PROCEEDINGS, 51st Workshop on Geothermal Reservoir Engineering
Stanford University, Stanford, California, February 9-11, 2026
SGP-TR-230

Surrogate Modeling for Geothermal Systems: Accelerating Optimization, History Matching, and Uncertainty Quantification

Zhouji LIANG, Junjie YU, Robin THIBAUT, Fangning ZHENG, Carl HOILAND, Ahinoam POLLACK

george@zanskar.us

Keywords: Machine Learning; Reservoir Simulation; Artificial Intelligence

ABSTRACT

High-fidelity geothermal reservoir simulators are essential for forecasting subsurface pressure and temperature distributions but are computationally prohibitive for rapid optimization, history matching, and uncertainty quantification. Surrogate modeling offers a practical alternative by learning simulator input–output mappings and delivering fast predictions that can be embedded in engineering decision loops. In this study, we present a systematic comparison of surrogate models for steady-state geothermal applications, evaluating convolutional encoder–decoder architectures (U-Net), neural operator approaches (FNO, U-FNO, and Fourier-MIONet). Models are assessed using decision-relevant criteria, including predictive accuracy, physics consistency, generalization to unseen geologic heterogeneity, data efficiency, and computational cost. Our results demonstrate that neural operator–based surrogates achieve strong predictive performance from limited training data while reducing inference time by orders of magnitude relative to full-physics simulation. We highlight key trade-offs between surrogate architectures and provide practical guidelines for selecting models that balance fidelity, robustness, and speed in geothermal optimization, history matching, and uncertainty quantification workflows.

1. INTRODUCTION

Geothermal energy provides a critical source of firm, baseload renewable power, supporting the decarbonization of electricity grids. Geothermal resource assessment and development rely on ensembles of coupled subsurface flow and heat-transport simulations to forecast spatiotemporal temperature evolution under uncertain geologic structure and properties (Pollack and Mukerji, 2019; Daniilidis, 2021). However, high-fidelity numerical models are often too computationally expensive to support workflows requiring thousands of forward runs. Surrogate modeling addresses this bottleneck by learning fast approximations to the simulator's input–output mapping.

Within scientific machine learning, several complementary model families have emerged for learning subsurface field-to-field mappings (Mao, 2025). Convolutional encoder–decoder architectures such as U-Net have become a workhorse for surrogate modeling due to their strong inductive bias for local spatial structure and multi-resolution feature extraction (Ronneberger et al., 2015). A second major direction is neural operator learning, which targets mappings between function spaces and is therefore well aligned with parameterized PDE solution operators and less tied to a specific discretization (Kovachki et al., 2023). The Fourier Neural Operator (FNO) is a prominent example, leveraging Fourier-domain representations to efficiently model global interactions in PDE surrogates (Li et al., 2020; Zheng et al., 2025). Several structured variants further improve accuracy and data efficiency; for instance, U-FNO combines U-Net-style multiscale pathways with Fourier layers to better capture both local and global behaviors in multiphysics settings (Wen et al., 2022). In addition, multiple-input operator learning approaches such as MIONet extend operator regression to problems with multiple heterogeneous inputs (Jin et al., 2022), and Fourier-MIONet incorporates Fourier/operator components to improve efficiency and has shown strong performance in large-scale porous-media flow settings (Jiang et al., 2024). In parallel, physics-informed approaches such as PINNs enforce governing equations through the training objective (Raissi et al., 2019; Karniadakis et al., 2021), and reduced-order methods (e.g., reduced-basis / non-intrusive RB) (Degen et al., 2022) remain an important physics-based route for many-query tasks like optimization and UQ. While physics-informed training is a promising direction, in this study we focus on data-driven field surrogates and operator learners as a practical baseline for fast emulation in geothermal decision workflows.

Geothermal surrogate modeling has grown rapidly, but much of the geothermal-specific ML literature has emphasized reduced-order or well/plant-level targets—e.g., forecasting production or power-related time series and using these fast predictors inside optimization loops—rather than full-field 3D state emulation. This emphasis is evident in geothermal ML reviews and trend analyses, as well as in representative studies using RNN/CNN–RNN models to predict production performance and support optimization (Okoroafor, 2022). At the same time, recent geothermal work has highlighted the promise of both data-driven and physics-based ML surrogates for coupled geothermal processes, while noting ongoing challenges around reliability and deployment (Degen et al., 2022).

These trends motivate a need for application-facing benchmarks that directly evaluate field surrogates under the conditions that matter for geothermal decision workflows: heterogeneous geology, broad parametric variability, and multi-time predictions. Comparative studies in subsurface ML argue that surrogate selection is often ad hoc because head-to-head evaluations across model families (with consistent metrics and cost accounting) remain limited. Here, we address this gap by systematically comparing representative field-to-field surrogates—U-Net and neural-operator approaches (FNO, U-FNO, Fourier-MIONet)—for predicting 3D temperature volumes from ensembles of coupled flow–heat simulations generated by a finite-element solver. We evaluate models using decision-relevant criteria,

including accuracy, generalization to unseen configurations, data efficiency, and computational cost, to provide practical guidance for deploying surrogates as fast forward models and as building blocks for future optimization and UQ workflows.

2. MODELS

We compare several representative deep-learning surrogate architectures for 3D coupled flow–heat simulations: a convolutional encoder–decoder (U-Net) and a set of operator-learning models (FNO, U-FNO, and Fourier-MIONet). U-Net serves as a widely used CNN baseline for field-to-field prediction, whereas FNO provides a canonical neural-operator baseline for learning parameterized PDE solution operators. We additionally include U-FNO and Fourier-MIONet as two structured extensions that target limitations commonly observed in subsurface surrogates: U-FNO augments Fourier layers with U-Net-style multiscale feature pathways to better capture mixed local–global behavior, while Fourier-MIONet extends operator learning to settings with multiple heterogeneous inputs and leverages Fourier operator components for efficiency. Although many architectural refinements exist within each family, we restrict attention to these representative models to (i) enable a controlled comparison across distinct inductive biases (convolutional vs operator learning, and multiscale vs multi-input operator variants), (ii) avoid subjective cherry-picking from a rapidly expanding design space, and (iii) reflect architectures that are already widely adopted in subsurface surrogate modeling and are therefore most relevant for geothermal decision workflows.

2.1 U-Net

U-Net is a convolutional encoder–decoder designed for dense, grid-based prediction. The encoder progressively downsamples the input representation to extract multi-scale features, while the decoder upsamples to produce an output field at the original resolution (Ronneberger et al., 2015) (**Figure 1.1**). Skip connections pass high-frequency information from encoder to decoder, which helps preserve sharp spatial structure and improves reconstruction quality for field-to-field regression.

2.2 Fourier Neural Operator (FNO)

The Fourier Neural Operator (FNO) is an operator-learning architecture that targets mappings between functions rather than fixed-dimensional vectors. FNO layers apply global convolutions parameterized in Fourier space: inputs are transformed to the spectral domain, multiplied by learned complex-valued weights for a limited set of modes, and transformed back to physical space, with pointwise nonlinearities between layers (**Figure 1.2**). This spectral parameterization allows FNO to efficiently model long-range spatial correlations and has been shown to be effective for learning PDE solution operators.

2.3 U-FNO

U-FNO augments the basic FNO with an explicit multi-resolution pathway inspired by encoder–decoder designs (**Figure 1.3**). In addition to Fourier layers that capture global interactions, U-FNO includes downsampling/upsampling stages and skip connections to better represent localized features that can be difficult for purely spectral operators to reconstruct. Intuitively, the architecture combines the global context of Fourier operators with the spatial detail retention of multi-scale feature fusion.

2.3 Fourier-MIONet

MIONet (Multiple-Input Operator Network) extends operator learning to settings where the system response depends on multiple heterogeneous inputs (e.g., different fields or external forcings) (Jin, 2022). It typically combines separate “branch” representations for each input and a shared “trunk” representation of the output coordinates, then merges them via structured composition (e.g., tensor-product style interactions) to produce the target field (**Figure 1.4**). Fourier-MIONet incorporates Fourier/spectral components to improve expressivity and efficiency, making it well-suited for high-dimensional operator regression when inputs include multiple parameter fields (Jiang, 2024). In our experiment's configuration, Fourier-MIONet offers a different way to handle multi-type input parameters. In this study, the time step information is used as a scalar input to the trunk in Fourier-MIONet.

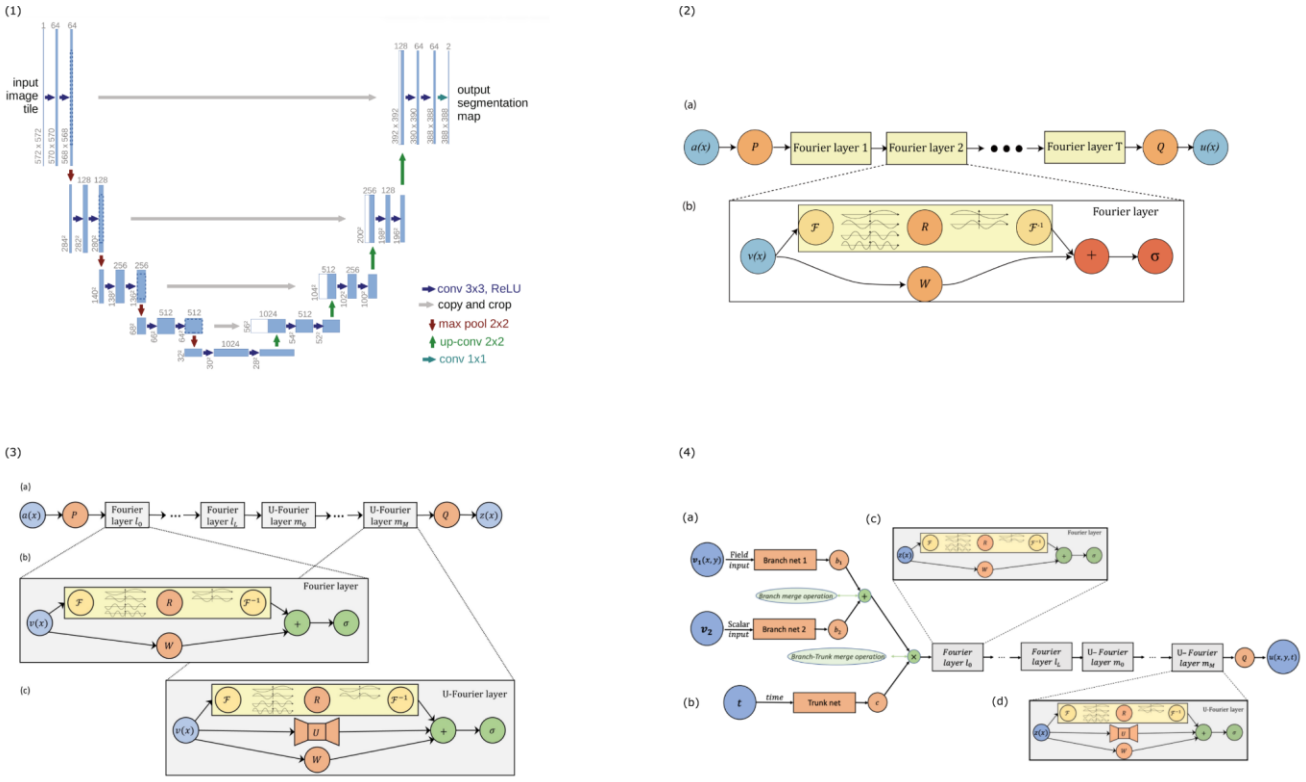

Figure 1: Surrogate model architectures compared in this study. (1) U-Net CNN encoder–decoder with skip connections (adapted from Ronneberger et al., 2015). (2) Fourier Neural Operator (FNO): lift–Fourier-layer stack–project; each Fourier layer applies FFT → learned spectral weights on truncated modes → inverse FFT, adds a pointwise linear term, and applies a nonlinearity (adapted from Li et al., 2020). (3) U-FNO: FNO augmented with U-Net-style multiscale pathways (down/up-sampling and skip connections) (adapted from Wen et al., 2022). (4) Fourier-MIONet: multi-input operator network with branch/trunk embeddings and Fourier/operator blocks for heterogeneous inputs (adapted from Jiang et al., 2024).

2. EXPERIMENT SETUP

In this study, we focus on fault- and fracture-dominated geothermal systems. The training data are generated using stochastic, high-fidelity simulations based on an anonymized real-world site. We define four distinct fault scenarios reflecting alternative geological evidence and conceptual models (shown in **Figure 2**); each scenario corresponds to a different number and configuration of fault surfaces, and the surface truncation order varies across scenarios. Fault properties and heat-flux boundary conditions are also treated stochastically. For each realization, parameters are sampled from the prior distribution and used to construct a geological model with a customized modeling workflow. The resulting model is then passed to commercial software to run coupled heat-and-fluid-flow simulations at high fidelity. Because uncertainty is introduced in fault geometry, fault and host-rock properties, and heat-flux boundary conditions, the training dataset spans a broad range of variability beyond rock properties alone, making the learning task substantially more challenging.

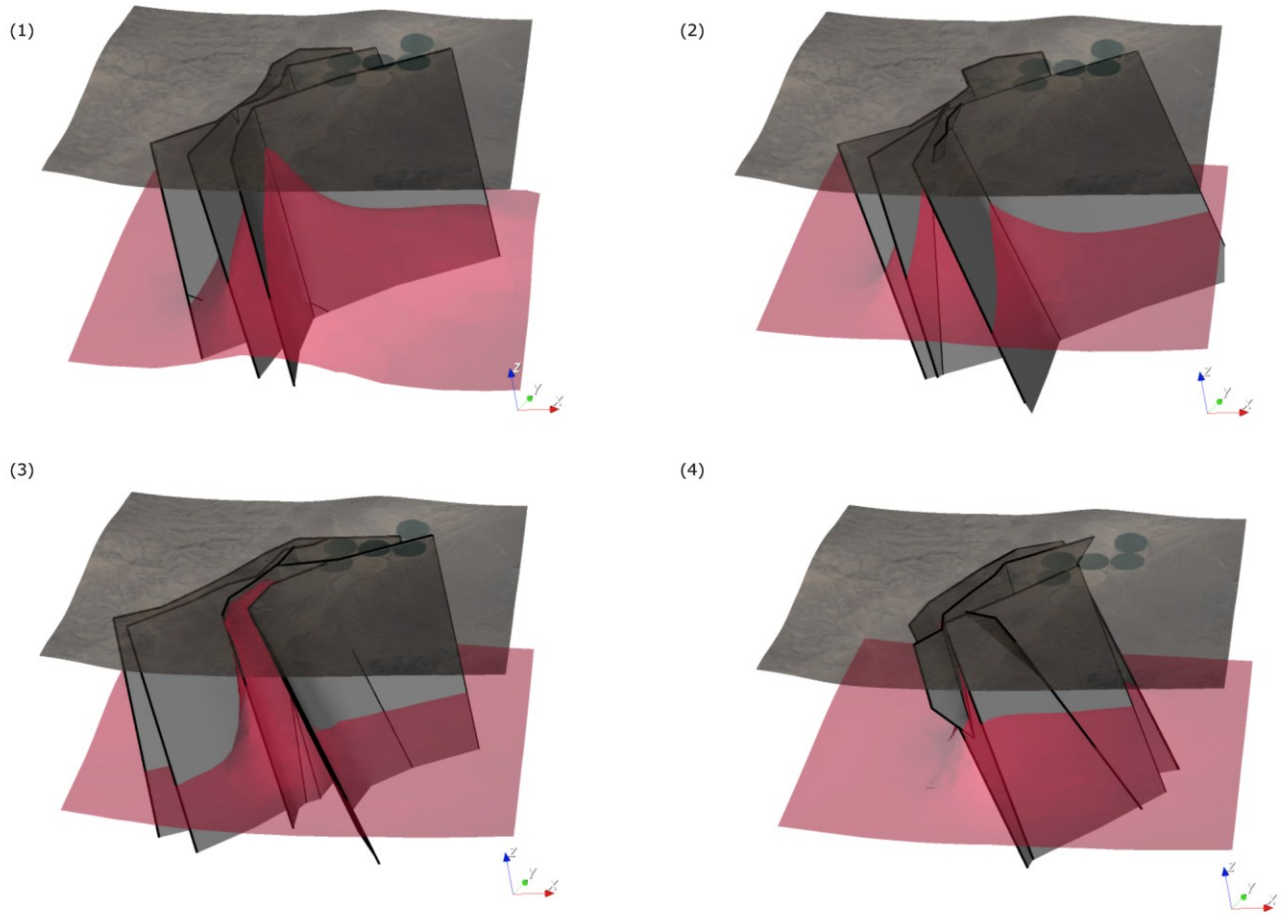

Figure 2: Random realization example of the 3D geological model from the four conceptual models. The colored surface represents the faults surfaces. The red surface at the bottom denotes the isosurface of 200 degC of the temperature field.

For each fault scenario, we generate 300 realizations by sampling from the prior distribution, resulting in 1,200 samples in total. Each high-fedility model takes between 30 minutes to two hours to generate. The surrogate model takes as input a set of spatial fields and fault-conditioned scalar parameters that together describe permeability structure and thermal boundary forcing. Specifically, the inputs include a 3D permeability field and a binary fault mask defined on the simulation grid. In addition, several fault-related parameters are imposed according to their physical definitions: fault thickness l_{th}, fault aperture factor ϵ, and the normal component of fault thermal conductivity k_{norm} are defined as fault-wide properties and are mapped to all grid cells belonging to the fault mask. Two basal anomaly parameters, σ_q (heat-flux anomaly amplitude) and ΔT_{deep} (deep temperature anomaly), are applied only along the bottom-fault region to represent localized deep forcing. Finally, the regional background basal heat flux q_{bg} is applied as a spatially uniform field across the entire domain. All parameters are therefore converted into a consistent set of input channels via rule-based mappings (fault-wide, bottom-fault–only, or domain-wide) and concatenated for learning. The model outputs the 3D temperature field at discrete time steps with fixed intervals, and all input/output fields are interpolated onto a common $64 \times 64 \times 64$ grid prior to training to ensure a consistent resolution across realizations (Example shown in **Figure 3** and **Figure 4**). Because of different simulation times, the output temperature field can have a different number of valid timesteps (max 8 timesteps). We construct a valid time mask to deal with the length difference. Time is treated as scalar input in Fourier-MIONet, and is treated as channel output for other models.

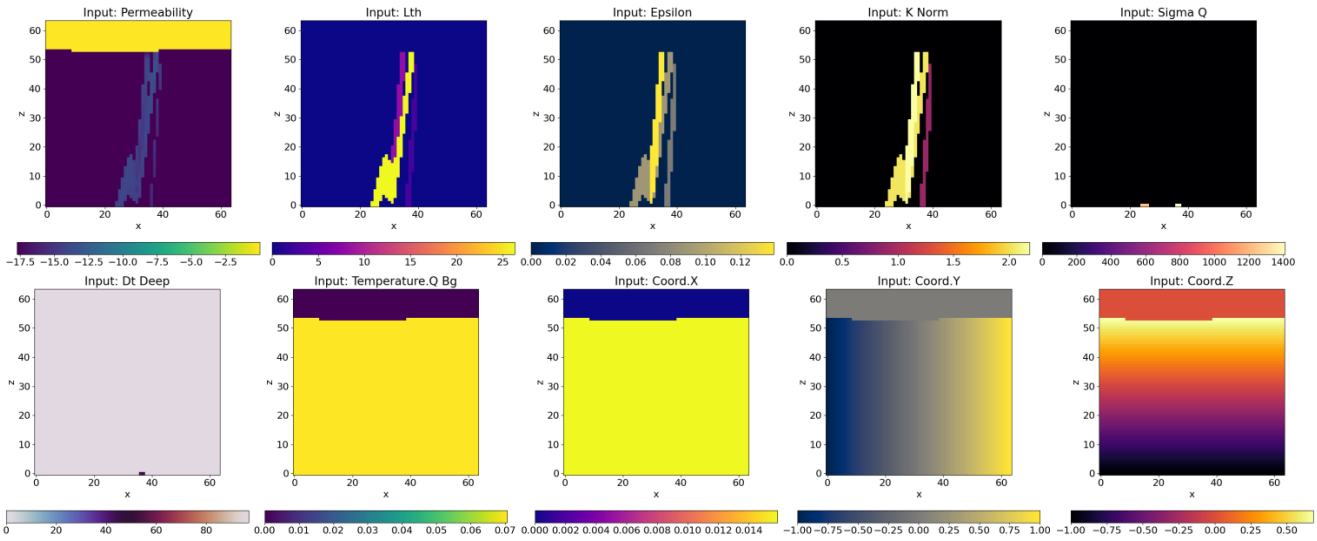

Figure 3: 2D slices of a random example of a 3D input data. Slices are taken from the mid-point in y direction. First row from left to right: Permeability, fault thickness l_{th} , fault aperture factor ϵ, and the normal component of fault thermal conductivity k_{norm} and heat-flux anomaly amplitude σ_q. Second row from left to right: deep temperature anomaly ΔT_{deep}, regional background basal heat flux q_{bg} and spatial coordinates. The top part of the model shows the effect of the topography mask.

We evaluate the surrogate under two complementary train/test protocols designed to probe both in-distribution performance and cross-scenario generalization. **Experiment 1 (multi-scenario training)** uses a random 80/20 split within each scenario: 80% of the samples from each of the four scenarios are pooled for training, and performance is evaluated on the held-out 20% from each scenario, reporting metrics per scenario and aggregated across all scenarios. **Experiment 2 (single-scenario training)** assesses transfer across conceptual models: the model is trained using 80% of the samples from one scenario only, and is then tested (i) on the remaining 20% from the same scenario to measure within-scenario generalization, and (ii) on all samples from the other three scenarios to quantify out-of-distribution performance under unseen fault configurations. All experiments use identical preprocessing and fixed hyperparameters to enable fair comparisons.

We apply an above-surface (air) mask to restrict computations to the physically meaningful subsurface domain. The mask is used consistently during training, validation, and testing, so losses and reported error metrics are evaluated only on subsurface cells (with above-surface voxels excluded).

3. RESULTS

For each architecture, we run two experiments described in Section 2 across the four conceptual fault scenarios. Within each experiment–scenario pair, models are trained on the training split under a fixed optimization setup (e.g., learning rate and L2 weight decay specified in advance). Held-out validation/test splits are used for evaluation. Training minimizes a relative error loss between predictions and ground truth following Wen et al. (2022): letting T_{true} denote the flattened true field and T_{pred} the flattened prediction, the loss L is defined as:

$$L(T_{true}, T_{pred}) = \frac{\|T_{true} - T_{pred}\|_2}{\|T_{true}\|_2}$$

where the operator $\|x\|_2$ denotes the l^2 norm of vector x.

3.1 Architecture Hyperparameters of each model

Because the model architectures differ, and because configurations are constrained by available hardware and the large dataset size, it is challenging to set up a strictly fair comparison across all models. We therefore adopted pragmatic, ad hoc configurations for each model. Detailed settings are provided in Table 1. All models are trained on a single 40GB A100 GPU on google cloud. Each model was trained for 50 epochs.

Table 1: Architecture Hyperparameters of each model

Model Type	Configuration		
U-Net	Depth	Embedding size	Conv blocks
	3	32	2

FNO	Depth		Embedding size		Modes	
	4		32		8	
U-FNO	Depth Fourier layers	Depth U-Fourier layers	Embedding size		Modes	
	3	3	32		8	
Fourier-MIONet	Depth Depth Fourier layers	Depth trunk MLP	Depth Fourier layers	Depth U-Fourier layers	Embedding size	Modes
	6	4	3	3	32	8

3.1 Same-Scenario Train–Test results

In the first experiment, each model is trained and evaluated on data from the same scenario. Example predictions on 2D slices are shown in **Figure 4.** Test errors for all architectures across the same-scenario experiments are summarized in **Figure 5**. Overall, all models perform well in this regime; in particular, FNO achieves relative errors below 0.025 for all scenarios. U-FNO and Fourier-MIONet both exhibit lower errors on average compared to the baseline U-NET.

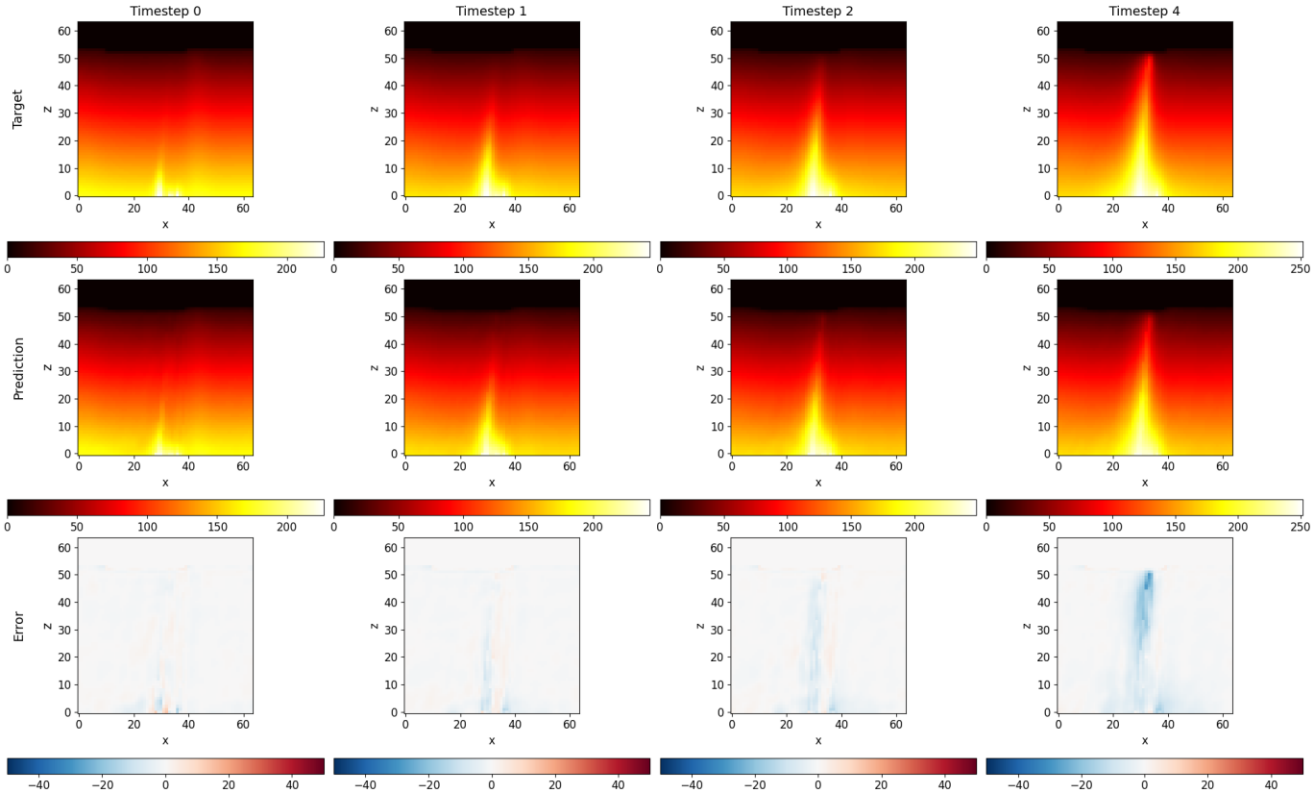

Figure 4: 2D slices of a random example of a 3D output data. The first row is ground truth; the second row is predictions (trained and tested on the same scenario with FNO); the third row is the error. Slices are taken from the mid-point in y direction. 4 valid timesteps are shown.

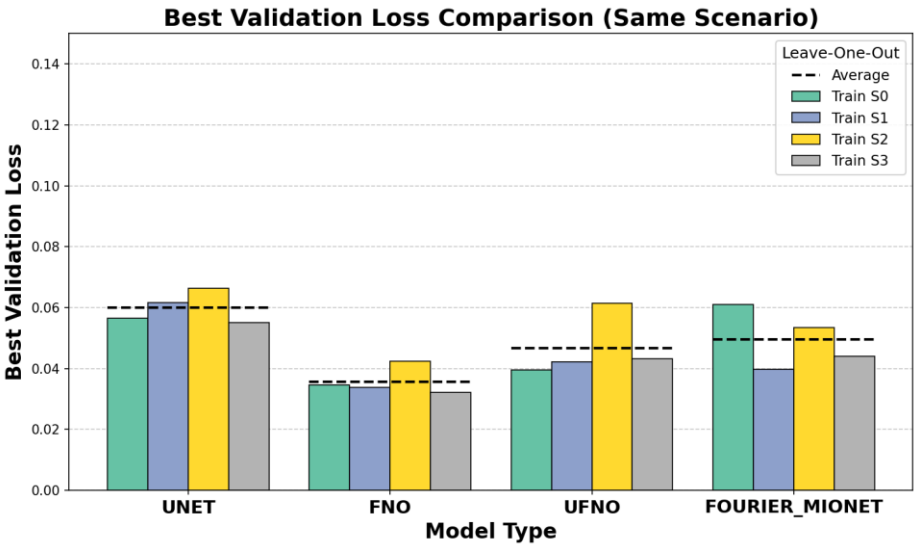

Figure 5: Summary of testing error for the experiment with training and testing data from the same scenario. Each bar is shown the error for an individual training and testing on the same scenario. The legend Si denotes the ith scenario is used in training and testing. i.e. Train S0 denotes training on scenario 0 data and test on scenarios 0

3.2 Cross-Scenario Train–Test Results (Train-on-One, Test-on-Others)

In the second experiment, we evaluate cross-scenario generalization using the same models trained in Section 3.1. Specifically, for each run, a model is trained on data from a single scenario and then tested not only on the held-out portion of that scenario, but also on all other scenarios. This setting is analogous to a leave-one-out evaluation over scenarios, and probes how well a model trained under one geological conceptualization transfers to unseen fault configurations and associated boundary and property variations.

Figure 6 summarizes the resulting test errors. Compared with the same-scenario results, errors generally increase under scenario shift, reflecting the distribution mismatch between training and testing data. Nevertheless, FNO remains the most robust, achieving relative errors of approximately 0.07 across unseen scenarios. Fourier-MIONet exhibits noticeably larger errors compared to other models than in the same-scenario experiment (**section 3.1)**, suggesting reduced generalization in this setting.

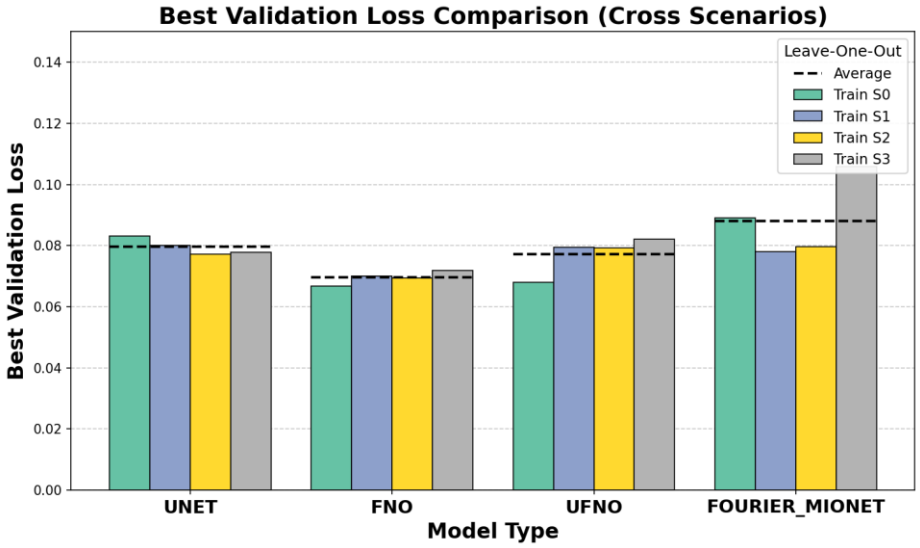

Figure 6: Summary of test errors for the cross-scenario experiment. Each bar reports the error for a model trained on a single scenario and evaluated on all remaining scenarios. In the legend, Si denotes that scenario i is used for training, while the other scenarios are used for testing. i.e. Train S0 denotes training on scenario 0 data and test on Scenarios 1,2,3

3.3 Combined-Scenario Train–Test results

In this experiment, we train each model on data pooled from all scenarios and evaluate it on a held-out test split drawn from the same combined dataset. This combined-scenario setting increases the diversity of fault geometries, material properties, and boundary conditions seen during training, and is intended to encourage the models to learn scenario-invariant relationships rather than overfitting to a single conceptual model.

Figure 7 summarizes the test errors for all architectures under combined-scenario training. Overall, training on the aggregated dataset improves robustness relative to the single-scenario setting when evaluated across heterogeneous conditions, while maintaining comparable accuracy on in-distribution samples. These results suggest that exposing models to multiple scenarios during training helps mitigate scenario-specific biases and yields more reliable performance when applied to realizations spanning a broader range of geological uncertainty. In this setting, FNO remains the top-performing model, with Fourier-MIONet achieving comparable accuracy.

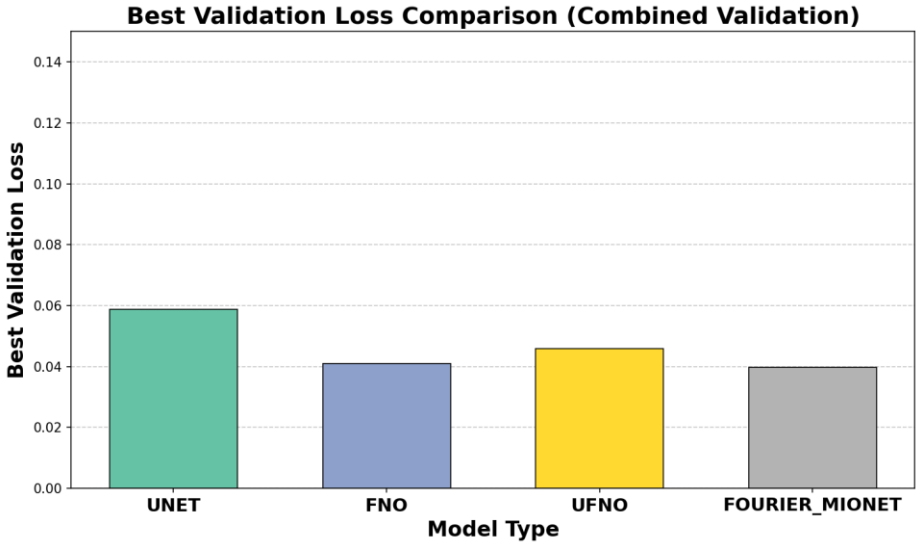

Figure 7: Summary of test errors for the combined-scenario train–test experiment. Each model is trained on data pooled from all scenarios (S1–S4) and evaluated on a held-out test split drawn from the same combined dataset. Bars report the test error for each model architecture under this setting.

4. CONCLUSION

This study presents, to our knowledge, the first comprehensive comparison of representative neural-operator surrogate architectures for 3D, time-dependent geothermal reservoir simulations. The results indicate that neural operators are promising surrogates for predicting steady-state temperature fields in fault- and fracture-dominated geothermal systems. In our experiments, the surrogate models achieve speedups of up to 10,000× relative to the high-fidelity coupled flow–heat simulations, while maintaining low prediction error. This reduction in computational cost enables workflows that are otherwise impractical, including large-scale uncertainty quantification (e.g., full Bayesian inference; Liang, 2023) and parameter optimization.

REFERENCES

Mao, S., Carbonero, A., & Mehana, M.. Deep learning for subsurface flow: A comparative study of U-net, Fourier neural operators, and transformers in underground hydrogen storage. Journal of Geophysical Research: Machine Learning and Computation, 2(1), (2025), e2024JH000401.

Pollack, A., & Mukerji, T. Accounting for subsurface uncertainty in enhanced geothermal systems to make more robust techno-economic decisions. Applied energy, 254, 113, (2019).

Daniilidis, A., Nick, H. M., & Bruhn, D. F. Interference between geothermal doublets across a fault under subsurface uncertainty; implications for field development and regulation. Geothermics, 91, 102041, (2021).

Ronneberger, O., Fischer, P., & Brox, T. , U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Cham: Springer international publishing, (2015).

Kovachki, N., Li, Z., Liu, B., Azizzadenesheli, K., Bhattacharya, K., Stuart, A., & Anandkumar, A. Neural operator: Learning maps between function spaces with applications to pdes. Journal of Machine Learning Research, 24(89), 1-97, (2023).

Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., & Anandkumar, A. Fourier neural operator for parametric partial differential equations. arXiv preprint arXiv:2010.08895, (2020).

Wen, G., Li, Z., Azizzadenesheli, K., Anandkumar, A., & Benson, S. M. U-FNO—An enhanced Fourier neural operator-based deep-learning model for multiphase flow. Advances in Water Resources, 163, 104180, (2022).

Jin, P., Meng, S., & Lu, L. MIONet: Learning multiple-input operators via tensor product. SIAM Journal on Scientific Computing, 44(6), A3490-A3514, (2022).

Jiang, Z., Zhu, M., & Lu, L. Fourier-MIONet: Fourier-enhanced multiple-input neural operators for multiphase modeling of geological carbon sequestration. Reliability Engineering & System Safety, 251, 110392, (2024).

Raissi, M., Perdikaris, P., & Karniadakis, G. E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. Journal of Computational physics, 378, 686-707, (2019).

Okoroafor, E. R., Smith, C. M., Ochie, K. I., Nwosu, C. J., Gudmundsdottir, H., & Aljubran, M. J. Machine learning in subsurface geothermal energy: Two decades in review. Geothermics, 102, 102401, (2022).

Liang, Z., Wellmann, F., & Ghattas, O. Uncertainty quantification of geologic model parameters in 3D gravity inversion by Hessian-informed Markov chain Monte Carlo. Geophysics, 88(1), (2023), G1-G18.

Degen, D., Cacace, M., & Wellmann, F. 3D multi-physics uncertainty quantification using physics-based machine learning. Scientific Reports, 12(1), 17491, (2022).

Zheng, F., Ma, M., Viswanathan, H., Pawar, R., Jha, B. and Chen, B., 2025. Deep Learning–Assisted Multiobjective Optimization of Geological CO2 Storage Performance under Geomechanical Risks. SPE Journal, pp.1-16.