

Finite-Volume-Aligned Physics Regularisation for Reservoir Pressure Surrogates under Distribution Shift: Robustness, Residual Monitoring, and Resolution-Transfer Limits

Franck-Hilaire Essiagne*, Kouassi Louis Kra, Moussa Camara

Laboratoire des Sciences Géographiques du Génie Civil et des Géosciences, Ecole Supérieure de Chimie du Pétrole et de l'Energie (ESCPE), Institut National Polytechnique Félix Houphouët-Boigny (INP-HB), Yamoussoukro, Côte d'Ivoire

Email: *franck.essiagne@inphb.ci

How to cite this paper: Essiagne, F.-H., Kra, K.L. and Camara, M. (2026) Finite-Volume-Aligned Physics Regularisation for Reservoir Pressure Surrogates under Distribution Shift: Robustness, Residual Monitoring, and Resolution-Transfer Limits. *Engineering*, 18, 367-396.

<https://doi.org/10.4236/eng.2026.1810021>

Received: August 8, 2026

Accepted: October 5, 2026

Published: October 8, 2026

Copyright © 2026 by author(s) and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

Fast reservoir surrogates are attractive for uncertainty quantification, optimisation, and repeated forecasting, yet data-only networks can lose accuracy and local conservation when geology or well controls leave the training distribution. This study isolates the marginal value of a finite-volume-aligned physics loss and then examines whether its residual can support trustworthy deployment. A 32×32 benchmark was generated for steady single-phase Darcy flow over heterogeneous permeability fields and variable injector-producer configurations. Identical 8089-parameter dilated convolutional networks were trained from 60 labelled simulations using pressure error alone (Data-60) or the same objective plus the normalized residual of the finite-volume stencil that generated the labels (PI-60). For the three-seed ensemble on fixed tests, PI-60 reduced mean relative pressure error by 5.3% on Test-ID and 10.4% on Fixed Compound-OOD (Test-OOD), with a 14.1% reduction on the independently regenerated Independent Combined-Shift stress test: P90 and P95 Fixed Compound-OOD errors decreased by 13.0% and 14.7%. The raw residual separated the designed ID and OOD populations extremely well (AUC = 0.997) but ranked case-wise OOD pressure error weakly (Spearman $\rho = 0.148$; worst-20%-error AUROC = 0.648). Stability-aware weighting materially improved this diagnostic role: a five-step Jacobi-preconditioned conjugate-gradient inverse-action score increased ρ to 0.382 and AUROC to 0.790, while the exact inverse-weighted energy score provided an analysis ceiling of $\rho = 0.466$ and AUROC = 0.838. In selective-simulation analysis, the preconditioned score reduced risk-coverage AURC by 4.6% relative to the raw residual; combining it with seed disagree-

ment did not help. Continuous severity sweeps further showed that PI-60 reduced integrated pressure error by 5.7% for shortening correlation length and 6.9% for increasing well rate, while reducing linear degradation slopes by 48.3% and 27.0%. Finally, a zero-shot $16 \times 16/32 \times 32/64 \times 64$ test exposed an important limit: the finite-volume reference showed decreasing adjacent-grid discrepancy and an empirical order near 2.38, whereas both learned surrogates became less coherent under refinement. PI-60 reduced the 32-to-64 discrepancy by 4.5% relative to Data-60 but did not restore convergence. These results distinguish fixed-grid finite-volume alignment from genuine discretization consistency and support residual monitoring only when stability, calibration, and resolution limits are made explicit. Batched CPU inference remained approximately 51 times faster than the reference finite-volume solve.

Keywords

Physics-Informed Machine Learning, Surrogate Modelling, Reservoir Simulation, Darcy Flow, Finite Volume, Distribution Shift, Residual Monitoring, Selective Simulation, Resolution Transfer

1. Introduction

Reservoir simulation underpins forecasting, uncertainty quantification, history matching, well-control optimization, and development planning. The repeated-evaluation burden is especially acute when geological ensembles and alternative control schedules must be propagated through a numerical model. Classical finite-volume formulations provide a transparent conservation framework and remain the reference for reservoir-flow calculations; well representation and grid-scale pressure interpretation also impose discretization-specific considerations [1]. Machine-learning surrogates seek to move most of this cost offline by learning a map from geological and operational inputs to high-dimensional states or production responses.

Scientific machine learning has evolved from differential-equation-constrained neural networks to operator-learning methods and discretization-aware hybrids. Physics-informed neural networks (PINNs) established the use of differential-equation residuals as soft constraints [2], while physics-constrained convolutional surrogates demonstrated that governing residuals could reduce dependence on labelled outputs in high-dimensional PDE problems [3]. Broader syntheses now distinguish physics-guided, physics-informed, physics-encoded, and neural-operator formulations and emphasize that the useful form of physics depends on discretization, optimization, and deployment purpose [4]-[9]. These advances are relevant to reservoirs because subsurface flow is governed by local conservation across strongly heterogeneous coefficients, a setting in which small state errors can coexist with physically inconsistent fluxes.

A substantial reservoir-surrogate literature now spans Bayesian convolutional encoder-decoder models, learned production forecasting, end-to-end simulator proxies, multiphase state surrogates, data assimilation, and time-varying controls [10]-[16]. Physics-constrained and physics-aware reservoir models have subsequently been developed for seepage, state and well-output prediction, porous-media flow, pressure management, neural operators, deep porous-flow surrogates, multiphase gradients, three-dimensional physics constraints, and mixed spatial/vector proxy inputs [17]-[25]. Recent work has extended this direction to physics-informed convolutional models for two-phase Darcy flow, domain-decomposed PINNs with sparse data, physics-informed data-driven flow models, and PointNet formulations driven by sparse observations [26]-[29]. These studies establish considerable potential, but they also make comparison difficult because architectures, labels, simulator fidelity, grids, and shift definitions often change simultaneously.

The most recent literature broadens the deployment target further. High-resolution carbon-storage surrogates, physics-informed spatio-temporal models, PINN reservoir simulators, nanoporous-flow predictors, multiphase PINNs, geological-CO₂ history-matching surrogates, improved neural operators, graph-network optimization surrogates, physics-constrained CO₂ models, control- and location-aware neural operators, transfer learning, feature-attention operators, layer-specific physics constraints, CNN-transformer optimization surrogates, large-scale inverse models, Fourier neural operators, well-aware PINNs, and differentiable multiphase simulators have all appeared from 2023 onward [30]-[47]. This rapid diversification creates a need for controlled evidence about what a physics term contributes when architecture and labelled data are held fixed.

This study therefore asks three linked questions. First, with the network, labelled examples, optimizer, grid, and finite-volume reference held fixed, how much does a finite-volume-aligned Darcy residual improve a low-data pressure surrogate? Second, is that benefit uniform across plausible geological and operational shifts, or does it preferentially protect failure regimes? Third, can the residual be transformed into a useful case-wise failure monitor, and do the apparent benefits survive a deliberate change of numerical resolution? The last question is essential because agreement with one discrete stencil is weaker than numerical consistency across a family of grids.

The study makes five contributions. i) It provides a matched three-seed Data-60 versus PI-60 ablation in which the only training difference is the finite-volume physics term. ii) It decomposes distribution shift into high-contrast, short-correlation, boundary-well, high-rate, and Independent Combined-Shift tests and adds continuous severity sweeps for the two shifts with the largest single-factor gains. iii) It quantifies upper-tail error and shows that the physics advantage concentrates in difficult OOD cases. iv) It separates broad shift detection from casewise failure ranking and tests raw, Jacobi-scaled, finite-iteration

preconditioned, and exact inverse-weighted residual scores through risk-coverage/selective-simulation analysis. v) It subjects the fixed-grid models to a common-scenario $16 \times 16/32 \times 32/64 \times 64$ resolution-transfer test, which falsifies the stronger claim that same-stencil physics alignment alone yields discretization-consistent surrogate convergence. The contribution is therefore a diagnostic and reproducible study of what discrete physics regularization does and does not buy.

2. Governing Problem and Numerical Benchmark

2.1. Dimensionless Single-Phase Pressure Problem

A two-dimensional, steady, incompressible single-phase pressure subproblem is considered on the unit square $\Omega = [0,1] \times [0,1]$. After nondimensionalization and omission of gravity, pressure satisfies

$$-\nabla \cdot (k(x) \nabla p(x)) = q(x), \forall x \in \Omega \quad (1)$$

where p is dimensionless pressure, $k > 0$ is scalar absolute permeability, and q is a balanced source-sink field. Constant-pressure boundaries are imposed on all four sides,

$$p(x) = 0, x \in \partial\Omega \quad (2)$$

The zero-boundary value defines the pressure datum and represents external constant-pressure support. The benchmark is deliberately dimensionless; reported pressure magnitudes are not field units. Each source map contains one injector and one producer with equal and opposite nominal total rates. Point sources are regularized with Gaussian kernels of standard deviation 1.15 grid cells to avoid a single-cell singular forcing and to maintain a consistent learning target at the selected resolution.

2.2. Finite-Volume Reference and Discrete Physics Operator

The governing equation is discretized on a uniform 32×32 Cartesian nodal grid using a node-centered two-point-flux control-volume formulation with harmonic interface averaging. For an interior node i , the discrete balance is

$$\sum_{j \in N(i)} T_{ij} (p_i - p_j) = q_i \quad (3)$$

where $N(i)$ is the four-neighbour von Neumann stencil and $T_{ij} = \frac{k_{ij}}{h^2}$ is the interface transmissibility. Harmonic averaging gives $k_{ij} = \frac{2k_i k_j}{k_i + k_j}$, with $h = \frac{1}{31}$.

Boundary pressures are fixed at zero; after elimination of the Dirichlet boundary degrees of freedom, the 30×30 interior linear system is symmetric positive definite and is solved using SciPy sparse linear algebra. The same assembled interior flux stencil, rather than an independently approximated continuous differential operator, is used to evaluate the learning residual. We refer to this property as finite-volume alignment: the training penalty and reference labels use the same

discrete conservative flux balance. It should not be confused with the stronger property of convergence across multiple grids or numerical schemes, which is tested separately in Section 4.10.

2.3. Geological and Control Sampling

Every permeability realization starts from an independent 32×32 array z_0 of standard-normal samples. The array is smoothed with `scipy.ndimage.gaussian_filter` using reflect boundary mode and a sampled Gaussian filter width c , then standardized to zero spatial mean and unit spatial standard deviation. The standardized field is multiplied by a sampled log-permeability scale σ_k and augmented by linear trends $t_x x + t_y y$, where x and y each span $[-1, 1]$ on the grid, $t_x \sim U[-0.35, 0.35]$, and $t_y \sim U[-0.25, 0.25]$. The final spatial mean is removed. Thus, the trend distributions are identical for all ID and shifted generators; the geology shifts change only σ_k and/or c . Exact ranges for every fixed and independently regenerated generator are given in **Table 1**.

Injector and producer row and column indices are sampled independently from the discrete uniform set $\{m, \dots, 31-m\}$, where m is the boundary-margin parameter in **Table 1** ($m = 5$ for the ID rule and $m = 2$ for the boundary-shift rule). A pair is accepted only when its Euclidean index-space separation exceeds $0.38 \times 32 = 12.16$; rejection sampling uses at most 100 attempts, after which the deterministic fallback is (m, m) and $(31-m, 31-m)$. Source magnitude is sampled independently from the corresponding continuous uniform range. Each well is represented by a Gaussian kernel with standard deviation 1.15 grid nodes; injector and producer kernels are normalized separately before forming $q(g_{\text{inj}} - g_{\text{prod}})$, and source values on the outer boundary are set to zero. The fixed train/validation/Test-ID/Fixed Compound-OOD dataset was generated once from NumPy RNG seed 20260805 in that order. The dataset partitions are summarized in **Table 2**, and the exact generator distributions are listed in **Table 1**.

Table 1. Exact generator parameter distributions. $U[a, b]$ denotes a continuous uniform distribution; all generators use $t_x \sim U[-0.35, 0.35]$ and $t_y \sim U[-0.25, 0.25]$.

Generator	σ_k	Gaussian Filter c (Grid Nodes)	Well Margin m	Source Magnitude q
ID: Train/Validation/Test-ID	$U[0.45, 0.85]$	$U[2.5, 5.0]$	5	$U[0.75, 1.20]$
Fixed Compound-OOD (Test-OOD)	$U[1.00, 1.35]$	$U[1.0, 2.2]$	2	$U[1.25, 1.65]$
Independent ID-Control	$U[0.45, 0.85]$	$U[2.5, 5.0]$	5	$U[0.75, 1.20]$
Independent High-Contrast	$U[1.00, 1.35]$	$U[2.5, 5.0]$	5	$U[0.75, 1.20]$
Independent Short-Correlation	$U[0.45, 0.85]$	$U[1.0, 2.2]$	5	$U[0.75, 1.20]$
Independent Boundary-Wells	$U[0.45, 0.85]$	$U[2.5, 5.0]$	2	$U[0.75, 1.20]$
Independent High-Rate	$U[0.45, 0.85]$	$U[2.5, 5.0]$	5	$U[1.25, 1.65]$
Independent Combined-Shift	$U[1.00, 1.35]$	$U[1.0, 2.2]$	2	$U[1.25, 1.65]$

Table 2. Dataset partitions and controlled stress-test design.

Split	Cases	Generator/Role
Training Pool	360	In-distribution; 60- and 300-label subsets drawn from the same pool
Validation	80	In-distribution; used for training monitoring only
Test-ID	120	Independent in-distribution cases
Fixed Compound-OOD (Test-OOD)	120	Fixed compound geology + well-control shift; independent 120-case test split
Factor-Shift Sets	6×120	6×120 independently regenerated sets: ID-control, high-contrast, short-correlation, boundary-wells, high-rate, independent combined-shift
Severity Sweeps	$2 \times 5 \times 120$	Common-scenario correlation-length and well-rate paths from ID to target OOD; locked models
Multi-Grid Test	80×3 Grids	Same continuum scenarios on 16×16 , 32×32 , and 64×64 ; zero-shot resolution transfer

2.4. Independent Factor-Shift Experiments

To determine which OOD factors drive the physics benefit, six additional 120-case sets were generated with the same finite-volume solver and independent deterministic random-number streams. They are named Independent ID-control, Independent High-contrast, Independent Short-correlation, Independent Boundary-wells, Independent High-rate, and Independent Combined-Shift, with generator seeds 20260820, 20261811, 20262802, 20263793, 20264784, and 20265775, respectively. Independent ID-control reproduces the ID generator; the next four sets change exactly one factor; Independent Combined-Shift applies all four shifted ranges jointly. Models were neither retrained nor tuned on these sets. Throughout the manuscript, the pre-existing fixed split generated under the master dataset RNG is called Fixed Compound-OOD (Test-OOD), whereas this separately regenerated stress-test set is called Independent Combined-Shift. The matched Data-60/PI-60 experimental workflow is summarized in **Figure 1**.

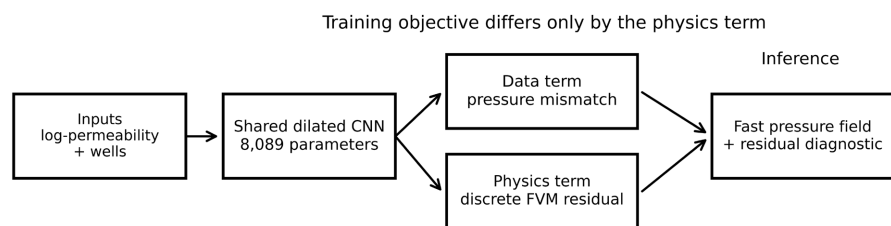


Figure 1. Controlled experimental design. Data-60 and PI-60 use the same permeability and well inputs, identical 8089-parameter dilated convolutional architecture, matched labelled cases, and matched training seeds. The only objective-level difference is the discrete finite-volume physics term.

3. Surrogate Model and Experimental Protocol

3.1. Network Architecture and Data Normalization

The surrogate is a compact fully convolutional network with seven 3×3 dilated convolutional layers of width 12 followed by a 1×1 output projection, for 8,089

trainable parameters. Dilation expands the receptive field without pooling so that the output remains aligned with the 32×32 nodal/control-volume grid. Inputs contain log-permeability and source-sink channels. For each training subset, log-permeability is normalized with that subset mean and standard deviation, the source map is divided by the maximum absolute source magnitude in that subset, and pressure is scaled only by $p_{\text{scale}} = \text{SD}(Y_{\text{train}})$, with no pressure-mean subtraction. The network predicts $\hat{p}_{\text{norm}}$ and every pressure field is de-normalized as $\hat{p} = p_{\text{scale}} \hat{p}_{\text{norm}}$ before the physics loss, residual diagnostics, and reported physical-space errors are evaluated.

Zero Dirichlet boundary values are enforced structurally inside every network forward pass. A fixed binary output mask sets the outermost row and column to exactly zero before either the data loss or physics loss is evaluated; boundary values are therefore neither free output degrees of freedom nor post-hoc overwritten predictions. For PI-60, the masked normalized pressure is multiplied by p_{scale} before the finite-volume residual is formed with the physical (de-normalized) log-permeability and source map. All residual diagnostics likewise use the de-normalized, boundary-masked pressure field and are evaluated on the 30×30 interior unknowns only.

The 360-case training pool was generated once, and label-count subsets were fixed by index before any test evaluation. Data-60 and PI-60 use cases 0 - 59, while Data-300 uses cases 0 - 299; hence the 60-case subset is nested inside the 300-case subset and is identical for all training seeds. The 60- and 300-label counts define the pre-fixed low-data and data-rich settings (a fivefold supervision contrast); they were not selected by inspecting test performance. Likewise, $\lambda = 0.03$ and the maximum epoch budgets (90 for Data-60/PI-60 and 35 for Data-300) are fixed configuration values in the reproducibility artifact: no λ sweep, epoch-budget search, or test-set hyperparameter tuning was performed. Every run completed its fixed epoch budget, and the retained checkpoint was selected using validation data only as the epoch with the minimum mean validation relative-L2 error. Retained epochs were 68, 60, and 65 for Data-60 seeds 11, 29, and 47; 83, 68, and 76 for PI-60; and 32 for Data-300 seed 101. Because λ was not adaptively tuned, the study does not claim that 0.03 is optimal; this is retained as a limitation. The fixed model, optimizer, training-budget, and monitoring settings are summarized in **Table 3**.

Table 3. Model and training settings.

Item	Setting
Grid	32×32 Cartesian nodal grid; 30×30 interior unknowns after Dirichlet elimination
Architecture	7 dilated 3×3 convolutions, width 12, plus 1×1 output projection
Trainable Parameters	8089
Low-Data Labels	60; fixed training cases 0 - 59; identical across seeds
Data-Rich Reference Labels	300; fixed cases 0 - 299; nested superset of the 60-case subset
Low-Data Epochs/Seeds	90-epoch budget; seeds 11, 29, 47; minimum-validation checkpoint retained

Continued

Data-Rich Reference	35-epoch budget; seed 101; minimum-validation checkpoint retained
Optimizer	AdamW, initial learning rate 2×10^{-3} , cosine schedule
Physics Weight	$\lambda = 0.03$; fixed before test evaluation; no test-based tuning
Software	Python, NumPy, SciPy, PyTorch; package supplied
Residual Monitors	raw NRMSE; Jacobi scaling; 5-step Jacobi-PCG; exact inverse-weighted analysis ceiling
Selective Simulation	coverage 0.20 - 1.00 in 0.05 increments; rejected cases delegated to FVM conceptually
Multi-Grid Evaluation	80 common scenarios on 16×16 , 32×32 , 64×64 ; trained 32-grid models held fixed

3.2. Data-Only and Physics-Regularized Objectives

For a batch of N labelled cases, the data-only model minimizes mean-square pressure error,

$$L_{data} = \frac{1}{N} \sum_{n=1}^N \|\hat{p}_n - p_n\|_2^2 \quad (4)$$

For PI-60, the predicted pressure is additionally evaluated with the same discrete finite-volume operator $A(k)$ used by the reference solver,

$$r_n = A(k_n) \hat{p}_n - q_n \quad (5)$$

and the physics loss normalizes the mean squared residual by source energy,

$$L_{phys} = \frac{\text{mean}(r^2)}{\text{mean}(q^2) + \varepsilon} \quad (6)$$

The combined objective is

$$L = L_{data} + \lambda L_{phys}, \lambda = 0.03 \quad (7)$$

For PI-60, the finite-volume residual is evaluated after de-normalization of the boundary-masked network output; the physical log-permeability and source map are passed to the residual operator without their network-input normalizations. The physics weight $\lambda = 0.03$ was fixed before any test-set evaluation rather than optimized on Test-ID or Test-OOD. The central comparison is therefore a controlled ablation of a pre-fixed finite-volume-aligned regularizer, not a test-set search for an optimal physics weight.

3.3. Evaluation Metrics and Statistical Inference

Primary accuracy is casewise relative L2 pressure error. The scientific hypothesis motivating the matched low-data ablation is directional: adding the finite-volume-aligned physics term should reduce pressure error. The two fixed-set pressure comparisons—Test-ID and Fixed Compound-OOD (Test-OOD)—are the co-primary inferential family. They were the core fixed-test comparisons in the study design, but the protocol was not externally preregistered. To avoid relying on a retrospective directional-testing assumption, the revision therefore uses two-sided paired Wilcoxon signed-rank tests for these co-primary comparisons. Family-wise type-I error across the two tests is controlled by Holm correction at

$\alpha = 0.05$. Peak relative error and normalized finite-volume residual are secondary outcomes.

For Data-60 and PI-60, the main fixed-test predictions are three-seed ensemble means. The case-bootstrap 95% confidence intervals and paired Wilcoxon tests act on the 120 test scenarios after the ensemble prediction has been formed. They therefore quantify test-case uncertainty conditional on this trained ensemble; they do not estimate variability over neural-network training. To expose training-run heterogeneity directly, matched Data-60 versus PI-60 effects are also reported separately for seeds 11, 29, and 47 in Supplementary **Table S1**. With only three training seeds, no population-level inference over random initialization is claimed. Data-300 has one seed and remains a descriptive reference.

Runtime was measured on CPU with batched neural inference and the same Python/SciPy finite-volume implementation. The speedup is an implementation-specific online comparison for this benchmark; it excludes model training and simulation-database generation and should not be transferred directly to commercial multiphase simulators.

3.4. Exploratory Robustness and Diagnostic Analyses

All factor-shift, upper-tail, baseline-difficulty, residual-monitoring, selective-simulation, continuous-severity, and multi-grid analyses are exploratory extensions beyond the two primary fixed-test pressure comparisons. *p*-values reported for these analyses are nominal and unadjusted and are interpreted descriptively rather than as confirmatory family-wise error-controlled tests. The exploratory analyses quantify error quantiles and exceedance rates, the relation between Data-60 baseline difficulty and PI-60 improvement, ID/OOD residual discrimination, case-wise failure ranking, three-seed disagreement, stability-weighted residuals, selective simulation, controlled shift-severity trajectories, and zero-shot multi-grid behaviour.

3.5. Stability-Weighted Residual Monitoring and Selective Simulation

For the fixed Test-ID and Fixed Compound-OOD (Test-OOD) cases, the Dirichlet boundary values were eliminated to form the interior symmetric positive-definite pressure operator A . For a surrogate prediction, the interior residual $r = A\hat{p} - b$ was evaluated in four norms. The existing raw score is the source-normalized residual root-mean-square. A Jacobi score weights each residual component by the inverse diagonal of A . A stronger approximate inverse-action score applies five steps of conjugate gradients with Jacobi preconditioning (PCG-5) to both r and b . Finally, an exact inverse-weighted score solves with A and is used only as an analysis ceiling because its cost is comparable in character to solving the pressure system itself. For any positive-definite inverse approximation B , the squared normalized score is

$$\eta B^2 = \frac{r^T B r}{b^T B b}, B \approx A^{-1} \quad (8)$$

For $B = A^{-1}$, the score is exactly the relative pressure error in the A-energy norm. This identity was verified numerically for every analysed case; the maximum relative discrepancy between the inverse-residual form and the direct energy-error form was 3.2×10^{-8} . The exact score therefore provides a principled ceiling against which cheap preconditioners can be judged, rather than a deployable monitor.

$$\eta A^{-12} = \frac{r^T A^{-1} r}{b^T A^{-1} b} = \frac{\|\hat{p} - p\|_A^2}{\|p\|_A^2} \quad (9)$$

Casewise monitoring quality is measured by Spearman correlation with relative L2 pressure error and AUROC for identifying the worst 20% of OOD pressure errors. Broad shift detection is evaluated separately by ID-versus-OOD AUROC. Selective simulation then ranks OOD cases from lowest to highest monitor score. For each surrogate coverage between 0.20 and 1.00, the retained-case mean and P90 pressure errors are reported; rejected cases are conceptually delegated to the finite-volume solver. Area under the risk-coverage curve (AURC) summarizes ranking quality. A combined monitor averages validation-calibrated empirical percentiles of PCG-5 score and three-seed disagreement; no pressure-error labels enter that combination.

3.6. Continuous Shift-Severity and Zero-Shot Multi-Grid Protocols

To replace binary OOD labels with controlled trajectories, 120 common scenarios were generated for each of two shift families at severities $s \in \{0, 0.25, 0.50, 0.75, 1\}$. For correlation-length shift (generator seed 20260910), each scenario draws one $u_c \sim U[0, 1]$, defines $c_{\text{ID}} = 2.5 + 2.5u_c$ and $c_{\text{OOD}} = 1.0 + 1.2u_c$, and uses $c(s) = (1-s)c_{\text{ID}} + sc_{\text{OOD}}$ while holding the same white-noise field, $\sigma_k \sim U[0.45, 0.85]$, trends, margin-5 wells, and ID-rate realization fixed. For well-rate shift (generator seed 20260911), each scenario draws $u_q \sim U[0, 1]$, defines $q_{\text{ID}} = 0.75 + 0.45u_q$ and $q_{\text{OOD}} = 1.25 + 0.40u_q$, and uses $q(s) = (1-s)q_{\text{ID}} + sq_{\text{OOD}}$ while holding geology and well locations fixed. Thus, within each common-scenario trajectory, only the targeted factor changes with severity.

The locked Data-60 and PI-60 ensembles are evaluated at every severity. For each scenario, integrated error is computed by trapezoidal integration over the five severity levels, and degradation slope is the ordinary least-squares coefficient from regressing that scenario's five relative-L2 errors on severity. Because every scenario uses the same five severity values, the mean of the 120 per-scenario OLS slopes equals the OLS slope of the plotted mean-error curve. Exploratory paired Wilcoxon tests and scenario bootstraps therefore act on the 120 per-scenario integrals or slopes, not on only five aggregate means.

A separate 80-scenario multi-grid test probes whether same-stencil physics alignment translates into numerical-resolution consistency. Each ID-like 32-grid permeability realization defines a smooth continuum interpolant evaluated on 16×16 , 32×32 , and 64×64 grids. Injector and producer coordinates are held fixed

in physical space. The Gaussian well source uses physical width $\frac{1.15}{31}$ and a reference-grid scaling that reproduces the original 32-grid source convention while maintaining a common continuum source density across resolutions. The same trained 32-grid convolutional parameters are then applied zero-shot at all three resolutions. Adjacent-grid discrepancies are evaluated after interpolation to the 64-grid physical coordinates. For a halving of grid spacing, the empirical adjacent-grid order is $\log_2(D_{16-32}/D_{32-64})$; positive values indicate shrinking cross-grid discrepancy under refinement. This is a deliberately stringent falsification test rather than a claim that the fixed-grid CNN was designed as a neural operator.

4. Results

4.1. Primary Matched Comparison: Accuracy and Conservation

PI-60 improved both pressure accuracy and discrete conservation on the fixed benchmark. On Test-ID, the three-seed ensemble mean relative L2 error decreased from 0.258 to 0.244, a 5.3% reduction. On Fixed Compound-OOD (Test-OOD), it decreased from 0.522 to 0.468, a 10.4% reduction. Mean normalized residual decreased from 1.056 to 0.722 in distribution (31.6%) and from 3.719 to 2.738 on Fixed Compound-OOD (26.4%). For the two co-primary pressure comparisons, the two-sided paired Wilcoxon p-values are 6.47×10^{-8} for Test-ID and 8.62×10^{-7} for Fixed Compound-OOD; Holm-adjusted values are 1.29×10^{-7} and 8.62×10^{-7} , respectively. Case-bootstrap intervals for the ensemble Data-60-minus-PI-60 difference exclude zero. These inferential quantities are conditional on the trained three-seed ensemble and characterize variation across test scenarios, not training-run uncertainty. The ensemble metrics are reported in **Table 4** and visualized in **Figure 2**.

Seed-matched effects show why that distinction matters. On Test-ID, mean relative-L2 reductions for seeds 11, 29, and 47 are 4.2%, 4.8%, and 7.7%. On Test-OOD, the corresponding reductions are 2.2%, 1.6%, and 36.1%. All three seed-specific mean differences favour PI-60, but the OOD effect magnitude is strongly heterogeneous, and the seed-47 contrast contributes materially to the ensemble advantage. Supplementary **Table S1** reports the seed-specific means and retained validation epochs. These are descriptive training-run effects; no inference over the population of possible training seeds is made.

Table 4. Primary ensemble performance. Values are casewise means with 95% bootstrap confidence intervals. Data-300 is a single-seed descriptive reference.

Model	Test-ID Rel. L2	Fixed Test-OOD Rel. L2	Test-ID Residual	Fixed Test-OOD Residual
Data-60	0.258 [0.242, 0.275]	0.522 [0.491, 0.556]	1.056 [1.007, 1.110]	3.719 [3.426, 4.034]
PI-60	0.244 [0.229, 0.261]	0.468 [0.445, 0.492]	0.722 [0.698, 0.748]	2.738 [2.506, 2.994]
Data-300	0.198 [0.184, 0.212]	0.509 [0.470, 0.552]	1.073 [1.002, 1.156]	4.064 [3.706, 4.426]

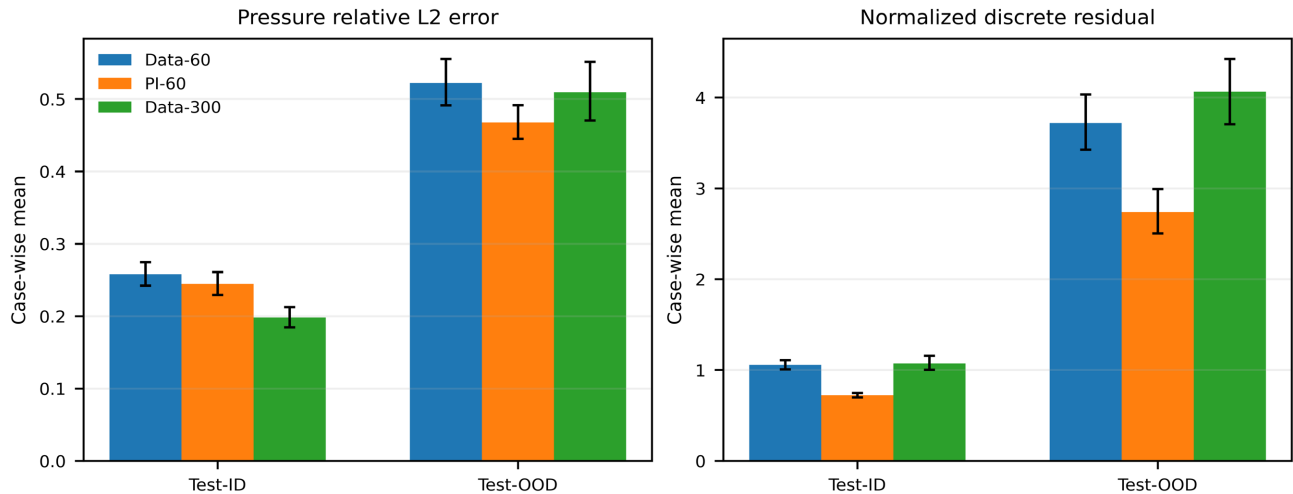


Figure 2. Primary pressure-error and conservation results for the fixed Test-ID and Fixed Compound-OOD (Test-OOD) sets. Error bars are case-bootstrap 95% confidence intervals. The low-data physics model improves both metrics, while the data-rich model primarily improves interpolation accuracy.

4.2. Paired Case Behaviour and Representative OOD Solution

The mean improvements are not driven by a small number of extreme cases. PI-60 has lower relative L2 error than Data-60 in 70% of both Test-ID and Test-OOD cases. The paired scatter also shows that the largest absolute gains tend to occur where the data-only baseline error is high. A representative OOD realization illustrates the same effect spatially: the physics-regularized field remains smooth at the reservoir scale while reducing localized errors near forcing and heterogeneity transitions. The residual term does not reproduce the finite-volume solution exactly, but it biases predictions toward flux-consistent fields. The paired case behaviour is shown in **Figure 3**, and representative pressure and absolute-error fields are shown in **Figure 4**.

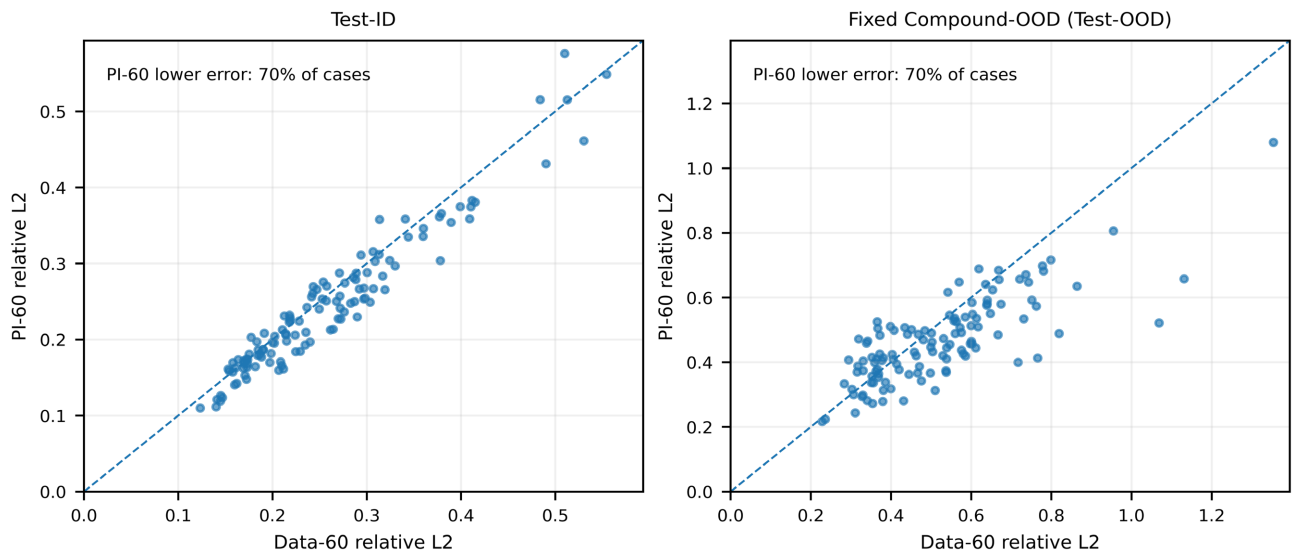


Figure 3. Casewise paired relative L2 error for Data-60 and PI-60. Points below the diagonal favour PI-60. PI-60 has lower error for 70% of cases in both Test-ID and the Fixed Compound-OOD (Test-OOD) set.

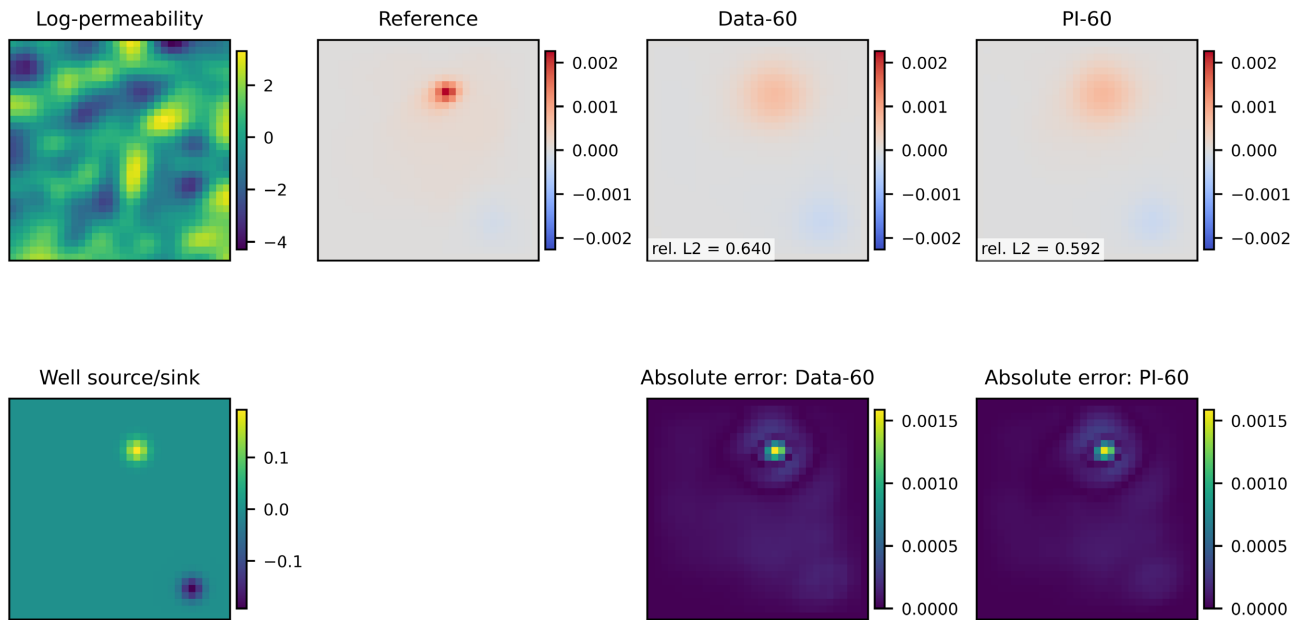


Figure 4. Representative Fixed Compound-OOD (Test-OOD) case selected near the median paired physics improvement. Top row: log-permeability, finite-volume reference pressure, Data-60 prediction, and PI-60 prediction. Bottom row: source-sink map and absolute-error maps. Colour limits are shared within comparable panels.

4.3. Online Runtime

All neural surrogates are substantially faster than the sparse finite-volume reference in the measured batched CPU regime. PI-60 requires 0.175 ms per case versus 8.873 ms for the finite-volume solve, corresponding to a 50.7x online speedup. Data-60 and Data-300 achieve 48.7x and 71.7x, respectively. The difference among neural variants is small relative to the solver gap; physics regularization changes training, not the online forward architecture. The measured online timings are summarized in [Table 5](#).

Table 5. Measured batched CPU online runtime.

Method	CPU ms/Case	Speedup vs FVM
Finite-Volume	8.873	1.0x
Data-60	0.182	48.7x
PI-60	0.175	50.7x
Data-300	0.124	71.7x

4.4. Shift Decomposition Reveals Factor-Specific Robustness and an Amplified Independent Combined-Shift Response

Sections 4.4 - 4.10 report exploratory analyses. Unless explicitly stated otherwise, their p-values are nominal and unadjusted.

The independent factor-shift experiments change the interpretation of the aggregate OOD result. Physics regularization does not provide a uniform percentage improvement across all perturbations. Relative L2 reduction is 5.2% on the independent

ID control, 2.7% under high permeability contrast, 8.9% under shortened correlation length, 2.9% for boundary-near wells, 8.3% for higher rates, and 14.1% under the Independent Combined-Shift stress test. In the Independent Combined-Shift set, PI-60 improves on 81.7% of cases and the bootstrap 95% interval for the mean Data-60-minus-PI-60 difference is [0.0587, 0.1007], with nominal exploratory one-sided Wilcoxon $p = 9.45 \times 10^{-14}$. The 14.1% Independent Combined-Shift gain is 1.59 times the strongest single-factor pressure gain (8.9%), indicating that the value of the physics regularizer is amplified when covariate shifts co-occur rather than being a fixed architecture-wide offset. Condition-wise metrics are reported in **Table 6**, and the corresponding pressure/residual reductions are summarized in **Figure 5**.

An important exception prevents an over-general claim. Boundary-well stress improves the mean relative L2 by 2.9% but slightly worsens the 90th-percentile error by 1.4%. High-contrast geology is also instructive: the residual improves by 50.1%, yet relative L2 improves by only 2.7%, and the peak-relative-error improvement is not significant (nominal exploratory $p = 0.349$). These cases show that better local conservation and better pressure accuracy are related objectives but not interchangeable outcomes. The conservation-accuracy decoupling across all six controlled conditions is visualized in **Figure 6**.

Table 6. Independent factor-shift experiments using locked trained models. Each condition contains 120 newly generated cases.

Condition	Data-60 Rel. L2	PI-60 Rel. L2	Reduction	PI Win Frac.	Residual Reduction
ID Control	0.2755	0.2613	5.2%	76.7%	30.3%
High Contrast	0.3848	0.3745	2.7%	57.5%	50.1%
Short Correlation	0.2962	0.2699	8.9%	67.5%	21.0%
Boundary Wells	0.3132	0.3042	2.9%	64.2%	30.1%
High Rate	0.2785	0.2554	8.3%	83.3%	30.9%
Independent Combined-Shift	0.5538	0.4758	14.1%	81.7%	32.1%

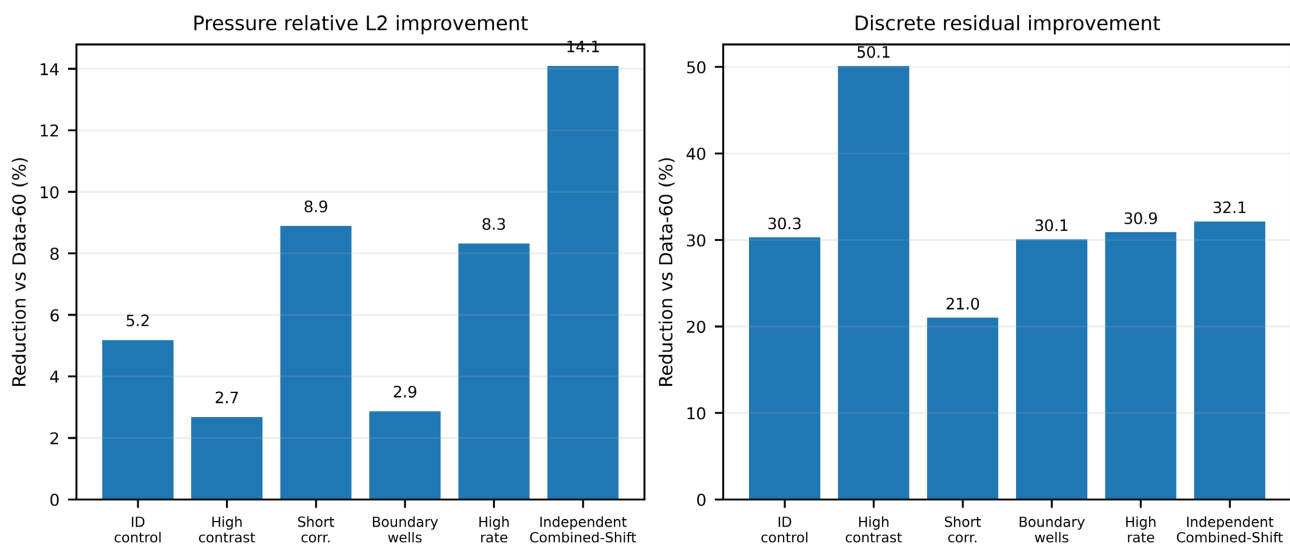


Figure 5. Shift-specific relative pressure-error and discrete-residual reductions of PI-60 relative to Data-60. The accuracy benefit is largest for Independent Combined-Shift, short correlation, and high rate, whereas the conservation benefit is largest for high-contrast geology.

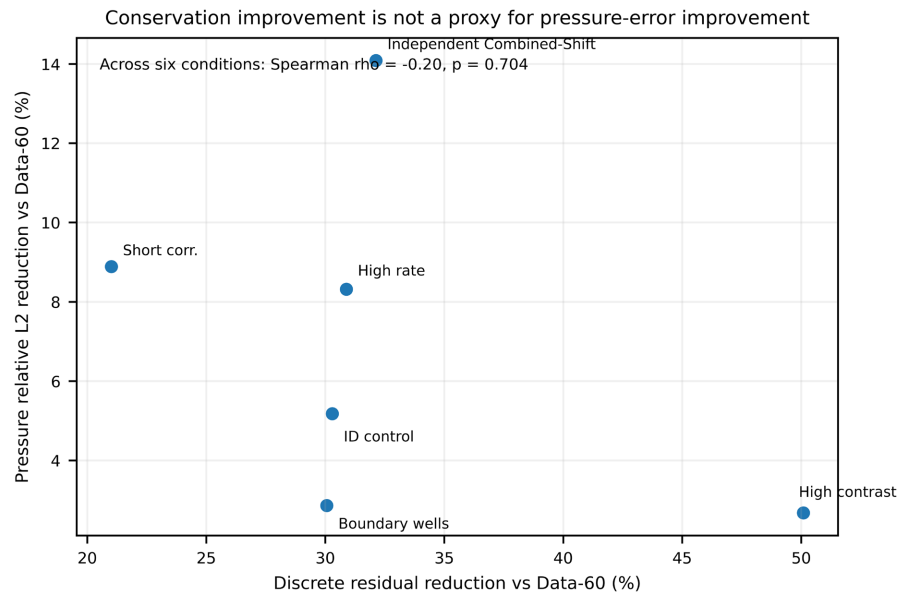


Figure 6. Conservation-accuracy decoupling across the six controlled conditions. High-contrast geology produces the largest residual reduction but only a small pressure-error reduction; Independent Combined-Shift produces the largest pressure-error gain with a more moderate conservation gain. Across the six conditions, the Spearman association is descriptive and weak ($\rho = -0.20$, $p = 0.704$).

4.5. Physics Regularization Preferentially Reduces Difficult-Case Tail Risk

Mean OOD error understates the distributional effect. For Data-60, the Test-ODD P90 and P95 relative L2 errors are 0.744 and 0.800; PI-60 reduces them to 0.647 and 0.682, corresponding to reductions of 13.0% and 14.7%. Using the Data-60 P90 value (0.744) as a fixed high-error threshold, 10.0% of Data-60 cases exceed the threshold by construction, whereas only 1.67% of PI-60 cases do so. The Data-60 P75 threshold is exceeded by 25.0% of Data-60 cases but only 14.17% of PI-60 cases.

The paired gain also scales with baseline difficulty. On Test-ODD, Data-60 baseline error and the improvement Data-60 minus PI-60 have Spearman $\rho = 0.644$ ($p = 2.15 \times 10^{-15}$). This is stronger than the corresponding Test-ID association ($\rho = 0.237$, $p = 0.0093$). The result supports a more precise claim than average robustness: within this benchmark, the discrete physics term behaves like a hard-case regularizer whose benefit becomes larger as the data-only prediction moves farther from the reference under distribution shift. The tail summary is reported in **Table 7**, and the full OOD error distribution and difficulty-benefit relation are shown in **Figure 7**.

Table 7. Tail distribution of case-wise relative L2 error.

Split/Model	Mean	P75	P90	P95	Maximum
Test-ID Data-60	0.258	0.299	0.380	0.418	0.554
Test-ID PI-60	0.244	0.282	0.359	0.385	0.576

Continued

Test-OOD Data-60	0.522	0.612	0.744	0.800	1.354
Test-OOD PI-60	0.468	0.534	0.647	0.682	1.080
Test-OOD Data-300	0.509	0.593	0.782	0.987	1.457

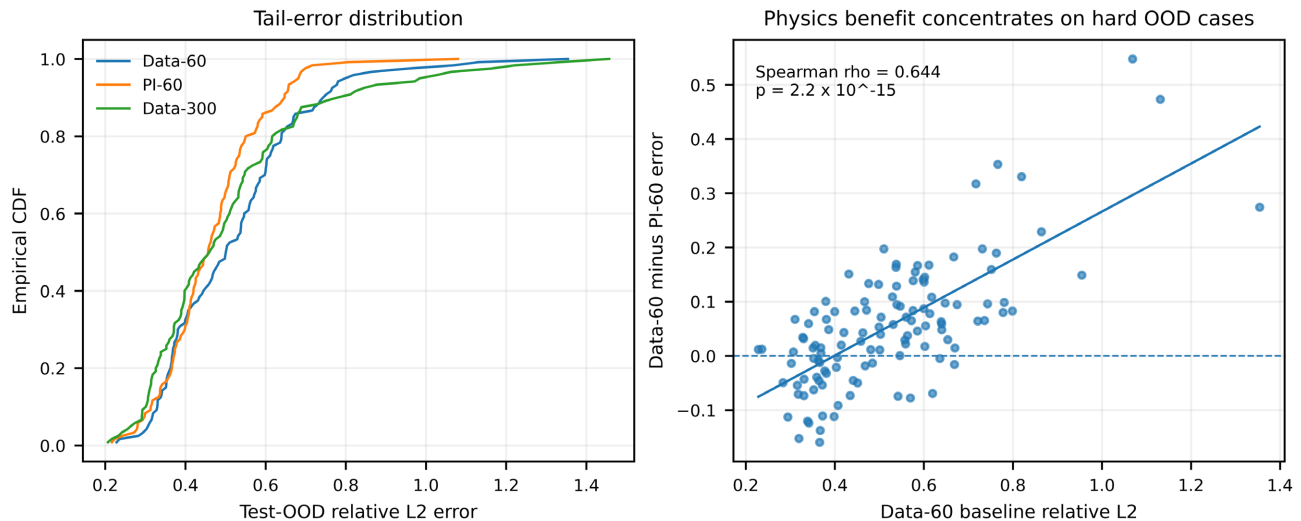


Figure 7. OOD upper-tail error behaviour. Left: empirical CDF of case-wise relative L2 error. Right: paired physics benefit versus Data-60 baseline difficulty, showing that the largest gains concentrate on harder OOD cases (Spearman $\rho = 0.644$, $p = 2.2 \times 10^{-15}$).

4.6. Residual Is a Strong Shift Watchdog but Not an Accuracy Certificate

The normalized discrete residual has two distinct diagnostic roles. First, it separates broad distribution regimes extremely well without access to reference pressure labels. For PI-60, classifying fixed Test-ID versus Test-OOD cases using residual alone yields $\text{AUC} = 0.997$; Data-60 and Data-300 yield $\text{AUC} = 0.995$ and 0.987 . This suggests an operational use in which unexpectedly large residuals trigger a warning, additional simulation, or fallback to the full solver.

Second, residual magnitude is not a calibrated within-regime error estimator. For PI-60, residual-error Spearman correlation is 0.396 on Test-ID but only 0.148 on Test-OOD, with $p = 0.108$ for the latter. Its AUC for identifying the worst 20% of PI-60 OOD pressure errors is only 0.648 . Thus, the residual can reliably indicate that the input population has shifted while still failing to rank individual OOD cases by true state error. The distinction matters for safe deployment: conservation violation is observable without labels, but a low residual should not be interpreted as proof that the pressure field is accurate.

Three-seed disagreement provides a complementary but similarly limited signal. Mean OOD disagreement decreases from 0.402 for Data-60 to 0.237 for PI-60, a 41.1% reduction, which indicates that the physics term stabilizes predictions across training seeds. However, PI-60 disagreement has essentially no relationship to OOD error ($\rho = -0.023$, $p = 0.804$; top 20%-error $\text{AUC} = 0.494$). In this experiment, neither residual nor ensemble spread alone is a sufficient casewise

error certificate; the residual is valuable primarily as a population-shift watchdog. The residual/disagreement diagnostics are reported in **Table 8**, and the ID-versus-ODD residual diagnostic is shown in **Figure 8**.

Table 8. Label-free residual and ensemble-disagreement diagnostics. “Residual OOD AUC” distinguishes Test-ID from Test-ODD using residual alone.

Model	ID Residual-Error ρ	OOD Residual-Error ρ	Residual OOD AUC	OOD Disagreement	Disagreement-Error ρ
Data-60	0.430	0.306	0.995	0.402	0.188
PI-60	0.396	0.148	0.997	0.237	-0.023
Data-300	0.475	0.312	0.987	n/a	n/a

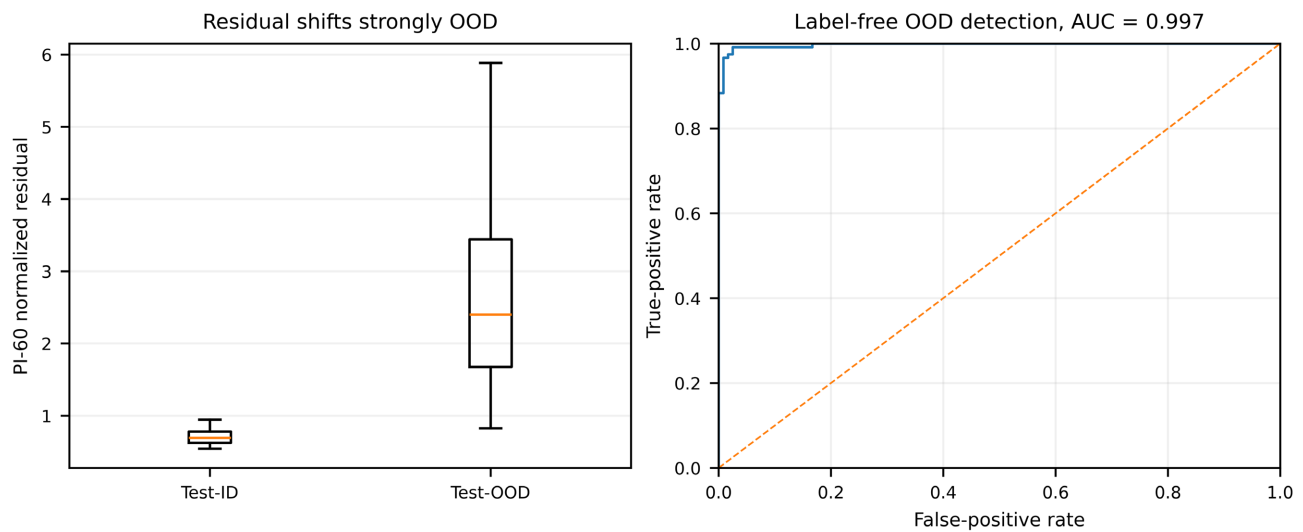


Figure 8. PI-60 residual as a label-free distribution-shift diagnostic. Left: residual distributions on ID and OOD cases. Right: receiver-operating-characteristic curve for discriminating the two populations using residual alone (AUC = 0.997). This does not imply case-wise pressure-error certification.

4.7. Stability-Aware Residual Weighting Improves Failure Ranking but Not Broad Shift Detection

The raw normalized residual remains an excellent detector of the designed population shift but is a weak casewise failure score. On PI-60 Test-ODD, its correlation with relative L2 error is $\rho = 0.148$ ($p = 0.108$), and its AUROC for the worst 20% of pressure errors is 0.648. Diagonal Jacobi scaling alone does not help ($\rho = 0.131$; AUROC = 0.624). In contrast, the five-step PCG inverse-action score increases ρ to 0.382 ($p = 1.71 \times 10^{-5}$) and worst-20%-error AUROC to 0.790. The exact inverse-weighted energy score reaches $\rho = 0.466$ and AUROC = 0.838, establishing that a substantial part of the raw-residual failure is attributable to the norm in which imbalance is measured rather than to the residual concept itself.

The improvement is specific to casewise ranking. ID-versus-ODD AUROC remains near saturation for every score: 0.997 for the raw residual, 0.999 for Jacobi scaling, and 0.996 for both PCG-5 and exact energy weighting. Thus, stability-aware weighting is unnecessary for recognizing this broad designed shift, but it is

materially more informative when the operational question is which OOD cases are most likely to be inaccurate. Monitor-level ranking metrics are reported in **Table 9** and compared graphically in **Figure 9**.

Table 9. PI-60 residual-monitor diagnostics on the fixed OOD set. Exact energy weighting is an analysis ceiling, not a cheap online monitor.

Monitor	OOD ρ	p	Worst-20% AUROC	ID/OOD AUROC
Raw FV	0.148	1.08e-01	0.648	0.997
Jacobi	0.131	1.54e-01	0.624	0.999
PCG-5	0.382	1.71e-05	0.790	0.996
Exact Energy	0.466	8.32e-08	0.838	0.996

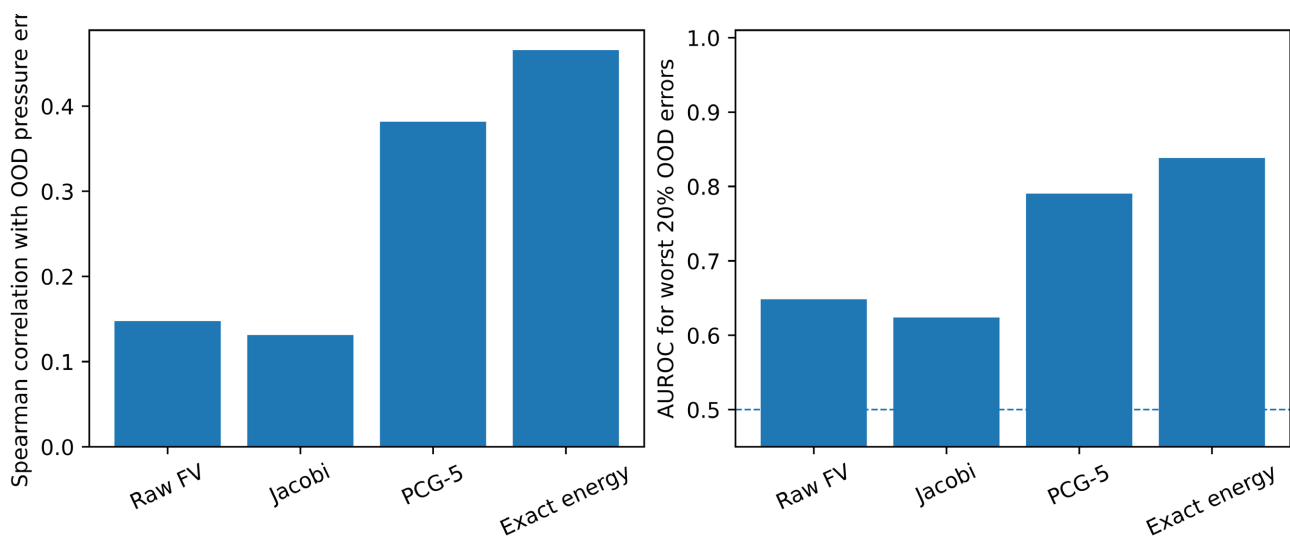


Figure 9. Stability-aware residual monitoring for PI-60 on Test-OOD. Left: rank association between each residual score and true pressure error. Right: AUROC for identifying the worst 20% of OOD pressure errors. The exact inverse-weighted energy score is a non-deployable analysis ceiling.

4.8. Selective Simulation Converts Residual Monitoring into an Explicit Risk-Coverage Trade-Off

Ranking cases by the monitor and delegating the highest-risk fraction to the finite-volume solver exposes a more operational distinction than AUROC alone. Across surrogate coverages from 20% to 100%, PCG-5 attains risk-coverage AURC = 0.430 compared with 0.450 for the raw residual, a 4.6% reduction; the exact energy score provides a ceiling of 0.423. At 50% coverage, the mean relative L2 error among retained PI-60 cases is 0.423 with PCG-5 versus 0.440 with the raw residual and 0.414 with exact energy weighting. The PCG-5 retained-case P90 is 0.540, compared with 0.576 for the raw residual.

The ensemble signal does not provide complementary information in this benchmark. Seed disagreement alone gives AURC = 0.468, and the label-free percentile combination of PCG-5 with disagreement gives 0.467, both worse than PCG-5 alone. This negative result is important: different uncertainty or physics

scores should not be assumed to improve reliability merely by being combined. Selective-simulation metrics are reported in **Table 10**, and the complete risk-coverage curves are shown in **Figure 10**.

Table 10. PI-60 selective-simulation performance. Lower AURC and retained-case error are better; *exact energy is an analysis ceiling.

Monitor	AURC	Mean @50%	P90 @50%	Mean @80%
Raw FV	0.450	0.440	0.576	0.461
PCG-5	0.430	0.423	0.540	0.448
Exact Energy*	0.423	0.414	0.514	0.445
Seed Disagreement	0.468	0.469	0.642	0.462
PCG-5 + Disagreement	0.467	0.462	0.648	0.469

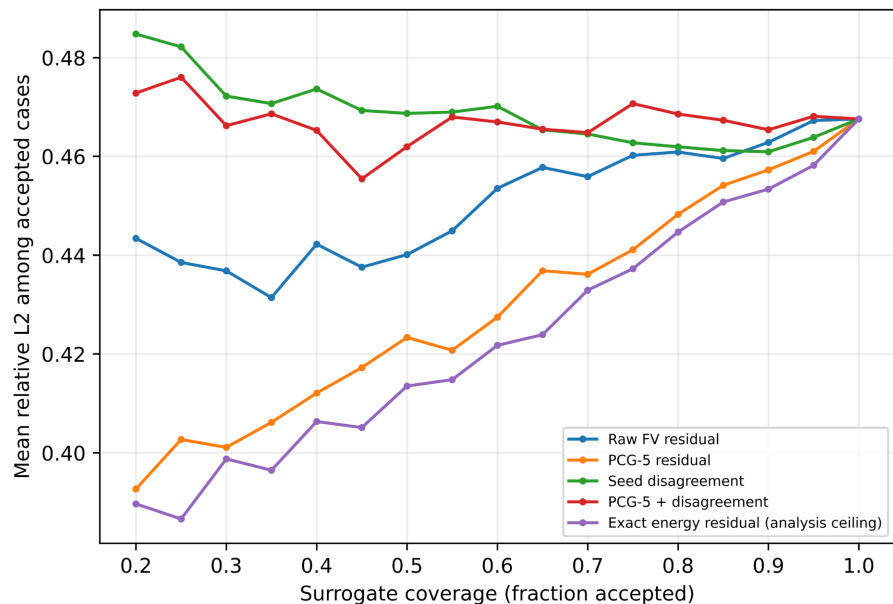


Figure 10. Selective-simulation risk-coverage curves for PI-60 on Test-OOD. Cases with the largest monitor score are delegated conceptually to the finite-volume solver; the vertical axis reports mean pressure error among cases retained by the surrogate. PCG-5 improves on the raw residual, while adding seed disagreement does not.

4.9. Continuous Severity Sweeps Show Flatter Degradation under Physics Regularization

The exploratory continuous sweeps confirm that the binary shift results are not artifacts of one selected OOD endpoint. Along the correlation-length path, mean Data-60 error changes from 0.2616 at severity 0 to 0.2926 at severity 1, whereas PI-60 changes from 0.2499 to 0.2649. The per-scenario integrated error is 5.7% lower for PI-60 (nominal paired $p = 1.78 \times 10^{-7}$), and the mean OLS degradation slope falls from 0.0312 to 0.0161, a 48.3% reduction (nominal $p = 0.00159$). The PI-60 curve is mildly non-monotone near the strongest short-correlation setting.

For increasing well rate, the five PI-60 mean errors plotted in **Figure 11** are

0.2419, 0.2417, 0.2445, 0.2491, and 0.2548 at severities 0, 0.25, 0.50, 0.75, and 1.00. Their ordinary least-squares slope is 0.01331; because the same severity design is used for every scenario, this equals the mean of the 120 per-scenario OLS slopes reported in **Table 11**. The corresponding Data-60 mean-curve slope is 0.01824, giving the reported 27.0% reduction. Integrated error is 6.9% lower for PI-60 (nominal paired $p = 8.92 \times 10^{-12}$), and the slope comparison has nominal $p = 1.36 \times 10^{-5}$. **Figure 11** has been regenerated with endpoint values and fitted slopes printed directly on the panels to make the numerical consistency explicit. The paired continuous-shift summary is reported in **Table 11**.

Table 11. Paired continuous-shift summaries over 120 common scenarios per family.

Shift Family	Integrated Error Reduction	Data Slope	PI Slope	Slope Reduction	p (Slope)
Correlation Length	5.7%	0.0312	0.0161	48.3%	1.59e-03
Well Rate	6.9%	0.0182	0.0133	27.0%	1.36e-05

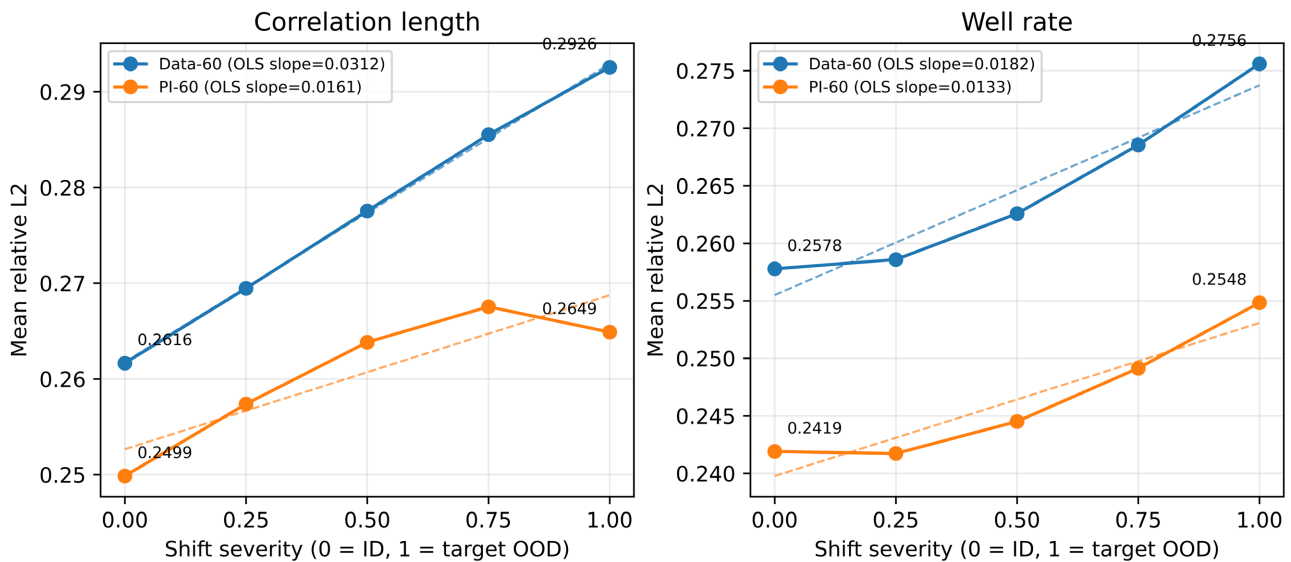


Figure 11. Mean relative pressure error as a continuous function of shift severity for shortened permeability correlation length and increased well rate. Severity 0 follows the ID generator and severity 1 reaches the target OOD generator; the same underlying scenarios are followed across each path. Numerical endpoints are printed beside the curves, and dashed lines show OLS fits; legends report the fitted slopes. For the well-rate PI-60 curve, the endpoint means are 0.2419 and 0.2548, and the fitted slope is 0.0133.

4.10. The Multi-Grid Falsification Test Rejects a Strong Discretization-Consistency Claim

The common-scenario multi-grid experiment produces a clear negative result. The finite-volume reference behaves as a convergent numerical sequence: mean adjacent-grid discrepancy falls from 0.0416 for 16-to-32 to 0.0080 for 32-to-64, corresponding to median empirical order 2.38. The learned surrogates do not follow that trend. Data-60 discrepancy increases from 0.871 to 1.638 and PI-60 from 0.864 to 1.563, giving negative median empirical orders of -0.897 and -0.861 . Same-grid pressure error is lowest at the trained 32×32 resolution and becomes

very large at both unseen resolutions, especially 64×64 .

PI-60 nevertheless remains modestly more coherent than Data-60 at the 32-to-64 transition: mean discrepancy is 4.54% lower, with a paired bootstrap interval for the absolute Data-minus-PI difference of $[0.0607, 0.0880]$ and nominal exploratory one-sided Wilcoxon $p = 1.44 \times 10^{-12}$. This statistically clear but practically small improvement does not restore convergence. The experiment therefore resolves the terminology issue: the present method is finite-volume-aligned on the training discretization, not discretization-consistent in the stronger multi-grid sense. The zero-shot multi-grid accuracy metrics are reported in **Table 12**, and the resolution-transfer comparison is shown in **Figure 12**.

Table 12. Zero-shot multi-grid accuracy of the unchanged 32-grid-trained ensembles.

Grid	Data-60 Rel. L2	PI-60 Rel. L2	Data Residual	PI Residual
16×16	0.921	0.922	0.983	0.983
32×32	0.250	0.237	0.987	0.706
64×64	1.614	1.538	5.236	4.992

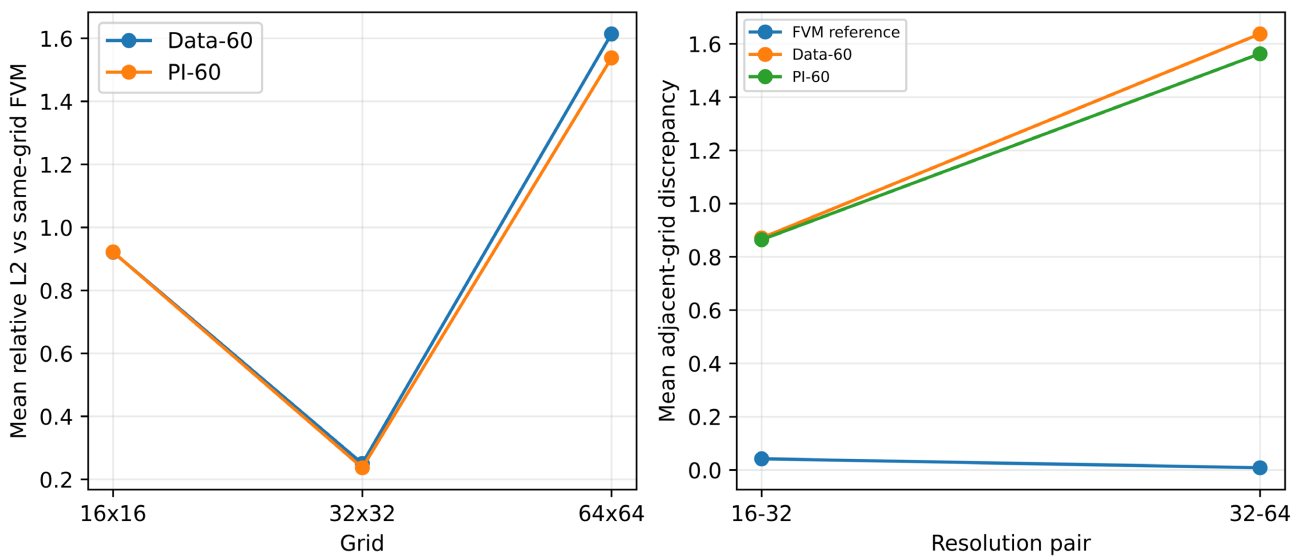


Figure 12. Zero-shot multi-grid falsification test. Left: same-grid pressure error for the unchanged 32-grid-trained surrogates. Right: adjacent-resolution discrepancy after mapping fields to common 64-grid coordinates. The finite-volume reference discrepancy shrinks under refinement, while both learned surrogates lose cross-grid coherence.

5. Discussion

5.1. What Finite-Volume-Aligned Physics Buys in a Low-Data Surrogate

The controlled ablation supports a narrow but useful conclusion: adding a finite-volume-aligned residual improves a compact pressure surrogate when labelled simulations are scarce, without changing the inference architecture. The largest direct effect is on local conservation, as expected, but pressure accuracy also improves. This behaviour is consistent with the broader physics-informed literature,

which views physics terms as inductive biases rather than substitutes for data [2]-[9]. The present formulation has a practical advantage over a separately differentiated continuous PDE residual because the learning penalty is evaluated with the same grid-level flux balance that defines the labels. The multi-grid test, however, establishes an equally important boundary: same-stencil alignment on one grid does not imply convergence when the numerical representation changes. We therefore use finite-volume-aligned, rather than discretization-consistent, to describe the demonstrated property.

The Data-300 reference shows why the result should not be framed as physics replacing simulation data. Five times more labels produce the best interpolation accuracy, but the single-seed Data-300 model does not match PI-60 conservation and performs worse than PI-60 on Independent Combined-Shift (0.528 versus 0.476 mean relative L2). Because Data-300 has only one training seed, that comparison is descriptive; nevertheless, it illustrates a plausible complementarity: representative data improve manifold coverage, while a residual can discourage locally implausible extrapolation. Recent reservoir and CO₂-storage surrogate studies increasingly combine architecture, transfer learning, operators, and physical structure [30]-[47], and the present ablation suggests that evaluating these components separately is essential for attributing gains.

5.2. Shift-Specific Robustness Is More Informative than One OOD Number

The factor-isolation experiments provide the strongest additional novelty beyond the initial manuscript. A single Fixed Compound-OOO (Test-OOO) score would imply that PI-60 provides about a 10% robustness gain. The decomposed tests reveal a more mechanistic picture: physics helps relatively little for contrast alone and boundary location alone, more for short spatial scales and stronger forcing, and most when these changes coincide. The Independent Combined-Shift advantage is not the arithmetic sum of single-factor improvements, and it should not be called causal synergy without a full factorial design. It is nevertheless an amplification: the 14.1% Independent Combined-Shift reduction is substantially larger than any single-factor reduction. This matters because real reservoir updates often change several covariates simultaneously rather than one at a time.

The high-contrast case also warns against using conservation improvement as a surrogate for solution accuracy. A 50.1% reduction in residual accompanies only a 2.7% reduction in pressure relative L2 and no significant peak-error gain. Across the six conditions, pressure-improvement percentage and residual-improvement percentage have a descriptive Spearman $\rho = -0.20$. With only six conditions, this correlation is not inferentially meaningful, but the geometric separation in **Figure 6** is enough to reject a monotonic one-to-one interpretation. A model can become more conservative without becoming proportionally more accurate in the chosen state norm.

5.3. Upper-Tail Protection and Degradation Rate Are More Informative than Mean Error Alone

For optimization or uncertainty workflows, rare large surrogate errors can be more damaging than small shifts in the mean. The present OOD P90 and P95 reductions, 13.0% and 14.7%, exceed the 10.4% mean reduction, and the fixed Data-60 P90 exceedance frequency falls from 10.0% to 1.67%. The strong positive association between baseline difficulty and physics benefit further indicates that the regularizer is not simply shrinking every prediction by a constant amount. The continuous severity sweeps reinforce this interpretation: PI-60 reduces the slope of error growth by 48.3% along the correlation-length path and 27.0% along the well-rate path. Reporting both tail statistics and degradation curves therefore reveals robustness properties that a single OOD mean cannot.

5.4. Residual-Based Monitoring: Useful Warning, Insufficient Certificate

The residual diagnostic reconciles statements that are often conflated in physics-informed ML. A governing-equation residual can be a powerful distribution-shift indicator while remaining a poor casewise estimator of state error. The new inverse-weighting experiment identifies conditioning as part of the explanation: the raw PI-60 residual has OOD $\rho = 0.148$ and worst-20%-error AUROC = 0.648, whereas five-step preconditioned inverse action increases these to 0.382 and 0.790. Exact inverse weighting improves them further, but at solve-like cost. Thus, the residual concept is more informative in an operator-aware norm, yet even the exact energy score is not a perfect predictor of relative L2 error because the two metrics weight error modes differently. Selective simulation makes the deployment consequence explicit: PCG-5 improves risk-coverage ranking over the raw residual, but the gain is moderate, and combining ensemble disagreement with PCG-5 makes performance worse. A sensible policy should therefore validate the monitor itself, use it to rank or abstain rather than to certify correctness, and retain a full-solver fallback.

5.5. Resolution-Transfer Failure Separates Discrete Alignment from Discretization Consistency

The three-grid experiment is deliberately a falsification test, and its negative result materially improves the interpretation of the study. The reference solver exhibits the expected shrinking discrepancy under refinement, whereas both convolutional surrogates are strongly tied to the 32-grid representation. Physics regularization reduces the fine-grid discrepancy by only 4.5% and cannot compensate for the change in physical receptive field, source representation, and learned grid-scale correlations. Consequently, a fixed-grid residual loss should not be advertised as mesh-independent or discretization-consistent. Achieving that stronger property likely requires multi-grid training, conservative transfer operators, resolution-capable architectures such as neural operators or graph models, and ex-

plicit cross-grid objectives; these are natural extensions but are not demonstrated here.

5.6. Relation to Recent Reservoir Surrogate Literature

Recent reservoir and geologic-storage studies increasingly emphasize neural operators, graph representations, spatio-temporal surrogates, well-aware constraints, transfer learning, and differentiable simulators [30]–[47]. The present study is deliberately smaller in physical scope but complementary in experimental focus. Its contribution is to isolate the marginal effect of a finite-volume-aligned conservation term and then stress-test that effect across controlled physical and numerical shifts. The findings suggest four reporting practices for future work: factorized and severity-graded OOD tests rather than only aggregate metrics; upper-tail statistics alongside means; explicit separation of residual as training objective, physical-consistency metric, shift detector, and case-wise failure monitor; and a multi-resolution falsification test before making claims of discretization invariance or consistency.

5.7. Limitations and Threats to Validity

The benchmark excludes saturation transport, relative permeability, capillarity, gravity, compressibility accumulation, faults, anisotropic tensors, non-Cartesian grids, phase behaviour, and Peaceman-style well-index coupling beyond the simplified source representation [1]. The steady elliptic pressure problem is an important reservoir-simulation component but is substantially simpler than a black-oil or compositional model. Reported error values and the 51x speedup must therefore not be transferred directly to commercial simulators or field assets.

The primary training grid contains 1024 nodal degrees of freedom before boundary elimination (900 interior unknowns), and sparse direct solution is already fast. Online acceleration depends on batching, hardware, and implementation overhead. Only CPU timing is reported; offline training, database generation, and residual-monitor costs are excluded from the headline speedup. The OOD generators are designed stress tests rather than probability distributions calibrated to a specific reservoir. Gaussian random fields omit channelized facies, fractures, faults, and conditioning to wells or seismic observations. The multi-grid experiment uses a smooth interpolant of the 32-grid permeability realization, so it tests representation transfer without introducing new subgrid geology; it is therefore a conservative resolution-transfer probe, not a complete multiscale geology study.

The factor-isolation extension is a one-factor-at-a-time design plus the Independent Combined-Shift condition, not a complete factorial experiment; interaction mechanisms therefore cannot be estimated causally. Data-300 has a single seed, and only three matched seeds were used for Data-60/PI-60; the seed-specific OOD effects are heterogeneous, so casewise ensemble inference must not be interpreted as training-seed uncertainty. The physics weight was not adaptively tuned and may not lie on the globally optimal accuracy-conservation Pareto front. Residual monitoring

is evaluated on one designed ID/OOD contrast and two controlled severity paths, so thresholds require validation under realistic geological populations and simulator noise. PCG-5 is a fixed-budget diagnostic choice rather than an optimized preconditioner. Finally, the multi-grid study is zero-shot: the CNN was trained only at 32×32 and no cross-grid consistency loss was used. Its failure therefore falsifies a strong claim for the present method but does not establish that multi-grid physics-informed operators cannot achieve numerical consistency.

6. Conclusions

A finite-volume-aligned residual improved a low-data convolutional reservoir pressure surrogate without changing its inference architecture or requiring additional labelled simulations. On the fixed benchmark, the three-seed PI-60 ensemble reduced mean relative L2 pressure error by 5.3% in distribution and 10.4% on Fixed Compound-OOD (Test-OOD), while reducing discrete mass-balance residual by 31.6% and 26.4%. Batched CPU inference remained about 51 times faster than the reference finite-volume solve.

The deeper analyses materially sharpen the finding. Physics benefit is shift-dependent rather than uniform: the independent factor tests show the largest single-factor gains for short spatial correlation and high rate, and the Independent Combined-Shift reaches 14.1%. Fixed Compound-OOD P90 and P95 errors fall by 13.0% and 14.7%, while continuous severity sweeps show that physics regularization reduces error-growth slopes by 48.3% for shortened correlation length and 27.0% for increased well rate. Raw residual remains a near-perfect broad shift detector ($AUC = 0.997$) but is a weak casewise error score; five-step preconditioned inverse action raises worst-20%-error AUROC from 0.648 to 0.790 and improves risk-coverage performance, whereas seed disagreement and its combination with the residual do not.

The multi-grid result provides the key qualification. The finite-volume reference converges across 16×16 , 32×32 , and 64×64 , but both fixed-grid-trained surrogates lose coherence under zero-shot refinement; PI-60 is only modestly better than Data-60. The demonstrated property is therefore finite-volume alignment on the training discretization, not general discretization consistency. The next validation step should combine transient multiphase physics, realistic well-index treatment and geological ensembles with resolution-capable architectures, conservative grid-transfer operators, multiple discretizations, and independently validated residual monitors. The central deployment lesson is correspondingly restrained: physics regularization can improve robustness and make residuals more useful, but trust requires testing the norm, the shift, and the numerical representation rather than treating a small residual as a universal certificate.

Data and Code Availability

The dataset, trained model weights, primary and extended evaluation tables, stability-weighted residual diagnostics, selective-simulation curves, continuous shift-

severity outputs, multi-grid outputs, figures, and Python source code are supplied with the reproducibility package accompanying this manuscript.

Ethics Statement

The study uses numerical data only and involves no human participants, animals, personal data, or proprietary field data.

Author Contributions

Franck-Hilaire Essiagne: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Visualization, and Writing Original Draft. **Moussa Camara:** Conceptualization, Methodology, Software, and Validation. **Kouassi Louis Kra:** Supervision, Resources, Writing, Review and Editing, Conceptualization, Methodology, Software, and Validation.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

- [1] Peaceman, D.W. (1978) Interpretation of Well-Block Pressures in Numerical Reservoir Simulation (Includes Associated Paper 6988). *Society of Petroleum Engineers Journal*, **18**, 183-194. <https://doi.org/10.2118/6893-pa>
- [2] Raissi, M., Perdikaris, P. and Karniadakis, G.E. (2019) 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. <https://doi.org/10.1016/j.jcp.2018.10.045>
- [3] Zhu, Y., Zabarar, N., Koutsourelakis, P. and Perdikaris, P. (2019) Physics-Constrained Deep Learning for High-Dimensional Surrogate Modeling and Uncertainty Quantification without Labeled Data. *Journal of Computational Physics*, **394**, 56-81. <https://doi.org/10.1016/j.jcp.2019.05.024>
- [4] Lu, L., Jin, P., Pang, G., Zhang, Z. and Karniadakis, G.E. (2021) Learning Nonlinear Operators via DeepONet Based on the Universal Approximation Theorem of Operators. *Nature Machine Intelligence*, **3**, 218-229. <https://doi.org/10.1038/s42256-021-00302-5>
- [5] Wang, S., Yu, X. and Perdikaris, P. (2022) When and Why Pinns Fail to Train: A Neural Tangent Kernel Perspective. *Journal of Computational Physics*, **449**, Article ID: 110768. <https://doi.org/10.1016/j.jcp.2021.110768>
- [6] Karniadakis, G.E., Kevrekidis, I.G., Lu, L., Perdikaris, P., Wang, S. and Yang, L. (2021) Physics-Informed Machine Learning. *Nature Reviews Physics*, **3**, 422-440. <https://doi.org/10.1038/s42254-021-00314-5>
- [7] Cuomo, S., Di Cola, V.S., Giampaolo, F., Rozza, G., Raissi, M. and Piccialli, F. (2022) Scientific Machine Learning through Physics-Informed Neural Networks: Where We Are and What's Next. *Journal of Scientific Computing*, **92**, Article No. 88. <https://doi.org/10.1007/s10915-022-01939-z>
- [8] Li, Z., Zheng, H., Kovachki, N., Jin, D., Chen, H., Liu, B., *et al.* (2024) Physics-Informed Neural Operator for Learning Partial Differential Equations. *ACM/IMS Journal of Data Science*, **1**, 1-27. <https://doi.org/10.1145/3648506>

- [9] Faroughi, S.A., Pawar, N.M., Fernandes, C., Raissi, M., Das, S., Kalantari, N.K., *et al.* (2024) Physics-Guided, Physics-Informed, and Physics-Encoded Neural Networks and Operators in Scientific Computing: Fluid and Solid Mechanics. *Journal of Computing and Information Science in Engineering*, **24**, Article 040802. <https://doi.org/10.1115/1.4064449>
- [10] Zhu, Y. and Zabaras, N. (2018) Bayesian Deep Convolutional Encoder-Decoder Networks for Surrogate Modeling and Uncertainty Quantification. *Journal of Computational Physics*, **366**, 415-447. <https://doi.org/10.1016/j.jcp.2018.04.018>
- [11] Jeong, H., Sun, A.Y., Lee, J. and Min, B. (2018) A Learning-Based Data-Driven Forecast Approach for Predicting Future Reservoir Performance. *Advances in Water Resources*, **118**, 95-109. <https://doi.org/10.1016/j.advwatres.2018.05.015>
- [12] Navrátil, J., King, A., Rios, J., Kollias, G., Torrado, R. and Cudas, A. (2019) Accelerating Physics-Based Simulations Using End-to-End Neural Network Proxies: An Application in Oil Reservoir Modeling. *Frontiers in Big Data*, **2**, Article 33. <https://doi.org/10.3389/fdata.2019.00033>
- [13] Mo, S., Zhu, Y., Zabaras, N., Shi, X. and Wu, J. (2019) Deep Convolutional Encoder-Decoder Networks for Uncertainty Quantification of Dynamic Multiphase Flow in Heterogeneous Media. *Water Resources Research*, **55**, 703-728. <https://doi.org/10.1029/2018wr023528>
- [14] Tang, M., Liu, Y. and Durlofsky, L.J. (2020) A Deep-Learning-Based Surrogate Model for Data Assimilation in Dynamic Subsurface Flow Problems. *Journal of Computational Physics*, **413**, Article ID: 109456. <https://doi.org/10.1016/j.jcp.2020.109456>
- [15] Jin, Z.L., Liu, Y. and Durlofsky, L.J. (2020) Deep-Learning-Based Surrogate Model for Reservoir Simulation with Time-Varying Well Controls. *Journal of Petroleum Science and Engineering*, **192**, Article ID: 107273. <https://doi.org/10.1016/j.petrol.2020.107273>
- [16] Wang, N., Zhang, D., Chang, H. and Li, H. (2020) Deep Learning of Subsurface Flow via Theory-Guided Neural Network. *Journal of Hydrology*, **584**, Article ID: 124700. <https://doi.org/10.1016/j.jhydrol.2020.124700>
- [17] Daolun, L., Luhang, S., Wenshu, Z., Xuliang, L. and Jieqing, T. (2021) Physics-Constrained Deep Learning for Solving Seepage Equation. *Journal of Petroleum Science and Engineering*, **206**, Article ID: 109046. <https://doi.org/10.1016/j.petrol.2021.109046>
- [18] Coutinho, E.J.R., Dall'Aqua, M. and Gildin, E. (2021) Physics-Aware Deep-Learning-Based Proxy Reservoir Simulation Model Equipped with State and Well Output Prediction. *Frontiers in Applied Mathematics and Statistics*, **7**, Article 651178. <https://doi.org/10.3389/fams.2021.651178>
- [19] Almajid, M.M. and Abu-Al-Saud, M.O. (2022) Prediction of Porous Media Fluid Flow Using Physics Informed Neural Networks. *Journal of Petroleum Science and Engineering*, **208**, Article ID: 109205. <https://doi.org/10.1016/j.petrol.2021.109205>
- [20] Sorokin, A.G., Pachalieva, A., O'Malley, D., Hyman, J.M., Hickernell, F.J. and Hengartner, N.W. (2024) Computationally Efficient and Error Aware Surrogate Construction for Numerical Solutions of Subsurface Flow through Porous Media. *Advances in Water Resources*, **193**, Article ID: 104836. <https://doi.org/10.1016/j.advwatres.2024.104836>
- [21] Wen, G., Li, Z., Azizzadenesheli, K., Anandkumar, A. and Benson, S.M. (2022) U-FNO—An Enhanced Fourier Neural Operator-Based Deep-Learning Model for Multiphase Flow. *Advances in Water Resources*, **163**, Article ID: 104180. <https://doi.org/10.1016/j.advwatres.2022.104180>

- [22] Shen, L., Li, D., Zha, W., Li, X. and Liu, X. (2022) Surrogate Modeling for Porous Flow Using Deep Neural Networks. *Journal of Petroleum Science and Engineering*, **213**, Article ID: 110460. <https://doi.org/10.1016/j.petrol.2022.110460>
- [23] Yan, B., Harp, D.R., Chen, B., Hoteit, H. and Pawar, R.J. (2022) A Gradient-Based Deep Neural Network Model for Simulating Multiphase Flow in Porous Media. *Journal of Computational Physics*, **463**, Article ID: 111277. <https://doi.org/10.1016/j.jcp.2022.111277>
- [24] Yan, B., Harp, D.R., Chen, B. and Pawar, R. (2022) A Physics-Constrained Deep Learning Model for Simulating Multiphase Flow in 3D Heterogeneous Porous Media. *Fuel*, **313**, Article ID: 122693. <https://doi.org/10.1016/j.fuel.2021.122693>
- [25] Zhang, K., Wang, X., Ma, X., Wang, J., Yang, Y., Zhang, L., *et al.* (2022) The Prediction of Reservoir Production Based Proxy Model Considering Spatial Data and Vector Data. *Journal of Petroleum Science and Engineering*, **208**, Article ID: 109694. <https://doi.org/10.1016/j.petrol.2021.109694>
- [26] Zhang, Z., Yan, X., Liu, P., Zhang, K., Han, R. and Wang, S. (2023) A Physics-Informed Convolutional Neural Network for the Simulation and Prediction of Two-Phase Darcy Flows in Heterogeneous Porous Media. *Journal of Computational Physics*, **477**, Article ID: 111919. <https://doi.org/10.1016/j.jcp.2023.111919>
- [27] Han, J., Xue, L., Wei, Y., Qi, Y., Wang, J., Liu, Y., *et al.* (2023) Physics-Informed Neural Network-Based Petroleum Reservoir Simulation with Sparse Data Using Domain Decomposition. *Petroleum Science*, **20**, 3450-3460. <https://doi.org/10.1016/j.petsci.2023.10.019>
- [28] Kazemi, M., Takbiri-Borujeni, A., Takbiri, S. and Kazemi, A. (2023) Physics-Informed Data-Driven Model for Fluid Flow in Porous Media. *Computers & Fluids*, **264**, Article ID: 105960. <https://doi.org/10.1016/j.compfluid.2023.105960>
- [29] Kashefi, A. and Mukerji, T. (2023) Prediction of Fluid Flow in Porous Media by Sparse Observations and Physics-Informed PointNet. *Neural Networks*, **167**, 80-91. <https://doi.org/10.1016/j.neunet.2023.08.006>
- [30] Wen, G., Li, Z., Long, Q., Azizzadenesheli, K., Anandkumar, A. and Benson, S.M. (2023) Real-Time High-Resolution CO₂ Geological Storage Prediction Using Nested Fourier Neural Operators. *Energy & Environmental Science*, **16**, 1732-1741. <https://doi.org/10.1039/d2ee04204e>
- [31] Bi, J., Li, J., Wu, K., Chen, Z., Chen, S., Jiang, L., *et al.* (2023) A Physics-Informed Spatial-Temporal Neural Network for Reservoir Simulation and Uncertainty Quantification. *SPE Journal*, **29**, 2026-2043. <https://doi.org/10.2118/218386-pa>
- [32] Liu, A., Li, J., Bi, J., Chen, Z., Wang, Y., Lu, C., *et al.* (2024) A Novel Reservoir Simulation Model Based on Physics Informed Neural Networks. *Physics of Fluids*, **36**, Article ID: 116617. <https://doi.org/10.1063/5.0239376>
- [33] Zhou, L., Sun, H., Fan, D., Zhang, L., Imani, G., Fu, S., *et al.* (2024) Flow Prediction of Heterogeneous Nanoporous Media Based on Physical Information Neural Network. *Gas Science and Engineering*, **125**, Article ID: 205307. <https://doi.org/10.1016/j.jgsce.2024.205307>
- [34] Zhang, J., Braga-Neto, U. and Gildin, E. (2024) Physics-Informed Neural Networks for Multiphase Flow in Porous Media Considering Dual Shocks and Interphase Solubility. *Energy & Fuels*, **38**, 17781-17795. <https://doi.org/10.1021/acs.energyfuels.4c02888>
- [35] Han, Y., Hamon, F.P., Jiang, S. and Durlowsky, L.J. (2024) Surrogate Model for Geological CO₂ Storage and Its Use in Hierarchical MCMC History Matching. *Advances in Water Resources*, **187**, Article ID: 104678.

- <https://doi.org/10.1016/j.advwatres.2024.104678>
- [36] Kadeethum, T., Verzi, S.J. and Yoon, H. (2024) An Improved Neural Operator Framework for Large-Scale CO₂ Storage Operations. *Geoenergy Science and Engineering*, **240**, Article ID: 213007. <https://doi.org/10.1016/j.geoen.2024.213007>
 - [37] Tang, H. and Durlofsky, L.J. (2024) Graph Network Surrogate Model for Subsurface Flow Optimization. *Journal of Computational Physics*, **512**, Article ID: 113132. <https://doi.org/10.1016/j.jcp.2024.113132>
 - [38] Zhao, M., Wang, Y., Gerritsma, M. and Hajibeygi, H. (2024) A Physics-Constraint Neural Network for CO₂ Storage in Deep Saline Aquifers during Injection and Post-Injection Periods. *Advances in Water Resources*, **193**, Article ID: 104837. <https://doi.org/10.1016/j.advwatres.2024.104837>
 - [39] Badawi, D. and Gildin, E. (2025) Neural Operator-Based Proxy for Reservoir Simulations Considering Varying Well Settings, Locations, and Permeability Fields. *Computers & Geosciences*, **196**, Article ID: 105826. <https://doi.org/10.1016/j.cageo.2024.105826>
 - [40] Cui, J., Sun, W., Jeong, H., Liu, J. and Zhou, W. (2025) Efficient Deep-Learning-Based Surrogate Model for Reservoir Production Optimization Using Transfer Learning and Multi-Fidelity Data. *Petroleum Science*, **22**, 1736-1756. <https://doi.org/10.1016/j.petsci.2025.02.014>
 - [41] Lin, R., Song, T. and Li, J. (2025) Feature Attention-Based Deep Neural Operator for Solving Seepage Flow Equations in Porous Media Reservoir Simulation. *Physics of Fluids*, **37**, Article ID: 067152. <https://doi.org/10.1063/5.0272866>
 - [42] Lin, J., Yan, X., Wang, E., Zhang, Q., Zhang, K., Liu, P., *et al.* (2025) A Layer-Specific Constraint-Based Enriched Physics-Informed Neural Network for Solving Two-Phase Flow Problems in Heterogeneous Porous Media. *Petroleum Science*, **22**, 4714-4735. <https://doi.org/10.1016/j.petsci.2025.07.008>
 - [43] Feng, Z., Yan, B., Shen, X., Zhang, F., Tariq, Z., Ouyang, W., *et al.* (2025) A Hybrid CNN-Transformer Surrogate Model for the Multi-Objective Robust Optimization of Geological Carbon Sequestration. *Advances in Water Resources*, **196**, Article ID: 104897. <https://doi.org/10.1016/j.advwatres.2025.104897>
 - [44] Chen, B., Yan, B., Aslam, B., Kang, Q., Harp, D. and Pawar, R. (2025) Deep Learning Accelerated Inverse Modeling and Forecasting for Large-Scale Geologic CO₂ Sequestration. *International Journal of Greenhouse Gas Control*, **144**, Article ID: 104383. <https://doi.org/10.1016/j.ijggc.2025.104383>
 - [45] Chandra, A., Koch, M., Pawar, S., Panda, A., Azizzadenesheli, K., Snippe, J., *et al.* (2025) Accelerating Porous Media Flow Simulations with Fourier Neural Operators: An Application to Geologic Storage of CO₂. *Advanced Theory and Simulations*, **9**, e00747. <https://doi.org/10.1002/adts.202500747>
 - [46] Walter, L., Kong, Q., Hanson-Hedgecock, S. and Vilarrasa, V. (2026) WellPINN: Accurate Well Representation for Transient Fluid Pressure Diffusion in Subsurface Reservoirs with Physics-Informed Neural Networks. *Water Resources Research*, **62**, Article ID: 3374164. <https://doi.org/10.1029/2025wr041404>
 - [47] Ur Rashid, H., Pachaliev, A. and O'Malley, D. (2026) Differentiable Multiphase Flow Model for Physics-Informed Machine Learning in Reservoir Pressure Management. *Scientific Reports*, **16**, Article No. 10345. <https://doi.org/10.1038/s41598-026-37063-3>

Appendix: Supplementary Seed and Training Diagnostics

Supplementary Table S1 reports the matched low-data effect for each training seed and the validation-selected checkpoint epochs. These values expose training-run heterogeneity; with only three seeds, they are descriptive and are not used for population-level inference over initialization. Supplementary Figure S1 then reports the full validation relative-L2 histories.

Table S1. Seed-specific Data-60 versus PI-60 relative-L2 effects and retained validation checkpoints.

Seed	Data Best Epoch	PI Best Epoch	ID Data	ID PI	ID Reduction	OOD Data	OOD PI	OOD Reduction
11	68	83	0.2638	0.2528	4.2%	0.5110	0.4998	2.2%
29	60	68	0.2968	0.2824	4.8%	0.5302	0.5219	1.6%
47	65	76	0.2811	0.2594	7.7%	0.8643	0.5524	36.1%

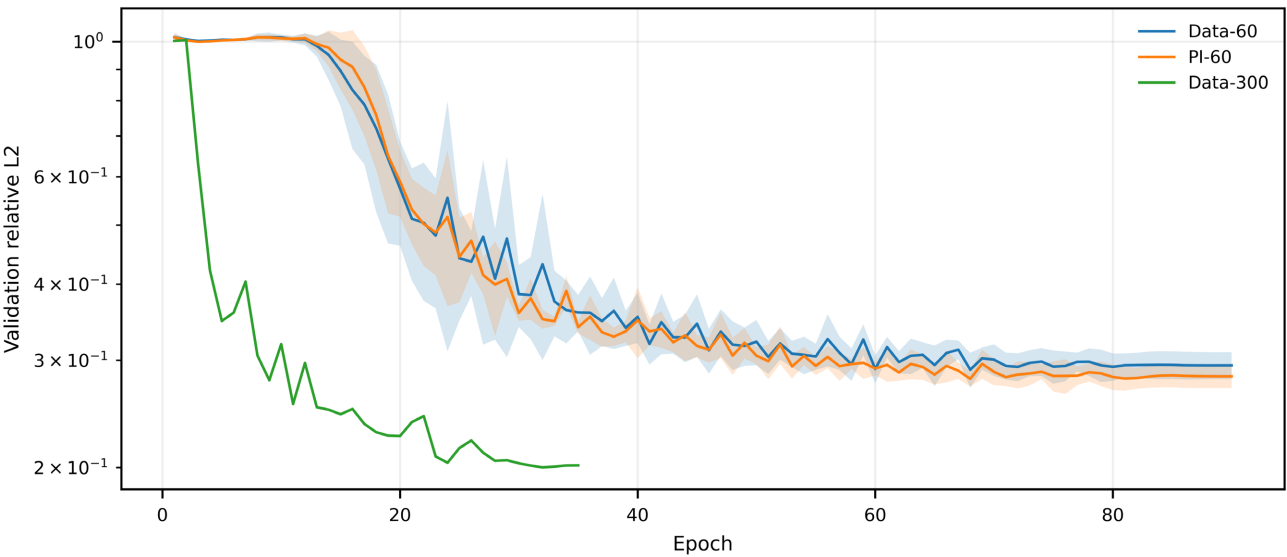


Figure S1. Validation relative L2 learning curves for Data-60, PI-60, and the Data-300 reference. Curves for low-data models show the mean across three seeds with one-standard-deviation bands.