Supplementary Information

Evaluating Microbial Corrosion Factors in a Synthetic Pipeline Benchmark and a Laboratory Calibration Case

Fei Zhao 1, Zhao Zhang 1, Xinyu Li 2, Rui Liu 2, Yangrui Zhang 2, Guoxi He 2,*

This supplement reports the recalculated fixed-seed diagnostics, reconstruction checks, ordinal examples and source-linked AH36 calibration. Synthetic results use the revised common-area generator and the fixed input scales in Table S16. The accompanying code supplies all generated records and row-level predictions.

S1. Fixed-seed benchmark and coverage diagnostics

The fixed-seed sample contains 265 segments (seed 42), including 27 active cases. Its random test contains 2 active cases, while the pipeline-3 holdout contains 24. Tables S1–S4 and Figures S1–S5 use the revised generator. The main comparison repeats the analysis across thirty independent draws.

The fixed presets are XGBoost with 400 depth-6 trees, learning rate 0.05 and row/column sampling 0.9; random forest with 400 trees and all features per split; BP-MLP with hidden layers (64,32), ReLU/Adam, alpha 0.0001 and maximum 5,000 iterations; and RBF SVR with C = 100, epsilon = 0.01 and gamma = scale. Applicable random seeds are 42. BP-MLP and SVR use the fixed feature scales in Table S16, training-only feature centering and training-only target standardization. The source code records every split.

At N = 120 cells/mL, X = 35 mg/L, pH = 5.46, and P = 13.75 MPa, Equation (2) gives fbac = 1.421297. Figure S1 reconstructs the equation over N = 0–10,000 cells/mL and X = 0–200 mg/L at that fixed pH and pressure. The factor ranges from 1.267850 to 1.860267; the upper clipping bound is inactive on this grid.

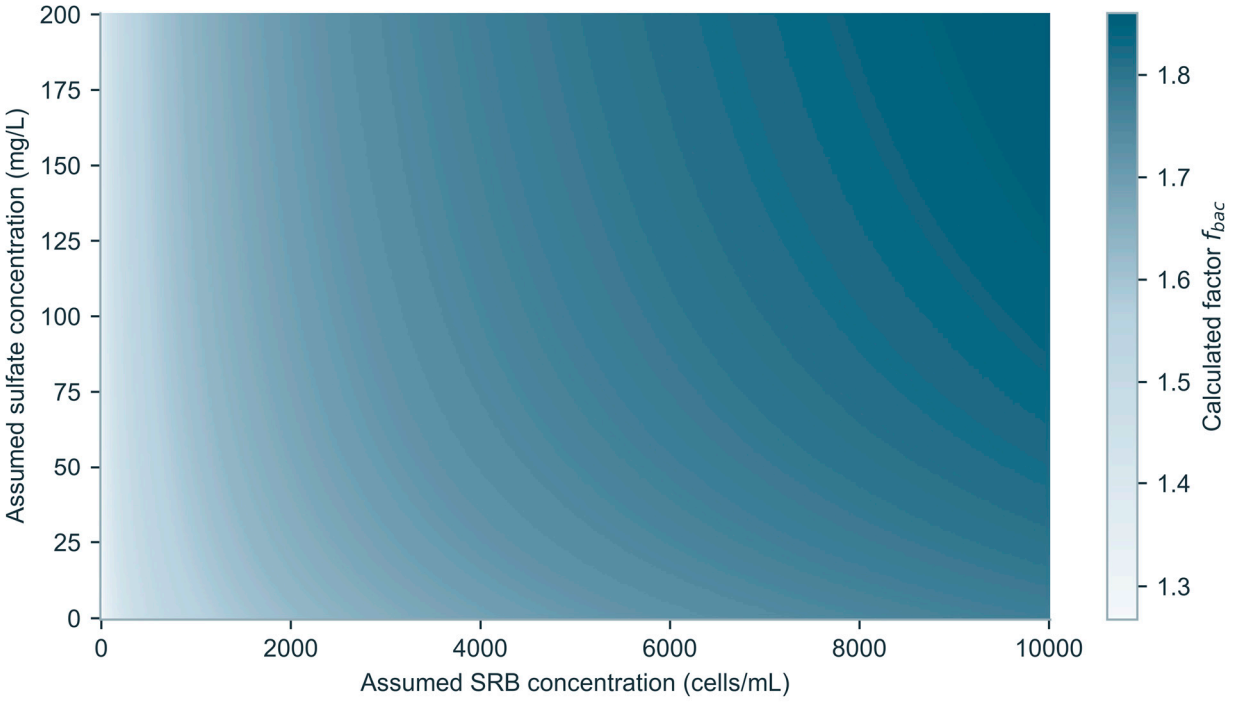


Figure S1. Response surface of the prescribed microbial factor in main Equation (2) at pH = 5.46 and P = 13.75 MPa.

Table S1 compares four fixed-preset regressors on the same 53-row revised random test. Random forest has the lowest RMSE in this realization. Figures S2 and S3 plot the corresponding exported metrics and prediction rows. The repeated grouped comparison in the main article evaluates the correction procedures across independent generator draws.

Table S1. Computed baseline performance on generated targets, n = 53. Units are defined in main Section 2.2.

Model	MSE	MAE	RMSE	R^2	MAPE (%)
XGBoost	0.00019756	0.010463	0.01406	0.9167	2.75
Random forest	0.00018551	0.010124	0.01362	0.9218	2.64
BP-MLP	0.00110123	0.021842	0.03318	0.5356	5.44
SVR	0.00216010	0.031467	0.04648	0.0890	7.89

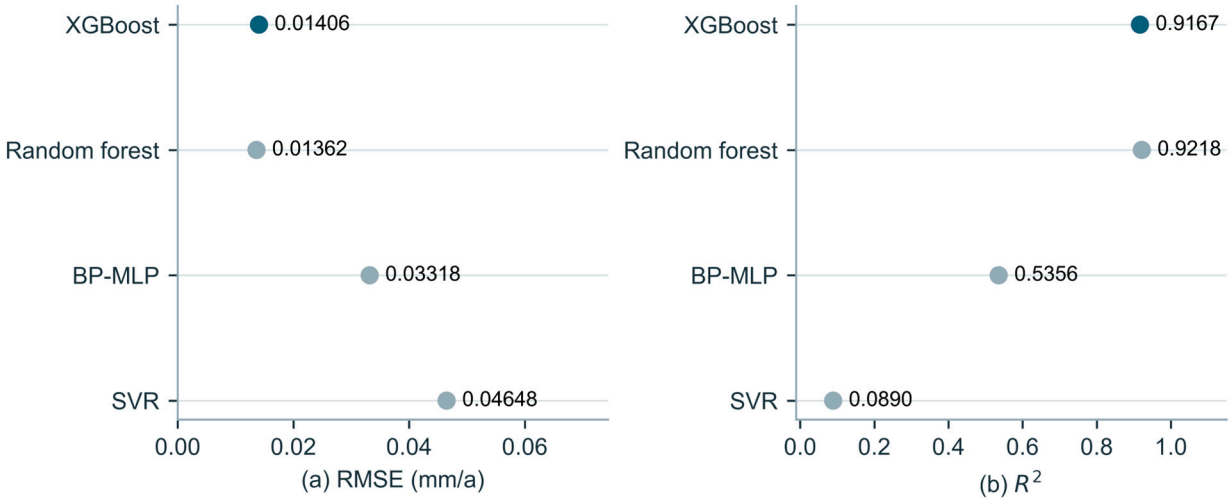


Figure S2. Computed RMSE and R^2 of the four regressors on the simulated random test, n = 53. RMSE labels use four significant figures, consistent with Table S1; the accompanying CSV files retain full numerical precision.

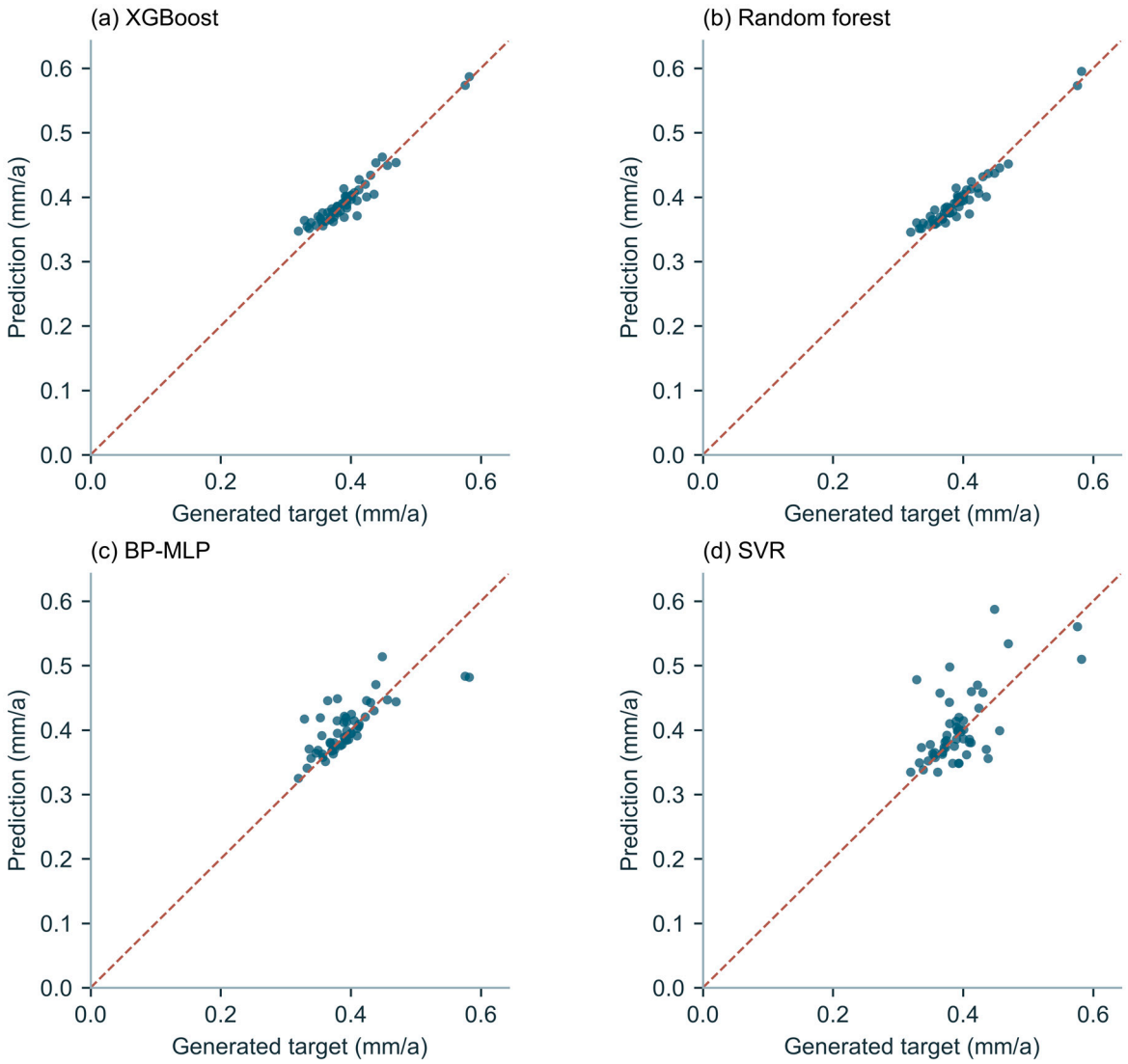


Figure S3. Predictions versus generated target rates for XGBoost, random forest, BP-MLP, and SVR on the same 53 simulated test rows. The dashed line is the 1:1 reference. Point coordinates come directly from exported predictions.

The random training set contains 25 active rows and the test contains 2. Direct-target and gated post-multiplication therefore differ on those active test rows. Table S2 retains all four reconstruction rules, using the generated total-rate target. The same factor is also supplied as a direct input in the main information-controlled comparison.

Table S2. Correction procedures on the revised 53-row random test, including 2 active rows. The factor is known from the generating scenario.

Prediction method	RMSE (mm/a)	MAE (mm/a)	R^2	MAPE (%)
Direct target	0.01406	0.010463	0.9167	2.75
Global post-factor	0.1751	0.172310	-11.9239	43.88
Gated post-factor	0.05783	0.021219	-0.4105	4.61
Factorized gate	0.01406	0.010609	0.9166	2.75

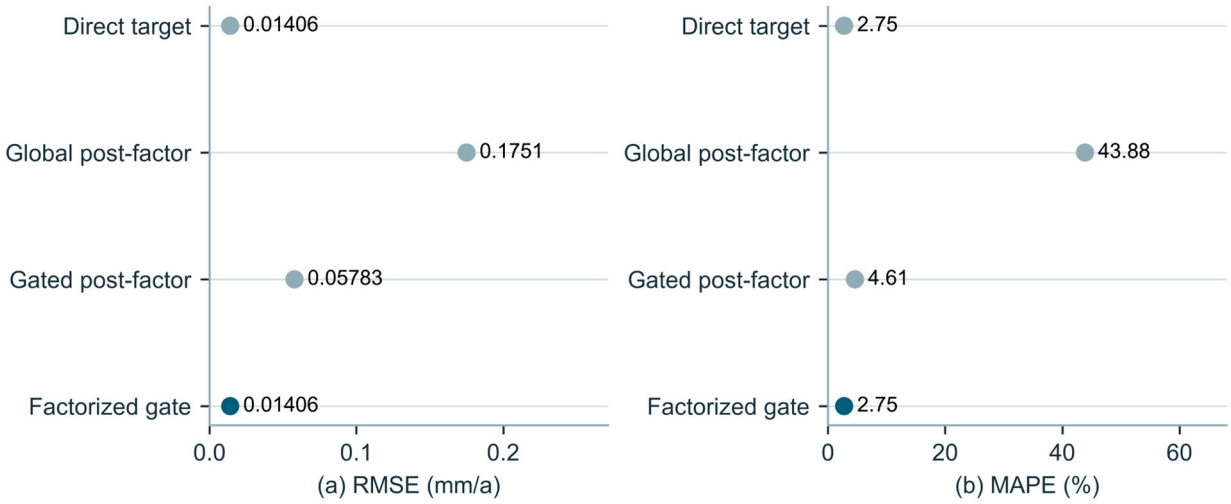


Figure S4. Recalculated correction errors on the revised random test. Direct and gated post-multiplication differ on active cases.

The fixed-seed threshold diagnostic evaluates 0.03, 0.05, 0.075, 0.10, 0.125 and 0.15 with the factorized fit and generated labels held fixed. It describes prediction-stage gate sensitivity. The complete thirty-draw grid is reported in Table S7, and the matched two-arm uncertainty analysis appears in Table S13.

The second fixed-seed fit trains on pipelines 1 and 2 (180 rows) and tests on pipeline 3 (85 rows). Training contains 3 active rows. Testing contains 61 inactive and 24 active rows. Pipeline and segment identities are disjoint between fitting and evaluation [1,2].

On the 85-row pipeline-3 test, Direct 15 RMSE is 0.04935 mm/year and factorized RMSE is 0.02398. Table S3 reports all metrics. The data archive additionally separates active and inactive errors. The main repeated protocol combines all three group folds within each of thirty generator draws.

Table S3. Fixed-seed results for artificial pipeline 3. Training uses 180 rows from the other two pipelines; testing uses 61 inactive and 24 active rows.

Prediction method	Training n	Test n	RMSE (mm/a)	R^2	MAPE (%)
Direct target	180	85	0.04935	0.7488	6.28
Factorized gate	180	85	0.02398	0.9407	4.32

S2. Slip distribution challenge

The separate evaluation-only challenge contains four artificial pipes of 50 rows each (seed 3042). Pipes 60–61 use S = 1 and pipes 62–63 use S = 8. All phase velocities, holdup, density, pressure and targets are generated jointly with that slip. The resulting sample has 100 inactive and 100 active rows. The fitted development models remain frozen. This experiment tests a prescribed change in phase slip and its associated covariate distribution.

Table S4 reports RMSE and R^2 for the active and inactive subsets and the complete 200-row challenge. Global and gated post-multiplication coincide on active rows. Figure S5 plots the errors from the same exported prediction rows.

Table S4. RMSE in mm/year and dimensionless R^2 under the jointly generated slip challenge. The 200 rows are evaluation-only records for frozen development fits.

Prediction method	Metric	Inactive / n = 100	Active / n = 100	All / n = 200
Direct target	RMSE	0.03500	0.07866	0.06088
Direct target	R^2	0.5613	0.2989	0.7243
Global post-factor	RMSE	0.1752	0.3050	0.2487
Global post-factor	R^2	-9.9860	-9.5422	-3.6015
Gated post-factor	RMSE	0.03500	0.3050	0.2171
Gated post-factor	R^2	0.5613	-9.5422	-2.5061
Factorized gate	RMSE	0.03451	0.05761	0.04749
Factorized gate	R^2	0.5734	0.6240	0.8323

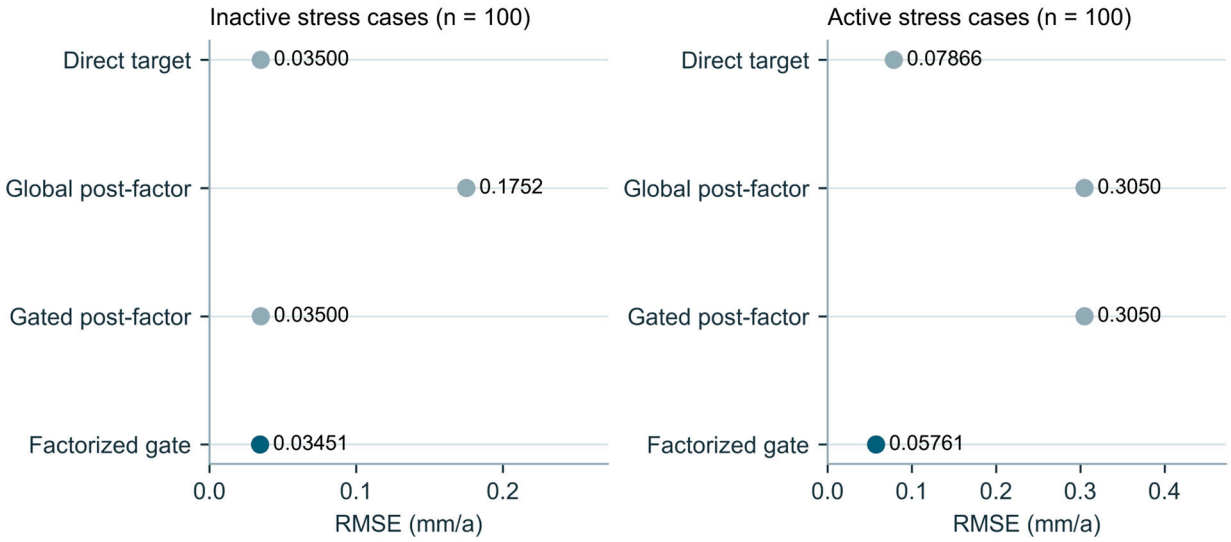


Figure S5. XGBoost errors on the slip-distribution challenge, with 100 inactive and 100 active cases. Each phase state is recalculated consistently with the selected slip ratio.

On active challenge rows, direct and factorized XGBoost RMSE are 0.07866 and 0.05761 mm/year. The complete R^2 values remain in Table S4. Negative values, where present, indicate larger error than the corresponding evaluation-subset mean reference.

The accompanying challenge file also contains BP-MLP and Ridge predictions, subset counts and metrics. Their fixed input divisors are identical to those in the nominal synthetic fits. The challenge uses independent pipeline identities and is reported separately from the repeated nominal evaluation.

S3. Complete reconstruction and sensitivity diagnostics

Tables S5–S9 retain the four original prediction procedures, including global and gated post-multiplication of total-rate predictions. These diagnostic controls quantify double application of a contribution already represented in the total-rate target. The main article separately provides direct models receiving the same known factor or raw

microbial descriptors. Reconstruction grids freeze models and targets; generative grids regenerate targets and refit every split. They answer different sensitivity questions.

Table S5. Pooled RMSE in mm/year across thirty revised generator draws. Random tests contain 394 active and 1,196 inactive rows; grouped tests contain 1,879 active and 6,071 inactive rows. Methods use identical rows within each protocol.

Protocol	Method	All	Inactive	Active
Random	Direct target	0.02131	0.01461	0.03443
Random	Global post-factor	0.2119	0.1746	0.2978
Random	Gated post-factor	0.1488	0.01461	0.2978
Random	Factorized gate	0.01583	0.01386	0.02069
Pipeline holdout	Direct target	0.04230	0.03006	0.06820
Pipeline holdout	Global post-factor	0.2127	0.1785	0.2973
Pipeline holdout	Gated post-factor	0.1469	0.03006	0.2973
Pipeline holdout	Factorized gate	0.03312	0.02981	0.04207

Table S6. Distribution of realization-level RMSE (mm/a) over thirty independent generator draws. IQR gives the 25th–75th percentiles; group folds are combined within each draw. These are descriptive synthetic distributions.

Protocol	Strategy	Median	IQR	Range
Random	Direct target	0.01829	0.01691–0.02049	0.01233–0.05307
Random	Global post-factor	0.2096	0.1989–0.2227	0.1746–0.2765
Random	Gated post-factor	0.1449	0.1106–0.1744	0.07470–0.2430
Random	Factorized gate	0.01540	0.01411–0.01768	0.01093–0.01954
Grouped	Direct target	0.04029	0.03286–0.04608	0.02487–0.07121
Grouped	Global post-factor	0.2051	0.1970–0.2259	0.1712–0.2675
Grouped	Gated post-factor	0.1395	0.1168–0.1644	0.05173–0.2167
Grouped	Factorized gate	0.03137	0.02563–0.03692	0.01708–0.05827

Table S7. Pooled pipeline-holdout RMSE in mm/a under prediction-stage perturbations. Column labels scale the excess factor fbac − 1; the original target and training factor remain fixed.

Threshold	Scale 0.5	Scale 1.0	Scale 1.5
0.050	0.08969	0.1417	0.2159
0.075	0.07695	0.1076	0.1656
0.100	0.05846	0.03312	0.05917
0.125	0.09723	0.09588	0.09838
0.150	0.1022	0.1022	0.1022

The main source-cell extraction yields thirty-two experimental coupon observations. Table S8 retains all original model/strategy coupon RMSE, including the diagnostic post-factors. Main Table 8 adds complete-interaction and saturation controls and primary treatment-replicate mean errors. Factor calibration uses training-replicate measurements exclusively.

Table S8. Out-of-replicate experimental RMSE (mm/a) for all learners and strategies, n = 32 coupon predictions. Factors use training measurements only; the full metrics for each held-out replicate and treatment subset are supplied.

Learner	Direct total	Global factor	Biotic factor	Factorized
Cell mean	0.003515	0.2172	0.2149	0.003515
XGBoost	0.004945	0.2105	0.2071	0.004126
Ridge	0.01428	0.1825	0.1645	0.004291

Table S9. Pooled XGBoost group-holdout RMSE (mm/a) after regenerating each scenario and refitting the models. Every row has 7,950 predictions from thirty draws; n(HL > H*) is a coverage count, not an inactive/active split for smooth gates. D, direct; G, global post-factor; H, gated post-factor; F, factorized gate.

Scenario	n(HL > H*)	D	G	H	F
Reference	1879	0.04230	0.2127	0.1469	0.03312
N half	1879	0.04009	0.1732	0.1107	0.03229
N one and half	1879	0.04714	0.2568	0.1886	0.03398
P half	1879	0.04081	0.1613	0.1113	0.03228
P one and half	1879	0.04501	0.2663	0.1861	0.03399
Threshold half	7090	0.05397	0.2565	0.2485	0.04140
Threshold one and half	4	0.02972	0.1784	0.02931	0.02931
Centered pressure	1879	0.04031	0.1139	0.08101	0.03149
No pressure term	1879	0.04026	0.1135	0.08070	0.03149
Smooth delta 0.005	1879	0.04440	0.2077	0.1230	0.03312
Smooth delta 0.020	1879	0.04268	0.2102	0.1005	0.03388

All evaluated factors remain below the upper clipping bound. The pressure alternatives isolate the contribution of the original reference offset, and the smooth alternatives examine transition-width dependence. All prespecified scenarios are retained in the comparison.

S4. Attribution of the revised synthetic model

Gain importance and Shapley additive explanations (SHAP) quantify attribution for the direct-target XGBoost model [3,4]. Native XGBoost pred_contribs returns fifteen contributions and a bias term [5], whose row sums are checked against regression predictions. Summary-plot colors encode each feature's value rescaled to the minimum and maximum of the test set, and vertical jitter separates points only. With correlated inputs, attribution depends on the fitted model and its treatment of dependence [6]; we interpret these outputs as descriptions of the fitted synthetic prediction function.

Liquid holdup, liquid superficial velocity and liquid flow rank highest by normalized gain in the revised direct-target XGBoost model (Figure S6). Figure S7 reports individual TreeSHAP contributions from the same fitted model.

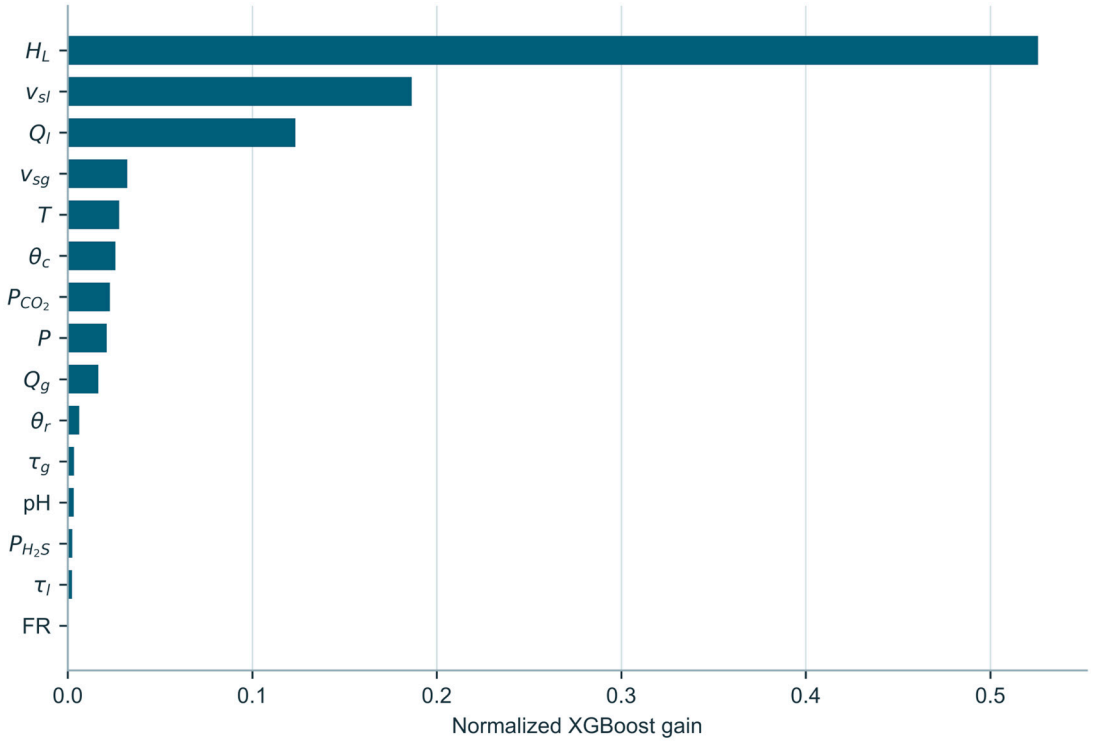


Figure S6. Normalized gain importance of the direct-target XGBoost model trained on simulated data. All fifteen covariates are shown.

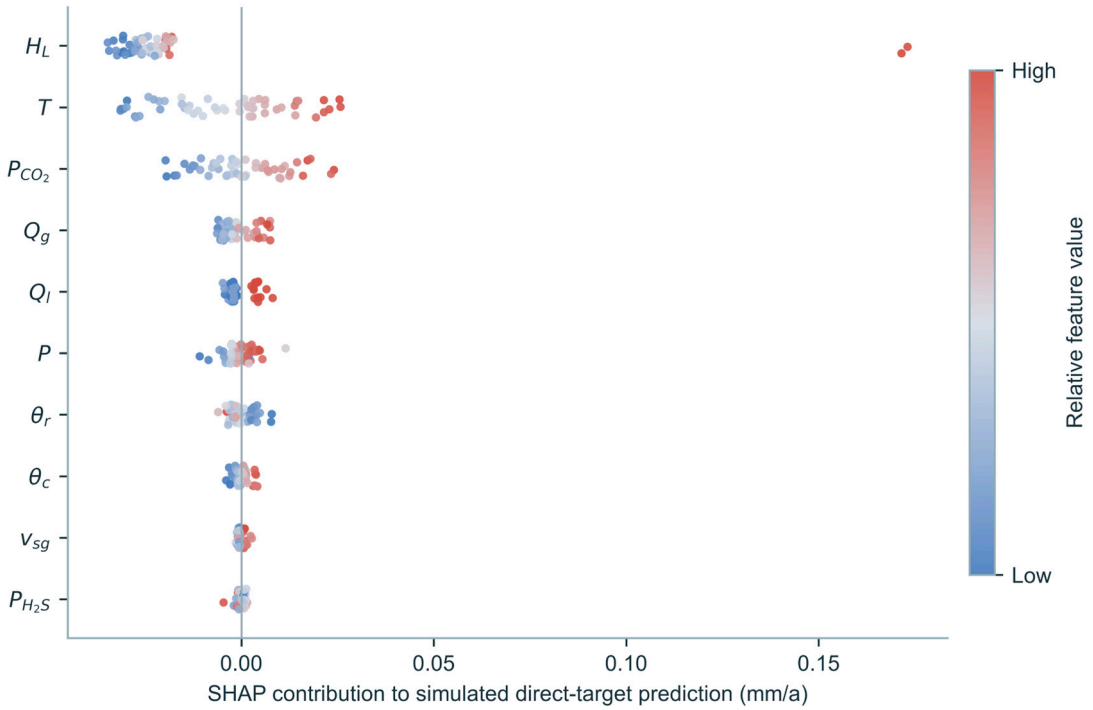


Figure S7. Native TreeSHAP contributions for the simulated direct-target model, showing the ten largest mean absolute contributions on the 53 test rows. Each feature contribution is in mm/a. Blue and red indicate low and high relative test-set feature values; display jitter has no effect on horizontal coordinates.

The maximum absolute difference between the contribution sum, including the bias term, and the XGBoost prediction was 5.36e-07 mm/year. The attribution describes the fitted function over the revised synthetic feature domain.

S5. Illustrative ordinal network and repeat inspection

Rate classes A–E and nominal wall-loss grades 1–5 use the boundaries in Table S10. The constant-rate proxy is L = 100 × predicted rate × assigned service age / assigned wall thickness. It represents a hypothetical trajectory at the current rate over the full assigned age. The ordinal matrix sums zero-based rate-class and loss-grade indices: scores 0–2 are low, 3–4 medium, 5–6 higher and 7–8 high, with all 25 cells in Figure S8. Both axes share the rate input, so the sum aggregates dependent severity descriptors. These labels define the reporting convention; probability–consequence risk and quantitative reliability require additional growth and defect-distribution models [7,8].

The proxy specifies a spatially averaged, constant-rate trajectory over the assigned service age. Localized pitting, protective corrosion-product scales, microbial incubation and changing flow or chemistry introduce spatial and temporal growth processes. Their assessment requires an appropriate rate history or growth model. Here, the constant-rate assumption is held fixed to evaluate the ordinal reporting rule.

Table S10. Exact class boundaries for the illustrative ordinal screening matrix. The loss quantity is a constant-rate proxy, and a in rate units denotes year.

Rate class	Rate interval (mm/a)	Loss grade	Nominal loss interval (%)
A	$v_{corr} < 0.13$	1	$L < 10$
B	$0.13 \le v_{corr} < 0.21$	2	$10 \le L < 15$
C	$0.21 \le v_{corr} < 0.31$	3	$15 \le L < 20$
D	$0.31 \le v_{corr} \le 0.38$	4	$20 \le L < 30$
E	$v_{corr} > 0.38$	5	$L \ge 30$

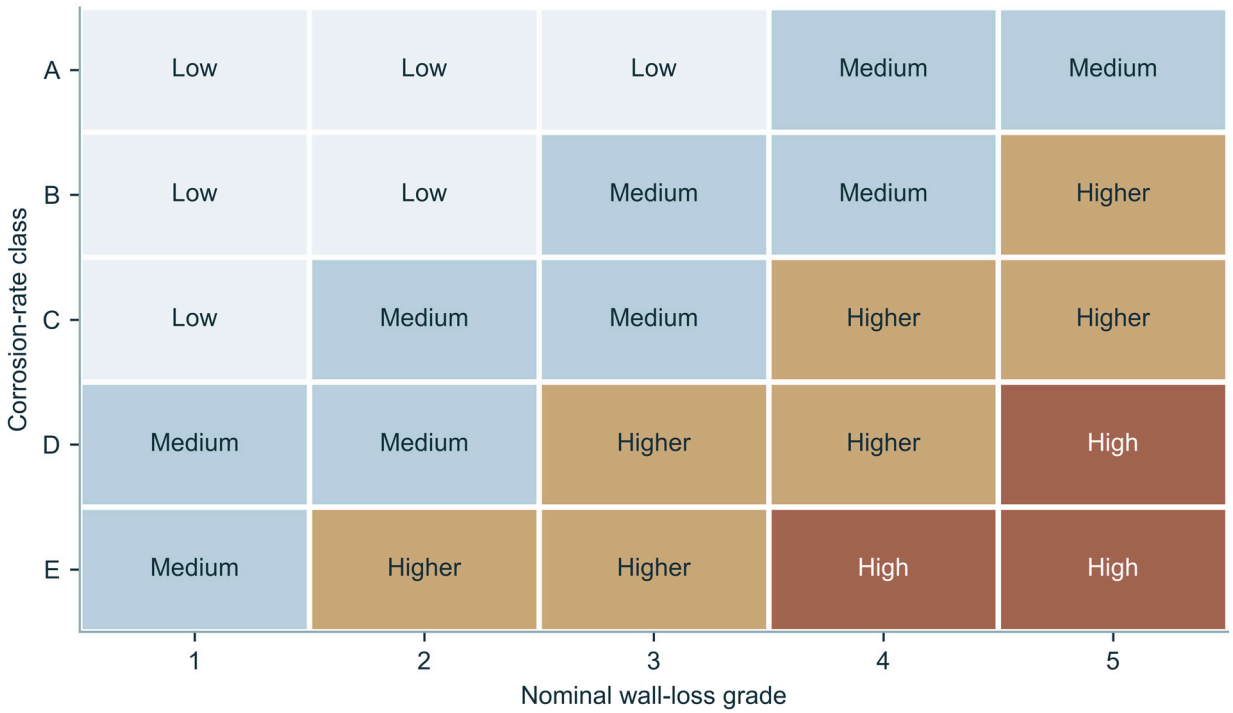


Figure S8. Complete illustrative ordinal matrix. Labels denote assigned screening severity under the sum rule.

The primary factorized model, fitted on the 212 random training rows, was frozen and applied to a separate network realization containing 15,154 segments in 56 artificial pipelines. Pipeline sizes were assigned from uniform weights in [0.7,1.3] using seed 2026, normalized to the total, rounded down, and completed by assigning remaining segments in pipeline order. Network covariates were generated with seed 1042. Pipeline and segment identifiers are local to each dataset realization.

The declared matrix places 0 network segments in Low, 0 in Medium, 47 in Higher and 15,107 in High, totaling 15,154. The class proportions follow from the revised generated rates, assigned service ages and wall thicknesses. This example checks the ordinal rule on the generated network; its class frequencies describe that synthetic realization.

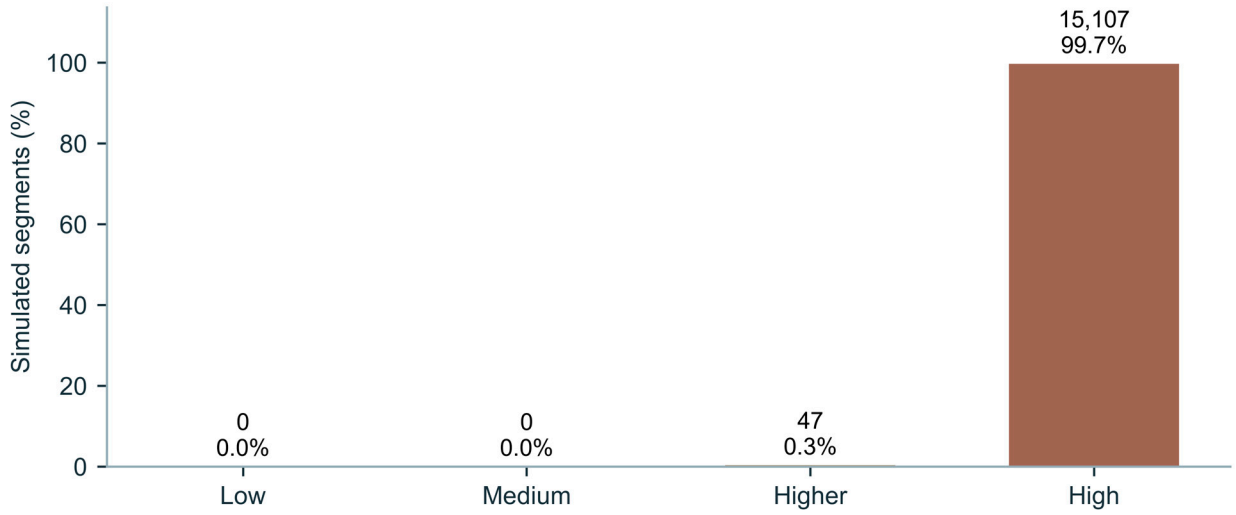


Figure S9. Class counts and percentages for 15,154 synthetic network segments. Risk labels refer to assigned ordinal screening severity.

A separate simulated repeat-inspection set uses seed 2042 and comprises 487 rows in artificial pipelines 57–59, with 163, 162 and 162 rows and assigned inspection intervals of four, five and six years. These pipeline identifiers are excluded from the development sample and network. Initial defect depths are assigned as D0 = 0.1 + 0.4 σ(10 θr) mm. Final depths D1 equal D0 plus the generated target rate multiplied by the assigned interval. Consequently, the derived rate in Equation (S1), with depths in millimetres and times in years, is algebraically the same noisy target as Equation (4):

$$v_{repeat} = \frac{D_1 - D_0}{t_1 - t_0} \quad \text{(S1)}$$

The repeat-inspection check uses exact matching and generated depths. Agreement is the confusion-matrix diagonal total divided by 487. Recall is calculated separately for each represented class; classes with zero support have undefined recall. The macro recall averages represented classes only. Rate correlation is Pearson r.

The revised repeat-inspection sample gives RMSE = 0.02269 mm/year, Pearson r = 0.9763, classification agreement 99.59% and macro recall 66.67% over the two represented classes. High accounts for 484 of 487 rows. Always assigning High would agree with 99.38% of these generated labels, so agreement is interpreted alongside support and recall (Table S11). Low and Medium have zero support. The check assesses numerical consistency on new synthetic records.

Table S11. Class support and recall on simulated repeat-inspection records. Evaluation uses pipeline identifiers held out from model fitting.

Generated class	Support n	Correctly classified	Recall (%)
Low	0	0	N/A
Medium	0	0	N/A
Higher	3	1	33.3
High	484	484	100.0

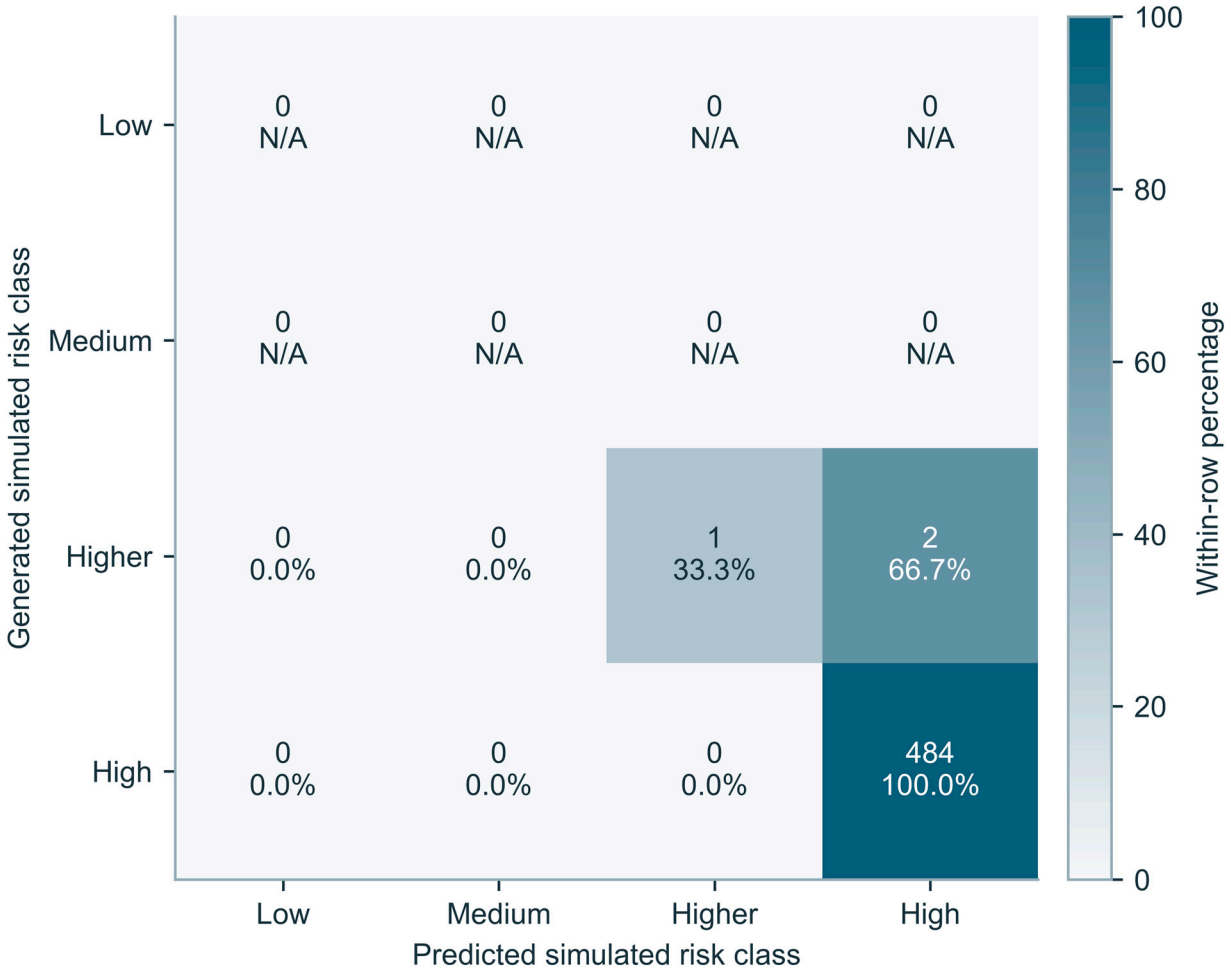


Figure S10. Confusion matrix for 487 revised synthetic repeat-inspection records. Rows are generated classes and columns predicted classes. Shading gives row percentages; N/A marks rows with zero support.

Both ordinal examples use the same generated rates and the constant-rate trajectory defined above. Reliability assessments also use corrosion growth, defect-depth distributions, residual strength and consequences [8]. Compressing scores into a few classes affects comparisons between rankings [7].

S6. Source-linked experimental data and extended controls

Measured observations are taken from Ivanovich et al. [9] using the authors' original Weight loss.xlsx workbook. The derived CSV preserves worksheet/cell identifiers, exposure duration, units and the original experimental descriptors. Published cached rates and independent source-mean/formula checks accompany the extraction. The author repository provides the original workbooks. The empirical_mic, submission_empirical_fairness and source-reproduction files document retrieval, extraction, calibration and complete model comparisons. Figure S11 retains the coupon-level diagnostic comparison; the main article reports the complete-interaction controls.

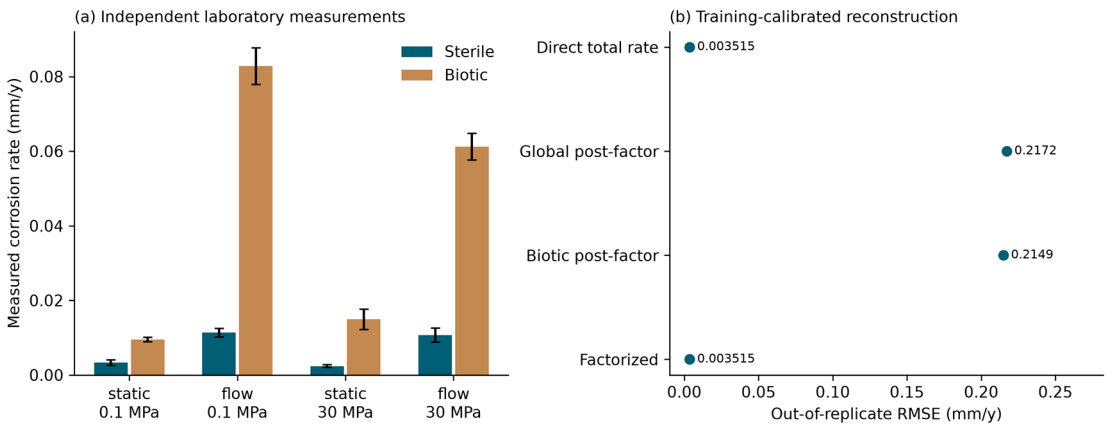


Figure S11. Author-generated plot from the open weight-loss workbook linked by Ivanovich et al. [9]. Panel (a) shows measured coupon means and sample SD across four coupons per treatment. Panel (b) shows out-of-replicate error for the saturated cell-mean reference under the four reconstruction rules. Source-cell identifiers and prediction rows accompany the analysis.

Cell-mean equivalence establishes direct and factorized prediction as equivalent representations of a saturated response model. The XGBoost and Ridge comparisons show how the fitted approximation structure affects the result. Calibrated context factors supply treatment interactions to the factorized Ridge model, while the direct main-effects specification in Table S8 represents additive effects. These comparisons connect measured factor calibration with the model-dependent behavior observed in the synthetic experiments.

S7. Additional factor and uncertainty analyses

The following tables use the revised common-area generator. Active cases satisfy the prescribed HL > 0.1 threshold. The same outer pipeline folds and seeds are retained for every procedure. Total-rate errors are in mm/year. Table S16 records the fixed feature scales. Code and row predictions accompany all comparisons.

Table S12. Paired direct-minus-factorized RMSE by independent generator draw and counts of wrong-band assignments. Order-statistic intervals have at least 95% coverage. The synthetic band cutoffs are 0.025, 0.13 and 0.25 mm/year.

Subset	Factor wins/losses	Median Δ	Interval for Δ	Sign p	Band errors D/F
Grouped active	29/1	0.01768	0.01208 to 0.02708	5.77e-08	0/0
Grouped inactive	16/14	0.0004653	-0.002038 to 0.001187	0.856	0/0
Grouped all	29/1	0.005691	0.004617 to 0.006731	5.77e-08	0/0
Random active	29/1	0.007688	0.005442 to 0.01420	5.77e-08	0/0
Random inactive	18/12	0.0005258	-0.0001554 to 0.001061	0.362	0/0
Random all	30/0	0.002782	0.001529 to 0.004086	1.86e-09	0/0

Table S13. Shared factor perturbations on grouped holdouts. In the frozen fits the altered factor enters both arms at test time. In refitted models it enters both arms at training and test time. The target stays fixed.

Fit	Relative error	Subset	Direct RMSE	Factorized RMSE
Frozen	±10%	Active	0.08455	0.07227
Frozen	±10%	Inactive	0.02951	0.02981
Frozen	±10%	All	0.04852	0.04373
Frozen	±20%	Active	0.08582	0.1242
Frozen	±20%	Inactive	0.02951	0.02981
Frozen	±20%	All	0.04905	0.06578
Refit	±10%	Active	0.06731	0.07627
Refit	±10%	Inactive	0.02961	0.03096
Refit	±10%	All	0.04172	0.04590
Refit	±20%	Active	0.06731	0.1391
Refit	±20%	Inactive	0.02961	0.03303
Refit	±20%	All	0.04172	0.07353

Table S14. Grouped RMSE for the revised nominal and independent microbial-channel generators. Each gives 1,879 active and 6,071 inactive predictions. Compare procedures within a generator because their targets differ.

Generator	Procedure	Active	Inactive	Pooled
Nominal	Direct 15 + g XGBoost	0.06099	0.02951	0.03930
Nominal	Factorized XGBoost	0.04207	0.02981	0.03312
Nominal	Direct 15 + g Ridge	0.07225	0.03282	0.04535
Nominal	Factorized Ridge	0.03959	0.02723	0.03061
Nominal	Direct log target	0.06821	0.02979	0.04216
Nominal	Log offset	0.04190	0.02992	0.03314
Nominal	Additive residual	0.04754	0.02945	0.03459
Nominal	Tuned direct 15 + g XGBoost	0.06166	0.02943	0.03950
Nominal	Tuned factorized XGBoost	0.04166	0.02904	0.03247
Independent	Direct 15 XGBoost	0.07096	0.02954	0.04308
Independent	Direct 15 + g XGBoost	0.06705	0.02978	0.04171
Independent	Factorized XGBoost	0.04012	0.02981	0.03254

Table S15. Same-factor Ridge RMSE for active cases by marginal factor coverage.

Training g range	Active n	Direct 15 + g	Factorized
Within	1741	0.06547	0.03929
Outside	138	0.1304	0.04313

Table S16. Fixed feature divisors for synthetic Ridge, BP-MLP and SVR. Centering means are estimated from each training set. The empirical calibration retains its separately specified training-only standardization.

Feature	Divisor in feature units
T (°C)	23.8
Qg (10^4 m^3/d)	4.24
Ql (m^3/d)	1173
P (MPa)	1
PH_2S (MPa)	0.0007
PCO_2 (MPa)	0.0004
vsg (m/s)	8.73
vsl (m/s)	0.209
HL	0.18
θc (rad)	0.05
θr (rad)	0.216
pH	0.75
FR	2
τg (Pa)	20
τl (Pa)	1
g when present	2

References

1. Wenger, S.J.; Olden, J.D. Assessing transferability of ecological models: An underappreciated aspect of statistical validation. *Methods Ecol. Evol.* **2012**, 3, 260–267. https://doi.org/10.1111/j.2041-210X.2011.00170.x.
2. Scikit-learn developers. Cross-validation: Evaluating estimator performance. Available online: https://scikit-learn.org/stable/modules/cross_validation.html (accessed 3 October 2026).
3. Lundberg, S.M.; Lee, S.-I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **2017**, 30, 4765–4774. https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.
4. Lundberg, S.M.; Erion, G.G.; Lee, S.-I. Consistent Individualized Feature Attribution for Tree Ensembles. *arXiv* **2018**, arXiv:1802.03888. https://doi.org/10.48550/arXiv.1802.03888.
5. XGBoost developers. Python API reference. Available online: https://xgboost.readthedocs.io/en/stable/python/python_api.html (accessed 3 October 2026).
6. Aas, K.; Jullum, M.; Løland, A. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artif. Intell.* **2021**, 298, 103502. https://doi.org/10.1016/j.artint.2021.103502.
7. Cox, L.A., Jr. What's Wrong with Risk Matrices? *Risk Anal.* **2008**, 28, 497–512. https://doi.org/10.1111/j.1539-6924.2008.01030.x.
8. Caleyo, F.; Velázquez, J.C.; Valor, A.; Hallen, J.M. Probability distribution of pitting corrosion depth and rate in underground pipelines: A Monte Carlo study. *Corros. Sci.* **2009**, 51, 1925–1934. https://doi.org/10.1016/j.corsci.2009.05.019.
9. Ivanovich, N.; Marsili, E.; Shen, X.; Messinese, E.; Marcos; Rajala, P.; Lauro, F.M. Exploring the impact of flow dynamics on corrosive biofilms under simulated deep-sea high-pressure conditions using bio-electrochemostasis. *Front. Microbiol.* **2025**, 16, 1540664. https://doi.org/10.3389/fmicb.2025.1540664.