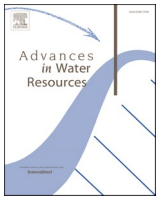
Advances in Water Resources 196 (2025) 104897

Contents lists available at ScienceDirect

Advances in Water Resources

journal homepage: www.elsevier.com/locate/advwatres



A hybrid CNN-transformer surrogate model for the multi-objective robust optimization of geological carbon sequestration

Zhao Feng [a,b], Bicheng Yan [c], Xianda Shen [a,b], Fengshou Zhang [a,b,*], Zeeshan Tariq [c], Weiquan Ouyang [a,b], Zhilei Han [c]

[a] Department of Geotechnical Engineering, College of Civil Engineering, Tongji University, Shanghai 200092, China
[b] Key Laboratory of Geotechnical and Underground Engineering of Ministry of Education, Tongji University, Shanghai 200092, China
[c] Physical Science and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia

ARTICLE INFO

Keywords:
Geological carbon sequestration
Surrogate model
Optimization
Deep learning
Transformers

ABSTRACT

The optimization of well controls over time constitutes an essential step in the design of cost-effective and safe geological carbon sequestration (GCS) projects. However, the computational expense of these optimization problems, due to the extensive number of simulation evaluations, presents significant challenges for real-time decision-making. In this paper, we propose a hybrid CNN-Transformer surrogate model to accelerate the well control optimization in GCS applications. The surrogate model encompasses a Convolution Neural Network (CNN) encoder to compress high-dimensional geological parameters, a Transformer processor to learn global patterns inherent in the well controls over time, and a CNN decoder to map the latent variables to the target solution variables. The surrogate model is trained to predict the spatiotemporal evolution of CO_2 saturation and pressure within 3D heterogeneous permeability fields under dynamic CO_2 injection rates. Results demonstrate that the surrogate model exhibits satisfactory performance in the context of prediction accuracy, computation efficiency, data scalability, and out-of-distribution generalizability. The surrogate model is further integrated with Multi-Objective Robust Optimization (MORO). Pareto optimal well controls are determined based on Non-dominated Sorting-based Genetic Algorithm II (NSGA-II), which maximize the storage efficiency and minimize the induced over-pressurization across an ensemble of uncertain geological realizations. The surrogate-based MORO reduces computational time by 99.99 % compared to simulation-based optimization. The proposed workflow not only highlights the feasibility of applying the CNN-Transformer model for complex subsurface flow systems but also provides a practical solution for real-time decision-making in GCS projects.

1. Introduction

Climate change has emerged as a critical issue that poses significant challenges to both natural ecosystems and human civilizations. The primary driver is the increasing emission of carbon dioxide (CO_2) into the atmosphere, which acts as a blanket, trapping heat and leading to a rise in global temperatures (Houghton, 2005). Among the various possible solutions, Geological Carbon Sequestration (GCS) stands out as a promising approach to reducing the concentration of atmospheric CO_2. The deployment of GCS involves capturing CO_2 from anthropogenic sources or directly from air, and then transporting it to a designated site where it is injected into deep subsurface geological formations, such as saline aquifers, depleted hydrocarbon reservoirs, and basalts (Aminu et al., 2017; Raza et al., 2022). These ubiquitous geological formations facilitate the large-scale and long-term storage of CO_2, providing a means for the sustainable energy transition to a low-carbon society (Ajayi et al., 2019).

The efficient and safe operation of GCS projects necessitates the identification of optimal engineering strategies, particularly well control schemes, as they exert a considerable influence on CO_2 plume migration and pressure buildup (Cihan et al., 2015; Li et al., 2019; Shamshiri and Jafarpour, 2012). Previous studies have applied numerical simulation along with sensitivity analysis to compare various scenarios of well schemes, offering some qualitative insights into the effects of wells controls on the overall storage performance (Li et al., 2019; Wang et al., 2023; Zhang et al., 2023). For quantitative optimization problems, numerous investigators couple numerical simulation with gradient-based or gradient-free optimization algorithms to determine

* Corresponding author at: Department of Geotechnical Engineering, College of Civil Engineering, Tongji University, Shanghai 200092, China.
E-mail address: fengshou.zhang@tongji.edu.cn (F. Zhang).

https://doi.org/10.1016/j.advwatres.2025.104897
Received 30 September 2024; Received in revised form 13 December 2024; Accepted 13 January 2025
Available online 17 January 2025

the optimal well controls. Shamshiri and Jafarpour (2012) combine the CO2STORE simulation module with the gradient-based BFGS quasi-Newton algorithm to optimize CO_2 injection rates, with the goal of maximizing both stored gas and sweep efficiency. Chen and Pawar (2020) utilize the Stochastic Simplex Approximate Gradient (StoSAG) to estimate the optimal well completions and controls to maximize the net present value (NPV) in a CO_2-EOR operation. Zou and Durlofsky (2023) use the gradient-free particle swarm optimization (PSO) and differential evolution (DE) for optimizing locations and time-varying injection rates, aiming at minimizing the mobile CO_2 fraction and maximizing the storage efficiency. Xie et al. (2024) integrate genetic algorithm (GA) with CMG-GEM simulator to obtain an optimal sequestration scheme based on the proposed uniformly increasing pressure metric. Nevertheless, these simulation-based optimization workflows rely heavily on the iterative calls of numerical simulations to evaluate the objective functions, resulting in a prohibitively high computation burden, especially when applied to large-scale multi-well fields. As an alternative, a surrogate model can be built to promptly approximate the input-output relation without compromising much accuracy, thereby serving as a replacement of the full-order numerical simulator to accelerate the optimization procedure.

Deep learning (DL) methods have been intensively investigated for data-driven surrogate modeling CO_2 multiphase flow in subsurface porous media, which can be roughly categorized into one-step models and sequential models. For one-step models, the prediction of reservoir responses during GCS is taken as an image-to-image regression task. Convolution Neural Networks (CNNs) (Mo et al., 2019a; Wen et al., 2021; Yan et al., 2022b), Variational Auto-Encoder (VAE) networks (Laloy et al., 2017; Mohd Razak et al., 2022), and Generative Adversarial Networks (GANs) (Stepien et al., 2023; Zhong et al., 2019) are adopted as efficient surrogate models to capture the high-dimensional spatiotemporal information in subsurface flow systems. In addition, neural operator-based models such as Fourier Neural Operator (FNO) (Wen et al., 2022; Yan et al., 2022a) and Deep Operator Network (DeepONet) (Diab and Al-Kobaisi, 2023; Jiang et al., 2024) are proposed as operator mappings in function spaces to capture flow states based on initial and boundary conditions. However, these one-step models lack the consideration of temporal dependencies and regard outputs at different time steps as independent identically distributed samples, which is inconsistent with computation process in physical system modeling. Another stream of research utilizes sequential modeling techniques to honor the causality of simulating physics systems. Autoregressive models are employed to iteratively generate predictions through temporal rollout, wherein each prediction serves as the input for the subsequent time step. This process is analogous to time-stepping schemes in numerical simulations, though it is susceptible to error accumulation (Mo et al., 2019b; Tang and Durlofsky, 2024). To address this limitation, Recurrent Neural Networks (RNNs), such as Long Short-Term Memory (LSTM) and its convolutional variant (ConvLSTM), are applied to characterize the sequence trajectories in a recurrent manner (Fan et al., 2024; Feng et al., 2024; Tang et al., 2022). The hidden states inherent in the architecture are stored and updated dynamically, representing the complex spatiotemporal evolution of state variables. However, RNN-based sequential models cannot fully exploit the parallel computing capability of modern GPUs, limiting their effectiveness in handling the long temporal spans typical in GCS processes. Moreover, RNN-based models are susceptible to gradient vanishing and exploding issues, which undermines their ability to effectively capture long-range dependencies.

The Transformer architecture, which is built purely on attention mechanism, offers a promising solution. Originally proposed by Vaswani et al. (2017), Transformers have greatly shaped the trend of Natural Language Processing (NLP) in the DL community (Brown et al., 2020; Devlin et al., 2019). The unprecedented success of Transformers in large language models has inspired researchers to employ them for modeling physical systems (Geneva and Zabaras, 2022; Hang et al., 2024; Hemmasian and Barati Farimani, 2023; Ovadia et al., 2024). Although there has been significant progress in the development of Transformer-based models for NLP and general physics modeling, further progress is needed to provide a more comprehensive model that takes into account the multi-modal features (e.g., high-dimensional input/output fields and time-varying engineering parameters) of GCS surrogate modeling.

In this paper, we develop a CNN-Transformer DL surrogate model, a hybrid end-to-end architecture that leverages the strengths of both CNNs and Transformers to capture local and global information effectively. Specifically, we design a CNN encoder module to compress the high-dimensional geological information into the latent space, a Transformer module to capture global dependencies in parallel through masked self-attention, and a CNN decoder module to map the latent variables to the target size of solutions. Inspired by the design of U-Net (Ronneberger et al., 2015), we add skip shortcuts between the encoder and the decoder to enhance intermediate information interaction. The Transformer architecture is only employed to process sequence data in the latent space, which greatly reduces the number of trainable parameters and increases the generalization of the surrogate model. The proposed CNN-Transformer model, compared to RNN-based models, exhibits satisfactory performance in terms of accuracy, efficiency, and generalizability. Furthermore, we integrate the trained CNN-Transformer model with the Multi-Objective Robust Optimization (MORO) framework to determine the Pareto optimal well controls considering the uncertainties in geological parameters. Our objectives are to maximize the storage efficiency while minimize the pressure perturbation effects caused by CO_2 injection. We also consider the realistic constraint of total storage capacity. The optimization problem is solved by the Non-dominated Sorting-based Genetic Algorithm II (NSGA-II) (Deb et al., 2002).

The rest of the paper proceeds as follows. In Section 2, we describe the surrogate modeling formulation and then provide the details of the proposed CNN-Transformer model. In Section 3, the surrogate model is applied for GCS in 3D heterogeneous geological formations with 4 injection wells under time-varying controls. The performance of the surrogate model is evaluated. Then, in Section 4, the surrogate model is integrated with the MORO framework to determine the Pareto optimal well controls based on an ensemble of uncertain geological realizations. In Section 5, we present further discussions on the surrogate model in terms of comparison with RNNs and generalizability. Finally, we summarize the main findings of this paper in Section 6. In the Appendices, we provide model architecture details, ablation study results, and attention map visualizations.

2. Hybrid CNN-Transformer surrogate model

2.1. Surrogate modeling formulation

In this paper, we consider a multi-phase (gaseous and aqueous) and multi-component (CO_2, H_2O, and NaCl) flow system in porous media, which mimics the conditions of CO_2 injection into deep saline aquifers. We are interested in the spatiotemporal evolution of gas saturation and pore pressure (referred to as saturation and pressure hereafter), as these two state variables characterize the behavior of injected CO_2 and are critical for assessing the capacity and stability of GCS projects (Bachu, 2015; Celia et al., 2015). The mass conservation for each component can be written in the general form Pruess et al. (1999):

$$\frac{\partial}{\partial t}\left(\phi\sum_{\beta}S_{\beta}\rho_{\beta}X_{\beta}^{\kappa}\right)-\nabla\cdot\left(K\sum_{\beta}X_{\beta}^{\kappa}\frac{k_{r,\beta}\rho_{\beta}}{\mu_{\beta}}\left(\nabla P_{\beta}-\rho_{\beta}g\right)\right)-\sum_{\beta}\rho_{\beta}X_{\beta}^{\kappa}q^{\kappa}=0, \tag{1}$$

where the first term is the fluid accumulation within rock pores, the second term is the mass flux described by multiphase Darcy's law, the last term is the source or sink term. The subscript β denotes the fluid

phase (gaseous phase g or aqueous phase *a*), and the superscript κ denotes the component. t is time, ϕ is the rock porosity, S_β is the phase saturation, ρ_β is the phase density, X_β^κ is the mass fraction of component κ in phase β, K is the rock permeability, $k_{r,\beta}$ is the relative permeability of phase β, μ_α is the phase viscosity, g is the gravitational acceleration, q^κ is the well volumetric flow rate of component κ.

Eq. (1) can be discretized using a finite volume scheme, and the state variables including pressure and saturation is calculated iteratively for each control volume at each time step. A single forward simulation is expressed as:

$$[\boldsymbol{P},\ \boldsymbol{S}] = f(\boldsymbol{m}, \boldsymbol{h}), \tag{2}$$

where f stands for the numerical simulation, $\boldsymbol{P},\ \boldsymbol{S} \in \mathbb{R}^{n_t \times n_z \times n_x \times n_y}$ denote the pressure and saturation at n_t time steps, respectively, n_z, n_x, n_y denote the dimensions along *z, x, y* directions, $\boldsymbol{m} \in \mathbb{R}^{n_z \times n_x \times n_y}$ indicates the given geological field, and $\boldsymbol{h} \in \mathbb{R}^{n_t \times n_w}$ represents the time-varying operating parameters of n_w perforated wells.

A surrogate model is regarded as a non-intrusive alternative to the numerical simulation. It is a nonlinear mapping from the input geological and engineering parameter space to the state variable solution space through statistical approximation. Machine learning and deep learning methods can be employed to establish the surrogate model, bypassing the need to solve governing equations iteratively. In this paper, we leverage deep neural networks for their universal approximation power and their capacity of handling high-dimensional problems. The forward simulation process is thus reformulated as follows:

$$[\boldsymbol{P},\ \boldsymbol{S}] \approx [\widehat{\boldsymbol{P}},\ \widehat{\boldsymbol{S}}] = \mathcal{NN}(\boldsymbol{m}, \boldsymbol{h}; \boldsymbol{\theta}), \tag{3}$$

where $\widehat{\boldsymbol{P}},\ \widehat{\boldsymbol{S}}$ are the approximated solutions, $\mathcal{NN}$ is the deep neural network parameterized by $\boldsymbol{\theta}$. For our specific case, $\boldsymbol{m}$ refers to the permeability field, and $\boldsymbol{h}$ denotes the dynamic CO_2 injection rates. It is important to note that the proposed methodology can be generalized to other scenarios where different geological or engineering parameters are involved. As Eq. (3) simulates the time-dependent process, autoregressive or sequence models are well-suited. Considering that autoregressive models suffer from error accumulation due to time rollout, this paper adopts sequence models, specifically Transformers.

Since the solutions for pressure and saturation exhibit distinct patterns (Tang and Durlofsky, 2024), two $\mathcal{NN}$ with the same structure but different trained weights and biases are used. Given a dataset of inputs $\mathscr{X} = \{(\boldsymbol{m}_1, \boldsymbol{h}_1), (\boldsymbol{m}_2, \boldsymbol{h}_2), \ldots, (\boldsymbol{m}_{N_s}, \boldsymbol{h}_{N_s})\}$ and outputs $\mathscr{Y}_P = \{\boldsymbol{P}_1, \boldsymbol{P}_2, \ldots, \boldsymbol{P}_{N_s}\}$, $\mathscr{Y}_S = \{\boldsymbol{S}_1, \boldsymbol{S}_2, \ldots, \boldsymbol{S}_{N_s}\}$, with N_s denoting the number of training samples, our goal is to develop two neural networks

$$\begin{aligned} \mathcal{NN}_P &: \ \mathscr{X} \rightarrow \mathscr{Y}_P, \\ \mathcal{NN}_S &: \ \mathscr{X} \rightarrow \mathscr{Y}_S. \end{aligned} \tag{4}$$

2.2. Neural network architecture

The neural network takes in geological and engineering parameters to predict the dynamics of state variables. The geological parameters and state variables are 3D spatial data, while the engineering parameters are time series data. As such, distinct modules are needed to address this multi-modal dataset. CNNs are particularly effective at capturing spatial hierarchies, making them suitable for modeling the complex relationships inherent in the Euclidean space. Accordingly, we utilize CNNs as the geological encoder and state variables' decoder. In addition, Transformers specialize in processing sequence data in parallel through an efficient attention mechanism. We employ Transformers to extract global patterns from time-varying engineering parameters. Therefore, by leveraging the strengths of both CNNs and Transformers, the proposed surrogate model integrates these structures into a hybrid end-to-end architecture. This subsection first introduces the core concept of Transformers, specifically the attention mechanism, and then provides a detailed description of the entire neural network.

2.2.1. Attention mechanism

The essence of attention mechanism lies in its ability to dynamically focus on different parts of the input according to their relevance to the task at hand. Transformers are a type of model that purely builds upon the attention mechanism, specifically scaled dot-product attention. Given an input sequence $\boldsymbol{X} \in \mathbb{R}^{t \times d}$ of length t and dimension d, it is first mapped to high-dimensional representations as query $\boldsymbol{Q}$, key $\boldsymbol{K}$, and value $\boldsymbol{V}$:

$$\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{W}_q,\ \boldsymbol{K} = \boldsymbol{X}\boldsymbol{W}_k,\ \boldsymbol{V} = \boldsymbol{X}\boldsymbol{W}_v, \tag{5}$$

where $\boldsymbol{Q} \in \mathbb{R}^{t \times d_k}$, $\boldsymbol{K} \in \mathbb{R}^{t \times d_k}$, $\boldsymbol{V} \in \mathbb{R}^{t \times d_v}$, and the linear transformation matrices $\boldsymbol{W}_q \in \mathbb{R}^{d \times d_k}$, $\boldsymbol{W}_k \in \mathbb{R}^{d \times d_k}$, $\boldsymbol{W}_v \in \mathbb{R}^{d \times d_v}$ are learned during model training.

Note that the query here is computed by the input itself as the key and value, thereby resulting in the so-called self-attention. There is also cross-attention, where the query is computed by a different input. We adopt the self-attention mechanism as it outperforms cross-attention based on our numerical experiments. The comparison between the two attention mechanisms can be found in Appendix B. The attention map is calculated with the dot product of the query and key:

$$\boldsymbol{A} = \mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right), \tag{6}$$

where $\boldsymbol{A} \in \mathbb{R}^{t \times t}$, Softmax is the normalized exponential function. The scaling factor $\sqrt{d_k}$ is used for magnitude normalization. Each row of the attention map has positive values that sum up to 1, representing the attention weights that the corresponding input element assigns to every other element in the sequence. The output is then obtained by multiplying the attention map $\boldsymbol{A}$ with value $\boldsymbol{V}$. The computation flowchart is illustrated in Fig. 1a. It should be noted that we have prior knowledge that only historical inputs impact the current output in physical systems. Therefore, to eliminate the influence of data at future time steps, a triangular matrix mask must be applied to the dot product before applying the Softmax function, as shown in Fig. 1b.

To further enhance the expressivity of the attention mechanism, multi-head attention is employed by projecting the input into h different subspaces and implementing the aforementioned scaled dot-product attention in parallel. Each subspace is referred to as an attention head. This enables the model to focus on different representations at different positions. The h outputs are then concatenated and once again projected to yield the final outcome. The procedure of multi-head attention is illustrated in Fig. 1c.

The operations of the attention mechanism in Transformers are all performed at the matrix level, thereby facilitating the parallel computing capabilities of modern GPUs. The contextual information along with long-range dependencies can be well captured by assigning attention scores to different positions in a sequence. The visualization of the attention maps for this paper's task is presented in Appendix C.

2.2.2. The CNN-Transformer

Given the multi-modal dataset, we design a hybrid CNN-Transformer neural network with three key modules: a CNN encoder module to compress the static geological parameters, a Transformer module to process the sequence data, and a CNN decoder module to generate the predictions. The overall architecture is illustrated in Fig. 2, with detailed description of each layer and the corresponding output shape provided in Appendix A.

The first module is the CNN encoder $\mathscr{G}$. It takes the static 3D geological parameters $\boldsymbol{m}$ as an input and compress it into a low-dimensional latent vector $\widetilde{\boldsymbol{m}} \in \mathbb{R}^{n_l}$, as follows:

$$\widetilde{\boldsymbol{m}} = \mathscr{G}(\boldsymbol{m}), \tag{7}$$

where n_l denotes the dimension of the latent space. $\mathscr{G}$ is composed of consecutive 3D convolutional blocks and downsampling blocks. Each

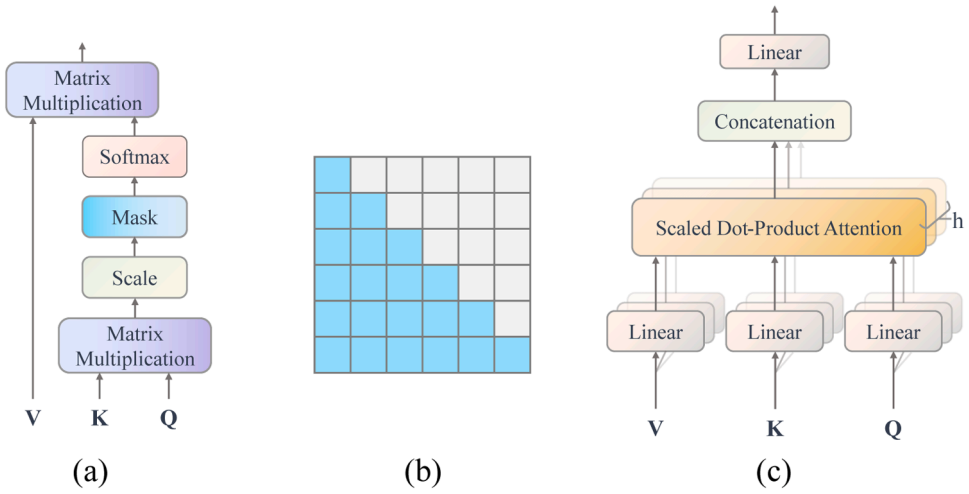


Fig. 1. (a) Scaled dot-product attention. (b) Mask of attention map. Grids highlighted in sky blue represent retained attention scores, while those in grey indicate masked-out scores. (c) Multi-head attention.

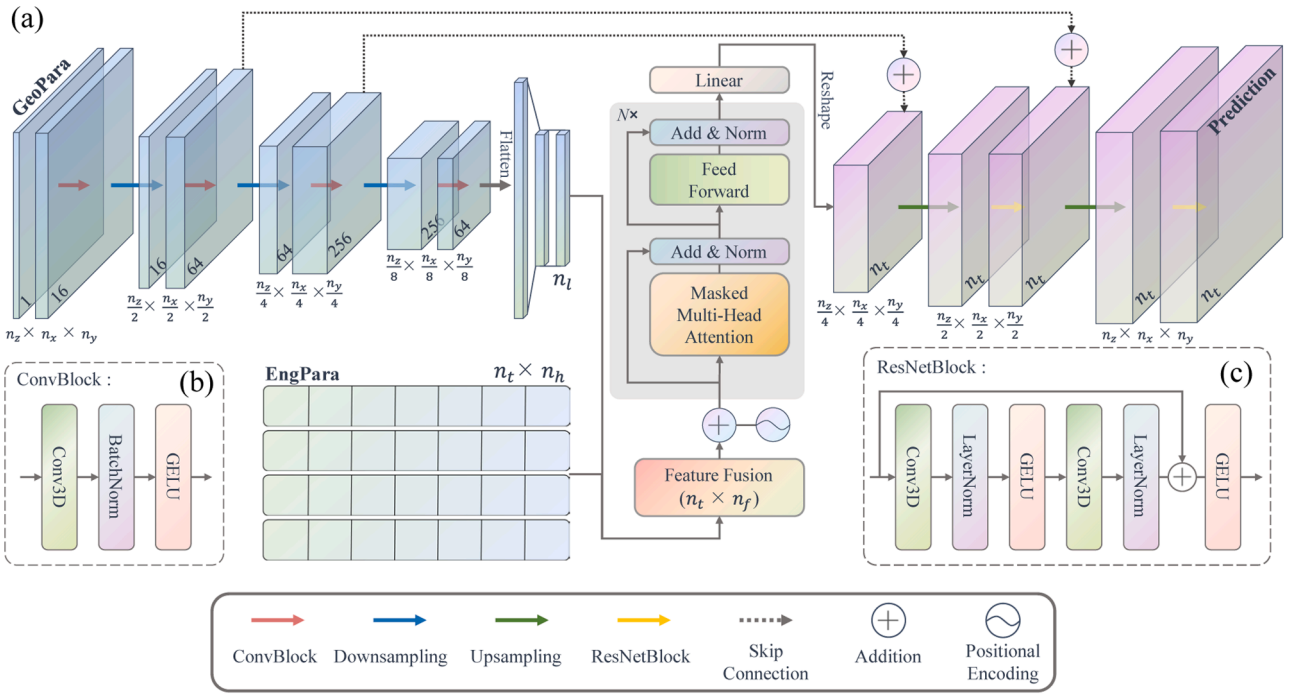


Fig. 2. (a) CNN-Transformer architecture. The static geological and time-varying engineering parameters are encoded to the latent space by the CNN encoder and MLP, respectively. Transformers are then applied to process the latent sequence, followed by the CNN decoder to generate predictions at all time steps simultaneously. Skip connections are used to enhance information flow between the encoder and decoder. The numbers at the corner of each convolutional block denote the channel. (b) Details of the convolutional blocks in the CNN encoder. (c) Details of the ResNet blocks in the CNN decoder.

convolutional block includes a 3D Convolutional (Conv3D) layer, a Batch Normalization (BatchNorm) layer, and a Gaussian Error Linear Units (GELU) (Hendrycks and Gimpel, 2023) activation function, as shown in Fig. 2b. The kernel size of Conv3D is 3 and the padding size is 1. The downsampling block is just a 3D convolutional layer with a kernel size of 3 and a stride of 2, which reduces the input dimension by a factor of 2. After passing through these convolutional blocks, the voxel data is flattened and processed by a Multi-Layer Perceptron (MLP) with one hidden layer to produce the final latent vector. This vector serves as a dense representation of the original geological field with a reduced number of parameters.

The second module is the Transformer processor $\mathscr{T}$. We first map the time-varying engineering parameters $\boldsymbol{h} \in \mathbb{R}^{n_t \times n_w}$ to a high-dimensional space $\widetilde{\boldsymbol{h}} \in \mathbb{R}^{n_t \times n_h}$, where n_h represents the number of features in this transformed space. The time-series records of n_w wells are mapped individually using an MLP with shared parameters. This approach allows the model to learn a consistent transformation across all wells, thereby enhancing generalization and improving computational efficiency. $\widetilde{\boldsymbol{m}}$ is conditioned at every time step to guide the Transformers to attend to specific geological settings. This can be achieved by broadcasting the original $\widetilde{\boldsymbol{m}} \in \mathbb{R}^{n_l}$ to $\widetilde{\boldsymbol{m}} \in \mathbb{R}^{n_t \times n_l}$. The geological and engineering information are then fused as:

$$\widetilde{\boldsymbol{z}} = \langle \widetilde{\boldsymbol{m}}, \widetilde{\boldsymbol{h}} \rangle, \tag{8}$$

where $\widetilde{\boldsymbol{z}} \in \mathbb{R}^{n_t \times n_f}$ is the latent representation with length n_t and

dimension n_f, $\langle \cdot \rangle$ is the fusion operation, which can either be addition, multiplication, or concatenation. We use addition to merge both geological and engineering information based on our ablation study. The comparison between different fusion operations is presented in Appendix B.

As Transformers do not inherently process sequence order, positional encodings are added to $\widetilde{\boldsymbol{z}}$ to provide information about the temporal positions of the data points. Following the approach proposed by Vaswani et al. (2017), we use sine and cosine functions with different frequencies to compute the positional encodings:

$$\begin{aligned} PE_{(pos,2i)} &= \sin\left(pos/10000^{2i/n_f}\right), \\ PE_{(pos,2i+1)} &= \cos\left(pos/10000^{2i/n_f}\right), \end{aligned} \tag{9}$$

where pos is the temporal position of a data point in the sequence (along the n_t dimension), and i is the index of a feature in the positional encoding (along the n_f dimension). As shown in the grey box of Fig. 2a, N layers of Transformers and a linear layer are used to process $\widetilde{\boldsymbol{z}}$. Each Transformer layer contains an attention layer as described in Section 2.2.1, a feed forward MLP, and two Layer Normalization (LayerNorm) layers with residual connections. The computation process can be regarded as:

$$\widetilde{\boldsymbol{z}} = \mathscr{T}(\widetilde{\boldsymbol{z}}). \tag{10}$$

We use the same notation before and after $\mathscr{T}$ processing as it does not change the dimension of the latent representation. The elements of $\widetilde{\boldsymbol{z}}$ have been updated by $\mathscr{T}$ through elaborate operations. Each row of $\widetilde{\boldsymbol{z}}$ is a vector representing the latent variable at the corresponding time step, which needs to be converted to a voxel before entering the 3D convolutional decoder layers. Here we enforce $n_f = \frac{n_z}{4} \times \frac{n_x}{4} \times \frac{n_y}{4}$ to facilitate the use of a simple reshape operation, resulting in the final $\widetilde{\boldsymbol{z}}$ having the shape $n_t \times \frac{n_z}{4} \times \frac{n_x}{4} \times \frac{n_y}{4}$.

The third module is the CNN decoder $\mathscr{D}$, which functions as the mapping from the latent variables to the state variables:

$$\widehat{\boldsymbol{o}} = \mathscr{D}(\widetilde{\boldsymbol{z}}), \tag{11}$$

where $\widehat{\boldsymbol{o}} \in \mathbb{R}^{n_t \times n_z \times n_x \times n_y}$ is either the predicted pressure or saturation. $\mathscr{D}$ consists of two layers of upsampling and ResNet (He et al., 2016) blocks. It is noteworthy that all the channels in $\mathscr{D}$ are fixed as n_t. This ensures that the output of each channel represents the state variable at a specific time step, thereby facilitating parallel prediction of temporal dynamics. The upsampling block is a sequential combination of Conv3D, LayerNorm, GELU, and 3D transpose convolution (TransConv3D). We use LayerNorm instead of BatchNorm in that the channel dimension corresponds to time steps. The upsampling block expands the size of the input voxel by a factor of 2. The details of the ResNet block are depicted in Fig. 2c. It is composed of two stacks of Conv3D, LayerNorm, and GELU with a shortcut connection.

Inspired by the skip connections in U-Net (Ronneberger et al., 2015) to preserve fine-grained spatial information, we add two shortcuts from $\mathscr{G}$ to $\mathscr{D}$. Instead of using concatenation operations as in the original U-Net paper, we utilize linear transformations with output dimensions of 1 and addition operations to transfer intermediate geological features. This approach is motivated by two considerations: First, the number of channels in $\mathscr{D}$ must remain fixed (i.e., n_t) while the static geological information needs to be broadcast to all the time steps; Second, linear transformations with learnable parameters can control the extent of intermediate information flow to $\mathscr{D}$. The effectiveness of the skip connections is demonstrated in Appendix B.

3. Surrogate model evaluation

3.1. Numerical experiment setup

In this paper, we consider the injection of supercritical CO_2 at a time-varying rate into a 3D Cartesian porous media. The size of the storage reservoir is 50m × 10km × 10km, discretized into 8 × 40 × 40 grids in the z, x, and y directions, respectively. There are four injection wells perforated through all layers, as shown in Fig. 3a. The bottom and top boundaries are assumed to be impermeable while the surrounding boundaries are set as open-flow conditions.

The geological model is taken as the heterogeneous permeability field $\boldsymbol{m}$, which follows a log-normal distribution, with a mean of 10 md, and a variance of 1 md. The correlation lengths are 6, 8, and 8 grids in the z, x, and y directions, respectively. We use the Gaussian covariance model from the open-source Python library GSTools (Müller et al., 2022) to generate 6500 random permeability fields. For each well in each geo-model, the CO_2 injection rates vary every six months, with values sampled from a uniform distribution between 4 and 8 kg/s. This corresponds to a storage capacity of 0.5–1.0 Mt/a of CO_2. An example is depicted in Fig. 3.

The reservoir is assumed to be located 1800 m below the surface. At the initial stage, it is fully saturated with brine with a salinity 170 g/L, and the temperature is set at a constant value of 90 °C. The pressure at the top layer is set as 19 MPa, and the initial pressure field is simulated to achieve gravitational equilibrium. The porosity is constant at 0.2 throughout the domain. These initial conditions are based on the In Salah onshore GCS project (Li and Laloui, 2016; White et al., 2014). We use the Van Genuchten capillary model and the Corey relative-permeability model, with parameters obtained from Tang et al. (2022). The details of the numerical model are presented in Table 1. The numerical simulation is performed using the well-established TOUGH3 simulator (Jung et al., 2017) on an Intel i9-13900K CPU. The results of the numerical simulation serve as the labels for supervised training of the surrogate model. The numerical model runs for 20 years, with pressure and saturation data output every six months, resulting in a sequence length of 40.

The dataset consists of 6500 samples and is divided into training, validation, and testing sets, with proportions of 80 %, 10 %, and 10 %, respectively. We apply min-max normalization for all inputs and outputs, except for saturation, which is already within the range of 0 to 1. We implement the surrogate model using the deep learning library PyTorch (Paszke et al., 2019). The neural network parameters are optimized using the AdamW optimizer (Loshchilov and Hutter, 2019) with a weight decay of 0.05. The loss criterion is the standard L^2 loss. The batch size is 10. The learning rate is set to 0.001 and automatically decreases when there is no improvement in validation loss. We train and validate the model for 800 epochs on an Nvidia GeForce RTX4090 GPU for pressure and saturation, respectively.

3.2. Evaluating metrics

The surrogate model is evaluated based on the extent of discrepancy between predictions and labels. As the magnitudes of saturation and pressure values differ, we use two distinct error metrics for these state variables. The magnitude of pressure is on the order of megapascals. To ensure the residual error is within a reasonable range, we use the Relative Root Mean Square Error (RRMSE). For each sample, the error of pressure at a specific time step t is calculated by:

$$\text{RRMSE}\ (\%) = \sqrt{\frac{1}{n_z n_x n_y} \sum_{i}^{n_z} \sum_{j}^{n_x} \sum_{k}^{n_y} \left(\frac{\widehat{P}^t_{i,j,k} - P^t_{i,j,k}}{P^t_{i,j,k}} \right)^2} \times 100, \tag{12}$$

where $\widehat{P}^t_{i,j,k}$ and $P^t_{i,j,k}$ are the pressure values of grid (i, j, k) at time t generated by the surrogate model and the numerical simulation, respectively.

Meanwhile, the range of saturation is between 0 and 1. Therefore, we use the standard RMSE as the evaluating metric. For each sample, the error of saturation at a specific time step t is calculated by:

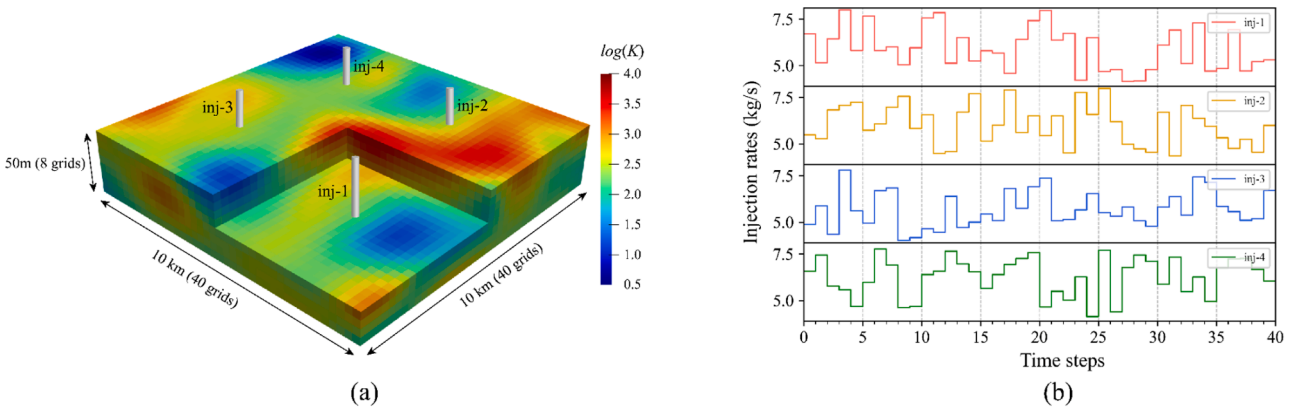


Fig. 3. The configurations of the numerical model. An example of (a) the 3D heterogeneous permeability field, and (b) the time-varying CO_2 injection rates.

Table 1
Parameters of the numerical model.

Reservoir parameters	Values
Domain size	$L_z = 50$m, $L_x = 10$km, $L_y = 10$km
Discretized grids	$n_z = 8$, $n_x = 40$, $n_y = 40$
Depth	1800 m
Brine salinity	170 g/L
Temperature	90 °C
Initial pressure at top	19 MPa
Porosity, ϕ	0.2
Permeability field, $\boldsymbol{m}$	$\mathcal{N}(10,\ 1)$ md
Engineering parameters	**Values**
Number of wells, n_w	4
Injection rates, $\boldsymbol{h}$	$\mathcal{U}(4,\ 8)$ kg/s
Relative permeability	**Values**
Irreducible aqueous saturation, S_{ar}	0.11
Residual gas saturation, S_{gr}	0.01
Exponential coefficient, n_a	4
Exponential coefficient, n_g	2
Capillary pressure	**Values**
Exponential coefficient, λ	0.254
Irreducible aqueous saturation, S_{ar}	0.11
Maximum capillary pressure, $P_{\max}$	12,500 Pa

$$\text{RMSE}\ (\%) = \sqrt{\frac{1}{n_z n_x n_y}\sum_{i}^{n_z}\sum_{j}^{n_x}\sum_{k}^{n_y}\left(\widehat{S}^t_{i,j,k} - S^t_{i,j,k}\right)^2} \times\ 100, \tag{13}$$

where $\widehat{S}^t_{i,j,k}$ and $S^t_{i,j,k}$ represent the saturation from the surrogate model and the numerical simulation, respectively. These two error metrics can be averaged over time steps or samples for evaluating purposes. We express RRMSE and RMSE in percentage form to magnify the values, thereby facilitating easier comparison.

Additionally, the non-dimensional coefficient of determination (denoted as R^2) is used for both pressure and saturation:

$$R^2 = 1 - \frac{\sum_i^{n_b}(y_i - \widehat{y}_i)^2}{\sum_i^{n_b}(y_i - \bar{y})^2}, \tag{14}$$

where n_b is the number of grid blocks being evaluated, y_i is the label, $\widehat{y}_i$ is the prediction, and $\bar{y}$ is the mean value. R^2 ranges from 0 to 1, with a higher value indicating a better fit.

3.3. Test set performance

Fig. 4 shows the evolution of L^2 loss during the training and validation stage for pressure and saturation. The loss values of both state variables converge to a plateau after 800 epochs, with some abrupt decreases due to the automatic adjustment of the learning rate. The ultimate minimum training and validation loss of saturation are 1.75×10^{-6} and 2.13×10^{-6} respectively, while those of pressure are 3.87×10^{-6} and 3.62×10^{-6}. This indicates that the predictions are very close to the labels. The minimum loss values of saturation are smaller than those of pressure in that there are many zero values in the saturation maps (e.g., the outer region of the CO_2 plume). Besides, due to the large weight decay value, no significant over-fitting issues are observed.

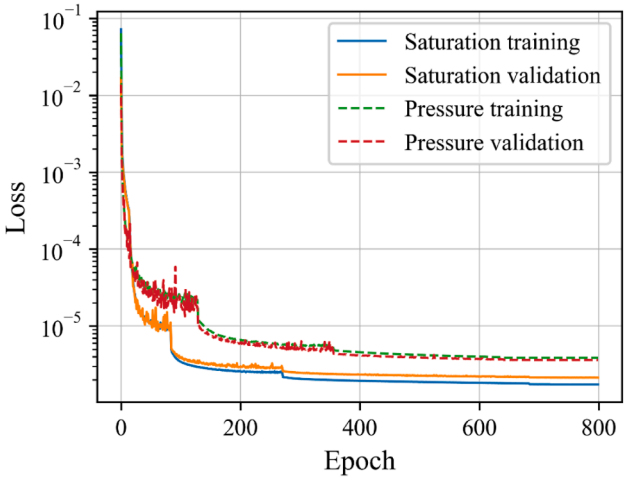


Fig. 4. Training and validation losses vs. epoch for pressure and saturation.

Fig. 5 displays the saturation fields of a random test sample at three selected time steps (i.e., 5, 10, and 20 years). The geological and engineering parameters of this test sample are shown in Fig. 3. The first row in Fig. 5 presents the results predicted by the surrogate model, the second row shows the results generated by the high-fidelity numerical simulation, and the third row displays the scatter plots of the corresponding saturation fields. Considering that the neural network might produce some values close to zero in the outer regions of CO_2 plume due to numerical issues, we only showcase the saturation values larger than 0.01 for visualization purposes. The variability of the CO_2 plume distribution around the four injection wells can be observed as a result of the heterogeneous permeability field and distinct injection schemes. As time progresses, the CO_2 plume moves upward and spreads across the top layer due to the buoyancy effect. From Fig. 5a to f, there is very close agreement between the surrogate predictions and the simulation outputs. RMSEs at the three time steps are 0.091 %, 0.100 %, and 0.148 %, respectively. The scatter plots in Fig. 5g–i further demonstrate the excellent performance of the surrogate model, with R^2 scores all greater

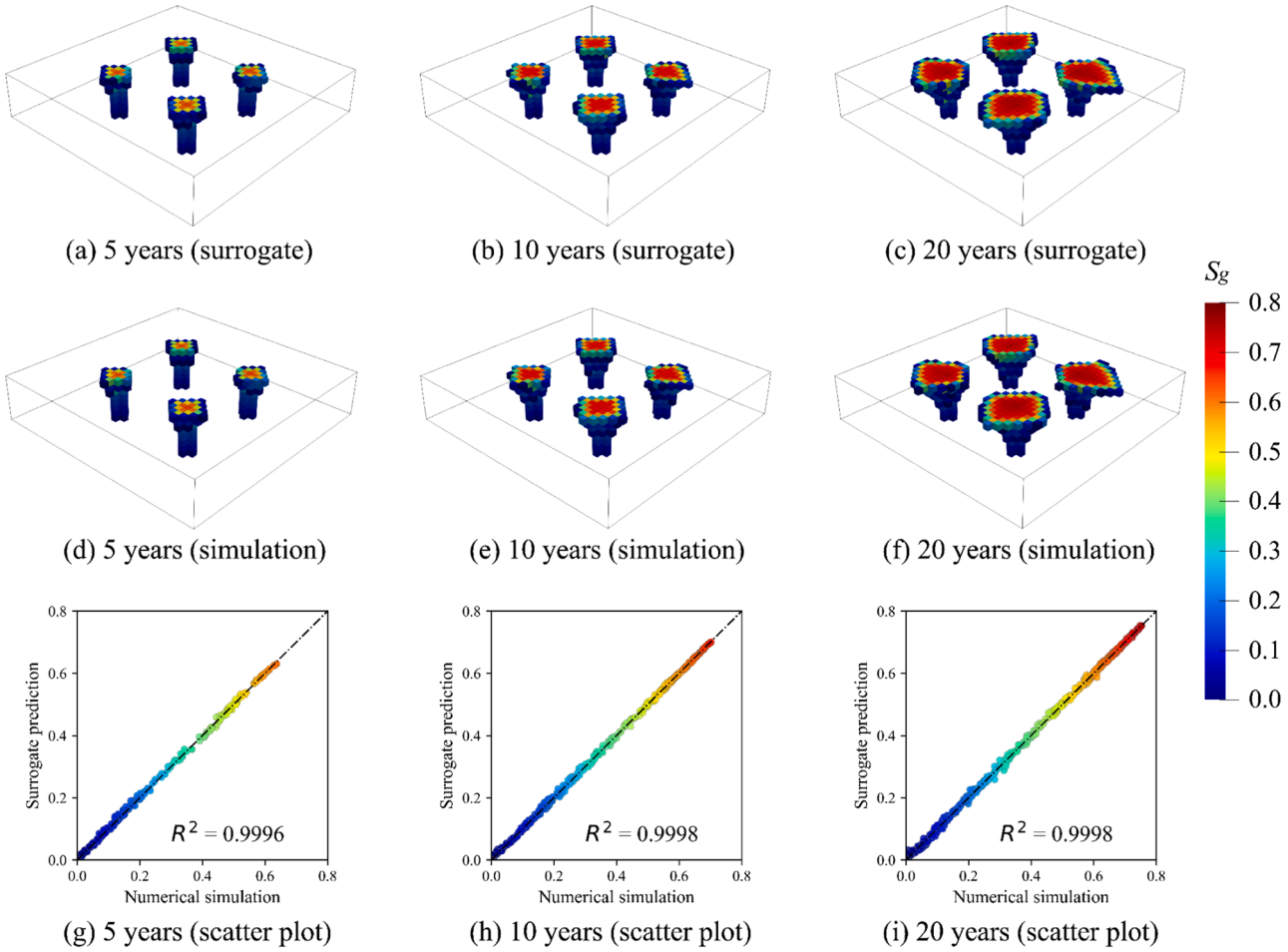


Fig. 5. Saturation fields from CNN-Transformer surrogate model (first row) and high-fidelity numerical simulation (second row) for a random test case at 5, 10 and 20 years. The third row is the scatter plots of the correspondence saturation fields, with point colors representing the values from the simulation.

than 0.9996.

The results of pressure for the same testing sample are illustrate in Fig. 6. The initial pressure in the reservoir is around 19 MPa, while the pressure builds up around the four wells and dissipates near the boundaries due to the open-flow conditions. The size of the over-pressurization zones around the four wells differs because of the complex interaction between the heterogeneous geo-model and the different well control parameters. The time-varying injection rates bring about complex patterns of expansion and contraction of the over-pressurization zone. The spatiotemporal evolution of the pressure fields is well captured by the surrogate model. The correspondence between the surrogate and simulation results is very satisfactory, as shown in Fig. 6a–f. The values of RRMSE at 5, 10, and 20 years are 0.118 %, 0.143 %, and 0.117 %, respectively. The scatter plots in Fig. 6g–i further proves the decent accuracy of the surrogate mode, with R^2 scores exceeding 0.9995. While there are minor discrepancies around the wells, visible as scatter points with large values in Fig. 6g–i, these differences are negligible considering the large pressure values (e.g., in megapascals).

In Fig. 7, multiple heterogeneous permeability fields with distinct CO_2 plume morphology along with the pressure distribution at the last time step (i.e., 20 years) are presented. Despite the diverse patterns of different realizations, the surrogate model can capture the long-term evolution of CO_2 saturation and pressure accurately. R^2 scores are higher than 0.99 for all realizations, and the error metrics are at low levels.

To evaluate the statistical correspondence across the whole test set, we compute the 10th, 50th and 90th percentile (denoted as P10, P50, and P90) results for saturation and pressure at the 40 time steps, as adopted in Han et al. (2024b) and Tang et al. (2022). We collect the data from four observation positions, which are offset by two grids from the injection wells in the x and y directions at the top layer, as shown in Fig. 8. The injection wells are not chosen to be observation positions because the values of the well perforation grids are usually high, making it difficult to observe clear trends for comparison purposes.

Results of saturation and pressure data are presented in Figs. 9 and 10. The solid curves denote the outputs from the numerical simulation, and the dashed curves represent the predictions generated by the surrogate model. The results of P10, P50, and P90 are shown in colors ranging from light to dark red. Interestingly, although the injection wells are located central symmetry, the responses near the wells are not identical, especially for saturation. This highlights the impact of heterogeneity in permeability and time-varying rates for CO_2 injection. Excellent agreement between the simulation and surrogate results is consistently observed for both saturation and pressure. It reveals that the CNN-Transformer surrogate model can produce satisfactory predictions for an ensemble of realizations. This capability enables us to employ the surrogate model in the MORO framework for optimization purposes.

4. Optimization with surrogate model

4.1. Multi-objective robust optimization formulation

Without loss of generality, a multi-objective optimization problem can be formulated as:

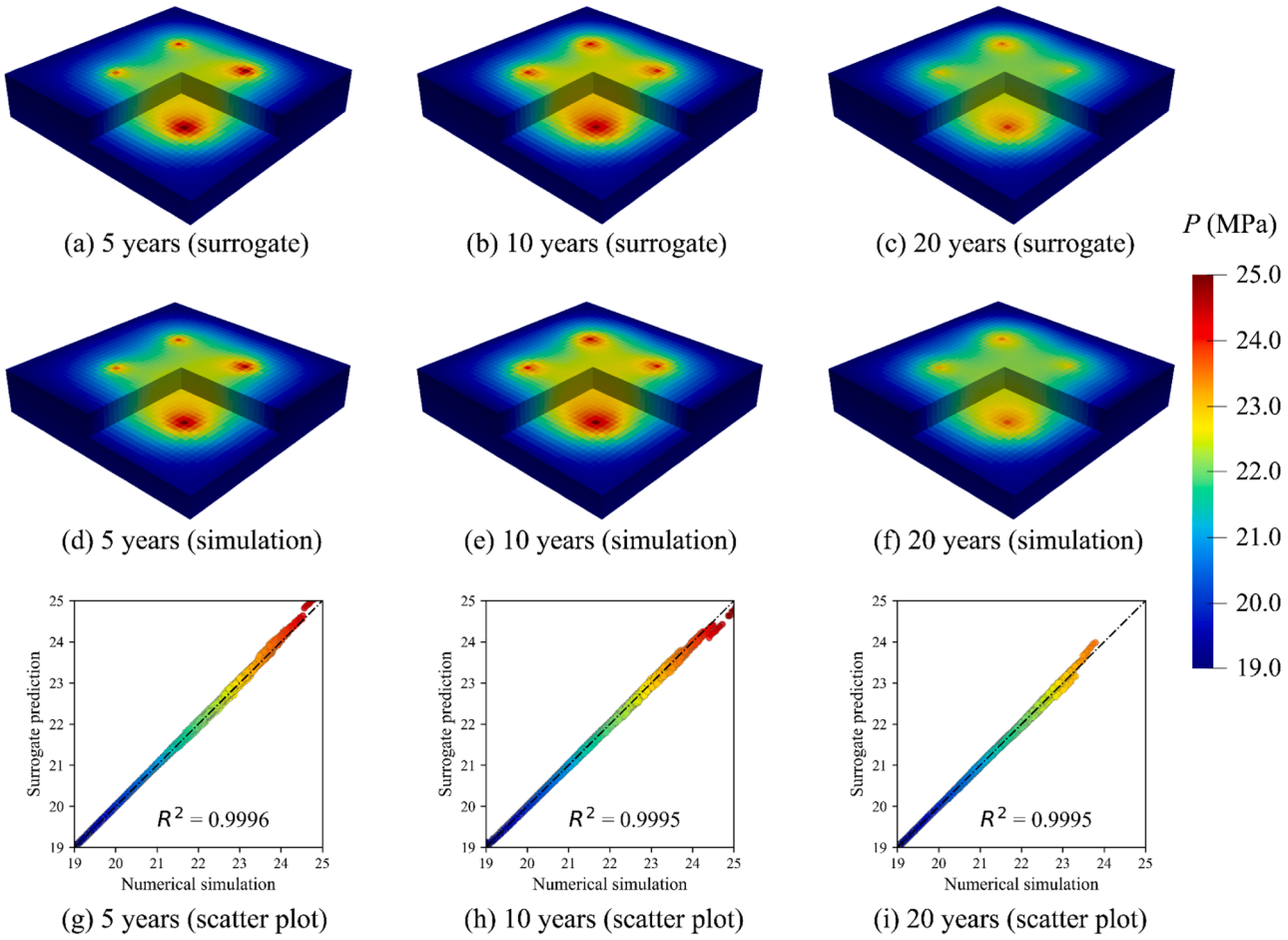


Fig. 6. Pressure fields from CNN-Transformer surrogate model (first row) and high-fidelity numerical simulation (second row) for a random test case at 5, 10 and 20 years. The third row is the scatter plots of the correspondence pressure fields, with point colors representing the values from the simulation.

$$\begin{aligned} &\min_{\boldsymbol{u}\in\boldsymbol{U}} J_i(\boldsymbol{u}) && i=1,2,\ldots,N_i, \text{ and } N_i \geq 2 \\ &\text{s.t. } g_j(\boldsymbol{u}) \leq 0 && j=1,2,\ldots,N_j, \\ &p_k(\boldsymbol{u}) = 0 && k=1,2,\ldots,N_k, \end{aligned} \tag{15}$$

where $\boldsymbol{u}$ is the decision variable vector to be optimized in the search space $\boldsymbol{U}$, J_i is the ith objective function, g_j is the jth inequality constraint, and p_k is the kth equality constraint. N_i, N_j, and N_k are the number of objective functions, inequality and equality constraints, respectively.

J_i can have variable magnitudes and may even be conflicting to each other. Therefore, in many cases, it is not feasible to convert this problem into a single objective problem by the weighted sum approach. Instead, our goal is to achieve the Pareto optimality. A decision variable $\boldsymbol{u}^*$ is Pareto optimal if $\boldsymbol{u}^* \in \boldsymbol{U}$ and there exists no $\boldsymbol{u} \neq \boldsymbol{u}^* \in \boldsymbol{U}$ such that 1) $J_i(\boldsymbol{u}^*) \leq J_i(\boldsymbol{u})$ for $i = 1, 2, \ldots, N_i$ and 2) $J_s(\boldsymbol{u}^*) < J_s(\boldsymbol{u})$ for at least one s (Censor, 1977). The set of Pareto optimal solutions forms the Pareto front, a boundary in the objective space beyond which any improvement in one objective results in a degradation of at least one other objective. The optimal decision variables can then be determined based on the trade-offs among the objectives.

In this paper, we aim to obtain the optimal time-varying operating parameters $\boldsymbol{h}$ considering the uncertainty of the geological model $\boldsymbol{m}$, thereby achieving the so-called robust optimization (Aliyev and Durlofsky, 2017; Yan et al., 2023). Each objective function J_i is calculated as the expectation for an ensemble of geological realizations, $\{\boldsymbol{m}_1, \boldsymbol{m}_2, \ldots, \boldsymbol{m}_{N_e}\}$, which can be expressed as:

$$\bar{J}_i(\boldsymbol{h}) = \frac{1}{N_e}\sum_{r=1}^{N_e} J_i(\boldsymbol{h}, \boldsymbol{m}_r), \tag{16}$$

where N_e is the total number of realizations, and $J_i(\boldsymbol{h},\boldsymbol{m}_r)$ is evaluated based on reservoir responses [$\boldsymbol{P}$, $\boldsymbol{S}$] computed from given [$\boldsymbol{h}$, $\boldsymbol{m}_r$]. Traditional optimization procedures require iteratively calling the numerical simulator to generate [$\boldsymbol{P}$, $\boldsymbol{S}$]. In contrast, once the surrogate model is trained, it can serve as an accurate and fast proxy of the numerical simulator. Hence, we propose to integrate the developed surrogate model with the MORO framework to speed up the decision-making process.

4.2. Objective function and optimization algorithm

The optimization problem involves the determination of the optimal injection scheme $\boldsymbol{h} \in \mathbb{R}^{40\times 4}$ such that 1) the storage efficiency is maximized, and 2) the induced over-pressurization is minimized.

The storage efficiency E is defined as the percentage of stored CO_2 volume within the reservoir pore volume (Bachu et al., 2007; Heath et al., 2014):

$$E = \frac{\sum_i^{n_z}\sum_j^{n_x}\sum_k^{n_y} S_{i,j,k} \times V_{i,j,k} \times \phi_{i,j,k}}{\sum_i^{n_z}\sum_j^{n_x}\sum_k^{n_y} V_{i,j,k} \times \phi_{i,j,k} \times \left(1 - S_{i,j,k}^{ar}\right)} \times 100, \tag{17}$$

where $S_{i,j,k}$ is the saturation value of grid index (i, j, k) at the last time step, $V_{i,j,k}$ is the grid volume, $\phi_{i,j,k}$ is the grid porosity, $S_{i,j,k}^{ar}$ is the grid irreducible aqueous saturation. The numerator in Eq. (17) denotes the gaseous CO_2 volume in the reservoir, and the denominator represents the total accessible pore volume of the reservoir. After 20 years of injection, the larger E is, the more amounts of CO_2 is stored, consuming the

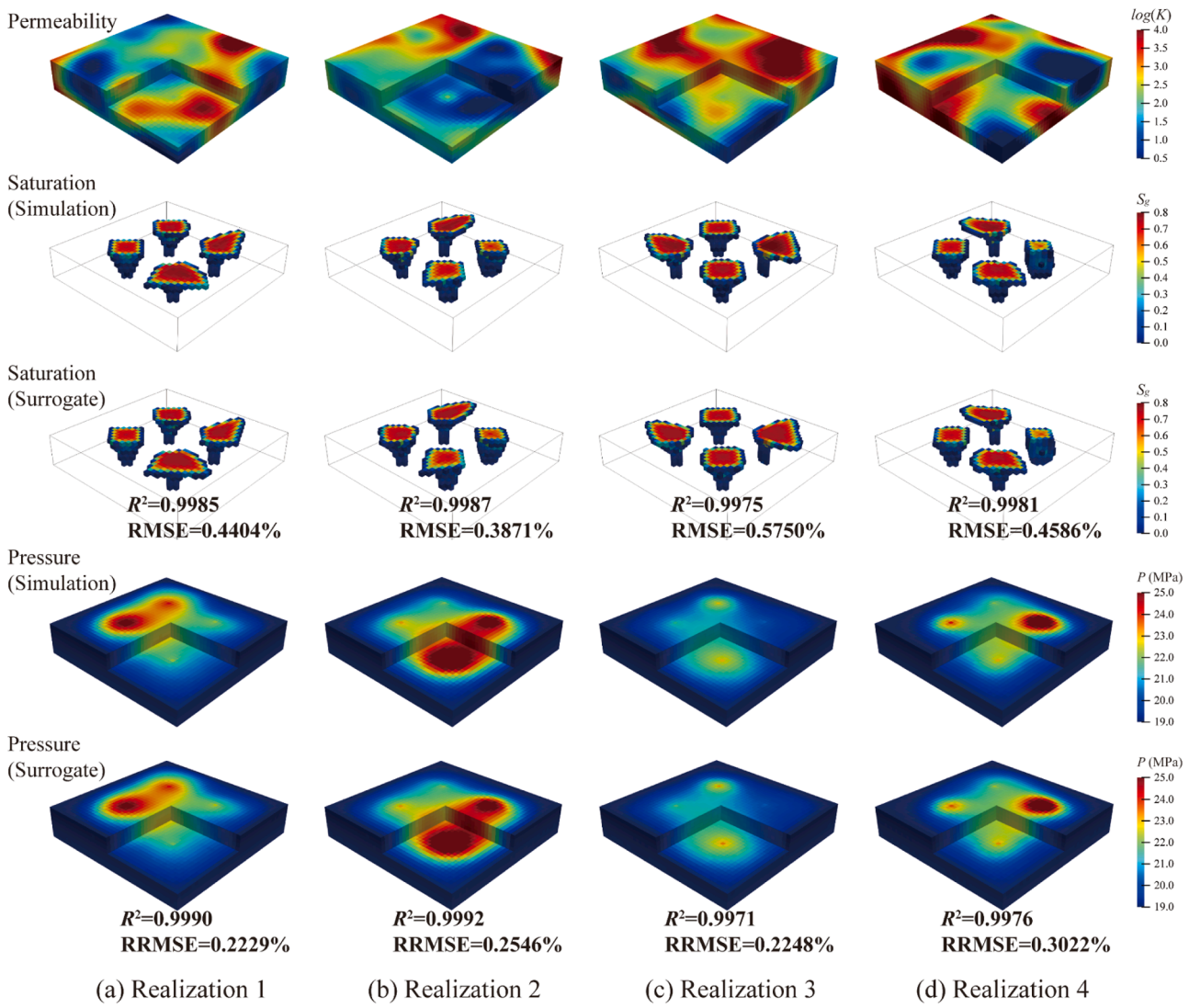


Fig. 7. Saturation and pressure fields for 4 test realizations at the last time step.

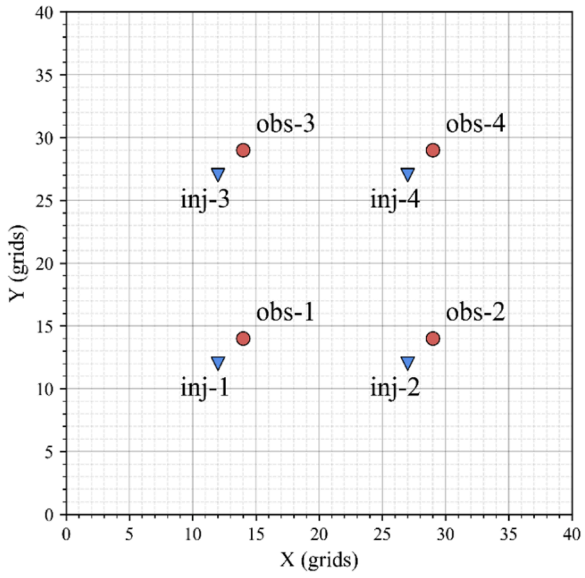


Fig. 8. Locations of the four injection wells (the blue inverted triangles) and the four observation positions (the red circles) at the top layer.

minor effects of solubility and mineral trapping are neglected (Raza et al., 2022). From a perspective of economic benefits, we would like to seek for optimal injection schemes which could maximize E.

However, there are some safety issues that should be considered during GCS projects. For example, the reservoir over-pressurization caused by fluid injection might bring about the risks of caprock integrity (Ju et al., 2021; Shukla et al., 2010), induced seismicity (Segall and Lu, 2015; Zoback and Gorelick, 2012), and surface deformation (Ramirez and Foxall, 2014; Rutqvist et al., 2010). For the purpose of pressure management, the effects of over-pressurization perturbation should be minimized. Considering that the reservoir pressure fluctuates dynamically due to the time-varying injection rates and would dissipate after well shut-in, we propose to minimize the peak over-pressurization values during the injection period to ensure the stable and safe operation of the GCS project. The maximum over-pressurization corresponding to one sample is calculated by:

$$\Delta P = \max_{i,j,k,t}\left(P^t_{i,j,k} - P_0\right), \tag{18}$$

where $P^t_{i,j,k}$ is the grid pressure value at time step t, and P_0 is the initial reservoir pressure.

Eqs. (17) and (18) are the two objective functions that we aim to maximize and minimize, respectively. It is obvious that they conflict

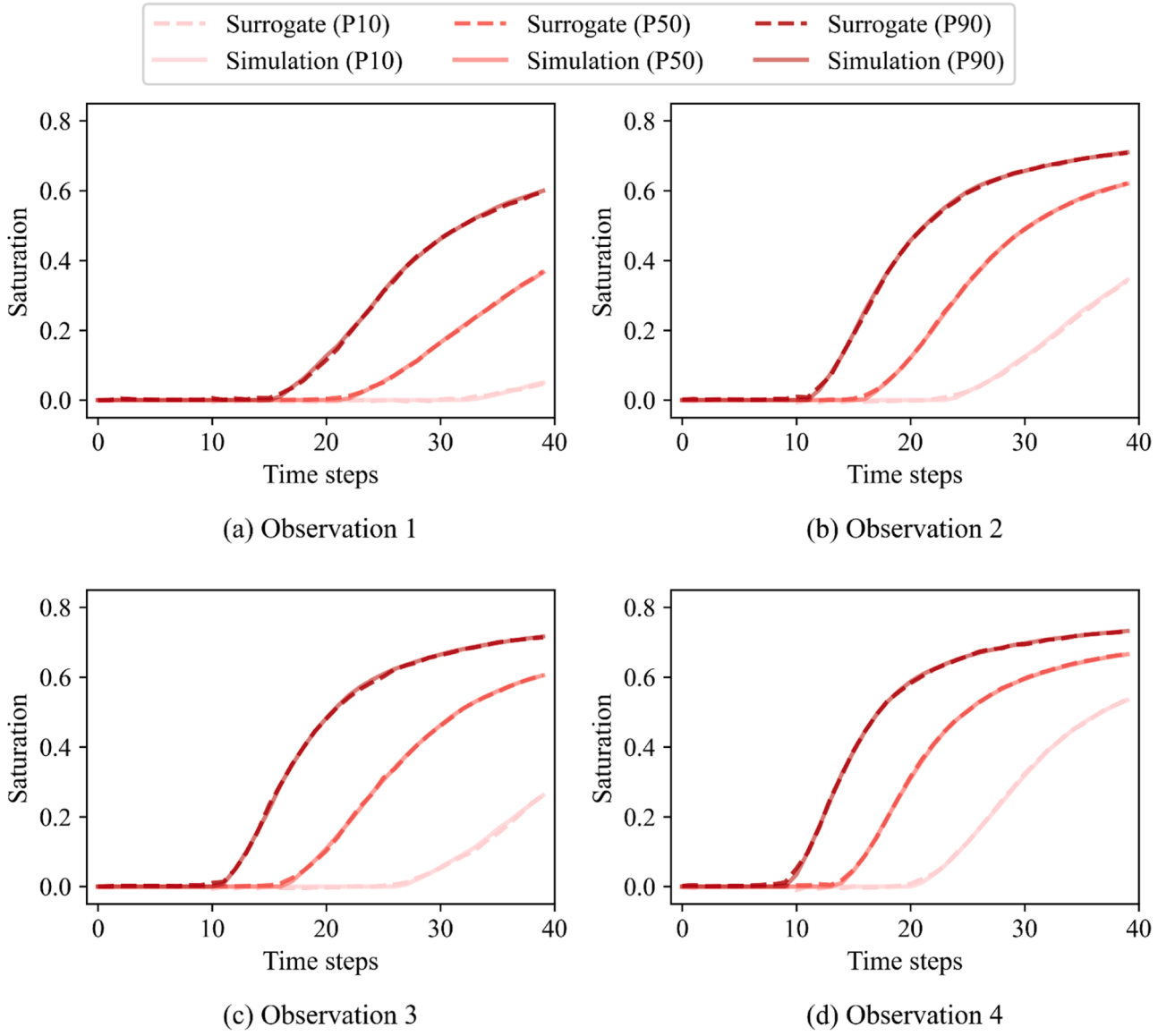


Fig. 9. Statistical results of saturation at the four observation positions (1–4) over the entire test set (650 samples). Dashed and solid curves represent the surrogate and simulation results, respectively. The P10, P50, and P90 curves are denoted by light, medium, and dark red, respectively.

with each other as an increase in the stored CO_2 amount comes at the cost of increased over-pressurization. In addition, we consider an inequality constraint for the injection rates to satisfy the storage target:

$$\frac{1}{40 \times 4} \sum_{t=1}^{40} \sum_{n=1}^{4} h_{t,n} \geq Q_{trg}, \tag{19}$$

where $h_{t,n}$ is the CO_2 rate of the n-th injection well at time step t, Q_{trg} is the predefined target value. This constraint ensures the mean injection rate of each well is greater than the target value Q_{trg}. Here we set Q_{trg} to be 6 kg/s, which corresponds to a storage capacity of 0.75 Mt/a.

We consider an ensemble realizations of permeability fields to formulate the MORO problem. Specifically, we take random selected 100 samples from the test set as a stochastic ensemble, and calculate the expectation of the objective functions over these samples. We accomplish the optimization task using the Non-dominated Sorting-based Genetic Algorithm II (NSGA-II) proposed by Deb et al. (2002). The algorithm follows the general outline of the classic genetic algorithm. The overall flowchart of the optimization process is illustrated in Fig. 11. The initial population is first generated by sampling from the search space, and each individual is a vector with 160 dimensions corresponding to the injection rates of the 4 wells across 40 time steps. The population is then evaluated based on the expected values of the objective functions over the 100 stochastic samples to find the Pareto optimality. It should be noteworthy to mention that we evaluate the objective functions by the trained CNN-Transformer surrogate model instead of the numerical simulation. Next, NSGA-II uses the binary tournament mating technique combined with the crowding distance sorting to select the individuals at the Pareto front. The crossover operation is then applied on the selected parent individuals to generate their offspring. Mutation can happen among the offspring with a predefined probability, which helps to increase the diversity of the population and reduce the risk of local optimality. The final population is a new generation, and the above procedures repeat until the termination condition is satisfied, such as reaching the maximum generation number. The algorithm is implemented based on the open-source library pymoo (Blank and Deb, 2020). The hyperparameters of the optimization algorithm are detailed in Table 2.

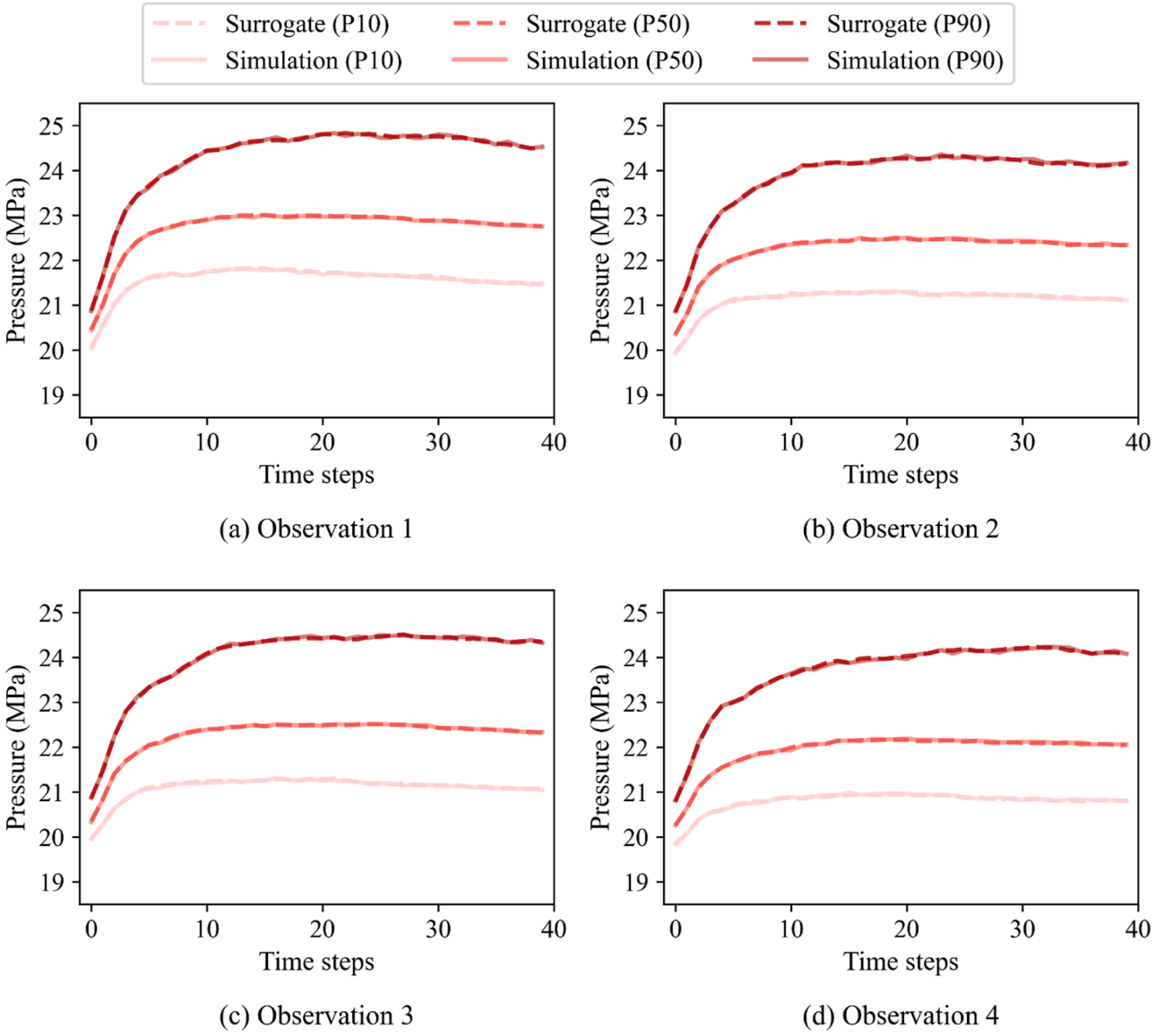


Fig. 10. Statistical results of pressure at the four observation positions (1–4) over the entire test set (650 samples). Dashed and solid curves represent the surrogate and simulation results, respectively. The P10, P50, and P90 curves are denoted by light, medium, and dark red, respectively.

4.3. Optimization results

The minimum value of over-pressurization and the maximum value of storage efficiency among the population are recorded as a function of generation, as depicted in Fig. 12a. After 300 generations, the optimization process terminates as the two objective functions show no further obvious improvement. Fig. 12b presents the ultimate Pareto front along with the historical individuals in the objective space. The individuals are initially generated randomly and move towards the Pareto front as optimization progresses. Each individual on the Pareto front denotes an achievable best solution (i.e., the optimal injection rates). One can select the optimal solution from the Pareto front based on the preferred trade-offs between the two objectives. In this paper, for comparison purposes, we choose the minimum (blue square), median (green star), and maximum (red triangle) level of the storage efficiency objective as three representative cases, as marked in Fig. 12b.

The optimized injection rates of the three representative cases are presented in Fig. 13. The three cases show distinct well control patterns due to the different levels of the storage efficiency. For the minimum E case in Fig. 13a, the time-varying CO_2 injection rates start with low values and fluctuate frequently with mean values around 6.05 kg/s, which is close to constrained rate threshold. For the medium E case in Fig. 13b, the rates initially fluctuate and then remain at high levels around 7.5 kg/s. The mean value of the 4 wells over time is 6.82 kg/s, which is higher than the minimum E case. Fig. 13c shows the injection rates for the maximum E case. It can be observed that the rates tend to stabilize near high values of 7.29 kg/s. The standard deviations are the lowest among the three cases, as the rates show no obvious variation. It should be noted that these well control parameters are applied to the statistical ensemble in the robust optimization scenario. The optimal solutions would be more general and realistic if the uncertainty in geological models is reduced, which could be accomplished by the history matching technique when field monitored data are available (Han et al., 2024a).

We compare the optimized injection rates with the constant lower limit (6 kg/s) to showcase the efficacy of MORO. We use the four types of injection schemes to calculate the two objectives for the 100 realizations. The histograms over the ensemble are presented in Fig. 14. The two rows show the results of the storage efficiency and over-pressurization objectives, respectively. Each column corresponds to an

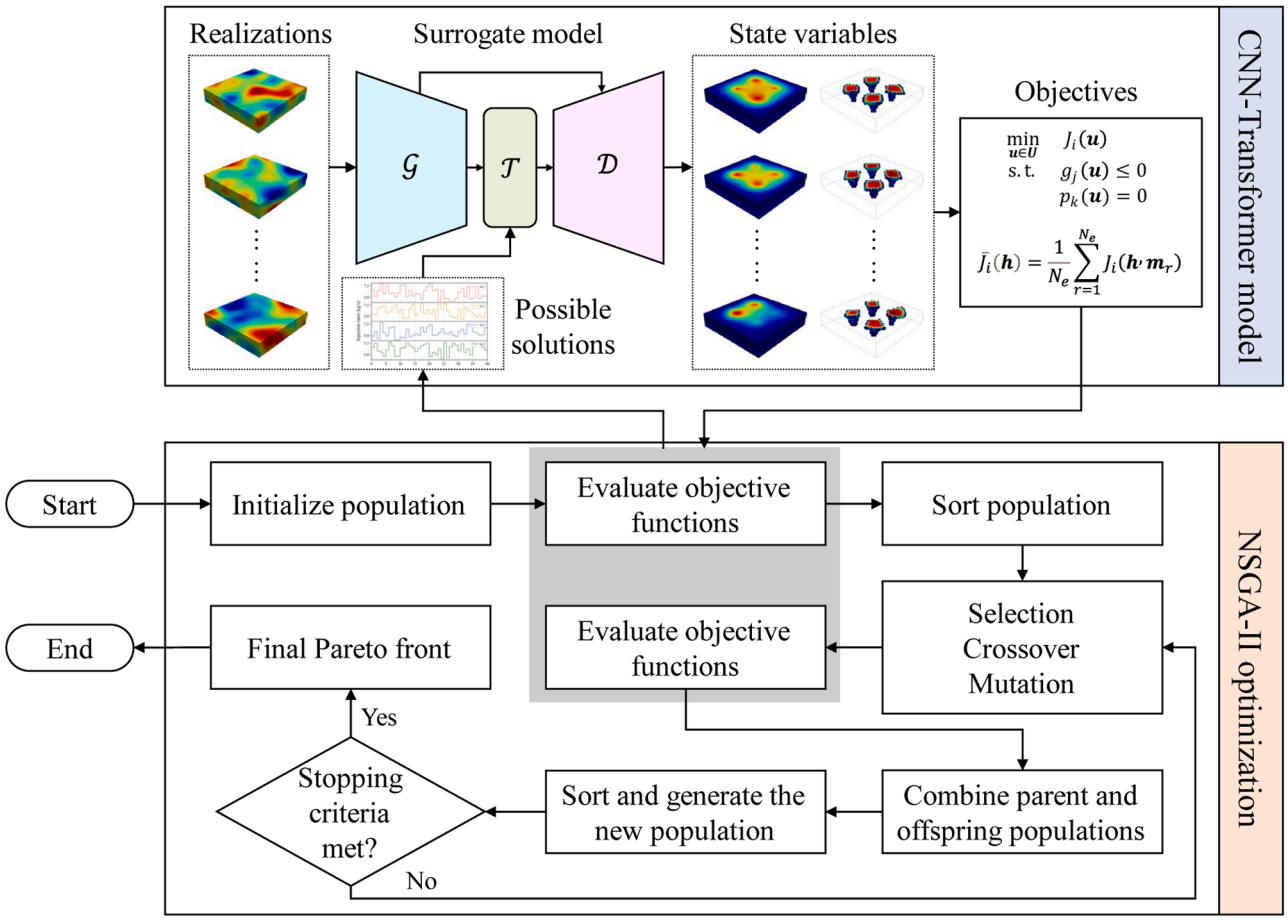


Fig. 11. Surrogate-integrated MORO framework. The processes of objective function evaluation (grey box) are accomplished by calling the trained CNN-Transformer surrogate model.

Table 2
Hyperparameters of the NSGA-II optimization algorithm.

Name	Value
Population size	350
Offspring size	200
Crossover method	Simulated binary
Crossover probability	0.9
Crossover index	15
Mutation method	Polynomial
Mutation index	20
Maximum generation	300

injection scheme case. For the storage efficiency objective, it can be observed that the constant rate case performs similar to the minimum *E* case, with a mean value 0.02 % higher than that of the minimum *E* case. The performance of the medium and maximum *E* case is much better, whose mean values are 2.86 % and 3.08 %, respectively. For the over-pressurization objective, the minimum *E* case exhibits the lowest level of pressure perturbation, with a mean value around 7.62 MPa. In contrast, the constant rate case induces an average over-pressurization of 8.77 MPa, which is 15 % larger than the minimum *E* case despite the two cases having similar mean injection rates. The medium *E* case is even slightly better than the constant rate case, with a mean over-pressurization 0.36 MPa lower. The maximum *E* case has the highest mean over-pressurization, with a value of 9.62 MPa. These results provide us with some insights into the efficacy of the optimized time-varying injection rates. For instance, the minimum *E* case achieves a similar storage efficiency target without inducing sever over-pressurization compared to the constant rate case. The medium *E* case improves storage efficiency while causes comparable over-pressurizations. The maximum *E* case, which has the highest storage efficiency and over-pressurization, can be adopted if the pressure perturbation is acceptable.

The optimized injection rates show distinct patterns compared to those in the training dataset, which are randomly sampled from a uniform distribution. To validate the optimization results, we compare the calculated objective functions over the 100 realizations based on both surrogate model and numerical simulation, as shown in Fig. 15. The mean and standard deviation values of storage efficiency calculated from surrogate model and numerical simulation are quite close for all the three representative injection scheme cases. The discrepancy for over-pressurization is also acceptable considering the megapascal magnitude of pressure. These results demonstrate that the CNN-Transformer surrogate model can well generalize to unseen well controls.

During each optimization generation, 350 individuals are evaluated over the 100 realizations. We run the optimization procedure by 300 generations, resulting in 10.5 million calls of the forward model. It would be impossible to use numerical simulation to accomplish this task, as each forward simulation takes around 5 min. Running the simulations in a sequential mode would take almost 99.89 years. In contrast, once trained, our proposed CNN-Transformer surrogate model can generate predictions within a second and can be evaluated in a batch mode. The total optimization process only takes 71 min when the surrogate model is employed. Therefore, integrating the surrogate model with the MORO framework reduces computational time by 99.99 %, achieving an acceleration of approximately 739,000 times.

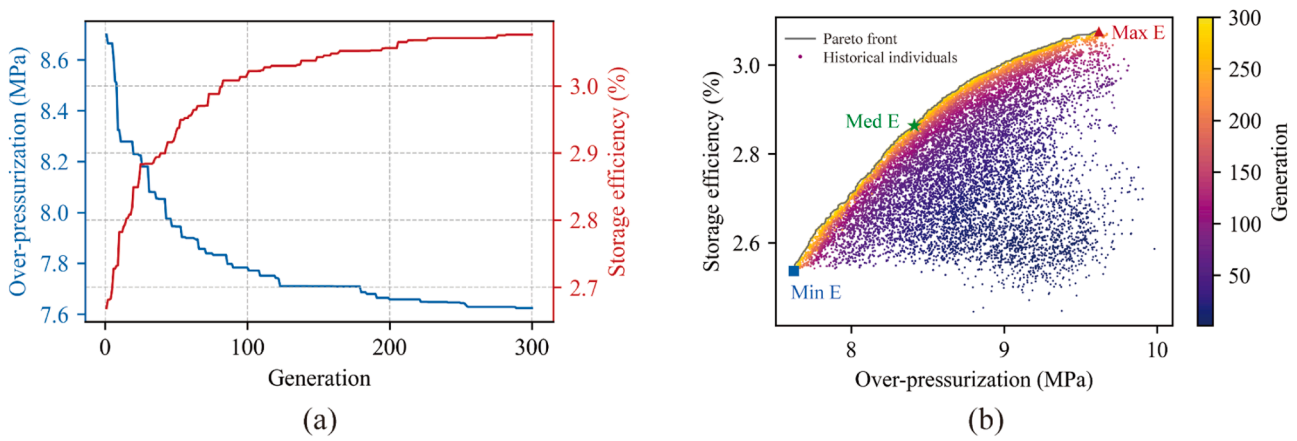


Fig. 12. (a) Best values of the objective functions vs. generation; (b) Evolution of historical individuals in the objective space and the final Pareto front. Three representative cases are marked as blue square (minimum *E*), green star (medium *E*), and red triangle (maximum *E*).

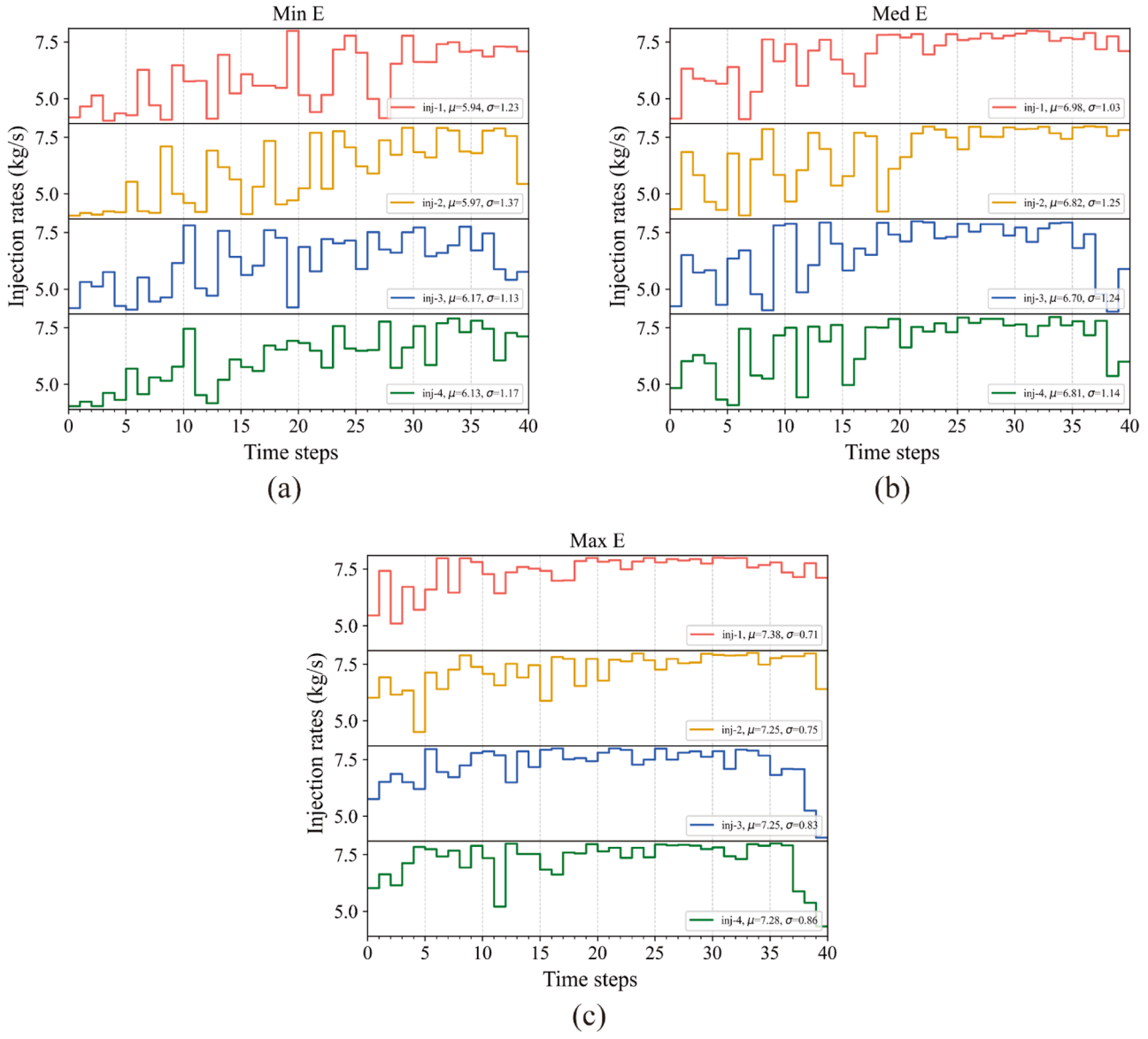


Fig. 13. Optimized CO_2 injection rates of the 4 wells for the three representative cases: (a) Minimum *E*; (b) Medium *E*; (c) Maximum *E*.

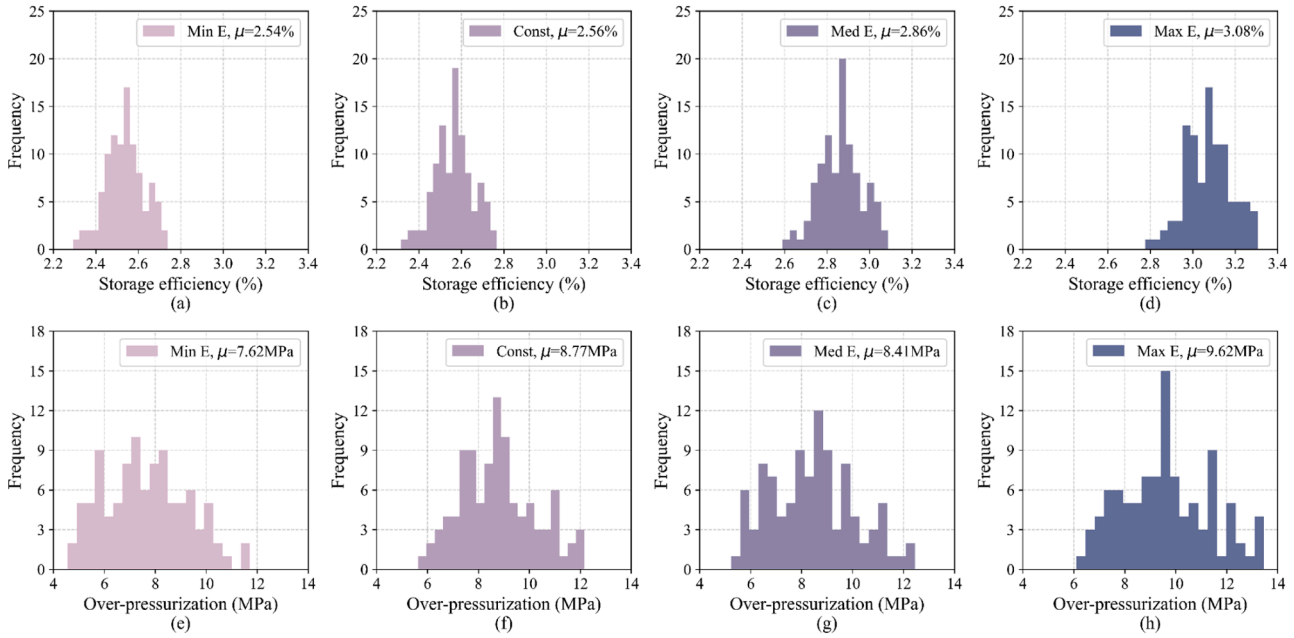


Fig. 14. Performance of the minimum *E* case, constant 6kg/s case, medium *E* case, and maximum *E* case (from left to right) over the 100 realizations. The upper row is the results for storage efficiency, and the lower row is the results for over-pressurization.

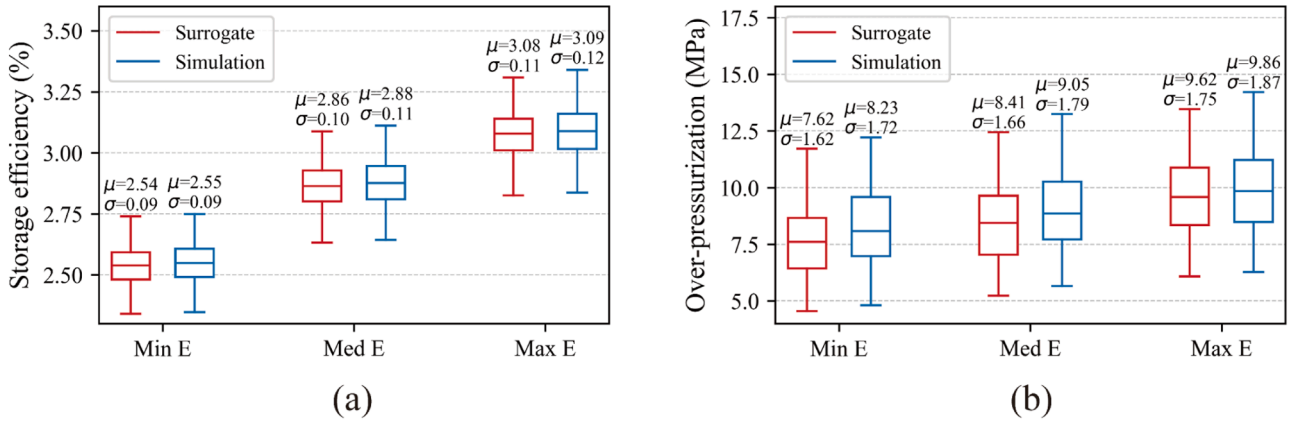


Fig. 15. Comparison of (a) storage efficiency and (b) over-pressurization for the minimum *E*, medium *E* and maximum *E* cases between surrogate model and numerical simulation.

5. Discussion

5.1. Comparison with RNNs

Traditional RNNs rely on hidden states or memory cells to store historical information. Typical examples are the Gated Recurrent Unit (GRU) (Cho et al., 2014) neural network which uses gating operations to modulate the information flow, and Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) neural network which introduces more complicate gating operations and another memory cell variable. However, RNNs usually struggle for capturing long-range dependencies of sequences due to the vanishment or explosion of the gradients (Bengio et al., 1994). Moreover, the forward and backward propagation are carried out step by step through time, making it unable to fully leverage the parallel processing capability of GPUs. We adopt Transformers (Vaswani et al., 2017) as the backbone of the surrogate model because they are one of the state-of-the-art sequence architectures that overcome the above drawbacks. To validate the outstanding performance of the proposed CNN-Transformer surrogate model, we replace the Transformer layers in the original neural network with an equivalent number of GRU or LSTM layers, resulting the hybrid CNN-GRU and CNN-LSTM models. For simplicity, we take these three models as Transformer, GRU, and LSTM in the following content.

The error comparison of the three models is illustrated in Fig. 16. The data corresponding to Transformer, GRU, and LSTM are denoted by red, yellow, and blue colors, respectively. Fig. 16a displays the RMSEs of the models every 5 time steps with boxes calculated over the whole test set. The errors of the models tend to increase as time progresses since the CO_2 plume becomes more complex. It can be observed that the error medians of Transformer are lower than those of RNNs throughout time. The overall error of saturation averaged over time is presented in Fig. 16b. The mean and standard deviation of the overall error for Transformer is 0.128 % and 0.027 %, while those for GRU and LSTM are 0.145 ± 0.026 % and 0.182 ± 0.035 % respectively. Transformer has the lowest mean value among the three models, though its standard deviation is slightly higher than that of GRU. The lower two subfigures in Fig. 16 show the results for pressure. The RRMSEs of pressure remain relatively stable throughout time. Again, the errors of Transformer are consistently lower than those of RNNs across the timeline, as shown in Fig. 16c. The mean value of pressure error averaged over time for

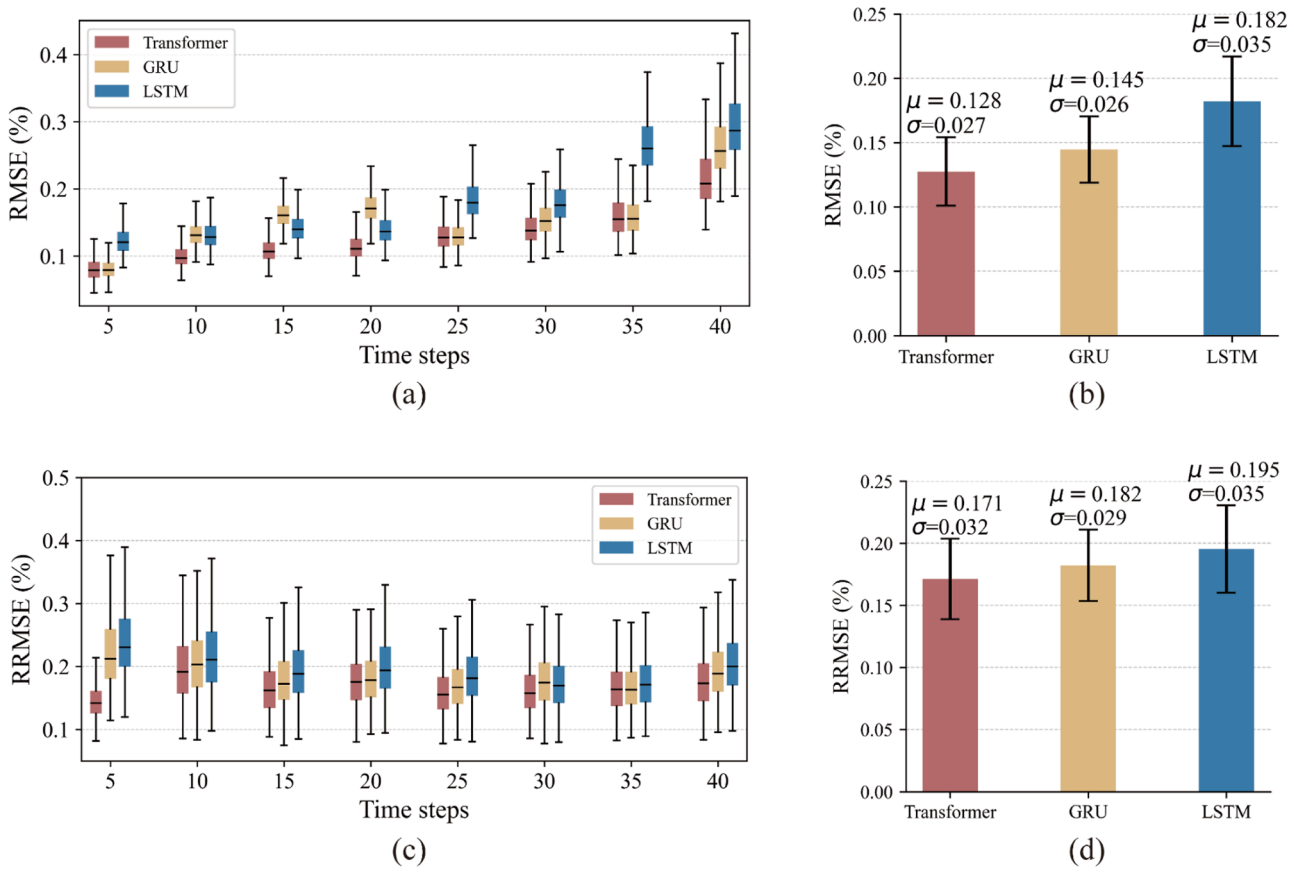


Fig. 16. Error comparison of Transformer (red), GRU (yellow), and LSTM (blue): (a) temporal error of saturation every 5 time steps, (b) overall error of saturation, (c) temporal error of pressure every 5 time steps, (d) overall error of pressure.

Transformer is 0.171 %, which is lower than GRU's 0.182 % and LSTM's 0.195 %. Transformer's standard deviation is 0.003 % higher than GRU's, but is still lower than that of LSTM. Therefore, Transformer outperforms GRU and LSTM with regard to prediction accuracy for both saturation and pressure.

Fig. 17a and b presents the scaling performance of the three models. As the size of the training dataset increases, the saturation error of the models generally decreases except for LSTM. This anomaly might be due to the limited model capacity of LSTM or an unsuitable training strategy, as the training hyperparameters are kept the same for all three models. Fine-tuning the LSTM model could potentially resolve this issue but is beyond the scope of this paper. For pressure, the error monotonically decreases with the increase of the training set size, as expected. The error of Transformer is consistently the lowest except when only 25 % data are used for training. This phenomenon can be attributed to the sparse inductive biases within the Transformer architecture, which makes it more flexible to capture complex long-range dependencies at the cost of requiring more training data. The decent scaling performance of Transformer would be further enhanced with access to larger numerical and experimental datasets.

Computational efficiency is compared in Fig. 17c and d. The Transformer model requires 165.33 min to simultaneously train saturation and pressure, whereas GRU and LSTM take 258.23 and 249.89 min, respectively. The training cost of Transformer is roughly 65 % of that of RNNs. The inference cost is calculated based on several forward runs in a batch mode on the GPU. The mean inference time of Transformer is 0.073 s, which is slightly less than that of LSTM (0.079 s) but notable less than that of GRU (0.101 s). Transformer can process the entire sequence at matrix level, which saves both training and inference costs. The gap in computational efficiency between Transformer and RNNs is expected to widen as sequence length increases.

These results highlight the superior performance of the proposed CNN-Transformer surrogate model compared to other RNN-based models in terms of prediction accuracy, data scalability, and computational efficiency.

5.2. *Generalization to out-of-distribution geological parameters*

Generalization measures the robustness of deep neural networks' predictions on unseen or even out-of-distribution scenarios, which are frequently encountered in real-world applications (Geirhos et al., 2020; Hendrycks et al., 2021). In the previous sections, we have demonstrated that the CNN-Transformer model can well generalize to testing realizations with parameters sampled from an identical distribution as the training dataset. To test the model performance on more challenging tasks, specifically out-of-distribution scenarios, we consider here more heterogeneous permeability fields, whose correlation lengths are 50 % smaller than the training samples. Four realization examples are depicted in Fig. 18.

Results for saturation and pressure are presented in Figs. 19 and 20. Outputs at the last time step for the four realizations are collected for comparison. In Fig. 19, we can observe that the CO_2 plume becomes much more irregular and non-uniform due to the increased geological heterogeneity. The surrogate model can still basically capture the morphology of the saturation map, though some discernible errors appear near the plume edge. R^2 scores of the four realizations range from 0.9594 to 0.9734, and RMSEs vary between 1.2145 % and 1.4954 %. The saturation prediction accuracy of out-of-distribution realizations is not as good as that of in-distribution ones, as the more complex saturation maps are never seen by the model at the training stage. In contrast, the results for pressure are more satisfactory, as shown in Fig. 20. The over-pressurization zone is well characterized by the

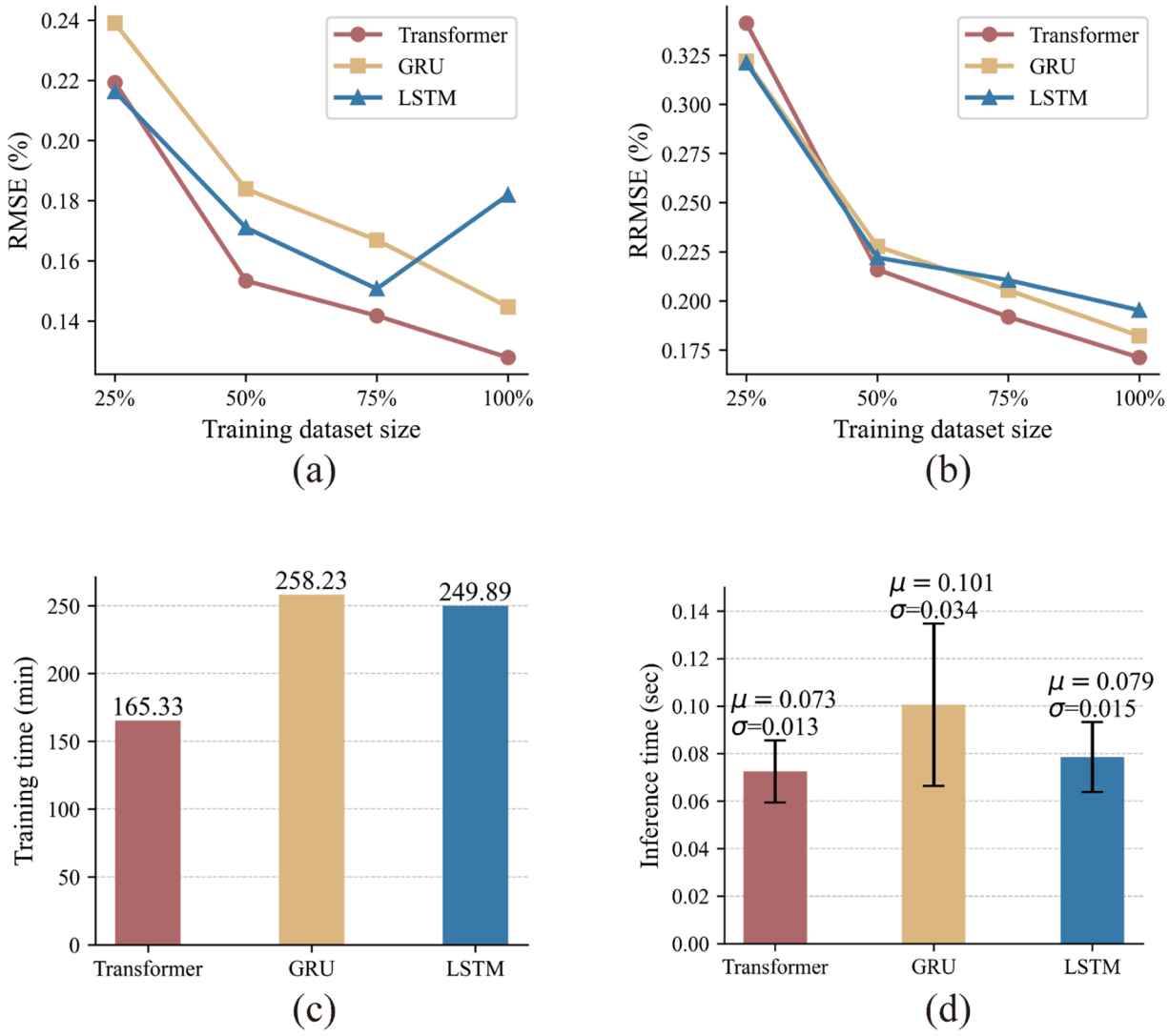


Fig. 17. (a) Error of saturation vs. training dataset size. (b) Error of pressure vs. training dataset size. (c) Training time cost of the models. (d) Inference time cost of the models.

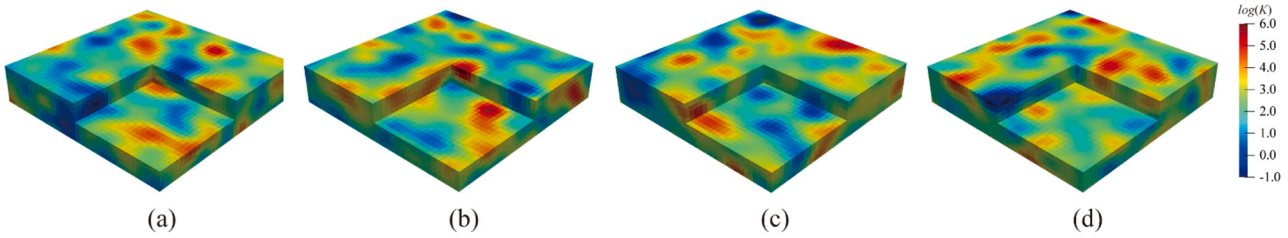


Fig. 18. Four out-of-distribution realizations of heterogeneous permeability fields with correlation lengths of 3, 4, and 4 grids in the *z*, *x*, and *y* directions, respectively.

surrogate model, with minor discrepancies around the injection wells. R^2 scores are all above 0.9788, and RMSEs are lower than 0.5267 %. Pressure results for out-of-distribution realizations are comparable to those for in-distribution samples in that pressure maps are continuous and exhibits similar patterns regardless of the geological heterogeneity.

In general, the intensified geological variability does not lead to significant degradation of the surrogate model's performance. The proposed CNN-Transformer model, which is trained on massive high-fidelity data, is capable of generalizing to out-of-distribution scenarios while maintaining an acceptable level of accuracy. It should be noteworthy that the CNN-Transformer model in this paper can be regarded as a pre-trained model, and is ready to transfer to other challenging and diverse tasks through transfer learning and fine-tuning techniques, as have been widely adopted in modern large language models (Brown et al., 2020; Devlin et al., 2019; Hendrycks et al., 2020).

6. Conclusions

In this paper, we develop a hybrid CNN-Transformer surrogate model to simulate the time-varying CO_2 injection into 3D heterogeneous

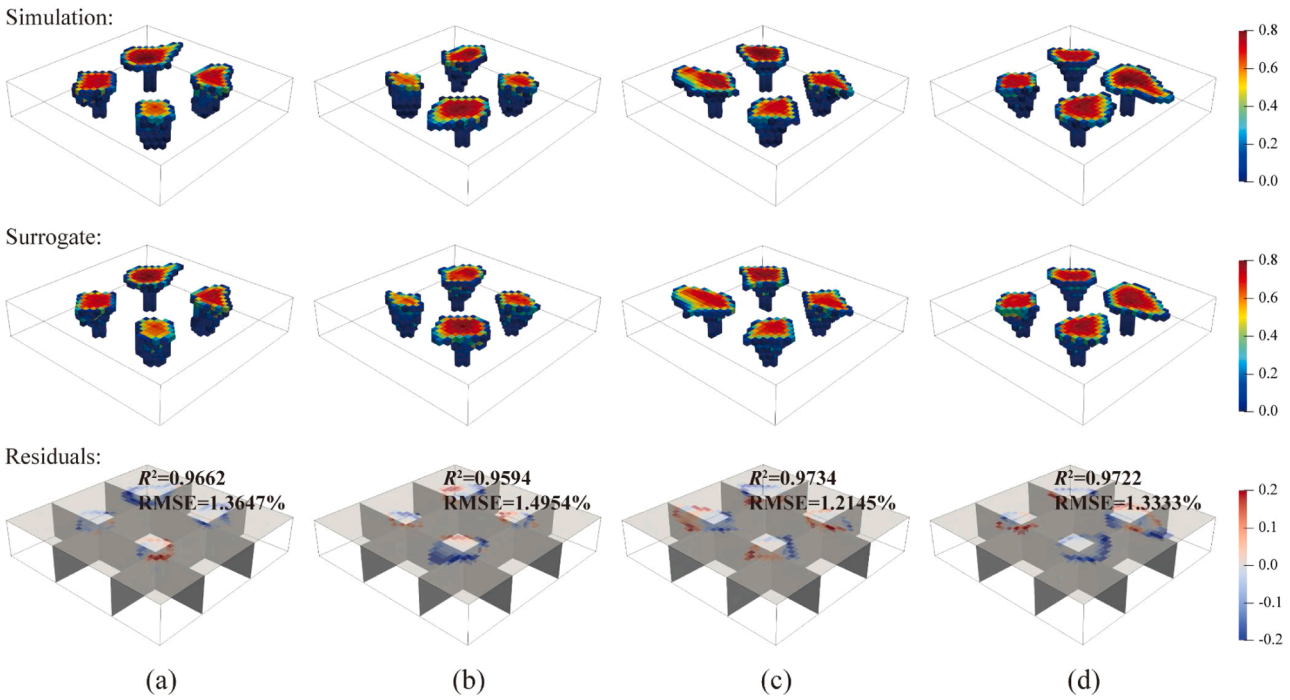


Fig. 19. Results of saturation at the last time step for the four out-of-distribution realizations (from left to right columns). The first row presents simulation data, the second row shows surrogate model predictions, and the third row displays the residuals between the simulation and surrogate model. The grey planes represent the clips of the top layer and the four wells.

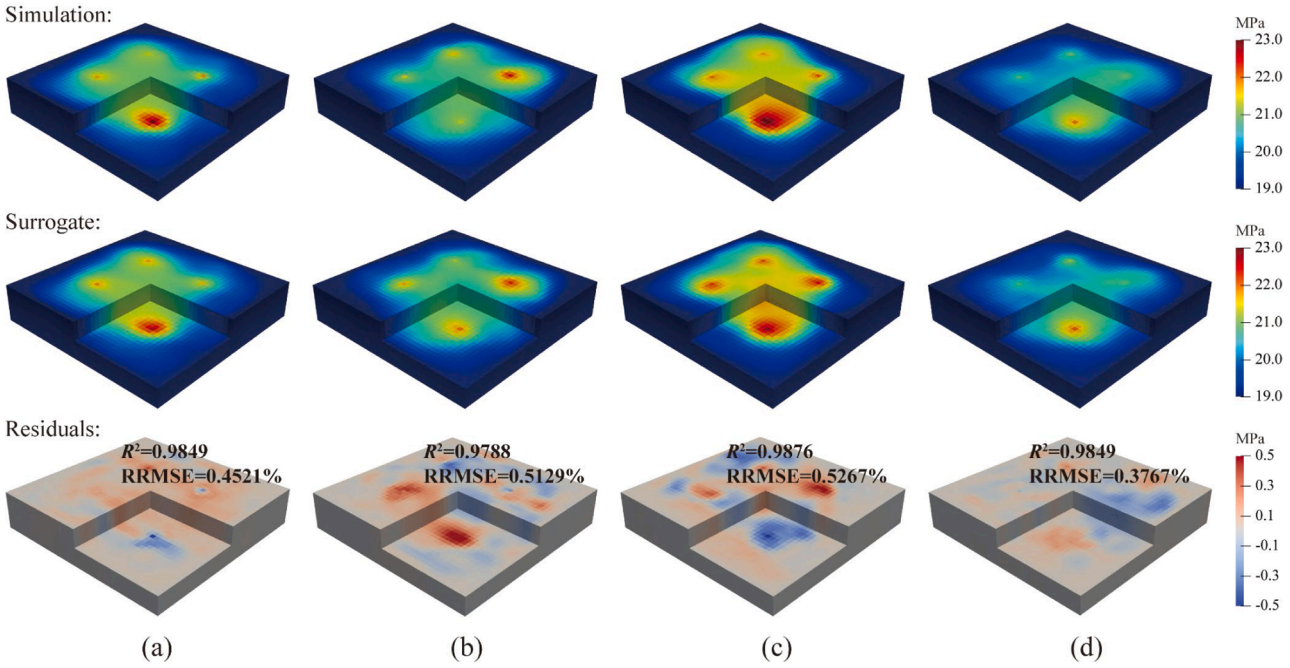


Fig. 20. Results of pressure at the last time step for the four out-of-distribution realizations (from left to right columns). The first row presents simulation data, the second row shows surrogate model predictions, and the third row displays the residuals between the simulation and surrogate model.

permeability fields. The surrogate model consists of three components: a CNN encoder to convert high-dimensional input geological parameters into latent space, a Transformer processor to capture global patterns of the dynamic well controls, and a CNN decoder to generate predictions at all time steps simultaneously. Skip connections are added between the encoder and the decoder to boost information flow. The proposed surrogate model is then integrated into the Multi-Objective Robust Optimization (MORO) framework to determine the Pareto optimal CO_2 injection schemes considering the uncertainties in permeability fields through the Non-dominated Sorting-based Genetic Algorithm II (NSGA-II). The multiple objectives are defined as maximize the storage efficiency and minimize the induced over-pressurization. The total storage capacity is considered as the nonlinear constraint. The main conclusions are as follows:

1) The CNN-Transformer surrogate model exhibits excellent prediction accuracy for both saturation and pressure. The statistical correspondence across the whole test set is also satisfactory, with P10, P50, and P90 percentiles for the surrogate and simulation model almost identical through time.
2) Compared to GRU- and LSTM-based models, the proposed CNN-Transformer model has the lowest error levels, with RMSE 0.128 $\pm$ 0.027 % for saturation, and RRMSE 0.171 $\pm$ 0.032 % for pressure. The performance of the CNN-Transformer model improves as the training dataset enlarges, showcasing better scaling capability.
3) The training cost of the CNN-Transformer model is roughly 65 % of that of GRU- and LSTM-based models. The inference speed is also faster. The CNN-Transformer model is more computational efficient.
4) The prediction accuracy of the CNN-Transformer model remains decent when generalizing to out-of-distribution permeability fields. For saturation, RMSE is lower than 1.495 %, and R^2 is higher than 0.959. For pressure, RRMSE is below 0.527 %, and R^2 is above 0.979.
5) The Pareto optimal injection schemes are determined. Three representative cases with different levels of storage efficiency show distinct injection pattern. As the storage efficiency improves, the extent of fluctuations decreases, and the mean injection rates increase.
6) The optimized injection schemes show better storage efficiency without inducing much over-pressurization compared to the constant rate scheme. The decision maker can decide the optimal scheme based on the trade-offs of the economic and safety objectives. The reliability of the optimization results is demonstrated by comparing with numerical simulations. Integrating CNN-Transformer with MORO reduces computational time by 99.99 %, achieving an acceleration of approximately 739,000 times.

CRediT authorship contribution statement

Zhao Feng: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Data curation. **Bicheng Yan:** Writing – review & editing, Validation, Investigation, Formal analysis, Conceptualization. **Xianda Shen:** Writing – review & editing, Validation, Formal analysis, Conceptualization. **Fengshou Zhang:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization. **Zeeshan Tariq:** Writing – review & editing, Validation, Formal analysis, Conceptualization. **Weiquan Ouyang:** Writing – review & editing, Validation, Investigation. **Zhilei Han:** Writing – review & editing, Validation.

Declaration of competing interest

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.

Acknowledgments

This work is supported by the National Key Research and Development Project, China (No. 2023YFE0110900) and National Natural Science Foundation of China (Nos. 42320104003, 42077247).

Appendix A. Neural network architecture details

Table A.1.

Table A.1
CNN-Transformer architecture. N denotes the batch size. Conv3D denotes the 3D convolutional layer with a kernel size of 3 × 3 × 3, a stride of 1, and a padding of 1, BN denotes the Batch Normalization layer, GELU denotes the Gaussian Error Linear Units activation function, LN denotes the Layer Normalization layer, Fusion denotes the addition of latent geological and engineering features, PE denotes positional encoding.

Module	Layer	Output shape
CNN Encoder	Input geological field	(N, 1, 8, 40, 40)
	ConvBlock (Conv3D / BN / GELU)	(N, 16, 8, 40, 40)
	Downsampling (Conv 3 × 3 × 3, stride 2, padding 1)	(N, 16, 4, 20, 20)
	ConvBlock (Conv3D / BN / GELU)	(N, 64, 4, 20, 20)
	Downsampling (Conv 3 × 3 × 3, stride 2, padding 1)	(N, 64, 2, 10, 10)
	ConvBlock (Conv3D / BN / GELU)	(N, 256, 2, 10, 10)
	Downsampling (Conv 3 × 3 × 3, stride 2, padding 1)	(N, 256, 1, 5, 5)
	ConvBlock (Conv3D / BN / GELU)	(N, 64, 1, 5, 5)
	Flatten	(N, 1600)
	MLP (Linear / GELU / Linear / LN)	(N, 200)
	Broadcast	(N, 40, 200)
Transformer processor	Input engineering parameters	(N, 40, 4)
	MLP (Linear / GELU / Linear / LN)	(N, 40, 200)
	Fusion / PE	(N, 40, 200)
	Attention / LN / MLP (Linear / GELU / Linear) / LN	(N, 40, 200)
	Attention / LN / MLP (Linear / GELU / Linear) / LN	(N, 40, 200)
	Linear	(N, 40, 200)
CNN Decoder	Reshape	(N, 40, 2, 10, 10)
	Upsampling (Conv3D / LN / GELU / TransConv3D)	(N, 40, 4, 20, 20)
	ResNet (Conv3D / LN / GELU / Conv3D / LN / GELU)	(N, 40, 4, 20, 20)
	Upsampling (Conv3D / LN / GELU / TransConv3D)	(N, 40, 8, 40, 40)
	ResNet (Conv3D / LN / GELU / Conv3D / LN / GELU)	(N, 40, 8, 40, 40)
	Output	(N, 40, 8, 40, 40)

Appendix B. Ablation study

Table B.1.

Table B.1
Five groups of ablation study for different model configurations. The trainable model parameters, RMSE for saturation, and RRMSE for pressure on the test set are presented. The base configuration denotes the model setup adopted in this paper. The first group compares the influence of the number of transformer layers. The second group studies the influence of the number of attention heads. The third group focuses on different types fusion methods for the geological and engineering features. In the concatenation configuration, the dimensions of the geological and engineering features are set to 100, so that the input dimension of transformer is kept unchanged. The fourth group presents the results for cross- and self-attention mechanism. For cross-attention, the key and value are computed by the geological feature, while the query is computed from the engineering feature. The fifth group shows the effects of the skip connections.

Group	Comparison type	Configuration	Parameters	RMSE of saturation (%)	RRMSE of pressure (%)
1	No. of layers	1	3,922,397	0.1438	0.1819
		2 (base)	4,286,272	**0.1275**	**0.1712**
		3	4,650,147	0.1418	0.1787
2	No. of attention heads	4	4,297,522	0.1735	0.1781
		8 (base)	4,286,272	**0.1275**	**0.1712**
3	Feature fusion	Concatenation	4,088,147	0.1350	0.1713
		Multiplication	4,286,272	0.1317	0.1729
		Addition (base)	4,286,272	**0.1275**	**0.1712**
4	Attention mechanism	Cross-attention	4,286,272	0.1387	0.1971
		Self-attention (base)	4,286,272	**0.1275**	**0.1712**
5	Skip connection	Without	4,286,272	0.1365	0.1842
		With (base)	4,286,272	**0.1275**	**0.1712**

Appendix C. Attention map visualization

To better visualize the attention scores at different positions, we apply an element-wise operation on the attention maps through log transformation and min-max normalization. For each attention score $a_{i,j}$, we transform it by:

$$a_{i,j} = 1 - \exp(-10 \times a_{i,j}),$$
$$a_{i,j} = \frac{a_{i,j} - \boldsymbol{A}_{\min}}{\boldsymbol{A}_{\max} - \boldsymbol{A}_{\min}}, \tag{C.1}$$

where e is Euler's number, $\boldsymbol{A}_{\min}$ and $\boldsymbol{A}_{\max}$ are the minimum and maximum values of the attention map $\boldsymbol{A}$.

The transformed attention maps are shown in Fig. C.1. The first self-attention layers of both saturation and pressure networks generally allocate even attention scores to all the historical positions, while the second layers show obvious position-varied patterns. We hypothesize that the first layer primarily captures shallow patterns, whereas the second layer learns more complex and contextual ones (Geva et al., 2021). Additionally, the second layer attention maps differ across different heads, as each head attend to distinct representations within its respective subspace.

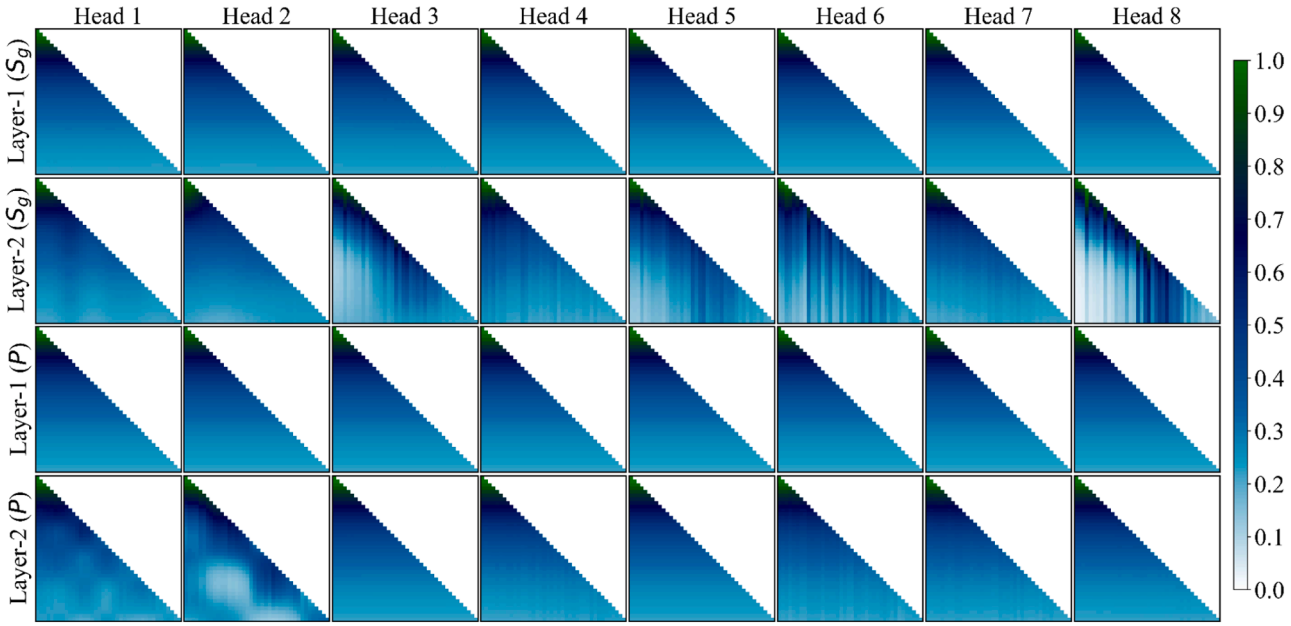


Fig. C.1. Attention maps visualization. The top two rows represent the first and second self-attention layer for saturation. The bottom two rows represent the first and second self-attention layer for pressure. The columns denote the eight attention heads.

Data availability

The codes and data for this work will be made publicly available upon acception. The codes and data used in this study are available at. https://github.com/fengzhao1239/CNN-Transformer.

References

Ajayi, T., Gomes, J.S., Bera, A., 2019. A review of CO_2 storage in geological formations emphasizing modeling, monitoring and capacity estimation approaches. Pet. Sci. 16 (5), 1028–1063. https://doi.org/10.1007/s12182-019-0340-8.

Aliyev, E., Durlofsky, L.J., 2017. Multilevel field development optimization under uncertainty using a sequence of upscaled models. Math. Geosci. 49 (3), 307–339. https://doi.org/10.1007/s11004-016-9643-0.

Aminu, M.D., Nabavi, S.A., Rochelle, C.A., Manovic, V., 2017. A review of developments in carbon dioxide storage. Appl. Energy 208, 1389–1419. https://doi.org/10.1016/j.apenergy.2017.09.015.

Bachu, S., 2015. Review of CO_2 storage efficiency in deep saline aquifers. Int. J. Greenh. Gas Control 40, 188–202. https://doi.org/10.1016/j.ijggc.2015.01.007.

Bachu, S., Bonijoly, D., Bradshaw, J., Burruss, R., Holloway, S., Christensen, N.P., Mathiassen, O.M., 2007. CO_2 storage capacity estimation: methodology and gaps. Int. J. Greenh. Gas Control 1 (4), 430–443. https://doi.org/10.1016/S1750-5836(07)00086-2.

Bengio, Y., Simard, P., Frasconi, P., 1994. Learning long-term dependencies with gradient descent is difficult. IEEE Trans. Neural Netw. 5 (2), 157–166. https://doi.org/10.1109/72.279181.

Blank, J., Deb, K., 2020. Pymoo: multi-objective optimization in Python. IEEE Access. 8, 89497–89509. https://doi.org/10.1109/ACCESS.2020.2990567.

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., … Amodei, D., (2020). Language Models are Few-Shot Learners (arXiv:2005.14165). arXiv. doi:10.48550/arXiv.2005.14165.

Celia, M.A., Bachu, S., Nordbotten, J.M., Bandilla, K.W., 2015. Status of CO_2 storage in deep saline aquifers with emphasis on modeling approaches and practical simulations. Water Resour. Res. 51 (9), 6846–6892. https://doi.org/10.1002/2015WR017609.

Censor, Y., 1977. Pareto optimality in multiobjective problems. Appl. Math. Optim. 4 (1), 41–59. https://doi.org/10.1007/BF01442131.

Chen, B., & Pawar, R. (2020). Joint optimization of well completions and controls for CO2 enhanced oil recovery and storage. *Day 2 Tue, September 01, 2020*, D021S027R003. https://doi.org/10.2118/200316-MS.

Cho, K., van Merrienboer, B., Bahdanau, D., & Bengio, Y. (2014). *On the properties of neural machine translation: encoder-decoder approaches* (arXiv:1409.1259). arXiv. http://arxiv.org/abs/1409.1259.

Cihan, A., Birkholzer, J.T., Bianchi, M., 2015. Optimal well placement and brine extraction for pressure management during CO_2 sequestration. Int. J. Greenh. Gas Control 42, 175–187. https://doi.org/10.1016/j.ijggc.2015.07.025.

Deb, K., Pratap, A., Agarwal, S., Meyarivan, T., 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans. Evol. Comput. 6 (2), 182–197. https://doi.org/10.1109/4235.996017.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: pre-training of deep bidirectional transformers for language understanding* (arXiv:1810.04805). arXiv. https://doi.org/10.48550/arXiv.1810.04805.

Diab, W., & Al-Kobaisi, M. (2023). *U-DeepONet: U-Net Enhanced Deep Operator Network for Geologic Carbon Sequestration* (arXiv:2311.15288). arXiv. https://doi.org/10.48550/arXiv.2311.15288.

Fan, M., Wang, H., Zhang, J., Hosseini, S.A., Lu, D., 2024. Advancing spatiotemporal forecasts of CO_2 plume migration using deep learning networks with transfer learning and interpretation analysis. Int. J. Greenh. Gas Control 132, 104061. https://doi.org/10.1016/j.ijggc.2024.104061.

Feng, Z., Tariq, Z., Shen, X., Yan, B., Tang, X., Zhang, F., 2024. An encoder-decoder ConvLSTM surrogate model for simulating geological CO_2 sequestration with dynamic well controls. Gas Sci. Eng. 125, 205314. https://doi.org/10.1016/j.jgsce.2024.205314.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A., 2020. Shortcut learning in deep neural networks. Nat. Mach. Intell. 2 (11), 665–673. https://doi.org/10.1038/s42256-020-00257-z.

Geneva, N., Zabaras, N., 2022. Transformers for modeling physical systems. Neural Netw. 146, 272–289. https://doi.org/10.1016/j.neunet.2021.11.022.

Geva, M., Schuster, R., Berant, J., & Levy, O., (2021). Transformer Feed-Forward Layers Are Key-Value Memories (arXiv:2012.14913). arXiv. http://arxiv.org/abs/2012.14913.

Han, Y., Hamon, F.P., Jiang, S., & Durlofsky, L.J. (2024a). Accelerated training of deep learning surrogate models for surface displacement and flow, with application to MCMC-based history matching of CO2 storage operations. arXiv preprint arXiv:2408.10717.

Han, Y., Hamon, F.P., Jiang, S., Durlofsky, L.J., 2024b. Surrogate model for geological CO_2 storage and its use in hierarchical MCMC history matching. Adv. Water Resour. 187, 104678. https://doi.org/10.1016/j.advwatres.2024.104678.

Hang, Z., Ma, Y., Wu, H., Wang, H., & Long, M. (2024). *Unisolver: PDE-Conditional Transformers Are Universal PDE Solvers* (arXiv:2405.17527). arXiv. http://arxiv.org/abs/2405.17527.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep Residual Learning for Image Recognition*. 770–778. https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

Heath, J.E., McKenna, S.A., Dewers, T.A., Roach, J.D., Kobos, P.H., 2014. Multiwell CO_2 injectivity: impact of boundary conditions and brine extraction on geologic CO_2 storage efficiency and pressure buildup. Environ. Sci. Technol. 48 (2), 1067–1074. https://doi.org/10.1021/es4017014.

Hemmasian, A., Barati Farimani, A., 2023. Reduced-order modeling of fluid flows with transformers. Phys. Fluids 35 (5), 057126. https://doi.org/10.1063/5.0151515.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., Song, D., Steinhardt, J., Gilmer, J., 2021. The many faces of robustness: a critical analysis of out-of-distribution generalization. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 8320–8329. https://doi.org/10.1109/ICCV48922.2021.00823.

Hendrycks, D., & Gimpel, K. (2023). *Gaussian Error Linear Units (GELUs)* (arXiv:1606.08415). arXiv. http://arxiv.org/abs/1606.08415.

Hendrycks, D., Liu, X., Wallace, E., Dziedzic, A., Krishnan, R., & Song, D. (2020). *Pretrained Transformers Improve Out-of-Distribution Robustness* (arXiv:2004.06100). arXiv. http://arxiv.org/abs/2004.06100.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

Houghton, J., 2005. Global warming. Rep. Prog. Phys. 68 (6), 1343–1403. https://doi.org/10.1088/0034-4885/68/6/R02.

Jiang, Z., Zhu, M., Lu, L., 2024. Fourier-MIONet: Fourier-enhanced multiple-input neural operators for multiphase modeling of geological carbon sequestration. Reliab. Eng. Syst. Saf. 251, 110392. https://doi.org/10.1016/j.ress.2024.110392.

Ju, X., Fu, P., Settegast, R.R., Morris, J.P., 2021. A coupled thermo-hydro-mechanical model for simulating leakoff-dominated hydraulic fracturing with application to geologic carbon storage. Int. J. Greenh. Gas Control 109, 103379. https://doi.org/10.1016/j.ijggc.2021.103379.

Jung, Y., Pau, G.S.H., Finsterle, S., Pollyea, R.M., 2017. TOUGH3: a new efficient version of the TOUGH suite of multiphase flow and transport simulators. Comput. Geosci. 108, 2–7. https://doi.org/10.1016/j.cageo.2016.09.009.

Laloy, E., Hérault, R., Lee, J., Jacques, D., Linde, N., 2017. Inversion using a new low-dimensional representation of complex binary geological media based on a deep neural network. Adv. Water Resour. 110, 387–405. https://doi.org/10.1016/j.advwatres.2017.09.029.

Li, C., Laloui, L., 2016. Coupled multiphase thermo-hydro-mechanical analysis of supercritical CO_2 injection: benchmark for the In Salah surface uplift problem. Int. J. Greenh. Gas Control 51, 394–408. https://doi.org/10.1016/j.ijggc.2016.05.025.

Li, C., Maggi, F., Zhang, K., Guo, C., Gan, Y., El-Zein, A., Pan, Z., Shen, L., 2019. Effects of variable injection rate on reservoir responses and implications for CO_2 storage in saline aquifers. Greenh. Gases 9 (4), 652–671. https://doi.org/10.1002/ghg.1888.

Loshchilov, I., & Hutter, F. (2019). *Decoupled Weight Decay Regularization* (arXiv:1711.05101). arXiv. http://arxiv.org/abs/1711.05101.

Mo, S., Zabaras, N., Shi, X., Wu, J., 2019a. Deep autoregressive neural networks for high-dimensional inverse problems in groundwater contaminant source identification. Water Resour. Res. 55 (5), 3856–3881. https://doi.org/10.1029/2018WR024638.

Mo, S., Zhu, Y., Zabaras, N., Shi, X., Wu, J., 2019b. Deep convolutional encoder-decoder networks for uncertainty quantification of dynamic multiphase flow in heterogeneous media. Water Resour. Res. 55 (1), 703–728. https://doi.org/10.1029/2018WR023528.

Mohd Razak, S., Jahandideh, A., Djuraev, U., Jafarpour, B., 2022. Deep learning for latent space data assimilation in subsurface flow systems. SPE J. 27 (05), 2820–2840. https://doi.org/10.2118/203997-PA.

Müller, S., Schüler, L., Zech, A., Heße, F., 2022. GSTools v1.3: a toolbox for geostatistical modelling in Python. Geosci. Model. Dev. 15 (7), 3161–3182. https://doi.org/10.5194/gmd-15-3161-2022.

Ovadia, O., Kahana, A., Stinis, P., Turkel, E., Givoli, D., Karniadakis, G.E., 2024. ViTO: vision transformer-operator. Comput. Methods Appl. Mech. Eng. 428, 117109. https://doi.org/10.1016/j.cma.2024.117109.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Chintala, S., 2019. PyTorch: an imperative style, high-performance deep learning library. Adv. Neural Inf. Process. Syst. 32. In: https://proceedings.neurips.cc/paper_files/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html.

Pruess, K., Oldenburg, C.M., & Moridis, G.J. (1999). *TOUGH2 User's Guide Version 2*. https://escholarship.org/uc/item/4df6700h.

Ramirez, A., Foxall, W., 2014. Stochastic inversion of InSAR data to assess the probability of pressure penetration into the lower caprock at In Salah. Int. J. Greenh. Gas Control 27, 42–58. https://doi.org/10.1016/j.ijggc.2014.05.005.

Raza, A., Glatz, G., Gholami, R., Mahmoud, M., Alafnan, S., 2022. Carbon mineralization and geological storage of CO_2 in basalt: mechanisms and technical challenges. Earth Sci. Rev. 229, 104036. https://doi.org/10.1016/j.earscirev.2022.104036.

Ronneberger, O., Fischer, P., & Brox, T. (2015). *U-Net: Convolutional Networks for Biomedical Image Segmentation* (arXiv:1505.04597). arXiv. http://arxiv.org/abs/1505.04597.

Rutqvist, J., Vasco, D.W., Myer, L., 2010. Coupled reservoir-geomechanical analysis of CO_2 injection and ground deformations at In Salah, Algeria. Int. J. Greenh. Gas Control 4 (2), 225–230. https://doi.org/10.1016/j.ijggc.2009.10.017.

Segall, P., Lu, S., 2015. Injection-induced seismicity: poroelastic and earthquake nucleation effects: injection induced seismicity. J. Geophys. Res. 120 (7), 5082–5103. https://doi.org/10.1002/2015JB012060.

Shamshiri, H., Jafarpour, B., 2012. Controlled CO_2 injection into heterogeneous geologic formations for improved solubility and residual trapping. Water Resour. Res. 48 (2). https://doi.org/10.1029/2011WR010455, 2011WR010455.

Shukla, R., Ranjith, P., Haque, A., Choi, X., 2010. A review of studies on CO_2 sequestration and caprock integrity. Fuel 89 (10), 2651–2664. https://doi.org/10.1016/j.fuel.2010.05.012.

Stepien, M., Ferreira, C.A.S., Hosseinzadehsadati, S., Kadeethum, T., Nick, H.M., 2023. Continuous conditional generative adversarial networks for data-driven modelling of geologic CO_2 storage and plume evolution. Gas Sci. Eng. 115, 204982. https://doi.org/10.1016/j.jgsce.2023.204982.

Tang, H., Durlofsky, L.J., 2024. Graph network surrogate model for subsurface flow optimization. J. Comput. Phys. 512, 113132. https://doi.org/10.1016/j.jcp.2024.113132.

Tang, M., Ju, X., Durlofsky, L.J., 2022. Deep-learning-based coupled flow-geomechanics surrogate model for CO_2 sequestration. Int. J. Greenh. Gas Control 118, 103692. https://doi.org/10.1016/j.ijggc.2022.103692.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30. In: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Wang, H., Kou, Z., Ji, Z., Wang, S., Li, Y., Jiao, Z., Johnson, M., McLaughlin, J.F., 2023. Investigation of enhanced CO_2 storage in deep saline aquifers by WAG and brine extraction in the Minnelusa sandstone, Wyoming. Energy 265, 126379. https://doi.org/10.1016/j.energy.2022.126379.

Wen, G., Li, Z., Azizzadenesheli, K., Anandkumar, A., Benson, S.M., 2022. U-FNO—An enhanced Fourier neural operator-based deep-learning model for multiphase flow. Adv. Water Resour. 163, 104180. https://doi.org/10.1016/j.advwatres.2022.104180.

Wen, G., Tang, M., Benson, S.M., 2021. Towards a predictor for CO_2 plume migration using deep neural networks. Int. J. Greenh. Gas Control 105, 103223. https://doi.org/10.1016/j.ijggc.2020.103223.

White, J.A., Chiaramonte, L., Ezzedine, S., Foxall, W., Hao, Y., Ramirez, A., McNab, W., 2014. Geomechanical behavior of the reservoir and caprock system at the In Salah CO_2 storage project. Proc. Natl. Acad. Sci. 111 (24), 8747–8752. https://doi.org/10.1073/pnas.1316465111.

Xie, Z., Cao, C., Zhang, L., Zhao, Y., Zhang, R., Li, J., Zhang, D., 2024. A new pressure management framework for CO_2 sequestration in deep saline aquifers based on genetic algorithm. Geoenergy Sci. Eng. 234, 212668. https://doi.org/10.1016/j.geoen.2024.212668.

Yan, B., Chen, B., Robert Harp, D., Jia, W., Pawar, R.J., 2022a. A robust deep learning workflow to predict multiphase flow behavior during geological CO_2 sequestration injection and post-injection periods. J. Hydrol. 607, 127542. https://doi.org/10.1016/j.jhydrol.2022.127542.

Yan, B., Gudala, M., Sun, S., 2023. Robust optimization of geothermal recovery based on a generalized thermal decline model and deep learning. Energy Convers. Manag. 286, 117033. https://doi.org/10.1016/j.enconman.2023.117033.

Yan, B., Harp, D.R., Chen, B., Pawar, R., 2022b. A physics-constrained deep learning model for simulating multiphase flow in 3D heterogeneous porous media. Fuel 313, 122693. https://doi.org/10.1016/j.fuel.2021.122693.

Zhang, H., Al Kobaisi, M., Arif, M., 2023. Impact of wettability and injection rate on CO_2 plume migration and trapping capacity: a numerical investigation. Fuel 331, 125721. https://doi.org/10.1016/j.fuel.2022.125721.

Zhong, Z., Sun, A.Y., Jeong, H., 2019. Predicting CO_2 plume migration in heterogeneous formations using conditional deep convolutional generative adversarial network. Water Resour. Res. 55 (7), 5830–5851. https://doi.org/10.1029/2018WR024592.

Zoback, M.D., Gorelick, S.M., 2012. Earthquake triggering and large-scale geologic storage of carbon dioxide. Proc. Natl. Acad. Sci. 109 (26), 10164–10168. https://doi.org/10.1073/pnas.1202473109.

Zou, A., Durlofsky, L.J., 2023. Integrated framework for constrained optimization of horizontal/deviated well placement and control for geological CO_2 storage. SPE J. 28 (05), 2462–2481. https://doi.org/10.2118/212228-PA.