPa injection pressure to achieve 9.249 MW heat flux but likely reducing lifetime due to aggressive thermal depletion) and Scenario 3 (using only 5.094 MPa to balance moderate heat flux of 4.126 MW with extended 39.5-year lifetime). For project developers prioritizing rapid return on investment and high power output, the high-pressure, high-flux configuration may be economically attractive despite shorter operational periods, particularly in contexts where high electricity prices or carbon credits justify aggressive extraction strategies. Conversely, operators seeking long-term resource sustainability and lower operational risk may prefer conservative injection pressures that preserve thermal resources and extend project life, accepting lower instantaneous power output in exchange for decades-long stable operation. The convergence of storage-focused scenarios (4-6) to similar optimal solutions ($\text{dP} \approx 5.1\text{-}5.3$ MPa, $\text{Lw} \approx 970$ m, $\text{sCO}_2 \approx 8$ Mt, $\text{extEn} \approx 5.8$ PJ) reveals a particularly valuable insight: when CO_2 sequestration objectives are incorporated into the optimization, the framework naturally identifies operating conditions that simultaneously maximize long-term energy extraction and carbon storage while minimizing injection requirements. This suggests that CPG systems designed with explicit carbon storage goals may inherently operate more efficiently and sustainably than those optimized solely for energy production. Decision-makers should carefully consider these trade-offs in the context of local regulatory requirements (e.g., carbon pricing mechanisms, renewable energy mandates), project economics (capital costs, operational expenses, revenue streams), reservoir characteristics (available storage capacity, thermal recovery potential), and stakeholder priorities (climate mitigation goals, energy security objectives, economic development targets) when selecting among the Pareto-optimal solutions identified by the multi-objective optimization framework.

I 1.6. Model validation strategy and generalization assessment

The current validation strategy, while demonstrating strong performance on the test dataset, exhibits several limitations that warrant acknowledgment and motivate future research directions. First, the single 80/20 train-test split, although commonly employed in ML studies, provides only a limited assessment of model generalization and may be sensitive to the particular random partitioning of the data. The absence of k -fold cross-validation represents a gap, as repeated random sampling and averaging across multiple train-test splits would provide more robust estimates of model performance variability and reduce the risk that reported metrics reflect fortuitous data partitioning rather than genuine predictive capability. Future studies should implement k -fold cross-validation (e.g., 5-fold or 10-fold) where the NUFT simulations are systematically divided into k subsets, with each subset serving once as a validation set while the remaining subsets are used for training, thereby enabling calculation of mean performance metrics and confidence intervals that more accurately characterize model reliability. Second, the study lacks online or streaming validation tests where the surrogate models would be evaluated on sequentially arriving new data points, which is particularly relevant for real-time reservoir management applications where CPG operators might need to continuously update predictions as new monitoring data becomes available during system operation. The absence of temporal validation is especially concerning given that CPG system behavior evolves over multi-decadal timescales, and surrogate models trained on early-time simulation data may not accurately

predict late-time performance as thermal depletion, CO₂ plume migration, and geomechanical changes alter system dynamics.

Third, all validation was conducted on synthetic data from a single geological setting (in North Oman) with specific reservoir properties (mean porosity 0.25, mean permeability 5.3 mD, 90°C initial temperature, 180 ppt brine salinity, 5° dip angle). The models' ability to generalize to held-out geological settings with substantially different reservoir characteristics remains completely untested, raising serious concerns about the practical applicability of the developed QML framework to real-world CPG projects in diverse geological contexts. CPG systems may be deployed in sedimentary basins with vastly different properties including highly fractured carbonate reservoirs with dual-porosity behavior, deep saline aquifers with heterogeneous permeability structures spanning orders of magnitude, depleted hydrocarbon fields with residual oil or gas saturation, or volcanic rock formations with complex thermal regimes. The surrogate models trained on the relatively homogeneous Oman reservoir simulations may fail catastrophically when applied to these alternative settings if the underlying physics or parameter sensitivities differ substantially. Prospective validation requiring collection of NUFT simulation data from multiple diverse geological settings (varying lithology, structural complexity, thermal gradients, fluid properties, well configurations) and evaluation of whether surrogate models trained on one setting can accurately predict performance in others would provide much more convincing evidence of practical utility. Until such cross-site validation is performed, the generalizability claims of the developed QML framework must be considered speculative, and practitioners should exercise caution when applying these models to CPG projects in geological settings that differ significantly from the North Oman training data.

I 1.7. Scalability to large-scale field applications

The case study's limitation to a relatively modest synthetic reservoir model (2000 m × 3000 m × 20 m domain) raises questions about how the QML surrogate modeling framework would scale to larger, more complex field-scale CPG applications that are more representative of commercial deployment scenarios. Real-world geothermal projects often involve large spatial domains and such large-scale models can require millions of computational cells and produce simulation runtimes measured in days or weeks per realization, making the generation of comprehensive training datasets through Latin Hypercube Sampling or other systematic parameter space exploration methods extremely computationally expensive. A key uncertainty is whether the current training dataset size would remain sufficient for complex field models with higher-dimensional parameter spaces, or whether substantially larger datasets would be required to adequately sample the expanded design space and capture the increased nonlinearity introduced by geological complexity. If training data requirements scale unfavorably with model complexity, the computational savings offered by QML surrogates could be partially or entirely offset by the increased cost of generating training data, potentially limiting practical applicability to small-to-medium scale reservoir models.

I 1.8. Positioning relative to classical surrogate modeling approaches

While the developed QML framework demonstrates excellent prediction accuracy and computational efficiency, establishing the true value of quantum-inspired architectures requires systematic comparison against classical ML alternatives. To provide initial benchmarking, this study compared the performance of QML models (QNN and HQER) against two widely-applied classical ML algorithms (GBR and GBT). Recent advances in classical ML have produced more sophisticated approaches specifically designed for scientific computing applications, including Physics-Informed Neural Networks (PINNs) that embed governing partial differential equations directly into the loss function to enforce conservation laws and improve extrapolation beyond training data, Fourier Neural Operators (FNOs) that learn solution operators in frequency space and have demonstrated remarkable accuracy for subsurface flow problems with orders of magnitude computational speedup, deep neural networks with advanced architectures incorporating convolutional layers for spatial pattern recognition, recurrent or transformer layers for temporal dynamics, and attention mechanisms for multi-scale feature extraction. Therefore, future work should implement comprehensive classical

baselines including deep neural networks with comparable parameter counts to the QML models, PINNs that explicitly enforce mass and energy conservation for CPG systems, and FNO architectures that may be particularly well-suited to the spatiotemporal nature of reservoir simulation data. Such comparisons should evaluate not only prediction accuracy but also training efficiency, inference speed, robustness to parameter variations outside the training distribution, and data efficiency (performance with limited training samples) to rigorously quantify any quantum advantage and identify the specific prediction tasks, dataset characteristics, or problem structures where quantum-enhanced approaches provide meaningful benefits that justify their additional implementation complexity over mature classical alternatives.

References

- Arthur, D., Date, P. Hybrid Quantum-Classical Neural Networks. In 2022 IEEE International Conference on Quantum Computing and Engineering (QCE), 2022: 49–55.
- Babaei, M., Govindan, R., Korre, A. et al. Calculation of pressure-and migration-constrained dynamic CO₂ storage capacity of the North Sea Forties and Nelson dome structures. *International Journal of Greenhouse Gas Control*, 2016, 53: 127-140.
- Beer, K., Bondarenko, D., Farrelly, T., et al. Training deep quantum neural networks. *Nature Communications*, 2020, 11(1): 808.
- Benesty, J., Chen, J., Huang, Y. et al., 2009. Pearson correlation coefficient. In *Noise reduction in speech processing*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2009: 1-4.
- Bengtsson, A., Vikstål, P., Warren, C., et al. Improved Success Probability with Greater Circuit Depth for the Quantum Approximate Optimization Algorithm. *Physical Review Applied*, 2020, 14(3): 034010.
- Buscheck, T. A., Sun, Y., Chen, M., et al. Active CO₂ reservoir management for carbon storage: Analysis of operational strategies to relieve pressure buildup and improve injectivity. *International Journal of Greenhouse Gas Control*, 2012, 6: 230–245.
- Chen, M., Reza Nikoo, M., Al-Maktoumi, A., et al. The impact of geological heterogeneity on coupled CO₂ storage and geothermal extraction in inclined reservoirs. *Journal of Hydrology*, 2023, 6 Farhi, E., Goldstone, J., Gutmann, S. A Quantum Approximate Optimization Algorithm. *arXiv*. 2014. <https://doi.org/10.48550/arXiv.1411.40281>: 128950.

- Fryda, T., LeDell, E., Gill, N., et al. h2o: R interface for the 'H2O' scalable machine learning platform (Version 3.44.0.3) [R package]. CRAN, 2024. <https://doi.org/10.32614/CRAN.package.h2o>
- Jolliffe, I. T., Cadima, J. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 2016, 374(2065): 20150202.
- Kursa, M. B., Rudnicki, W. R. Feature Selection with the Boruta Package. *Journal of Statistical Software*, 2010, 36: 1–13.
- Liang, Y., Peng, W., Zheng, Z.-J., et al. A hybrid quantum–classical neural network with deep residual learning. *Neural Networks*, 2021, 143: 133–147.
- Majumder, A., Lewis, D., Jayashankar, A., et al. Variational Quantum Circuits for Multi-Qubit Gate Automata. *arXiv*. 2024. <https://doi.org/10.48550/arXiv.2209.00139>
- Moridis GJ, Pruess K. TOUTH simulations of the updegraff's set of fluid and heat flow problems. Tech rep, Lawrence Berkeley Laboratory, 1992.
- Nitao, J. J. Reference Manual for the NUFT Flow and Transport Code, Version 3.0. 1998, 1(605).
- Pandis, N. The chi-square test. *American Journal of Orthodontics and Dentofacial Orthopedics*, 2016, 150(5): 898–899.
- Pearson, K. VII. Note on regression and inheritance in the case of two parents. *proceedings of the royal society of London*, 1895, 58(347-352): 240-242.
- Plackett, R. L. Karl Pearson and the Chi-Squared Test. *International Statistical Review / Revue Internationale de Statistique*, 1983, 51(1): 59–72.
- Rajabi, M. M., Chen, M., Bozorgpour, A., et al. Stochastic Techno-economic Analysis of CO₂-circulated Geothermal Energy Production in a Closed Reservoir System. *Geothermics*, 2021, 96: 102202.
- Ridgeway, G., Edwards, D., Kriegler, B., et al. gbm: Generalized boosted regression models (Version 2.2.2) [R package]. CRAN, 2024. <https://doi.org/10.32614/CRAN.package.gbm>

- Schuld, M., Sinayskiy, I., & Petruccione, F. An introduction to quantum machine learning. *Contemporary Physics*, 2015, 56(2): 172–185.
- Tudisco, A., Marchesin, A., Zamboni, M., et al. Evaluating Angle and Amplitude Encoding Strategies for Variational Quantum Machine Learning: their impact on model's accuracy. *arXiv*, 2025.
- Willmott, C., Robeson, S., Matsuura, K. A refined index of model performance. *International Journal of Climatology*, 2012, 32.
- Yeo, I., Johnson, R. A. A new family of power transformations to improve normality or symmetry. *Biometrika*, 2000, 87(4): 954–959.
- Zhao, G., Mu, L., Wu, X., 2012. The Application of Horizontal Wells in Effectively Developing Thinly-Bedded Carbonate Reservoirs of Block 5, Oman. *Search and Discovery Article #41020*, AAPG Annual Convention and Exhibition, Long Beach, California, April 22-25, 2012
- Zhou, Q., Birkholzer, J. T., Tsang, C. F., et al. A method for quick assessment of CO₂ storage capacity in closed and semi-closed saline formations. *International Journal of Greenhouse gas control*, 2008, 2(4): 626-639.