Scientific Research Publishing

Open Journal of Geology, 2026, 16(11), 848-883
https://www.scirp.org/journal/ojg
ISSN Online: 2161-7589
ISSN Print: 2161-7570

Hydraulic Topology as a Predictive Law for Zero-Shot Generalization in Learned Reservoir Dynamics

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

Laboratoire des Sciences Géographiques du Génie Civil et des Géosciences, École Supérieure de Chimie, du Pétrole et de l'Énergie (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) Hydraulic Topology as a Predictive Law for Zero-Shot Generalization in Learned Reservoir Dynamics. *Open Journal of Geology*, **16**, 848-883.
https://doi.org/10.4236/ojg.2026.1611040

Received: October 17, 2026
Accepted: November 22, 2026
Published: November 25, 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

Learned reservoir simulators can be accurate on familiar geological ensembles, but their reliability on genuinely unseen hydraulic structures remains difficult to assess without target simulation labels. We evaluated a normalized, five-port Kron-reduced rock-transmissibility operator as a pre-simulation predictor of forecasting error in 600 controlled two-dimensional reservoirs spanning four discovery and two external geological families. Across reduced-basis regressors, hydraulic-response distance was more strongly associated with multi-time pressure-saturation prediction error than geological distance (Spearman $\rho = 0.546$ - 0.594 versus 0.350 - 0.395). Six-family cross-validation of the error predictor achieved $R^2 = 0.373$ - 0.470, but that analysis used error labels from each external family when it served as a training fold; a separate, genuinely blind calibration trained only on discovery-family labels and achieved external $R^2 = 0.168$ - 0.281. On 200 independent external reservoirs, three ResCNN runs gave a reservoir-aggregated $\rho = 0.454$ (95% reservoir-bootstrap CI 0.339 - 0.558), and three FNO runs gave $\rho = 0.486$ (0.375 - 0.582). Across 40 paired bottleneck interventions, maximal barrier severity increased mean RB-Ridge error by 72.1% (95% paired-bootstrap CI 47.0 - 103.3%) and RB-MLP error by 116.2% (96.7% - 137.7%). Nonlinear Corey-flow robustness used the original rock-transmissibility operator computed before dynamic saturation simulation, avoiding target-trajectory leakage. Under matched standardized-Euclidean distances, the Kron fingerprint retained stronger blind external calibration than the 46-component composite, although the composite outperformed it in one discovery cross-validation comparison. The susceptibility correction was not consistently beneficial. The results support a bounded, physically interpretable hydraulic-response domain-of-validity indicator rather than a universal topology law.

Keywords

Hydraulic-Response Distance, Out-of-Distribution Forecasting, Kron Reduction, Petroleum Reservoir Simulation, Zero-Shot Validation, Clustered Bootstrap

1. Introduction

Fast learned surrogates are now credible components of subsurface simulation workflows. Physics-informed networks, convolutional reduced-order models, graph neural networks and neural operators have demonstrated accurate prediction of pressure, saturation and related quantities for classes of porous-media problems [1]-[12]. U-FNO, CCSNet and recent neural-operator reservoir proxies show that learned mappings can span substantial parameter variability and can reduce inference cost by orders of magnitude [5]-[7]. Graph models provide a complementary route for faulted and fractured media because the discretization itself can be treated as a graph rather than as a fixed image [8] [9] [13]. These advances change the practical question. The limiting issue is no longer whether a surrogate can be accurate on a test set; it is whether one can know, before trusting a forecast, that an unseen reservoir is sufficiently similar to the training domain in the variables that actually control flow.

Most out-of-distribution (OOD) assessments are still defined in parameter space: permeability statistics, well controls, geometry, resolution or latent embeddings. Recent mathematical work on neural operators makes clear that OOD risk is a distinct object from interpolation error and requires its own analysis [12] [14]. In subsurface flow this problem is especially acute because small geological edits can have disproportionately large dynamic consequences when they create or remove high-conductivity bridges, seal faults, reconnect fractures or change source-to-sink bottlenecks. Conversely, two visually dissimilar permeability maps may generate nearly indistinguishable interwell pressure responses if their effective hydraulic connection structure is equivalent.

Physical-network theory provides the missing language. An oriented finite-volume graph has an incidence matrix B; edge conductances define a weighted Laplacian $L = BWB^{\mathrm{T}}$; Schur-complement or Kron reduction removes internal degrees of freedom while preserving the terminal input-output relation; and effective resistance and cut structure quantify complementary aspects of communication and bottlenecks [15]-[18]. Flow diagnostics provide a reservoir-engineering analogue by describing time of flight, well allocation and dynamic heterogeneity [19]-[22]. The uploaded deep-research synthesis motivating this study made an important distinction: topology alone is not a complete law of dynamics. Topological constraints determine the admissible grammar of flow, while constitutive physics, edge weights and boundary forcing determine the realized sentence. The strongest hypothesis is therefore not that graph adjacency predicts surrogate er-

ror, but that a weighted hydraulic response operator defines a physically meaningful domain of validity for a frozen learned model.

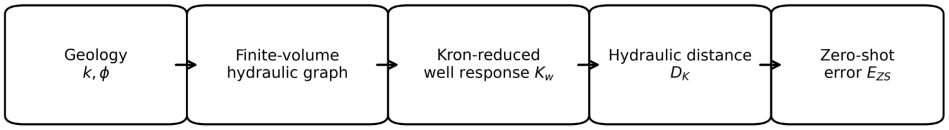
Notation at first use: B denotes the oriented cell-face incidence matrix, W the diagonal face-conductance matrix and L the corresponding hydraulic graph Laplacian; B^{T} denotes its transpose. These objects are defined formally in Section 2.1. "Kron reduction" denotes elimination of unobserved interior cell pressures while preserving the terminal pressure-flow relation.

This paper tests that hypothesis rather than assuming it. We construct a deliberately controlled two-dimensional discovery campaign, because a claim of a predictive law should first survive falsification under interpretable conditions before being extrapolated to field-scale three-dimensional black-oil models. Six geology families are generated with overlapping marginal permeability statistics but different connectivity mechanisms. Four are used for family-held-out discovery; two are never used to define the law and serve as blind external families. We compare three reduced-order surrogate classes, then stress the relation with convolutional and Fourier deep models, nonlinear mobility, controlled bottleneck interventions and a doubled spatial resolution.

The study makes seven contributions, subject to the methodological qualifications detailed below. First, it formalizes weighted five-port hydraulic-response equivalence through the normalized Kron-reduced operator K_w. Second, it compares hydraulic and geological distance across multiple surrogate classes. Third, matched bottleneck interventions provide dose-response evidence, quantified with base-reservoir clustering. Fourth, deep runs, nonlinear simulations and an independently retrained 32×32 benchmark test robustness; none is a 3-D field test. Fifth, D_K ranks high-error forecasts for selective simulation within the investigated cases. Sixth, a breakpoint near $D_K \approx 2.1$ - 2.3 is exploratory, not an independently confirmed transition. Seventh, the proposed susceptibility correction does not improve calibrated error prediction. We distinguish surrogate-family holdouts from the calibration-model family holdouts and report blind external validation separately.

The result is deliberately bounded. We do not claim that D_K is a universal scalar law for all reservoir physics. We show that, within a controlled family of finite-volume source-to-sink flow problems, the distance between well-conditioned response operators is a reproducible organizing variable for zero-shot surrogate failure. That is the level of claim appropriate for a fundamental Advances in Water Resources study: a mechanistic, falsifiable relationship that can subsequently be tested under richer multiphase, three-dimensional and field-scale conditions. Figure 1 summarizes the proposed physical error-prediction workflow.

Geological fields are converted to transmissibility-weighted flow networks and reduced to the response seen at the five wells. Zero-shot error is tested against distance from the hydraulic-response manifold represented by training data, rather than against image similarity alone.



Candidate law: $\log E_{ZS} = a_M + \Gamma_M \log(1 + D_K)$

Test: does a physics-derived boundary-response distance outperform geological similarity, flow diagnostics and richer composite topology descriptors on unseen geology families?

Controlled Phase-I study: 600 base realizations + 200 intervention cases + 180 nonlinear robustness cases; strict family-held-out and blind external testing.

Figure 1. Study concept.

2. Theory and Hypotheses

2.1. Finite-Volume Flow as a Weighted Hydraulic Network

For phase α in an isothermal porous medium, mass conservation and Darcy flow are written in standard form as

$$\frac{\partial(\phi\rho_\alpha S_\alpha)}{\partial t} + \nabla\cdot(\rho_\alpha \boldsymbol{u}_\alpha) = q_\alpha, \quad \boldsymbol{u}_\alpha = -\boldsymbol{K}\frac{k_{r\alpha}(S_\alpha)}{\mu_\alpha}(\nabla p_\alpha - \rho_\alpha g \nabla z). \tag{1}$$

here α identifies the fluid phase (water, w, or oil, o); φ is pore-volume fraction (porosity); ρ_α is phase density; S_α is phase saturation; t is time; $\boldsymbol{u}_\alpha$ is the Darcy (superficial) velocity; and q_α is the phase mass-source term per bulk volume (positive for injection). $\boldsymbol{K}$ is the absolute permeability tensor, $k_{r\alpha}$ is phase relative permeability, μ_α is dynamic phase viscosity, p_α is phase pressure, g is gravitational acceleration and z is elevation. The phase saturations satisfy $S_w + S_o = 1$. This general mass-conservation statement is specialised to the incompressible, gravity-free conditions used in the primary numerical campaign.

A cell-centred finite-volume discretization converts Equation (1) into accumulation plus signed face fluxes. For neighbouring cells i and j, the rock contribution is a transmissibility T_{ij}, obtained here from the harmonic mean of cell permeabilities. In the primary controlled experiment mobility is constant, so the pressure system is linear; in the nonlinear robustness experiment, T_{ij} is multiplied by total mobility $\lambda_t(S)$. The discrete graph is therefore physical rather than purely combinatorial: cells are nodes, communicating faces are edges and edge weights are hydraulic conductances.

$$L_H = BWB^{\mathrm{T}}, \quad W = \mathrm{diag}(w_e), \quad L_H p = q. \tag{2}$$

here L_H is the rock-transmissibility-weighted discrete hydraulic Laplacian; B is the signed cell-face incidence matrix (cells × internal faces); W is the diagonal matrix of non-negative edge conductances w_e; the superscript T denotes transpose; p is the vector of cell pressures; and q is the vector of net cell source/sink flow rates under the chosen sign convention. diag(·) constructs a diagonal matrix. In the primary calculation, w_e is the harmonic face transmissibility T_{ij} (unit-mobility rock graph); the nonlinear robustness score uses the same pre-trajectory

rock-only weights. A reference pressure or imposed well pressures fixes the Laplacian gauge.

Equation (2) is simultaneously a graph Laplacian and the gauge-dependent discrete Darcy operator. This identity matters because it provides exact conservation structure, a response geometry and a family of derived quantities without asking a learned model to infer them statistically [15] [16].

2.2. Kron-Reduced Well-Response Equivalence

Partition pressure degrees of freedom into terminal/well nodes b and interior cells i. The weighted Laplacian is block-partitioned as

$$L = \begin{bmatrix} L_{bb} & L_{bi} \\ L_{ib} & L_{ii} \end{bmatrix}. \tag{3}$$

Subscripts b and i denote the selected five terminal/well cells and the remaining interior cells, respectively. The four blocks L_{bb}, L_{bi}, L_{ib} and L_{ii} therefore describe terminal-terminal, terminal-interior, interior-terminal and interior-interior couplings in the partitioned Laplacian L.

Eliminating interior nodes with no external injection gives the Schur complement

$$K_w = L_{bb} - L_{bi} L_{ii}^{-1} L_{ib}. \tag{4}$$

here K_w is the Kron-reduced well-response operator (a 5 × 5 symmetric matrix in this study); the interior block is inverted only after imposing the terminal-pressure conditions. The matrix maps specified terminal pressure changes to the corresponding terminal flow changes under the linear rock-only pressure model. Its entries depend on permeability, grid connectivity and the fixed well-port configuration; it is not an unweighted topological invariant.

K_w is a Dirichlet-to-Neumann-type operator: prescribed well pressures map to terminal fluxes. Internally different reservoirs can therefore be hydraulically equivalent at the ports if their reduced operators are close, even when their permeability images, fault layouts or channel geometry are not. This motivates a response-equivalence hypothesis: surrogate transfer should track distance in K_w more closely than distance in the raw geological image. The hypothesis is stronger than a visual-similarity argument and weaker and more defensible than claiming that full internal topology is identifiable from well response.

2.3. Response Distance and Competing Hydraulic Descriptors

For each realization, K_w is normalized by its trace to remove trivial conductance scale. The upper triangular entries of the five-port operator form a 15-component response vector. Training-only standardization is applied, and D_K for a test realization r is the mean Euclidean distance to its five nearest training vectors:

$$D_K(r) = \frac{1}{5} \sum_{s \in \mathcal{N}_5(r)} \left\| \mathbf{z}_K(r) - \mathbf{z}_K(s) \right\|_2. \tag{5}$$

here $D_K(r)$ is the dimensionless Kron-response distance of target reservoir r

from the surrogate-training ensemble; s indexes a training reservoir; the five-neighbour set contains the five nearest training cases; $\boldsymbol{z}_K(r)$ is the 15-entry upper-triangular vector of the trace-normalised five-port K_w after training-only feature standardisation; and the double vertical bars denote the Euclidean norm. Only the current training fold supplies the standardisation means, variances and neighbour set. In particular D_K is a scalar distance, whereas K_w is a matrix.

A richer hydraulic descriptor D_H was also constructed from K_w entries, pairwise effective resistances, non-zero reduced-operator eigenvalues, directional cut-capacity proxies and flow diagnostics. D_H used shrinkage-Mahalanobis distance to the same training manifold. This composite score was not privileged in the analysis; rather, one purpose of the study was to test whether the extra descriptors add value beyond Equation (5).

$$R_{\text{eff}}(a,b) = (\boldsymbol{e}_a - \boldsymbol{e}_b)^{\mathsf{T}} \operatorname{pinv}(K_w)(\boldsymbol{e}_a - \boldsymbol{e}_b). \tag{6}$$

here $R_{\text{eff}}(a,b)$ is effective hydraulic resistance between terminal ports a and b; e_a and e_b are the corresponding five-dimensional unit coordinate vectors; pinv(·) denotes the Moore-Penrose pseudoinverse; and superscript T denotes transpose. The relation applies to the passive connected linear network represented at the terminals; it is used as a supplementary descriptor, not as an exact representation of nonlinear two-phase transport.

For comparison, a flow-diagnostic distance D_{FD} used time-of-flight quantiles, producer time of flight, well-allocation fractions, flux entropy and a Gini/Lorenz-type throughflow heterogeneity measure [19]-[22]. A geological distance D_{geo} combined low-dimensional log-permeability representations with marginal and gradient statistics. These predictors create a deliberately difficult benchmark for D_K: the proposed law must outperform both static geology and established flow-response summaries.

2.4. Stronger Susceptibility Hypothesis and Its Falsification Target

The deep-research synthesis suggested a stronger perturbation argument. If $L' = L + \Delta L$, the resolvent identity yields

$$(L+\Delta L)^{-1} - L^{-1} = -(L+\Delta L)^{-1}\,\Delta L\,L^{-1}. \tag{7}$$

here L is a gauge-fixed, invertible pressure operator on the relevant pressure subspace; $L' = L + \Delta L$ is its perturbed counterpart; ΔL is the operator change caused by modified hydraulic connections; and superscript −1 indicates the inverse on that subspace. The identity motivates pressure sensitivity but is not a theorem bounding prediction error of any trained neural surrogate.

On the zero-mean subspace of a connected symmetric graph, $\|L^{+}\|_2 = 1/\lambda_2$. Equal operator changes can therefore have unequal physical effects near weakly connected states. We encoded this idea in a normalized susceptibility χ_H and a dimensionless extrapolation number $\Pi_H = \Theta_H \chi_H$, where Θ_H was the composite hydraulic distance normalized by the 95th percentile of within-training dis-

tance. Importantly, Π_H was prespecified as a stronger candidate, not treated as a guaranteed discovery. A top-tier test requires the possibility that the simpler K_w geometry wins.

$$\Pi_H = \Theta_H \chi_H, \quad \Theta_H = \frac{D_H}{Q_{0.95}\left(D_H^{\mathrm{ID}}\right)}. \tag{8}$$

here Π_H is the proposed susceptibility-corrected extrapolation number; Θ_H is the composite-fingerprint distance D_H divided by Q0.95 (DHID), the 95th percentile of within-training hydraulic distances; χ_H is the dimensionless, training-normalised susceptibility estimate; and the superscript ID means in-distribution. The numerical experiment does not find consistent predictive improvement from Π_H over the simpler D_K.

2.5. Operational Law and Generalization Elasticity

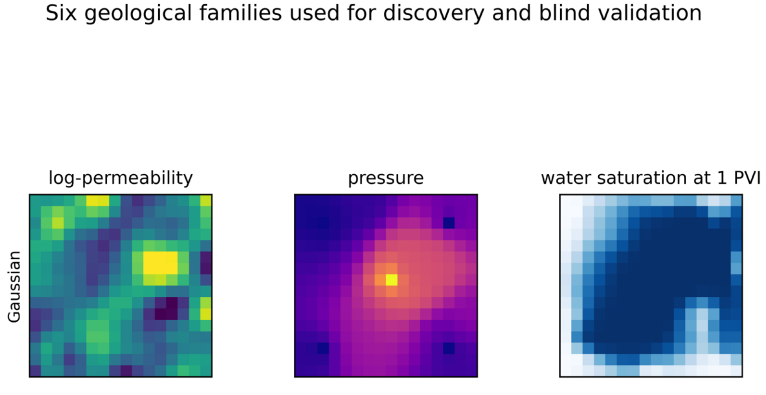
The principal empirical relationship is expressed as a bounded-domain scaling rather than a universal theorem:

$$\log E_{\mathrm{ZS}} = a_M + \Gamma_M \log\left(1 + D_K\right) + \varepsilon_M. \tag{9}$$

here E_{ZS} is the non-negative out-of-family surrogate prediction error evaluated at a prespecified forecast horizon; a_M is the model-class-specific regression intercept; Γ_M is the hydraulic-generalisation elasticity (slope) for surrogate class M; D_K is the five-neighbour Kron-response distance; and ε_M is an unexplained statistical residual, not a numerical stabiliser or measurement error estimate. This regression is an empirical model conditional on the tested simulation regime, not a universal physical law.

Γ_M describes degradation as hydraulic extrapolation increases and is conceptually distinct from average test error. A model may be accurate in-distribution yet have a large Γ_M and fail rapidly once the well-response geometry leaves training support. A second operational task is failure detection: D_K should rank cases likely to exceed an error tolerance even without access to target labels.

Figure 2 shows representative permeability, pressure and water-saturation fields for the six geological families. The log-permeability fields were clipped and recentered so that mean conductance alone did not identify family membership.



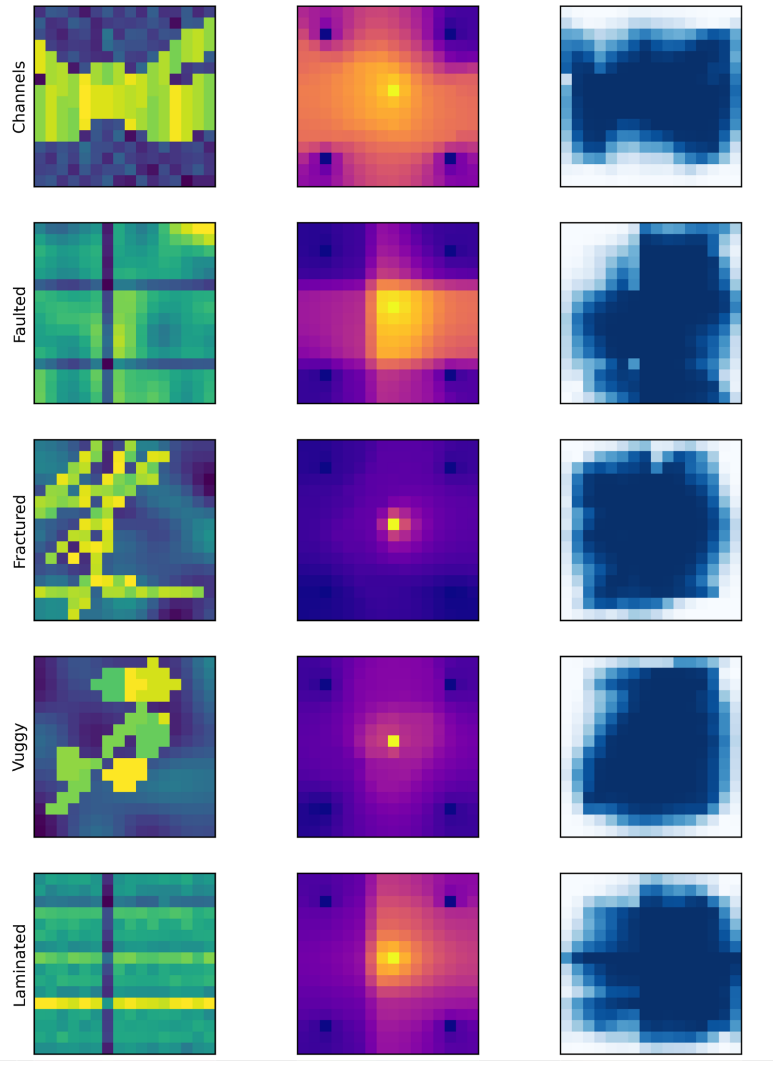


Figure 2. Representative realizations from the six geological families used in the controlled study.

3. Materials and Methods

3.1. Controlled Phase-I Design and Reproducibility Principle

The primary experiment was intentionally small enough to permit complete control and repeated falsification. The grid contained 16 × 16 finite-volume cells with constant porosity $\varphi = 0.20$. A central injector and four near-corner producers formed five terminal ports. Injector pressure was fixed at unity and producer pressures at zero in dimensionless units. Rock transmissibilities used harmonic face averaging on unit geometry. Five report times were stored at 0.10, 0.25, 0.50, 0.75 and 1.00 injected pore volumes (PVI). This normalization removes arbitrary clock-time scaling and makes transport responses comparable across conductance levels.

The symbols φ, p and PVI respectively denote porosity, dimensionless pressure and injected pore-volume units in this numerical setup. One PVI corresponds to cumulative injected volume divided by total pore volume, not to one day. The five well

cells are the prescribed terminal ports used to form the reduced response matrix.

The primary flow campaign used a unit-mobility conservative displacement. Pressure was solved exactly for each realization and saturation was then advanced by an explicit conservative upwind transport update with $f_w = S$. This is not presented as a full petroleum-reservoir model; it is a theorem-oriented discovery system in which connectivity effects can be isolated. The nonlinear robustness campaign later replaced this with Corey mobilities $k_{rw} = S^2$ and $k_{ro} = (1-S)^2$, $\mu_w = 0.5$ and $\mu_o = 2.0$, with pressure recomputed at every transport step. Gravity and capillary pressure were intentionally excluded from Phase I so that a failure of the proposed law could not be attributed to confounding physics.

In this primary benchmark, $f_w = S$ denotes the linear water fractional-flow closure, with S the water-saturation state (equivalent to S_w in the two-phase notation). This simplified transport law should not be confused with Corey fractional flow in the nonlinear robustness campaign.

3.2. Geological Ensembles

Six families were generated with distinct connectivity mechanisms. Gaussian fields varied variance, correlation scale and weak trends. Channelized fields contained one to three sinuous high-conductivity channels with optional interruptions. Faulted fields superposed one to three low-conductivity barriers on correlated backgrounds, with occasional transmissive gaps. Fractured fields contained several high-conductivity line features over a lower-permeability matrix. Vuggy fields combined high-conductivity inclusions with sparse connecting fractures. Laminated fields used horizontal, vertical or diagonal layering with optional cross-barriers. Log permeability was clipped and recentered so that overall mean conductance did not trivially identify family membership.

One hundred realizations were generated per family (600 total) with immutable seeds. Gaussian, channelized, faulted and fractured systems formed the 400-realization discovery domain. Vuggy and laminated systems were withheld as blind external families. The four discovery families were also evaluated in leave-one-family-out mode, producing four strict topology-family shifts. All time frames from a realization remained in the same split. Table 1 provides the corresponding numerical summary.

Table 1. Geological ensemble and split roles.

Family	Role	N	Primary connectivity variation
Gaussian	Discovery	100	correlation scale, variance, weak trend
Channelized	Discovery	100	sinuosity, width, branching, interruptions
Faulted	Discovery	100	barrier number, orientation, transmissive gaps
Fractured	Discovery	100	fracture count, length, orientation, intersections
Vuggy	Blind external	100	vug density and connectivity, with sparse connecting fractures
Laminated	Blind external	100	layer orientation + barrier contrast

3.3. Hydraulic, Flow-Diagnostic and Geological Fingerprints

For the nonlinear campaign, D_K was computed before transport from rock-only harmonic transmissibilities: form $L_{\text{rock}} = B\,\text{diag}\left(T_{ij}\right)B^{\text{T}}$ Kron—reduce it to K_w normalize K_w by its trace, and standardize using discovery—family rock operators. The calculation uses permeability, grid and prescribed well ports only; it does not use simulated saturation, state-dependent mobility, pressure recomputed during Corey displacement, or future well responses. Thus, the reported nonlinear D_K is a presimulation rock-response fingerprint; a mobility-weighted future-state distance was not tested.

For every realization, the full cell Laplacian, the five-port K_w operator, pairwise effective resistances, nonzero K_w eigenvalues, directional cut-capacity proxies and flow diagnostics were computed from the reference solution. Time of flight was calculated on the directed positive-flux graph using edge travel costs proportional to pore volume divided by flux. Producer allocation was obtained from the terminal flux vector. Flux entropy quantified how diffusely the total flux was distributed across edges, while a Gini-like statistic summarized concentration of cell throughflow.

D_K used only normalized K_w entries. D_H used a 46-component composite hydraulic fingerprint and Ledoit-Wolf shrinkage covariance in a training-standardized Mahalanobis metric. D_{FD} used standardized Euclidean distance over the flow-diagnostic vector. D_{geo} used a 20-component PCA of log permeability augmented by mean, variance, quantiles, directional gradient statistics and lag-one spatial correlations. Every scaler, PCA and covariance estimate was fitted only on the corresponding training set.

Descriptor definitions: D_H is the 46-entry composite hydraulic-fingerprint distance (originally shrinkage-Mahalanobis); D_{FD} is the flow-diagnostic distance; D_{geo} is the geology-derived distance; and D_K is the Euclidean five-neighbour distance from the 15-entry Kron vector. The matched-estimator sensitivity analysis applies the same Euclidean five-neighbour rule to both 15-entry and 46-entry features.

3.4. Learned Surrogate Classes

Reproducible reduced-basis training specification (verified against archived source scripts): The discovery/family-fold model takes the first 40 principal components of log permeability (or the lesser available dimension) plus t and t^2, and predicts the first 80 pressure-saturation state principal components (or the lesser available dimension); input and output PCA and StandardScaler are fitted only on training cases. The RB-Ridge regularization is $\alpha = 1.0$. RB-MLP uses ReLU hidden layers (96, 96), L2 parameter $\alpha = 10^{-4}$, Adam, initial learning rate 10^{-3}, a maximum of 120 training iterations, early stopping with a random 12% validation fraction drawn from training TIME-SAMPLE ROWS, and patience n_iter_no_change = 8 (original run random_state = 1016 in discovery folds, 123 externally). RB-ExtraTrees uses 100 trees, max_features = 0.8, min_samples_leaf = 2, and ran-

dom_state = 1016 (discovery) or 123 (external). These hyperparameters were fixed in the archived scripts, not optimized using external cases. Caution: RB-MLP internal early-stopping validation is row-based, not grouped by reservoir; its validation rows can share a reservoir with fitting rows. This does not leak held-out family outcomes but can favour optimistic within-discovery stopping and should be improved in future training.

Training notation: PCA denotes principal-component analysis; RB denotes reduced-basis regression; α denotes the regularisation strength in the Ridge/MLP software specification (not the phase index α in Equation (1)); and t^2 is squared report time. The literal software argument names in monospace, such as n_iter_no_change and random_state, are code identifiers not manuscript variables or typographic subscripts.

The first suite was deliberately architecture-diverse but computationally lightweight. Log-permeability fields were compressed by training-only PCA, time was appended as an input and pressure-plus-saturation states were compressed by output PCA. Three regressors were then fitted in the reduced space: ridge regression (RB-Ridge), a two-hidden-layer multilayer perceptron with 96 neurons per layer (RB-MLP), and an ensemble of 100 extremely randomized regression trees (RB-ExtraTrees). Decoding restored full 16 × 16 pressure and saturation fields before error evaluation. These models are not proposed as state-of-the-art surrogates; they provide independent functional forms for testing whether the same physical distance predicts their failures.

Deep architectures (verified against train_deep.py) used identical discovery data and 60 reservoir-level discovery validation cases sampled with NumPy seed 20261016; every time sample from a selected validation reservoir stayed in validation. ResCNN had width 24 and five residual blocks; FNO had width 16, six Fourier modes and four spectral layers. Each network received two channels (standardized discovery log permeability and report time) and predicted pressure and saturation through two sigmoid outputs. Training used Adam (learning rate 2×10^{-3}; weight decay 10^{-5}), batch size 64, mean-squared-error loss, at most 18 epochs, validation every epoch, patience five epochs, and checkpoint retention at the lowest discovery-validation MSE. Results were stored under run labels 101, 202 and 303. Reproducibility qualification: the archived program seeded stochastic generators AFTER model instantiation, so the labels do not guarantee deterministic weight initialization. They represent three separately executed model fits; this limitation is now explicit, and the supplementary rerun protocol seeds BEFORE constructing the model. The archived results were not replaced by unreported retraining.

ResCNN means residual convolutional neural network; FNO means Fourier neural operator; MSE means mean-squared prediction error; Adam is the optimisation algorithm. The numerical run labels 101, 202 and 303 index saved fits and are not additional independent reservoirs.

A second suite tested architecture transfer more aggressively. A residual convolutional network (ResCNN; width 24, five residual blocks) and a Fourier neural op-

erator (FNO; width 16, six Fourier modes and four spectral layers) were trained on all 400 discovery realizations and evaluated blindly on the 200 external vuggy/laminated cases. Three independent seeds (101, 202 and 303) were used for each architecture. The deep models therefore test whether the hydraulic law survives a substantially different representation, not whether one model is globally superior.

3.5. Zero-Shot Protocol and Error Metrics

Terminology correction: each learned model predicts the state at a requested report time from permeability and time, rather than being stepped recursively from its own previous forecast. E_{roll} is retained as the data-file variable name for integrated MULTI-TIME FORECAST ERROR, not an autoregressive rollout-stability diagnostic. No result in this manuscript establishes repeated-step error propagation or a long-horizon predictability theorem.

Zero-shot means that the target geology family contributes no state trajectory to training, no fine-tuning, no target-specific normalization and no hyperparameter selection. For the four discovery families, each family was held out in turn and the remaining three formed training. For the two external families, all four discovery families were used for training and the external data remained blind. Distances were recomputed relative to the exact training manifold of each experiment.

Pressure and saturation were evaluated at each report time. The principal endpoint was integrated rollout error

$$E_{\mathrm{roll}} = \frac{1}{N_t}\sum_{n=1}^{N_t}\frac{E_p(t_n)+E_S(t_n)}{2}. \tag{10}$$

here E_{roll} denotes the dimensionless mean error across $N_t = 5$ report times t_n; E_p and E_S are, respectively, the normalised pressure and water-saturation errors at each time. The subscript "roll" is retained for consistency with the archived CSV field; the models predict each report time directly and are not autoregressively rolled out.

$$E_x(t_n) = \frac{\left\|\hat{x}(t_n)-x(t_n)\right\|_2}{\left\|x(t_n)\right\|_2+\epsilon},\ x\in\{p,S_w\}. \tag{11}$$

here x represents either the pressure field p or the water-saturation field S_w; $\hat{x}$ is the surrogate-predicted field; $\|\cdot\|_2$ is the Euclidean vector norm over grid cells; t_n is report time n; and $\epsilon > 0$ is a small denominator stabiliser. E_p and E_S are obtained by selecting $x = p$ and $x = S_w$, respectively. All metrics compare values at the same time, on the same grid and with the same reference simulator output.

Pressure error E_p, saturation error E_S and final-time combined error were retained as secondary outcomes. The main analyses use reservoir realizations as independent units; time steps are never treated as independent replicates.

3.6. Controlled Bottleneck Interventions

Association alone cannot establish mechanism. Forty base cases were sampled from Gaussian and channelized families and a near-spanning low-permeability

barrier containing a two-cell gap was imposed. Barrier log-permeability decrement took severity values 0, 1.5, 3.0, 4.5 and 6.0, creating 200 paired intervention cases. Global permeability statistics changed far less than source-to-sink communication. For each severity, D_K, change in the simulator response relative to the unperturbed base case and surrogate rollout error were measured. This creates a controlled dose-response test of bottleneck amplification.

3.7. Nonlinear and Resolution Robustness

For nonlinear reduced-basis models, the archived analysis instead retained 35 log-permeability input PCA components and 70 state-output PCA components; it used Ridge $\alpha = 1.0$, MLP (96, 96), max_iter = 150 with early_stopping = True and n_iter_no_change = 8 (other scikit-learn defaults, including validation_fraction = 0.1, initial learning rate = 10^{-3} and L2 $\alpha = 10^{-4}$), and 120 ExtraTrees with min_samples_leaf = 2. All nonlinear PCA and scalers were fitted on the 120 discovery-family cases. The nonlinear scripts did not conduct a prospective hyperparameter search on blind families.

To test whether the result was an artifact of linear fractional flow, a separate 180-case nonlinear campaign was generated: 30 realizations per family, with Corey relative permeabilities $k_{rw} = S^2$ and $k_{ro} = (1-S)^2$, $\mu_w = 0.5$, $\mu_o = 2.0$, and pressure recomputed during IMPES-style transport. The four discovery families (120 cases) trained the surrogate; the vuggy and laminated families (60 cases) remained blind. Separately, a 32 × 32 campaign generated 20 realizations per family (120 total) with the same primary unit-mobility physics; 80 discovery cases trained reduced-basis surrogates and 40 external cases tested resolution robustness.

3.8. Statistical Analysis

Two nested prediction problems are distinguished. Level 1 (reservoir surrogate): for each discovery holdout, the model is fitted to the other three discovery families and predicts the omitted family; in the external experiment it is trained on all four discovery families and predicts vuggy/laminated reservoirs without target labels or fine-tuning. Level 2 (error predictor): six-family leave-one-family-out cross-validation refits a one-dimensional error-distance regression on FIVE families, including error labels from vuggy or laminated reservoirs when they are in the calibration folds. Accordingly, six-family R^2 measures error-model cross-family transfer but is NOT an untouched external evaluation. Only the separate blind external calibration is strictly external: the error predictor is fitted exclusively to discovery-family out-of-fold error labels, then evaluated ONCE on the 200 external cases. Across both levels, geometry fingerprints, feature sets (15 Kron and 46 composite entries), distance definitions, $k = 5$ nearest-neighbour rule, and descriptor selection were fixed before inspecting blind external forecasting errors. Fold-dependent scalers, PCA bases and covariance estimates are fitted exclusively on the respective training cohort. No external-family error labels enter the blind fit.

Spearman rank correlation quantified monotonic association between each dis-

tance and E_{roll}. Bootstrap confidence intervals resampled reservoir realizations, not time frames. Predictor quality was assessed by a six-family leave-one-family-out regression of $\log E_{\text{roll}}$ using standardized $\log(1+D)$ inputs. Blind external calibration fitted the error-distance relation on family-held-out discovery cases and scored predictions only on the two external families. Failure detection was evaluated using AUROC for $E_{\text{roll}} > 0.4$. Selective-simulation risk-coverage curves ranked cases by a score and measured mean retained error as high-score cases were progressively routed away from the surrogate.

Statistical notation: ρ denotes Spearman rank correlation in Results (as distinct from phase density ρ_α in Equation (1)); R^2 denotes the out-of-sample coefficient of determination, which can be negative; AUROC is area under the receiver-operating-characteristic curve; and CI denotes confidence interval. The binary high-error event is $E_{\text{roll}} > 0.4$ under the study's chosen error scale.

Statistical independence and uncertainty were evaluated at the reservoir level. For the deep experiments, each external case has three run-specific error values; we report per-run correlations and a prespecified reservoir-level association using the MEAN error across the three runs ($N = 200$ independent cases, not 600). Percentile 95% intervals use 3,000 case-cluster bootstrap resamples, retaining all three errors together for every sampled reservoir. For interventions, all five severity levels for each of 40 base reservoirs are retained together in 4,000 paired cluster resamples; uncertainty is reported for the difference in errors and the ratio of severity-6 to severity-0 group means (not the mean of individual percentage changes). Exploratory matched-estimator comparisons use an identical training-standardized Euclidean five-neighbour distance for the 15-entry Kron and 46-entry hydraulic-composite fingerprints. The full results appear in Appendix B. Case-level intervals condition on the three archived networks and do not estimate uncertainty across the broader population of possible training seeds.

An exploratory segmented regression searched for a change point in $\log E_{\text{roll}}$ versus $\log(1+D_K)$. Because the breakpoint was not prespecified and was estimated from the same data, it is reported as hypothesis-generating rather than confirmatory. The susceptibility formulation Π_H was also treated as a falsification target: if it did not improve prediction over D_K, the simpler response law would be preferred.

Breakpoint notation: the fitted transition in D_K is an exploratory, data-selected change point, not a prespecified physical critical value. The susceptibility score Π_H is evaluated as an alternative error predictor, not asserted as a governing identity. Figure 3 compares response-equivalent and geology-equivalent counterexamples.

Reservoir B is much farther from anchor A in geological space than reservoir C, yet B is closer in Kron-response space and its full simulator response is correspondingly closer. This directly illustrates why geological appearance and dynamic transferability are not interchangeable. Figure 4 contrasts hydraulic and geological distance against simulated response differences.

Simulator-response distance increases more coherently with Kron-response

distance than with geological distance. The relationship is strongest for pressure and combined pressure-transport response; saturation alone retains stronger dependence on local geological arrangement.

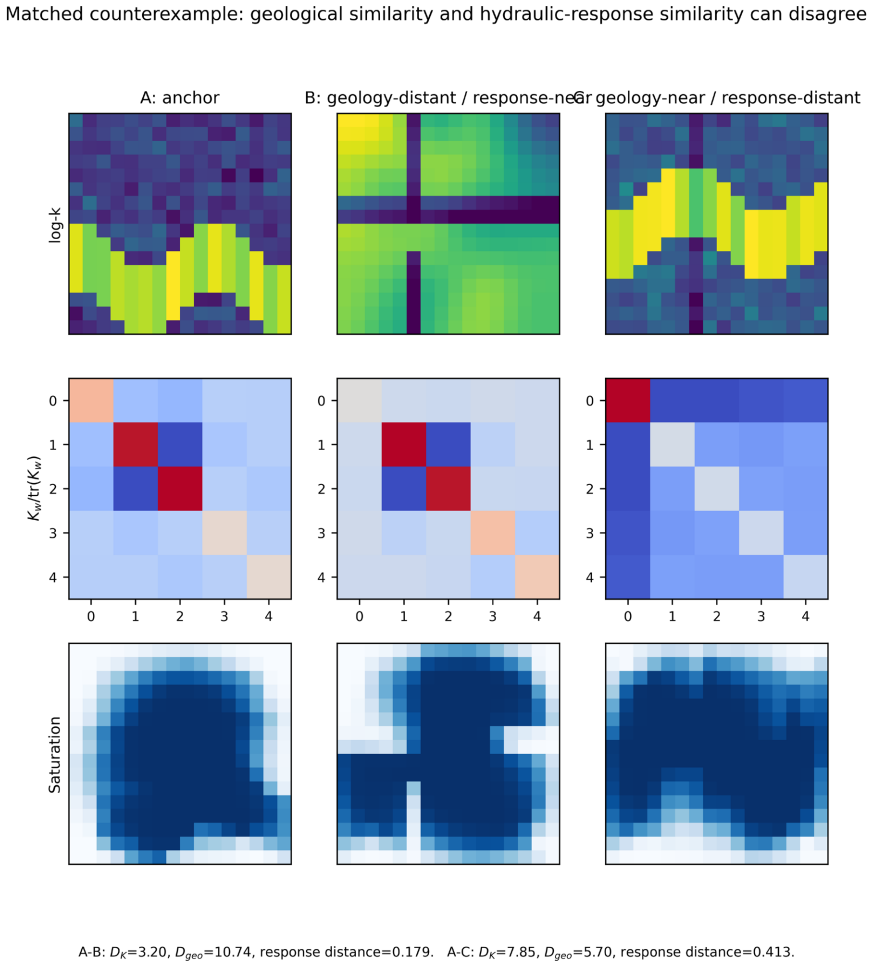


Figure 3. Matched response-equivalence counterexample.

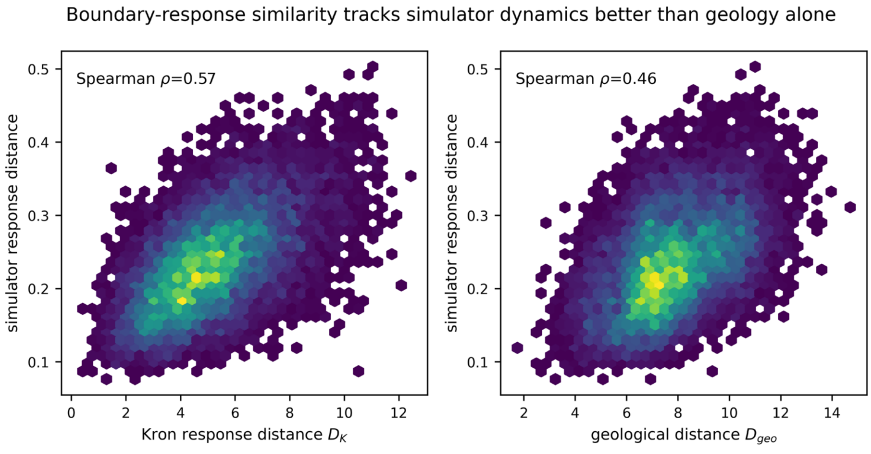


Figure 4. Pairwise response-equivalence analysis across randomly sampled discovery pairs.

4. Results

4.1. Zero-Shot Error Varies Systematically by Hydraulic Family

The family-held-out experiments generated 1,200 zero-shot predictions across the three reduced-basis surrogates, with a further 600 blind external predictions. Faulted systems were the most difficult discovery family for every model: mean E_{roll} was 0.465 for RB-Ridge, 0.538 for RB-MLP and 0.622 for RB-ExtraTrees. Fractured systems were markedly easier (0.269 - 0.315), despite visually complex permeability fields. This inversion of visual complexity and predictive difficulty is the first indication that raw image complexity is not the controlling variable. On blind external families, vuggy errors were 0.281, 0.322 and 0.332, while laminated errors were 0.318, 0.392 and 0.410 for Ridge, MLP and ExtraTrees, respectively. Table 2 provides the corresponding numerical summary.

Table 2. Mean integrated zero-shot rollout error by geology family.

Test family	RB-Ridge	RB-MLP	RB-ExtraTrees
gaussian	0.285	0.388	0.378
channels	0.312	0.344	0.364
faulted	0.465	0.538	0.622
fractured	0.269	0.314	0.315
vuggy	0.281	0.322	0.332
laminated	0.318	0.392	0.410

4.2. Kron Response Is a Stronger Organizing Variable Than Geology

Across all six families, D_K showed a consistent monotonic relation with zero-shot error for every reduced-basis model. Spearman ρ was 0.546 for RB-Ridge, 0.566 for RB-MLP and 0.594 for RB-ExtraTrees. Geological distance was substantially weaker at 0.350, 0.395 and 0.379, respectively. The correlation advantage of D_K over D_{geo} was therefore +0.196, +0.171 and +0.215. Reservoir-level bootstrap intervals for the difference excluded zero in all three cases (approximately 0.13 - 0.26 for Ridge, 0.11 - 0.24 for MLP and 0.15 - 0.28 for ExtraTrees).

The key observation is not merely that D_K correlates with error. D_H, D_{FD} and D_{geo} were all computed from information that could plausibly encode family identity, yet D_K remained the most consistently useful single quantity. For ExtraTrees, for example, $\rho\left(D_K, E_{\mathrm{roll}}\right) = 0.594$ compared with 0.568 for the 46-component composite D_H, 0.496 for flow diagnostics and 0.379 for geology. The simplest boundary-response representation thus retained more transferable information than a richer manually assembled hydraulic fingerprint. Figure 5 relates multi-time forecast error to hydraulic-response distance.

Integrated zero-shot error increases monotonically with distance from the training manifold in normalized Kron-reduced well-response space for ridge, MLP and ExtraTrees surrogates. Black curves show the fitted log-log response and

Γ is hydraulic generalization elasticity.

4.3. Cross-Family Prediction of Failure Favours Response Geometry

A stronger test asks whether the distance predicts errors in an entire family excluded from the error model itself. Six-family leave-one-family-out prediction of log error gave R^2 = 0.373, 0.379 and 0.470 from D_K for Ridge, MLP and ExtraTrees. Geological distance gave only 0.042, 0.117 and 0.133. The composite hydraulic descriptor reached 0.376, 0.288 and 0.267, while flow diagnostics alone were weak or unstable. Combining D_K with geological and flow-diagnostic distances raised R^2 to 0.448, 0.406 and 0.487, respectively. Thus, geology contains some complementary information, but most of the transferable single-predictor signal is already present in K_w.

The susceptibility model was explicitly falsified in this test. A model using Θ_H and χ_H produced negative cross-family R^2 for all three surrogate classes, indicating that the perturbation-bound intuition did not translate into a calibrated predictor in this controlled regime. The result narrows the theory: algebraic connectivity is physically meaningful, but multiplying an empirical support distance by a susceptibility factor overcorrects the observed surrogate errors here. Table 3 provides the corresponding numerical summary.

Figure 6 compares cross-family error-prediction accuracy.

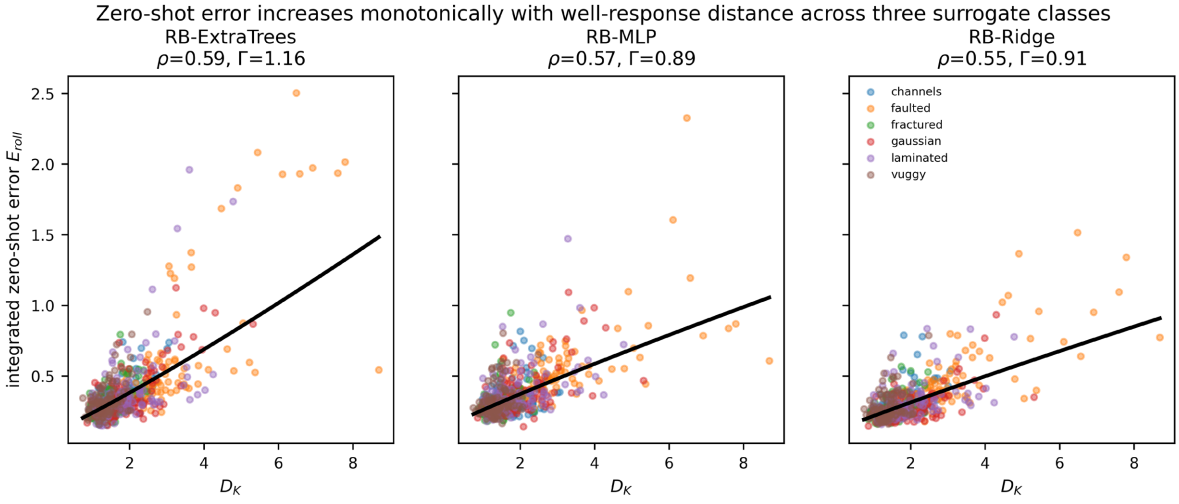


Figure 5. Primary law test.

Table 3. Six-family leave-one-family-out predictive R^2 for log rollout error. Negative values indicate failure to generalize the error model.

Predictor	RB-Ridge	RB-MLP	RB-ExtraTrees
Geology	0.042	0.117	0.133
Flow diagnostics	0.092	0.025	−0.084
Hydraulic composite	0.376	0.288	0.267

Continued

Kron response	0.373	0.379	0.470
Combined	0.448	0.406	0.487
Susceptibility	−1.139	−1.533	−1.822

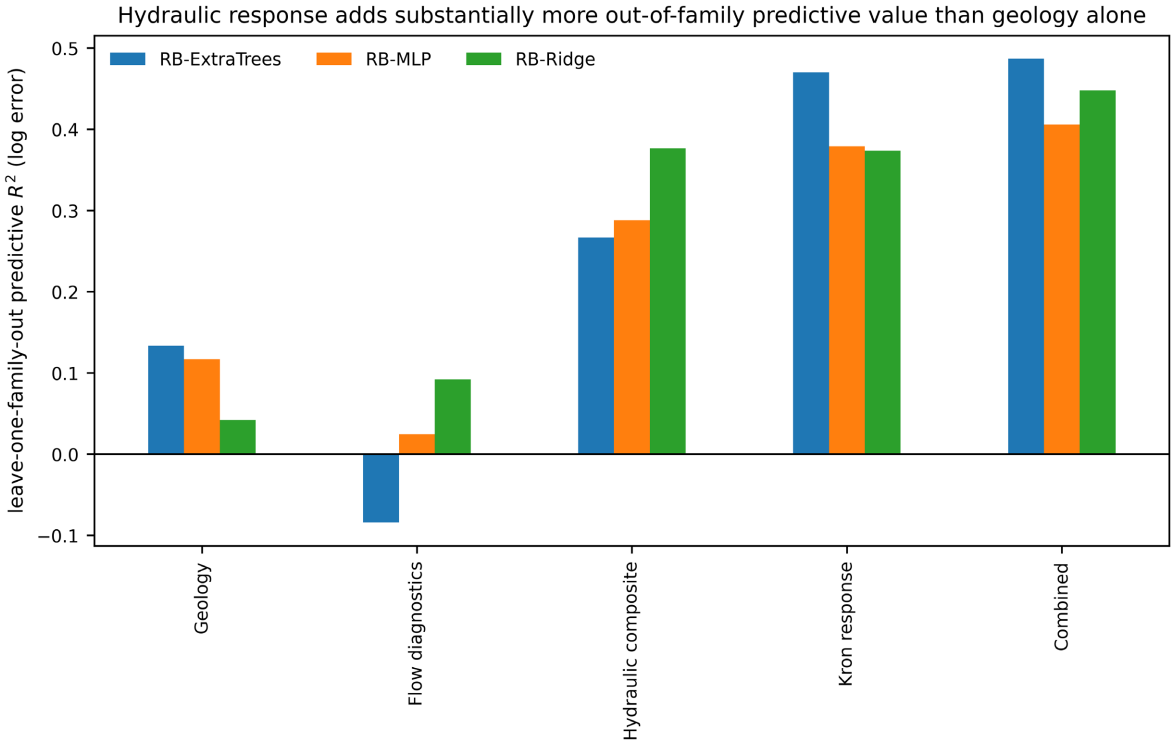


Figure 6. Out-of-family error-prediction comparison.

The Kron-reduced response distance provides much higher leave-one-family-out predictive R^2 than geology or flow diagnostics alone. A combined model adds only a modest gain, while the susceptibility formulation fails to generalize.

4.4. Blind External Validation Confirms a Domain-of-Validity Signal

The most important calibration test fitted each one-dimensional error law using only the four discovery-family experiments and then predicted error for vuggy and laminated reservoirs. D_K achieved external log-error R^2 of 0.168 for Ridge, 0.223 for MLP and 0.281 for ExtraTrees. D_H gave 0.170, 0.194 and 0.242, while D_{geo} was essentially non-predictive (−0.032, 0.023 and −0.039). These values are modest in absolute terms as expected for a one-dimensional predictor of complex model error but they are positive under a genuine topology-family shift and consistently superior to image-based distance.

This result matters conceptually. The law is not fitted after seeing the blind geology. A response geometry estimated from the frozen target reservoir, without its future state trajectory, carries measurable information about how far a learned

simulator is likely to miss. That is precisely the function required of a pre-simulation trust indicator.

4.5. Response Equivalence Is Observable at the Simulator Level

Across randomly sampled pairs from the 400 discovery reservoirs, pairwise distance between complete simulator responses correlated more strongly with D_K than with D_{geo} (Spearman $\rho \approx 0.56$ versus ≈ 0.43 in the full pair sample). The advantage was particularly clear for pressure response, where the reduced operator directly controls communication. Saturation distance showed a smaller advantage and, in one decomposition, retained slightly stronger association with geological distance. This is physically consistent: K_w captures terminal pressure response exactly in the linearized problem, whereas transport fronts depend on where conductance is placed internally even when terminal maps are similar.

A matched counterexample makes the distinction concrete. Taking a channelized anchor A, a faulted reservoir B had $D_K = 3.203$ but $D_{geo} = 10.739$ and a combined simulator-response distance of 0.179. A second channelized reservoir C looked geologically closer ($D_{geo} = 5.702$) yet was farther in response space ($D_K = 7.848$) and had more than twice the simulator-response distance (0.413). The result is not an assertion that K_w uniquely determines transport; it demonstrates that well-response equivalence can dominate image resemblance for dynamic similarity.

4.6. Controlled Bottleneck Intervention Provides Mechanistic Evidence

The barrier experiment produced a monotonic dose response. At severity 0, mean D_K was 1.30. Increasing the barrier decrement to 1.5, 3.0, 4.5 and 6.0 raised D_K to 1.49, 1.85, 2.25 and 2.42. Over the same sequence, the normalized simulator-response change increased from effectively zero to 0.046, 0.105, 0.153 and 0.175. RB-MLP error increased from 0.244 to 0.528 (+116%), and RB-Ridge error increased from 0.269 to 0.463 (+72%).

A paired reanalysis (Appendix B, Table B2) resamples the 40 base reservoirs with ALL five severity levels held together; severities are repeated measurements, not new independent reservoirs. At severity 6 relative to severity 0, RB-Ridge error increased by 0.194 in absolute integrated error (95% paired-bootstrap CI 0.129 - 0.271), equivalent to a +72.1% ratio-of-means increase (47.0% - 103.3%). RB-MLP increased by 0.284 (0.240 - 0.334), or +116.2% (96.7% - 137.7%). Mean D_K increased by 1.126 (approximately 95% CI 0.92 - 1.37). Both confidence intervals for absolute increases exclude zero. These paired intervals support a population-level mean response across the sampled base reservoirs, not uniform deterioration for every individual realization.

Because the intervention altered a source-to-sink bottleneck in matched base reservoirs, this experiment addresses a major confound of observational correlations: geology family. The progressive change in D_K and surrogate error occurs within the same underlying realization as the bottleneck is strengthened. It there-

fore supports a mechanistic interpretation in which response-space extrapolation, rather than label membership alone, amplifies zero-shot failure.

The perturbations are controlled dose-response interventions on the SAME simulated base reservoirs, and the paired bootstrap preserves that design. Nevertheless, weakening the barrier changes local conductance and fluxes as well as graph-response features, so the experiment provides mechanistic intervention evidence rather than identification of D_K as a unique causal mediator; independent matched perturbations would be required for that stronger conclusion. Figure 7 shows the paired bottleneck intervention severity series.

Increasing barrier severity systematically increases Kron-response distance, physical response change and surrogate error. Each trajectory originates from a matched base reservoir, providing a dose-response mechanism rather than a cross-family association alone.

4.7. Exploratory Evidence for a Finite Extrapolation Regime

Single-slope models describe the global monotonic relation, but segmented regression suggested a change in degradation rate. Estimated breakpoints occurred around D_K = 2.09 for Ridge and MLP and D_K = 2.31 for ExtraTrees. Adding a post-break slope improved AIC by approximately 18.0 for Ridge, 3.2 for MLP and 16.0 for ExtraTrees relative to a single straight line. The additional high-distance log-slope was approximately 0.63, 0.34 and 0.67, respectively.

These breakpoints are exploratory: they were optimized on the same sample and should not be interpreted as a universal critical threshold. Their importance is hypothesis generation. They suggest that future 3-D confirmation should prespecify a response-domain boundary and test whether risk accelerates once D_K exceeds the upper support of the training manifold. The proposed rule for fixing the threshold from discovery-domain data alone, together with the other criteria for independent 3-D confirmation, is specified in Appendix A. Figure 8 presents the exploratory segmented-response analysis.

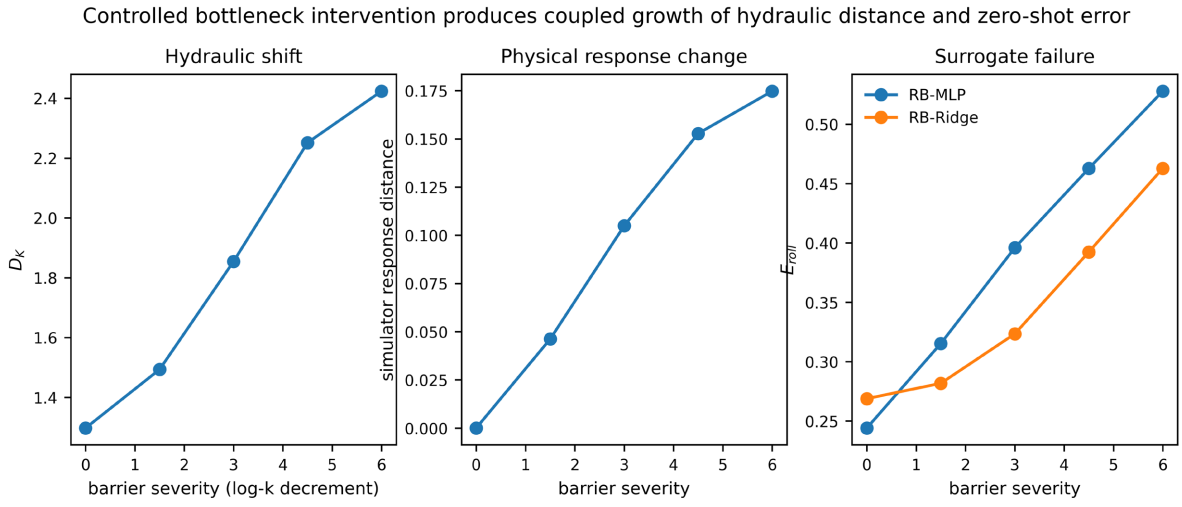


Figure 7. Controlled bottleneck interventions.

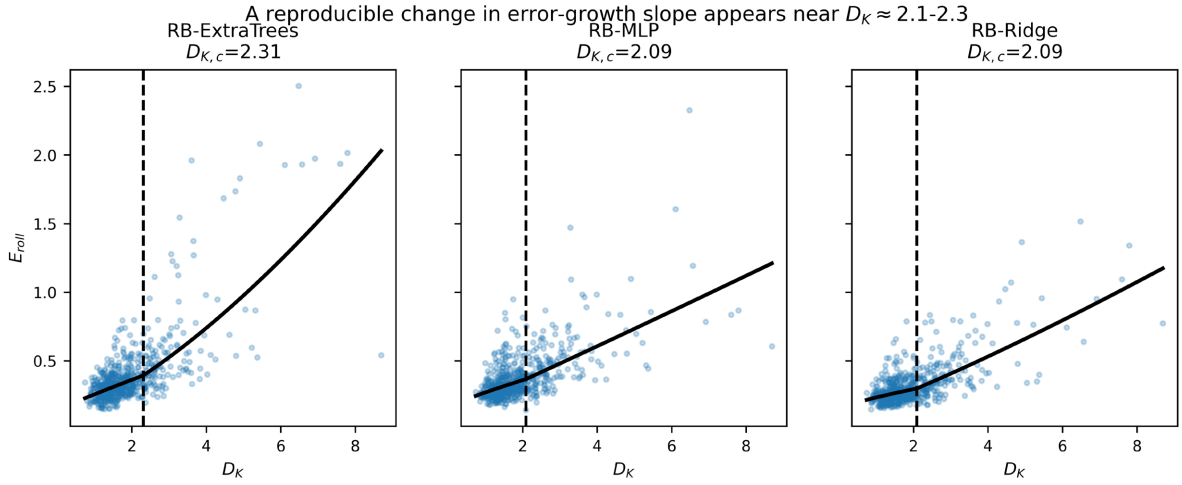


Figure 8. Exploratory segmented fits. Several surrogate classes show accelerated error growth after $D_K \approx 2.1$ - 2.3.

Because breakpoints were estimated post hoc, the figure motivates a confirmatory threshold experiment rather than establishing a universal critical value.

4.8. The Law Survives Deep ResCNN and FNO Blind Tests

The deep-model experiment asked whether the relationship survives architectures that do not use the reduced-basis regression representation. Across three independent seeds and the two blind external families, ResCNN mean E_{roll} was 0.325 on vuggy and 0.392 on laminated cases. FNO errors were higher in this deliberately data-limited cross-family setting: 0.621 and 0.746. The point of this comparison is not to rank architectures globally; it is that both exhibit the same response-distance trend.

Reservoir-level reanalysis (Appendix B, Table B1) corrects the dependence created by treating 200 external cases × three runs as 600 independent points. Spearman correlations between D_K and each case's MEAN three-run error were ρ = 0.454 (95% case-cluster bootstrap CI 0.339 - 0.558) for ResCNN and 0.486 (0.375 - 0.582) for FNO. All three per-run associations were positive: ResCNN ρ = 0.456/0.445/0.411 for runs 101/202/303; FNO 0.480/0.484/0.458. The corresponding mean-error correlation advantages over D_{geo} were +0.246 (95% CI +0.101 to +0.387) and +0.248 (+0.100 to +0.387). The originally reported pooled correlations 0.437 and 0.475 remain descriptive and have not been used as independent-sample inferential statistics. These CIs establish case-level consistency conditional on three fits but cannot reliably estimate the distribution of training-seed effects from only three runs.

Across the 200 unique external reservoirs, the historical pooled case-run correlations were 0.437 (ResCNN) and 0.475 (FNO), but the same cases were evaluated three times; we treat these pooled values as descriptive only. For cluster-aware inference and run-specific estimates see Appendix B, Table B1. Original pooled geological correlations 0.196 and 0.228 are likewise descriptive.

Figure 9 illustrates the deep-model blind external associations.

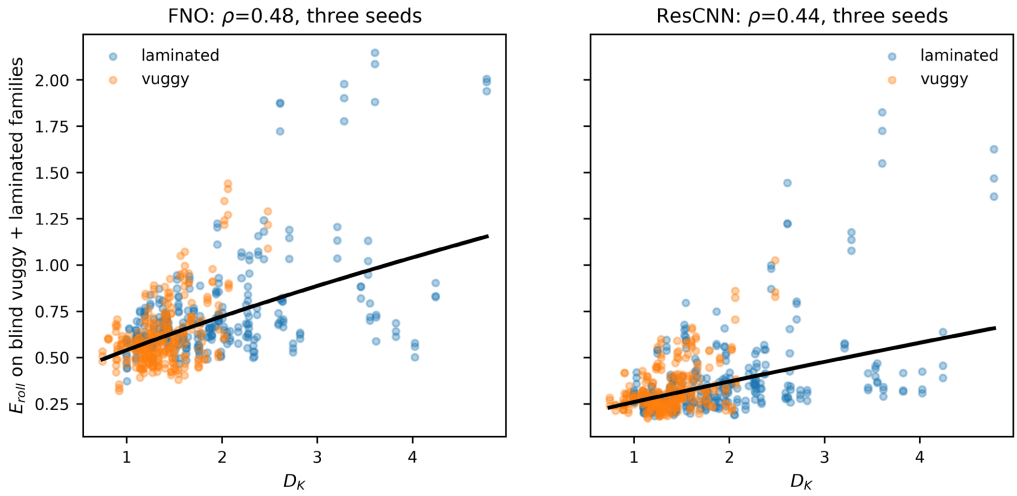


Figure 9. Blind external validation with deep architectures over three independent seeds.

ResCNN and FNO have different absolute error levels, but both show increasing failure with Kron-response distance and much weaker dependence on geological distance.

4.9. Nonlinear Two-Phase and Resolution Robustness

The nonlinear Corey experiment provides the most important physics stress test in Phase I. On the 60 blind external nonlinear cases, mean E_{roll} for laminated reservoirs was 0.267 (Ridge), 0.356 (MLP) and 0.302 (ExtraTrees); vuggy errors were 0.229, 0.313 and 0.284. Despite state-dependent mobilities and pressure recomputation, D_K remained the strongest hydraulic predictor: Spearman ρ = 0.533 for Ridge, 0.443 for MLP and 0.645 for ExtraTrees. The corresponding D_H values were 0.489, 0.390 and 0.622. Flow-diagnostic correlations were lower, particularly for Ridge and MLP (≈0.20).

Doubling spatial resolution to 32 × 32 did not erase the ordering. In the blind vuggy/laminated test, RB-Ridge obtained mean E_{roll} = 0.266 with $\rho\left(D_K, E_{\text{roll}}\right) = 0.716$; RB-MLP obtained mean E_{roll} = 0.402 with ρ = 0.830. These correlations are higher than in the primary 16 × 16 campaign, indicating that the response metric is not tied to one particular grid size. This is resolution robustness, not a claim of continuum convergence, because the model was retrained at the new resolution. Figure 10 summarizes the nonlinear two-phase robustness results.

Kron-response distance remains the strongest or co-strongest hydraulic predictor on blind vuggy and laminated cases, demonstrating that the primary relation is not an artifact of the linear fractional-flow benchmark. Table 4 provides the corresponding numerical summary.

4.10. Hydraulic Distance Can Act as a Trust Score

A distance law becomes operationally useful if it can rank risk without target labels. Defining high error as $E_{\text{roll}} > 0.4$, D_K produced AUROC = 0.835 for Ridge,

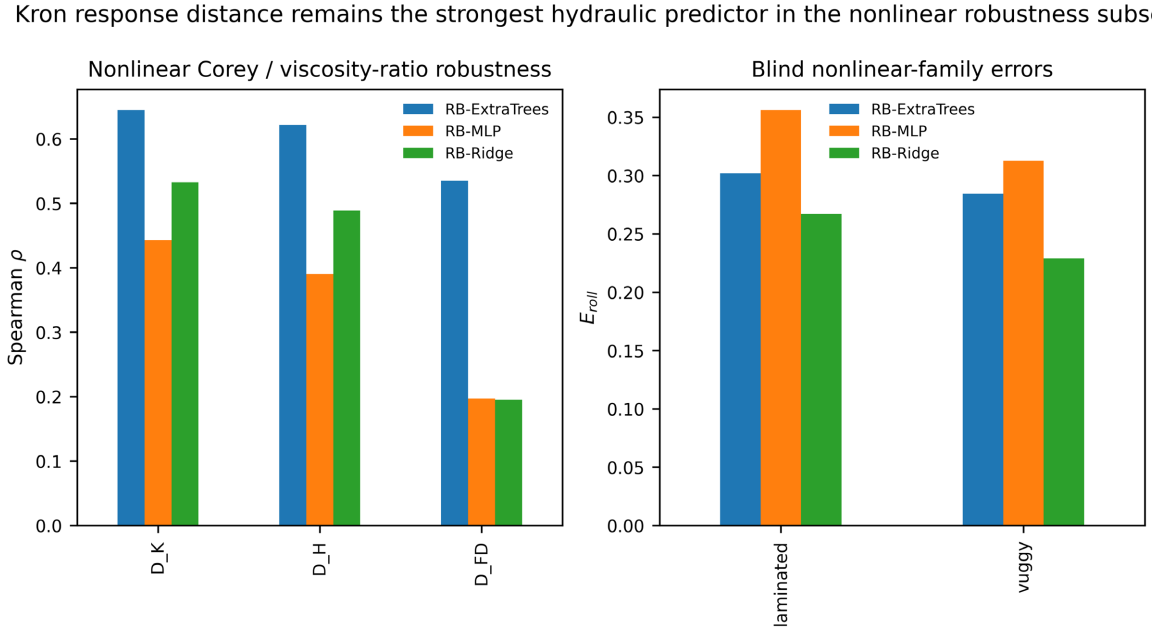


Figure 10. Nonlinear two-phase robustness with Corey relative permeabilities and viscosity ratio $\mu_o/\mu_w = 4$.

Table 4. Independent architecture, physics and resolution stress tests.

Robustness test	Blind sample	Main D_K result	Comparator/error
Deep ResCNN, blind external	200 reservoirs; 3 runs	$\rho = 0.454$ [0.339, 0.558]	D_{geo} : $\rho = 0.208$
Deep FNO, blind external	200 reservoirs; 3 runs	$\rho = 0.486$ [0.375, 0.582]	D_{geo} : $\rho = 0.238$
Nonlinear Corey: RB-Ridge	60 external	$\rho = 0.533$	D_H : 0.489
Nonlinear Corey: RB-MLP	60 external	$\rho = 0.443$	D_H : 0.390
Nonlinear Corey: ExtraTrees	60 external	$\rho = 0.645$	D_H : 0.622
32 × 32: RB-Ridge	40 external	$\rho = 0.716$	mean $E = 0.266$
32 × 32: RB-MLP	40 external	$\rho = 0.830$	mean $E = 0.402$

0.788 for MLP and 0.821 for ExtraTrees. D_{geo} achieved only 0.684, 0.724 and 0.690. D_H was second-best for MLP and ExtraTrees, while D_{FD} was close to D_K for Ridge but weaker for the other models. These values indicate that response distance is informative not only in regression but also in binary trust decisions.

Risk-coverage analysis leads to the same conclusion. When cases are sorted from low to high D_K and high-distance cases are progressively rejected, the mean error of retained predictions rises more slowly than when sorting by geology. The normalized area under the risk-coverage curve over 20 - 100% coverage was 0.271 versus 0.289 for Ridge, 0.323 versus 0.342 for MLP and 0.316 versus 0.348 for ExtraTrees for D_K versus D_{geo}. The difference is operationally meaningful: a model can accept low-D_K cases and route high-D_K cases to a full simulator or to an enriched training workflow.

Figure 11 shows selective-simulation risk as a function of retained coverage.

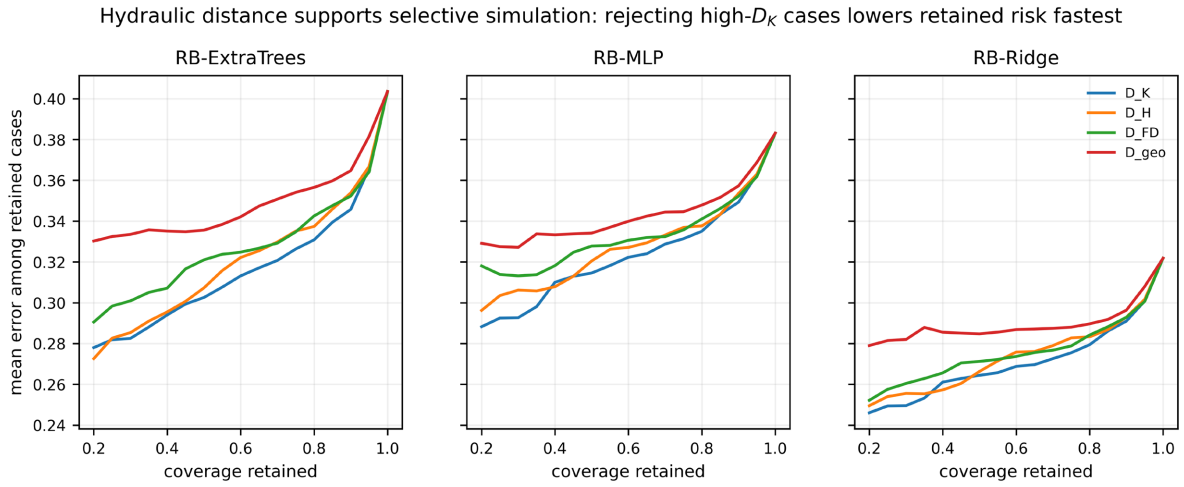


Figure 11. Selective-simulation risk-coverage curves.

For this controlled study, ranking forecasts by D_K supports its use as a candidate pre-simulation risk-ranking score.

4.11. The Strongest Initial Hypothesis Was Not Supported

The initial theory proposed that hydraulic extrapolation should be multiplied by a susceptibility factor derived from the smallest non-zero reduced-operator eigenvalue. The data do not support that formulation. Across Ridge, MLP and ExtraTrees, $\rho(\Pi_H, E_{\text{roll}})$ was 0.408, 0.420 and 0.480 consistently lower than the corresponding D_K correlations of 0.546, 0.566 and 0.594. Six-family cross-validation with Θ_H and χ_H was strongly negative, indicating poor calibration across topology families.

Matched metric analysis (Appendix B, Table B3) separates descriptor CONTENT from distance ESTIMATION. For both the 15-entry K_w and 46-entry composite fingerprint, we fit training-only StandardScaler transformations, then average the five smallest Euclidean distances, without covariance-based weighting. Spearman $\rho(D_K, E_{\text{roll}})$ was 0.546/0.566/0.594 across Ridge/MLP/ExtraTrees; the equal-metric composite gave 0.500/0.511/0.548. Six-family error-predictor R^2 was 0.374/0.379/0.470 for Kron and 0.404/0.322/0.390 for the composite: the composite is BETTER in the Ridge discovery cross-validation comparison, while Kron is better for MLP and ExtraTrees. In the STRICTLY blind external calibration, Kron achieved R^2 = 0.167/0.222/0.281 and the equal-metric composite 0.157/0.191/0.226. We therefore qualify any unconditional claim of Kron superiority: it is consistently superior in blind external evaluation and aggregate rank association, but NOT on every discovery-fold model-metric combination.

This falsification is a central new finding, not a weakness. It shows that the response operator already contains the relevant conditioning information for this five-port configuration; explicitly reweighting by algebraic-connectivity susceptibility double-counts or distorts it. Similarly, the 46-component D_H composite did not improve on D_K despite incorporating effective resistance, spectra, cut

proxies and flow diagnostics. The validated law is therefore simpler than the original deep-research conjecture: distance in terminal response geometry is the primary organizing variable in the controlled domain.

Figure 12 contrasts failure detection with the susceptibility correction. In the left panel, D_K achieves the highest AUROC for $E_{\text{roll}} > 0.4$. In the right panel, the susceptibility-corrected Π_H does not improve rank association. Table 5 summarizes these results.

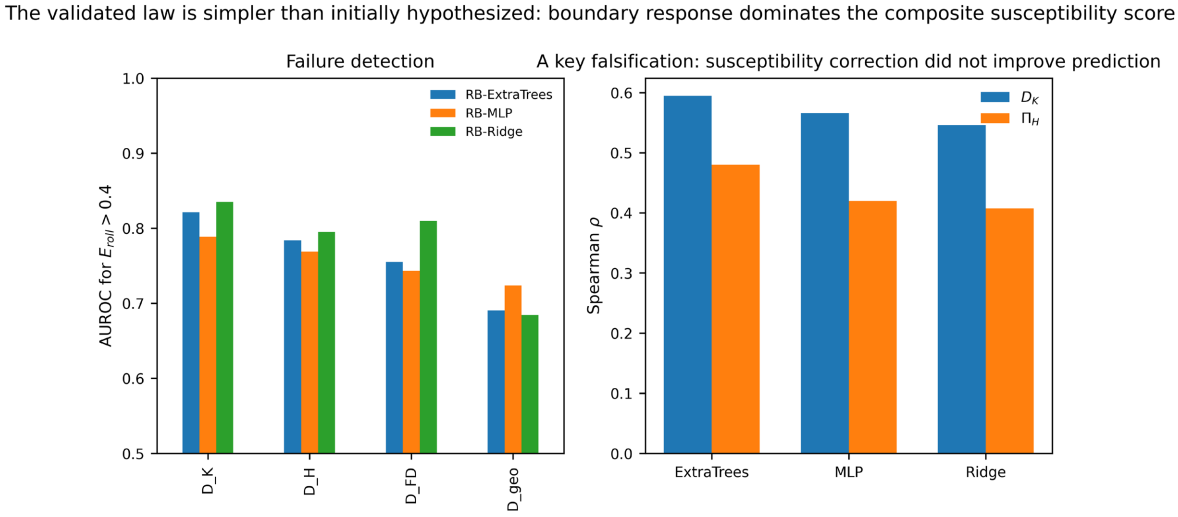


Figure 12. Simplification and falsification.

Table 5. Discovery ledger and falsification status.

ID	Hypothesis/discovery	Outcome	Evidence
F1	D_K predicts zero-shot error better than D_{geo}	Supported	ρ advantage +0.17 to +0.22; bootstrap CI excludes 0
F2	D_K predicts unseen-family error	Supported	six-family CV R^2 0.37 - 0.47; external R^2 0.17 - 0.28
F3	Bottleneck edits change D_K and prediction error	Mechanistic dose-response supported	D_K +87%; error +72 - 116% at max severity
F4	Law is architecture-specific	Rejected	relation persists in Ridge/MLP/trees/ ResCNN/FNO
F5	Law is linear-physics artifact	Rejected within Phase I	nonlinear Corey subset $\rho = 0.44$ - 0.65
F6	Susceptibility correction Π_H improves prediction	Rejected	Π_H weaker than D_K; negative cross-family calibration
F7	Single universal critical D_K exists	Not established	exploratory breakpoint ≈ 2.1 - 2.3; needs prespecified confirmation

5. Discussion

5.1. What the Validated Law Actually Says

The strongest defensible conclusion is narrower and more useful than the original slogan "topology predicts generalization". Within the controlled finite-volume domain, zero-shot failure is organized by distance between the well-conditioned hydraulic response of the target and the response manifold represented during training. The relevant object is therefore not graph topology in the strict combinatorial sense. It is a topology-induced, conductance-weighted boundary operator. This distinction reconciles physical-network theory with reservoir engineering: adjacency constrains possible communication, conductance weights set resistance, and the Schur complement summarizes what the wells actually experience.

That interpretation explains why D_K can outperform D_{geo}. Geological distance treats changes in regions with little dynamic influence similarly to changes on source-to-sink backbones. K_w automatically integrates the network and weights every internal pathway by its contribution to the terminal response. A large pixelwise change can be dynamically benign if alternative pathways compensate; a visually small edit can be severe if it seals a bridge or reconnects a compartment. The matched-pair and intervention experiments show both cases.

5.2. Relation to Flow Diagnostics and Porous-Media Topology

Flow diagnostics already provide powerful reduced descriptions of reservoir communication [19]-[22]. The present results should be read as complementary, not as a claim that classical diagnostics are obsolete. D_{FD} carried real predictive signal and, for Ridge failure detection, approached D_K. The distinction is that K_w is a compact linear boundary-response operator available directly from the pressure system, whereas time-of-flight and allocation depend on a specific driving condition and directed flux field. The data suggest that terminal response geometry is the more stable cross-family coordinate in this experiment, while flow diagnostics can add task-specific transport information.

At pore scale, max-flow/min-cut and topological descriptors have been linked to permeability and phase connectivity [18] [23] [24]. Our barrier experiment provides a reservoir-scale analogue: as a near-spanning barrier is strengthened, the well-response operator moves away from its training support and model error increases. Yet the composite D_H did not outperform D_K, warning against indiscriminately aggregating graph descriptors. A physically rich descriptor is not automatically a better OOD coordinate.

5.3. Why the Negative Susceptibility Result Matters

The perturbation bound involving $1/\lambda_2$ remains mathematically correct for linearized operator sensitivity, but a bound is not necessarily a calibrated error law for a learned surrogate. Three reasons can explain the observed failure of Π_H. First, K_w already embeds the same conditioning information through its eigenstructure, so multiplying by χ_H may double-count weak connectivity. Second,

λ_2 summarizes the single slowest reduced mode, whereas prediction error can depend on several modes and on model-specific representation. Third, the surrogate error is not identical to the physical state perturbation bounded by the resolvent identity. This distinction between physically valid sensitivity bounds and empirically useful OOD scores is important for scientific machine learning more broadly.

5.4. A New Way to Compare Surrogate Architectures

Conventional benchmark tables rank models by average RMSE. Equation (9) suggests an additional axis: hydraulic generalization elasticity Γ_M. In the primary campaign Γ was approximately 0.91 for Ridge, 0.89 for MLP and 1.16 for ExtraTrees. In the deep external test Γ was ≈0.88 for ResCNN and ≈0.72 for FNO. These numbers should not be over-interpreted across experiments because training budgets differ, but the concept is useful: two surrogates with similar in-domain error can have very different rates of degradation with response-space extrapolation. Future neural-operator benchmarks should report error as a function of physical distance from training support, not only as a single test-set average.

5.5. Operational Implication: Selective Simulation Instead of Blind Trust

The AUROC and risk-coverage results convert the law into a deployment principle. Before accepting a learned forecast, compute the target reservoir's reduced well-response operator under a cheap pressure solve, standardize it with frozen training statistics, and evaluate D_K. Low-distance cases can proceed through the surrogate; high-distance cases can trigger a full simulator, active-learning acquisition or ensemble uncertainty analysis. This architecture does not require D_K to be a perfect error certificate. It requires only that risk be ranked well enough to allocate expensive simulation selectively. Recent operator-learning work has independently emphasized uncertainty-aware deployment [14]; D_K provides a complementary physics-derived score that does not require retraining an uncertainty head.

5.6. Why the Result Is Potentially Transferable beyond Petroleum Reservoirs?

Equation (4) is not specific to hydrocarbons. Schur-complement response operators arise in groundwater flow, geothermal systems, CO_2 storage, electrical networks and other diffusion-dominated media. Recent subsurface neural operators and physics-constrained graph models are already moving toward heterogeneous geology, fractures and coupled processes [10] [11] [13] [24]-[26]. The present study suggests a general experimental strategy for those domains: define the response operator seen at controlled/observed ports, measure distance from training support in that operator space, and test whether it organizes OOD error. The specific distance and ports will change; the principle of response-conditioned gen-

eralization may not.

6. Limitations and Scope of the Predictive Law

The law is validated only within a controlled two-dimensional finite-volume campaign. The primary grid is 16 × 16 and the 32 × 32 study is a resolution robustness test, not a field-scale validation.

The blind families are synthetic vuggy and laminated generators. They are structurally distinct from the discovery families, but they do not reproduce the full statistical complexity of a real carbonate, fluvial or faulted field.

The nonlinear robustness study includes state-dependent relative permeability and an adverse viscosity ratio, but not gravity segregation, capillary pressure, compressibility, hysteresis, PVT variation, geomechanics or three-phase flow.

K_w is tied to the selected terminal/well configuration. If wells move, the response operator changes; this is a feature for deployment but means that D_K is a well-conditioned similarity, not an intrinsic geology metric.

No single scalar can fully determine saturation-front evolution. Pairwise analysis showed that internal geology remains important for transport even when terminal pressure response is similar.

The exploratory D_K breakpoint must be retested with a prespecified threshold in a new dataset. Calling it a universal phase transition would exceed the evidence.

The deep models were intentionally modest and data limited. Their purpose was architecture transfer of the law, not establishing a state-of-the-art FNO benchmark.

Additional reproducibility limitations: the original public reproducibility ZIP contained analysis scripts and summary CSVs but omitted the input dataset.npz and the nonlinear per-family trajectory arrays needed for complete re-execution. This revision includes a regenerated fixed-seed 600-case primary data archive and separately documents the remaining absence of the nonlinear trajectory arrays, rather than claiming that all experiments can be rebuilt from the prior package without rerunning simulations. The code also applies a row-level rather than reservoir-grouped stopping split for the reduced-basis MLP and seeds deep-model initializations only after construction; both issues are disclosed and designated for correction in future independent replication.

A true top-tier field confirmation should next freeze the Phase-I definition of D_K and test it without modification on 3-D SPE10/Norne-type models, variable well schedules, black-oil or CO_2/H_2 physics and an independent simulator such as OPM Flow. The proposed confirmation protocol including unchanged hydraulic-distance definitions, an independently held-out geology and well pattern, and external-simulator testing is detailed in Appendix A.

These limitations bound the claim rather than negate it. The study establishes a mechanistic and repeatedly stress-tested Phase-I law: well-response distance is a strong predictor of zero-shot difficulty under controlled heterogeneous flow. The next scientific question is whether the same frozen coordinate remains predictive

when the missing physics and scale are introduced.

7. Conclusions

This study asked a deliberately difficult question: can one predict, from physics rather than target labels, when a learned reservoir model will fail on a reservoir it has never seen? Across 600 primary realizations, strict family-held-out tests, two blind geology families, controlled bottleneck perturbations, three reduced-order surrogate classes, three-seed ResCNN/FNO experiments, nonlinear Corey flow and a doubled grid resolution, the answer is consistently affirmative within the controlled study domain. The most robust predictor is not geological-image distance and not a heavily engineered topology score. It is the distance D_K between normalized Kron-reduced well-response operators and the training response manifold.

D_K correlated with rollout error at $\rho = 0.546$ - 0.594 for the primary surrogate suite, provided six-family held-out predictive $R^2 = 0.373$ - 0.470, retained positive calibration on blind vuggy/laminated families, and detected high-error forecasts with AUROC = 0.788 - 0.835. Strengthening a controlled bottleneck moved D_K from 1.30 to 2.42 while surrogate error rose by 72% - 116%. The relation persisted under deep architectures, nonlinear mobility and 32×32 resolution. These independent tests support a response-conditioned domain-of-validity interpretation rather than a family-label artifact.

Equally important, the research falsified its stronger initial conjecture. A 46-component hydraulic descriptor and a susceptibility-corrected extrapolation number did not outperform D_K. This negative result simplifies the theory: the terminal response operator appears to be a sufficient first order organizing variable for the tested regime. We therefore propose the bounded empirical law $\log E_{ZS} = a_M + \Gamma_M \log(1 + D_K) + \varepsilon$ as a reproducible way to report surrogate degradation with hydraulic extrapolation. The law should now be frozen and tested in 3-D, field-scale and richer multiphase systems. If it survives those confirmatory stages, hydraulic response geometry could become a practical pre-simulation trust criterion for learned subsurface models. The tests required before extending this claim beyond Phase I are explicitly outlined in Appendix A.

Data and Code Availability

The study provides source scripts, summary-level CSVs, the fixed-seed regenerated 600-case primary dataset, matched-estimator features and clustered-bootstrap analyses in the companion reproducibility archive. The nonlinear per-family trajectory NPZ outputs were not present in the archived package; they must be reconstructed by running the supplied nonlinear simulation scripts. The 3-D and field-scale validations have not been performed.

Acknowledgements

The authors acknowledge the open scientific-computing and porous-media re-

search communities whose published methods motivated the response-operator and zero-shot validation framework. This research received no external funding.

Author Contributions

Franck-Hilaire Essiagne: Conceptualization, Methodology, Software, Formal analysis, Writing—original draft, Visualization. Kouassi Louis Kra: Conceptualization, Supervision, Validation, Writing—review and editing. Moussa Camara: Investigation, Resources, Validation.

Conflicts of Interest

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

References

[1] 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

[2] He, Q., Barajas-Solano, D., Tartakovsky, G. and Tartakovsky, A.M. (2020) Physics-informed Neural Networks for Multiphysics Data Assimilation with Application to Subsurface Transport. *Advances in Water Resources*, **141**, Article 103610. https://doi.org/10.1016/j.advwatres.2020.103610

[3] 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 107273. https://doi.org/10.1016/j.petrol.2020.107273

[4] Tang, M., Liu, Y. and Durlofsky, L.J. (2021) Deep-Learning-Based Surrogate Flow Modeling and Geological Parameterization for Data Assimilation in 3D Subsurface Flow. *Computer Methods in Applied Mechanics and Engineering*, **376**, Article 113636. https://doi.org/10.1016/j.cma.2020.113636

[5] Wen, G., Hay, C. and Benson, S.M. (2021) CCSNet: A Deep Learning Modeling Suite for CO_2 Storage. *Advances in Water Resources*, **155**, Article 104009. https://doi.org/10.1016/j.advwatres.2021.104009

[6] 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 104180. https://doi.org/10.1016/j.advwatres.2022.104180

[7] 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 105826. https://doi.org/10.1016/j.cageo.2024.105826

[8] Ju, X., Hamon, F.P., Wen, G., Kanfar, R., Araya-Polo, M. and Tchelepi, H.A. (2024) Learning CO_2 Plume Migration in Faulted Reservoirs with Graph Neural Networks. *Computers & Geosciences*, **193**, Article 105711. https://doi.org/10.1016/j.cageo.2024.105711

[9] Al Kobaisi, M., Zhang, W., Diab, W. and Hajibeygi, H. (2026) Learning Multi-Phase

Flow and Transport in Fractured Porous Media with Auto-Regressive and Recurrent Graph Neural Networks. *Earth Energy Science*, **2**, Article 100035. https://doi.org/10.1016/j.ees.2025.11.002

[10] Chen, H., Xue, L., Liu, L., Zou, G., Han, J., Dong, Y., *et al.* (2025) Physics-Informed Graph Neural Network for Predicting Fluid Flow in Porous Media. *Petroleum Science*, **22**, 4240-4253. https://doi.org/10.1016/j.petsci.2025.06.007

[11] Nunez, A., Misra, S. and Falola, Y. (2025) Transfer Learning for Geological Carbon Storage Forecasting Using Neural Operator. *Advances in Water Resources*, **199**, Article 104948. https://doi.org/10.1016/j.advwatres.2025.104948

[12] Lara Benitez, J.A., Furuya, T., Faucher, F., Kratsios, A., Tricoche, X. and de Hoop, M.V. (2024) Out-of-Distributional Risk Bounds for Neural Operators with Applications to the Helmholtz Equation. *Journal of Computational Physics*, **513**, Article 113168. https://doi.org/10.1016/j.jcp.2024.113168

[13] Gudala, M. and Yan, B. (2025) Graph Neural Network for Multi-Physics Geothermal Simulation with Discrete Fracture Network. *Advances in Water Resources*, **205**, Article 105062. https://doi.org/10.1016/j.advwatres.2025.105062

[14] Winovich, N., Daneker, M., Lu, L. and Lin, G. (2026) Active Operator Learning with Predictive Uncertainty Quantification for Partial Differential Equations. *Journal of Computational Physics*, **555**, Article 114791. https://doi.org/10.1016/j.jcp.2026.114791

[15] van der Schaft, A.J. and Maschke, B.M. (2013) Port-Hamiltonian Systems on Graphs. *SIAM Journal on Control and Optimization*, **51**, 906-937. https://doi.org/10.1137/110840091

[16] Dorfler, F. and Bullo, F. (2013) Kron Reduction of Graphs with Applications to Electrical Networks. *IEEE Transactions on Circuits and Systems I: Regular Papers*, **60**, 150-163. https://doi.org/10.1109/tcsi.2012.2215780

[17] Sugiyama, T. and Sato, K. (2023) Kron Reduction and Effective Resistance of Directed Graphs. *SIAM Journal on Matrix Analysis and Applications*, **44**, 270-292. https://doi.org/10.1137/22m1480823

[18] Armstrong, R.T., Lanetc, Z., Mostaghimi, P., Zhuravljov, A., Herring, A. and Robins, V. (2021) Correspondence of Max-Flow to the Absolute Permeability of Porous Systems. *Physical Review Fluids*, **6**, Article 054003. https://doi.org/10.1103/physrevfluids.6.054003

[19] Møyner, O., Krogstad, S. and Lie, K. (2015) The Application of Flow Diagnostics for Reservoir Management. *SPE Journal*, **20**, 306-323. https://doi.org/10.2118/171557-pa

[20] Lie, K.-A. (2019) Flow Diagnostics. In: *An Introduction to Reservoir Simulation Using MATLAB/GNU Octave*, Cambridge University Press, 477-517. https://doi.org/10.1017/9781108591416.018

[21] Watson, F., Krogstad, S. and Lie, K. (2021) The Use of Flow Diagnostics to Rank Model Ensembles. *Computational Geosciences*, **26**, 803-822. https://doi.org/10.1007/s10596-021-10087-6

[22] Shook, G.M. and Mitchell, K.M. (2009) A Robust Measure of Heterogeneity for Ranking Earth Models: The F-PHI Curve and Dynamic Lorenz Coefficient. *SPE Annual Technical Conference and Exhibition*, New Orleans, 4-7 October 2009, 2593-2605. https://doi.org/10.2118/124625-ms

[23] Omrani, H. and Omrani, H. (2026) A Topological and Manifold Framework for Porous Media Modeling: A Review of Engineering Applications and Mathematics. *Next Chemical Engineering*, **2**, Article 100023.

https://doi.org/10.1016/j.nxcen.2026.100023

[24] Zhao, Q. and Xiao, H. (2026) An End-to-End Differentiable, Graph Neural Network-Embedded Pore Network Model for Permeability Prediction. *Journal of Hydrology*, **675**, Article 135602. https://doi.org/10.1016/j.jhydrol.2026.135602

[25] Li, T., Zou, Y., Zou, S., Chang, X., Zhang, L. and Deng, X. (2025) Learning to Solve PDEs with Finite Volume-Informed Neural Networks in a Data-Free Approach. *Journal of Computational Physics*, **530**, Article 113919. https://doi.org/10.1016/j.jcp.2025.113919

[26] Xue, Z., Bi, J., Ma, H., Sun, Z. and Chen, Z. (2026) A Robust Physics-Constrained Neural Operator Framework for Efficient Geothermal Resource Development. *Nature Communications*, **17**, Article No. 6929. https://doi.org/10.1038/s41467-026-73183-0

Nomenclature and Abbreviations

Symbols are defined at first use and collected here for cross-reference. Table A1 lists reservoir-flow and finite-volume variables; Table A2 defines hydraulic operators and distance measures; and Table A3 summarizes surrogate-modelling and statistical quantities. Subscripts identify physical phases, wells, descriptor classes and statistical outcomes; they are not literal underscore characters.

Table A1. Reservoir flow and finite-volume variables.

Symbol/ abbreviation	Definition and usage
α; w, o	Fluid-phase index; subscripts w and o denote water and oil.
φ	Porosity (pore-volume fraction), dimensionless; fixed at 0.20 in primary tests.
ρ_α	Mass density of phase α; distinct from rank-correlation ρ.
S_α; S_w, S_o	Saturation of phase α; water and oil saturations with $S_w + S_o = 1$.
t; t_n; N_t	Physical time; nth report time; number of report times (five in Phase I).
p_α; p	Phase pressure and discrete cell-pressure vector; report-time field p.
u_α	Darcy/superficial velocity of phase α.
q_α; q	Phase mass source per bulk volume; discrete net cell sources/sinks (sign as specified).
K; k; $\log k$	Permeability tensor; scalar rock permeability; its natural logarithm.
$k_{r\alpha}$; k_{rw}; k_{ro}	Relative permeability of phase α, water, and oil, respectively.
μ_α; μ_w; μ_o	Dynamic viscosities of phase α, water and oil.
g; z	Gravitational acceleration; elevation coordinate.
T_{ij}; w_e	Rock transmissibility across cells i and j; hydraulic face-conductance weight.
$\lambda_t(S_w)$	Total mobility as a function of water saturation; varies during nonlinear displacement.
f_w; PVI	Water fractional-flow function; dimensionless injected pore volumes.

Table A2. Hydraulic operators and distances.

Symbol/ abbreviation	Definition and usage
B; W	Signed cell-edge incidence matrix; diagonal edge-conductance matrix.
L; L_H; L_{rock}	Weighted Laplacian; hydraulic Laplacian; rock-only Laplacian before multiphase evolution.
L_{bb}; L_{bi}; L_{ib}; L_{ii}	Terminal/interior partition blocks of L (b: ports; i: interior).
K_w; $K_{w,\text{rock}}$	Kron-reduced well-response operator; its rock-transmissibility-only version.
$\text{tr}(K_w)$; z_K	Trace of K_w; its 15 upper-triangular, trace-normalised, standardised features.
r; s; $\mathcal{N}_5(r)$	Target reservoir; training reservoir; five nearest training reservoirs.

Continued

D_K; D_H	Five-neighbour Kron-response distance; 46-feature composite hydraulic distance.
D_{FD}; D_{geo}	Flow-diagnostic fingerprint distance; geological-property fingerprint distance.
$R_{eff}(a,b)$; e_a; e_b	Terminal effective resistance; unit port vectors for ports a and b.
λ_2; χ_H	Second-smallest non-zero relevant Laplacian eigenvalue; normalised susceptibility.
Θ_H; Π_H	Hydraulic training-support extrapolation factor; susceptibility-corrected score.
$Q_{0.95}$; ID	95th-percentile operator; in-distribution (training-domain) reference.
ΔL; L'	Perturbation of weighted Laplacian; perturbed operator $L + \Delta L$.
$\|\cdot\|_2$; $(\cdot)^T$; $(\cdot)^+$	Euclidean norm; transpose; Moore-Penrose pseudoinverse.

Table A3. Surrogate modelling and statistical measures.

Symbol/ abbreviation	Definition and usage
E_{ZS}; E_{roll}	Zero-shot prediction error; mean five-time pressure-saturation forecasting error.
E_p; E_S; E_x	Normalized pressure, saturation and generic state error.
x; $\hat{x}$	Reference physical field (p or S_w); corresponding surrogate prediction.
M; a_M; Γ_M; ε_M	Surrogate class; regression intercept; hydraulic generalization elasticity; regression residual.
ε	Positive numerical denominator stabilizer; distinct from ε_M.
ρ_{Sp}	Spearman rank correlation coefficient; distinct from phase density ρ_a.
R^2; AUROC; CI	Coefficient of determination; area under ROC curve; confidence interval.
PCA; RB; ResCNN; FNO	Principal-component analysis; reduced-basis; residual CNN; Fourier neural operator.
α_{reg}	Model regularization hyperparameter; not a fluid phase when used in training settings.
N; n; k_{NN}	Independent reservoirs; indexed samples/times; number of nearest neighbours ($k_{NN} = 5$).
MP; OOD; LOFO	Multiphase; out-of-distribution; leave-one-family-out.

Appendix A. Prespecified Confirmation Criteria for the Next 3-D Stage

The present study motivates a confirmatory 3-D campaign, but the Phase-I law must be frozen before that campaign. We recommend prespecifying the following: (i) K_w normalization and $k = 5$ neighbour rule unchanged; (ii) D_K threshold derived only from the current discovery domain; (iii) no reweighting by λ_2 unless registered as a separate hypothesis; (iv) primary endpoint defined before 3-D simulation; (v) at least one geology family and one well pattern fully excluded from training; and (vi) independent simulation with OPM Flow or another full-physics engine. A successful confirmation would require positive error-distance association, incremental value beyond flow diagnostics and calibration of failure probability on the external family.

Appendix B. Reservoir-Clustered Uncertainty and Matched Distance-Estimator Analyses

All new confidence intervals are percentile bootstrap intervals. Deep-run errors are first aggregated to one case-level mean and resampled by case, with all three run values retained together. Intervention resampling uses the 40 base reservoirs as units, with all five barrier severities kept together. No time frame or model run is counted as an independent reservoir.

Table B1. Deep-network case-level and per-run Spearman correlations (N = 200 unique reservoirs; 3000 clustered resamples).

Model	Case-mean ρ [95% CI]	Run 101/202/303 ρ	$\Delta\rho$ vs geology [95% CI]
ResCNN	0.454 [0.339, 0.558]	0.456/0.445/0.411	0.246 [0.101, 0.387]
FNO	0.486 [0.375, 0.582]	0.480/0.484/0.458	0.248 [0.100, 0.387]

Table B2. Paired intervention changes from severity 0 to severity 6 (N = 40 base reservoirs; 4000 clustered resamples). Percentage increase is the ratio of group mean errors, not the mean of base-specific percentage changes.

Model	ΔE_{roll} [95% CI]	% increase [95% CI]	ΔD_K [95% CI]
RB-Ridge	0.194 [0.129, 0.271]	72.1 [47.0, 103.3]	1.126 [0.92, 1.37]
RB-MLP	0.284 [0.240, 0.334]	116.2 [96.7, 137.7]	1.126 [0.92, 1.37]

Table B3. Equal-estimator sensitivity: both fingerprints use training-standardized Euclidean five-neighbour distance. Entries show Ridge/MLP/ExtraTrees. Six-family error-model CV uses external-family labels in calibration folds; blind external R^2 does not.

Metric	ρ across 600 cases	Six-family LOFO R^2	Blind external R^2
15-entry Kron D_K	0.546/0.566/0.594	0.374/0.379/0.470	0.167/0.222/0.281
46-entry composite Euclidean	0.500/0.511/0.548	0.404/0.322/0.390	0.157/0.191/0.226
46-entry composite Mahalanobis	0.522/0.530/0.568	0.376/0.287/0.265	0.170/0.194/0.242

Reproduction protocol: starting from run_study.py, regenerate dataset.npz using the immutable family seeds; for each discovery fold and each external split, refit both training-only StandardScalers before applying Euclidean $k = 5$; join case IDs (400 - 599) to deep-run CSVs, compute per-run Spearman and a case-level correlation on the three-run average; bootstrap unique case IDs. For interventions pivot by base ID and severity, resample entire base trajectories; see the accompanying analysis script and machine-readable output.