J. Hydrol. Hydromech., 74, 2026, 2, 212–223
https://doi.org/10.2478/johh-2026-0015
©2026 Zeroual Abdelatif and Fourar Ali, published by Slovak Academy of Sciences.
This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International Licence.

Journal of Hydrology and Hydromechanics

Geometric shift–aware uncertainty quantification for discharge coefficient prediction of triangular planform weirs

Zeroual Abdelatif[1,2*], Fourar Ali[3]

[1] Department of Hydraulic, Faculty of Sciences and Applied Sciences, University of Oum El Bouaghi, Algeria.
[2] Sustainable Development and Environmental Protection Laboratory (SDEPL), University of Oum El Bouaghi, Algeria.
[3] Laboratory of Applied Research in Hydraulics, Department of Hydraulic, Faculty of Technology, University of Batna 2, Batna, Algeria.
[*] Corresponding author. E-mail: abdelatif.zeroual@univ-oeb.dz

Abstract: Reliable estimation of the discharge coefficient *Cd* is central to sharp-crested weir hydraulics, yet data-driven models are commonly evaluated with random train-test splits that can overstate generalization when experiments are structured by discrete geometries. Using a reconstructed run-level dataset for triangular planform weirs (n=123) spanning vertex angles $\theta \in \{30^o, 60^o, 90^o, 120^o, 150^o, 180^o\}$, this study benchmarks tabular regression under (i) random 80/20 splitting, which represents interpolation within a mixed-geometry dataset, and (ii) a leave-one-angle-out (LOAO) protocol that enforces geometric domain shift. Dimensionless predictors are derived from Buckingham-π reasoning and formulated to remain numerically stable as $\theta \to 180°$. Under random splitting, gradient boosting achieves high accuracy, whereas LOAO reveals substantially weaker generalization, with the largest errors near the $\theta = 180°$ regime transition. For uncertainty quantification, global split conformal intervals are near-nominal under random splits but severely under-cover under LOAO, indicating that calibration residuals are not transferable across geometries. CV-pooled residual calibration restores LOAO coverage at the cost of substantially wider intervals. Finally, combining k-nearest-neighbor distance-based out-of-distribution detection with a two-tier interval policy yields intermediate operating points that better balance reliability and sharpness when deploying beyond the calibration envelope. Overall, the results indicate that apparently strong interpolation performance can conceal important geometric extrapolation failures, and that safe deployment on novel weir geometries requires explicitly shift-aware uncertainty quantification.

Keywords: Triangular planform weir; Discharge coefficient; Domain shift; Conformal prediction; Out-of-distribution detection; Selective prediction.

1 INTRODUCTION

Reliable estimation of the discharge coefficient *Cd* is fundamental to sharp-crested weir hydraulics for discharge measurement, diversion design, and operational water-level control. In practical rating relations, *Cd* aggregates deviations from ideal flow assumptions, including approach non-uniformity, streamline curvature, boundary-layer effects, local separation, and aeration, into a single coefficient linking discharge and head. As a result, uncertainty in Cd directly affects discharge capacity estimates and flow accounting.

Classical empirical or semi-analytical formulations remain widely used, yet their transferability often degrades when geometry or operating conditions depart from calibration regimes, a concern that becomes acute for nonstandard planforms where geometric changes can alter contraction and loss mechanisms. The continuing emphasis on case-specific calibration in recent studies underscores that *Cd* is not universally transferable across configurations without explicit treatment of regime differences (Haghbin and Sharafati, 2022).

Dimensional analysis offers a principled route to reduce scale dependence and improve comparability across conditions. For free-surface flows in which inertia–gravity balance dominates, Froude similarity motivates nondimensional predictor groups that compress hydraulic and geometric influences into interpretable variables. However, nondimensionalization does not guarantee invariance when a change in geometry induces a qualitatively different flow structure, and care is also required in feature construction near geometric limits where standard trigonometric forms can become numerically ill-conditioned (Buckingham, 1914).

In parallel, machine learning (ML) has gained increasing traction for predicting discharge coefficients, owing to its ability to approximate complex nonlinear relationships without imposing a predefined functional form. Recent studies report strong point-prediction performance across a range of complex weir configurations, including optimized predictive models for hydraulic weirs and applications to semicircular labyrinth weirs (Belaabed et al., 2025; Asgharzadeh-Bonab et al., 2025). More broadly, review papers highlight the rapid expansion of ML-based approaches in modelling hydraulic coefficients and flow-control structures (Haghbin and Sharafati, 2022; Seyedian et al., 2023). While these data-driven models often demonstrate high interpolation accuracy, they also raise important questions regarding their generalization capability beyond the calibration domain. In response, uncertainty-aware evaluation is beginning to emerge in the weir literature, emphasizing the need to assess not only predictive accuracy but also the reliability of model outputs for engineering decision-making (Seyedian and Kisi, 2024).

A key methodological issue is that model performance is typically reported under random train-test splits, which largely measure interpolation within a pooled dataset. Experimental hydraulics datasets are frequently structured by discrete geometries and controlled test protocols, so mixing all geometries in both training and testing can leak geometry-

specific structure and yield overly optimistic estimates of generalization. For triangular planform weirs, the vertex angle θ naturally defines a geometry domain, motivating a geometry-held-out evaluation in which all runs at one angle are withheld during training to explicitly test deployment on an unseen planform regime.

This geometry-shift setting also raises a reliability challenge for uncertainty quantification (UQ). Conformal prediction provides distribution-free prediction intervals with finite-sample marginal coverage under exchangeability assumptions, but under geometric domain shift the calibration residual distribution may not transfer, leading to under-coverage (Angelopoulos and Bates, 2023; Tibshirani et al., 2019; Barber et al., 2023). Practical remedies include more conservative calibration strategies, but these can widen intervals substantially, motivating additional mechanisms (such as out-of-distribution (OOD) detection and selective prediction) to manage the reliability–sharpness trade-offs associated with empirical coverage and interval width in a deployment-relevant way.

Accordingly, this study establishes a geometric shift-aware benchmarking framework for Cd prediction of triangular planform weirs. A run-level dataset is reconstructed from Kumar et al. (2011), a stable dimensionless feature representation is derived, tabular regressors are evaluated under both random splitting and leave-one-angle-out (LOAO), and conformal uncertainty-quantification strategies are benchmarked under geometry shift. A kNN-based out-of-distribution (OOD) detector is then combined with a two-tier interval policy to support safer use beyond the calibration envelope.

The objectives of the study are to reconstruct a run-level dataset suitable for benchmark evaluation, derive a dimensionless predictor set that remains numerically stable at θ = 180°, quantify geometry-driven generalization loss under LOAO relative to random splits, compare conformal prediction variants under geometric shift, and evaluate OOD-aware two-tier selective prediction for safer deployment. The central hypotheses are that LOAO reveals materially lower performance than random splits, that baseline conformal prediction under-covers under LOAO, that conservative residual pooling can restore empirical coverage at the cost of wider intervals, and that OOD-aware two-tier policies improve deployment safety in out-of-support regimes.

2 MATERIALS AND METHODS

2.1 Data source and hydraulic geometry

The dataset used in this benchmark was reconstructed from the experimental series reported by Kumar et al. (2011) and comprises n=123 run-level observations for sharp-crested triangular planform weirs with vertex angles $\theta \in \{30°,60°,90°,120°,150°,180°\}$. The experiments were conducted in a horizontal rectangular channel of width B≈0.28 m. In the original notation of Kumar et al. (2011), crest height and upstream head over the crest were denoted by *w* and *h*, respectively. For notation consistency throughout the present manuscript, these quantities are written here as *P* and *H*. Fig.1 defines the hydraulic and geometric variables adopted in this study and clarifies this notation harmonization.

For each vertex angle θ, the crest length L follows from the planform geometry as

$$L = \frac{B}{\sin(\theta/2)} \tag{1}$$

Under free-flow conditions, discharge over the triangular planform weir is related to the discharge coefficient *Cd* by:

$$Q = \frac{2}{3} C_d \sqrt{2g} L H^{3/2} \tag{2}$$

where Q is discharge, g is gravitational acceleration, L is crest length, and H is upstream head over the crest. Equivalently, the discharge coefficient can be written as

$$C_d = \frac{3Q}{2\sqrt{2g}\, L\, H^{3/2}} \tag{3}$$

The approach depth is defined as:

$$y = P + H \tag{4}$$

The reconstructed dataset, although modest in size, provides dense hydraulic coverage across six discrete vertex-angle regimes. Because all observations originate from a single experimental campaign, the observed shifts in *Cd* can be interpreted primarily as geometric effects rather than as inter-laboratory artifacts. At the same time, the use of a single source is an important limitation that constrains generalizability beyond the present benchmark.

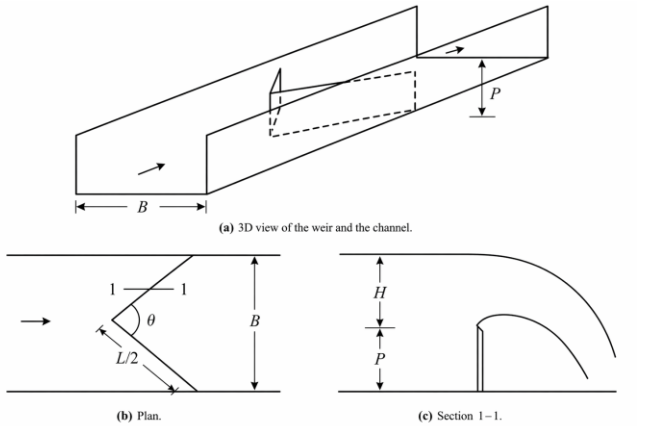


Fig. 1. Definition sketch of the sharp-crested triangular planform weir and notation used in the present study.
Note: (a) 3D view of the weir and the channel; (b) plan view showing channel width B, half crest length L/2, and vertex angle θ; and (c) Section 1-1 showing crest height P and upstream head over the crest H. Relative to Kumar et al. (2011), the notation is harmonized as P ≡ w and H ≡ h. The total crest length is L = 2(L/2).

2.2 Dimensional analysis and dimensionless predictors

The overall methodology adopted in this study is summarized in Fig. 2. To ensure scale independence and improve interpretability, the raw hydraulic and geometric variables were mapped to a set of dimensionless predictors derived from Buckingham-π reasoning. The mean approach velocity is written as

$$V = \frac{Q}{B\, y} \tag{5}$$

and the Froude number is then defined as:

$$Fr = \frac{V}{\sqrt{g\, y}} \tag{6}$$

The main stable predictor set retained for modelling is {*Fr, L/P, P/B, (Bsinθ)/L, y/[Bsin(θ/2)], θ/180*}. This formulation was adopted to preserve hydraulic interpretability while remaining numerically well behaved near θ =180°. In particular, the alternative group *y/(Bsinθ)* becomes singular at θ =180°, whereas *y*/[*Bsin(θ/2)*] remains finite because sin(θ/2)=sin(90∘)=1. Fig.3 later shows that this geometry-dependent structure is also reflected in the hydraulic response of *Cd*.

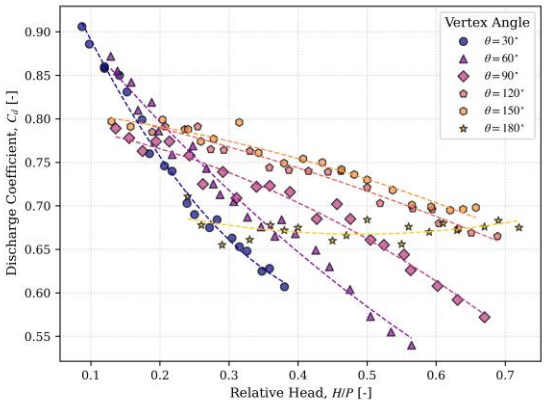


Fig.3. Variation of the discharge coefficient Cd with the relative head ratio H/P for the studied vertex angles.
Note: The Fig. 3 summarizes the hydraulic response of the reconstructed dataset across θ ∈ {30°, 60°, 90°, 120°, 150°, 180°}. It highlights the strong dependence of Cd on both vertex angle and relative head, with lower-angle labyrinth configurations showing higher Cd at low H/P and stronger degradation as H/P increases.

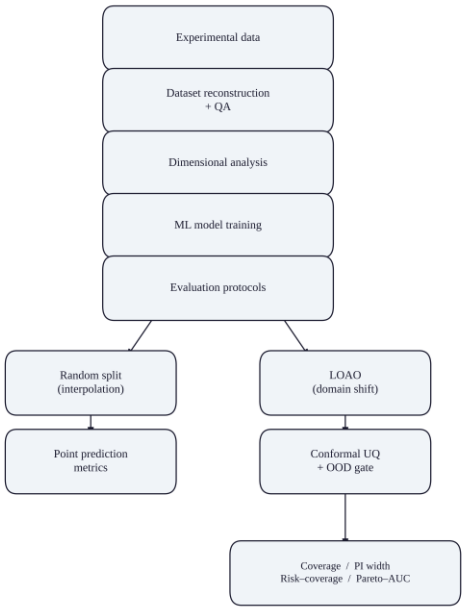


Figure 2. Workflow of dataset reconstruction, modelling, evaluation, and uncertainty quantification.
Note: The workflow summarizes the main stages of the study: reconstruction and quality assurance of the experimental dataset, dimensional analysis and feature construction, ML model training, evaluation under random-split and leave-one-angle-out (LOAO) protocols, and uncertainty quantification using conformal prediction with an OOD-aware two-tier policy. Final outputs include point-prediction metrics, empirical coverage and prediction-interval width, risk-coverage analysis, and Pareto/AUC summaries.

2.3 Machine-learning models

We evaluated four supervised regression approaches with increasing functional flexibility to predict the discharge coefficient *Cd* from the proposed dimensionless predictors. All models were implemented in Python using scikit-learn, and a fixed random seed (random_state = 42) was adopted for all stochastic components to ensure reproducibility.

Ridge regression served as a linear baseline with *ℓ2* regularization (Hoerl and Kennard, 1970), providing a reference for assessing whether *Cd* varies approximately linearly with the dimensionless groups. Random forest (RF) regression was included as a nonparametric ensemble of decision trees (n_estimators = 100) to capture nonlinearities and interactions among predictors (Breiman, 2001). Gradient boosting regression (GBR) served as the primary tabular model; the histogram-based implementation was adopted with early stopping, learning_rate = 0.1, max_depth = 5, and a maximum of 100 boosting iterations, reflecting its strong suitability for structured tabular data with complex response surfaces (Friedman, 2001). Finally, k-nearest neighbors (kNN) was not treated as a primary predictive model, but rather as a distance-based mechanism for out-of-distribution (OOD) detection, with Euclidean distances computed in standardized feature space.

For the random-split evaluation, performance was summarized over 15 repeated random 80/20 splits. This design was adopted as a practical compromise between variance reduction and computational simplicity, while reducing dependence on any single partition and preserving a clear distinction between interpolation-based performance and LOAO-based extrapolation performance.

2.4 Evaluation protocols

Two complementary evaluation protocols were used. Under random splitting, the dataset was partitioned randomly into 80% training and 20% testing, so that both sets contained mixed geometries; this protocol measures interpolation within the pooled dataset. Under leave-one-angle-out (LOAO) evaluation, all runs corresponding to one vertex angle were withheld for testing while the model was trained on the remaining angles; this protocol imposes a strict geometric domain shift and directly tests extrapolation to a novel planform regime.

2.5 Uncertainty quantification framework

Predictive reliability was assessed using conformal prediction intervals (PIs) for the discharge coefficient *Cd*. For a target miscoverage level α, the objective is to construct intervals that achieve empirical coverage close to $1 - \alpha$ while remaining sufficiently sharp for practical use.

In the present study, uncertainty quantification is examined under two distinct settings. Under random splitting, the training, calibration, and test subsets are sampled from similar mixtures of vertex angles, so the residual distribution is relatively stable. Under leave-one-angle-out (LOAO) evaluation, however, the test set corresponds to a withheld geometry, which introduces a geometry-driven domain shift. In this case, calibration residuals obtained from the observed angles may no longer transfer well to the unseen angle, and prediction intervals that appear reliable under interpolation can become severely under-covering under deployment on a novel planform.

To study this effect, we compare four conformal calibration strategies: (i) global split conformal, (ii) Mondrian conformal,

(iii) kernel-conditional conformal, and (iv) CV-pooled residual calibration. Global split conformal provides a simple baseline using one residual distribution for all calibration points. Mondrian and kernel-conditional variants attempt to adapt interval width to local structure in the feature space. CV-pooled calibration is more conservative and pools residual information across folds in order to improve robustness under geometric shift.

For each evaluation protocol, the available data are divided into a proper training subset used to fit the point predictor and a calibration subset used only for interval construction. In the random-split setting, the dataset is first divided into train and test sets, and the training portion is then split again into proper-train and calibration subsets. Under LOAO, one angle is held out for testing and the remaining angles are partitioned into proper-train and calibration subsets within each fold.

To further reduce the risk of overconfident intervals on weakly supported inputs, conformal prediction is combined with a k-nearest-neighbor (kNN) out-of-distribution (OOD) detector in standardized feature space. Under this two-tier policy, inputs classified as in-support receive the baseline conformal interval, whereas flagged inputs are routed to a conservative interval tier. In engineering terms, this means that the framework does not report a narrow uncertainty bound when a queried geometry lies outside the calibration envelope, but instead enlarges the interval to reflect the additional epistemic uncertainty associated with geometric extrapolation.

The main text focuses on the practical interpretation of these uncertainty-quantification strategies in terms of coverage reliability, interval sharpness, and deployment safety. The formal definitions of the nonconformity score, finite-sample conformal quantile, two-tier routing rule, and AUC-based summary metrics are provided in Appendix A.

2.6 OOD detection and two-tier selective prediction

To manage predictions on inputs that are weakly supported by the training distribution, conformal prediction was combined with an out-of-distribution (OOD) detector based on k-nearest-neighbor (kNN) distances in standardized feature space. Predictors were standardized using the mean and standard deviation computed on the proper training set only, and the same transformation was then applied to calibration and test points within each fold.

The use of kNN distances is well suited to the present study because the predictor space is low-dimensional and physically interpretable, which allows geometric separation between supported and weakly supported configurations to be assessed in a transparent way. In this setting, large distances indicate that a queried test point lies far from the training distribution and therefore may correspond to a geometry or hydraulic regime that is insufficiently represented in calibration.

An OOD score $s(x)$ was computed for each test point, and a threshold τ was used to classify the point as either in-support or out-of-distribution. Under the two-tier policy, accepted points received the baseline conformal prediction interval, whereas OOD-flagged points were routed to a conservative interval tier based on CV-pooled calibration. This policy was designed to reduce the risk of reporting overly sharp uncertainty bounds for geometries lying outside the calibration envelope.

The practical interpretation is straightforward. When a queried geometry is well represented by the available calibration data, the framework returns a relatively narrow interval appropriate for routine deployment. When the queried geometry is weakly supported, the framework switches to a wider conservative interval in order to reflect the additional epistemic uncertainty associated with geometric extrapolation. In this way, the two-tier policy translates statistical shift detection into a deployment-oriented safety mechanism.

The corresponding performance trade-offs are summarized using acceptance rate, accepted-set error, accepted-set interval properties, and overall two-tier coverage and width. These quantities are reported across a range of OOD thresholds in order to identify operating points that balance interval sharpness, empirical reliability, and practical deploy ability under leave-one-angle-out evaluation.

Because the detector is applied in a low-dimensional and physically interpretable feature space, Euclidean distance remains meaningful in the present benchmark. The method should nevertheless be interpreted as a practical support diagnostic rather than as a full density model.

3 RESULTS

3.1 Hydraulic behaviour across vertex-angle regimes

The reconstructed experimental data show that the discharge coefficient Cd depends strongly on both the vertex angle θ and the relative head ratio H/P. As shown in Fig.3, the more acute triangular planforms exhibit the highest values of Cd at low relative heads, where lateral contraction and streamline convergence enhance the effective discharge capacity. As H/P increases, however, this advantage progressively decreases, which is consistent with stronger nappe interaction and increasing hydraulic interference within the labyrinth-like geometry.

By contrast, the $\theta = 180°$ configuration behaves as the linear-weir limit. Its discharge coefficient is lower at small relative heads, but its hydraulic response remains more stable as H/P increases. The convergence of the 150° and 180° regimes at higher H/P values suggests that the influence of θ weakens as the flow approaches a less contracted overflow pattern. At the same time, the low-head region preserves strong angle sensitivity, which makes extrapolation across vertex-angle regimes particularly challenging when one angle is entirely withheld from training.

3.2 Point prediction: interpolation versus geometric extrapolation

Point-prediction performance was first evaluated under repeated random 80/20 splits, which represent interpolation within a mixed-geometry dataset. Under this setting, the gradient boosting regressor (GBR) achieved the strongest overall performance, with low MAE and RMSE and high R^2 (Table 2A). These results indicate that the proposed stable dimensionless predictor set captures the dominant variation in Cd when all geometry regimes are represented in both training and testing.

The interpretation changes substantially under leave-one-angle-out (LOAO) evaluation. When one vertex-angle regime is withheld entirely for testing, overall predictive accuracy decreases markedly (Table 2B), showing that interpolation performance overestimates deployment performance on a novel geometry. The angle-wise diagnostics further show that the largest errors occur for the θ = 180°-fold (Table 2C). This behaviour is hydraulically plausible and consistent with Fig.3, since the approach to the linear-weir limit corresponds to a qualitative change in flow structure. In other words, the model is required to extrapolate across a regime boundary rather than merely interpolate within a continuous experimental cloud.

Taken together, these results confirm that random-split validation is insufficient for assessing practical generalization across planform changes. In the present benchmark, geometry-driven extrapolation is the dominant source of predictive degradation, and the strongest failure is associated with the transition toward the $\theta = 180°$ overflow regime.

3.3 Conformal prediction under geometric shift

The uncertainty-quantification results reveal the same contrast between interpolation and extrapolation. Under random splitting, conformal prediction intervals behave reasonably well because calibration and test cases are drawn from similar mixtures of geometries. Under LOAO, however, the calibration residual distribution no longer transfers reliably to the withheld geometry, and the baseline global split conformal method becomes strongly under-covering (Table 3).

At $\alpha = 0.05$, the Baseline_GlobalSplit method yields empirical coverage far below the nominal level, despite producing relatively narrow intervals. A similar pattern is observed at $\alpha=0.10$. This means that intervals that appear acceptably sharp under interpolation can become misleadingly optimistic when the model is applied to a novel planform. In practical hydraulic terms, the uncertainty bounds are too narrow precisely in the regime where extrapolation risk is greatest.

Conditional variants such as Mondrian and kernel-conditional conformal provide only limited recovery of coverage. Although they adapt interval construction to local structure in the feature space, they do not fully resolve the geometry-driven residual shift induced by LOAO. The implication is that local conditioning alone is not sufficient when the withheld geometry differs qualitatively from those represented in calibration.

3.4 Restoring reliability through conservative calibration

A more robust recovery of empirical coverage is obtained with CV-pooled residual calibration. At both $\alpha = 0.05$ and $\alpha = 0.10$, the CV-pooled strategy restores coverage to values much closer to the intended reliability level, but only by substantially increasing mean interval width (Table 3). This behaviour reflects an explicit trade-off between sharpness and safety: the intervals become wider because the method acknowledges the additional uncertainty associated with unsupported geometries.

From a hydraulic standpoint, this is a more credible outcome than severe under-coverage. If the withheld geometry corresponds to a regime transition, then a wider interval is preferable to an apparently precise but unreliable bound. The CV-pooled strategy should therefore be interpreted as introducing a conservative reliability margin when the calibration residuals are not transferable across vertex-angle regimes.

Where empirical coverage is reported, the corresponding values should be interpreted together with their uncertainty bounds rather than as exact constants. For this reason, Tables 3 and 4 report 95% Wilson confidence intervals for the main empirical coverage estimates.

Table 2. Point-prediction accuracy under random splits and leave-one-angle-out (LOAO) evaluation.

Panel A: Random-split summary (Stable feature set).						
Model	MAE (mean)	MAE (std)	RMSE (mean)	RMSE (std)	R2 (mean)	R2 (std)
GBR	0.0119	0.0021	0.0155	0.0031	0.9400	0.0302
Mean	0.0557	0.0075	0.0686	0.0097	-0.0898	0.1695
RF_small	0.0127	0.0021	0.0176	0.0031	0.9257	0.0257
Ridge	0.0234	0.0049	0.0302	0.0056	0.7899	0.0432
Panel B: LOAO overall performance (Stable feature set).						
Model	MAE	RMSE	R2	n		
GBR	0.0364	0.0459	0.5687	123		
RF_small	0.0354	0.0460	0.5663	123		
Mean	0.0582	0.0717	-0.0515	123		
Ridge	0.0523	0.0840	-0.4431	123		
Panel C: Angle-wise LOAO errors for GBR (Stable).						
θ [deg]	n	MAE	RMSE			
30	20	0.0567	0.0627			
60	23	0.0376	0.0430			
90	22	0.0340	0.0419			
120	20	0.0266	0.0293			
150	20	0.0272	0.0344			
180	18	0.0694	0.0791			

Note: Panel A reports random 80/20 split performance using 15 repeated random splits; values are shown as mean ± standard deviation. Panel B reports overall LOAO performance across all test cases. Panel C reports angle-wise LOAO errors for the gradient boosting regressor (GBR). Error metrics are reported in *Cd* units, except R^2.

Table 3. Conformal prediction performance under LOAO with 95% Wilson confidence intervals for empirical coverage.

Method	α	Coverage [95% CI]	Mean PI width
Baseline_GlobalSplit	0.05	0.496 [0.409, 0.583]	0.075
Adaptive_kNN_k=20	0.05	0.537 [0.449, 0.622]	0.083
Mondrian_2x2	0.05	0.545 [0.457, 0.630]	0.077
Kernel_m=2	0.05	0.553 [0.465, 0.638]	0.082
CVPooled_Global	0.05	0.919 [0.857, 0.955]	0.212
Baseline_GlobalSplit	0.10	0.407 [0.324, 0.495]	0.057
Adaptive_kNN_k=20	0.10	0.455 [0.370, 0.543]	0.065
Mondrian_3x2	0.10	0.504 [0.417, 0.591]	0.072
Kernel_m=0.75	0.10	0.439 [0.354, 0.527]	0.059
CVPooled_Global	0.10	0.911 [0.847, 0.949]	0.182

Note: Empirical marginal coverage and mean prediction-interval (PI) width are reported for conformal methods under leave-one-angle-out (LOAO) testing at miscoverage levels α = 0.05 and α = 0.10. Coverage values are pooled empirical proportions over the evaluated LOAO test cases, and bracketed intervals are 95% Wilson binomial confidence intervals. Mean PI width is reported in Cd units. Under LOAO-induced geometric shift, empirical coverage should be interpreted as realized performance rather than as a formal finite-sample guarantee.

3.5 Selective prediction with OOD-aware two-tier routing

To manage the coverage-sharpness trade-off more flexibly, the benchmark combines conformal prediction with out-of-distribution detection and a two-tier interval policy. Under this framework, each test point is first evaluated by the kNN-based OOD detector. Inputs judged to be in-support receive the baseline interval, whereas flagged inputs are routed to a conservative interval tier.

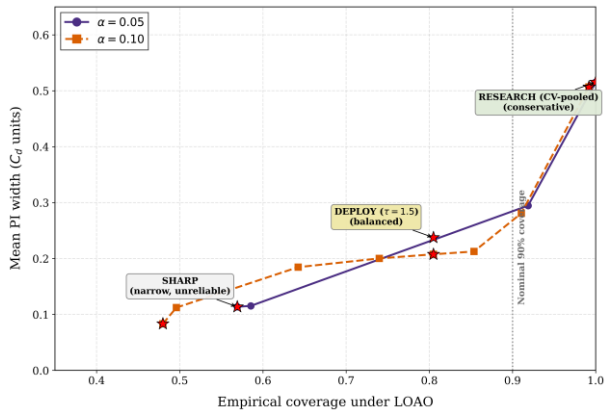


Fig.4. Coverage-width trade-off under LOAO with representative operating points for two-tier uncertainty quantification.
Note: Curves show the global coverage-width envelope for α = 0.05 and α = 0.10 across the evaluated OOD-threshold grid. Starred markers indicate representative operating points selected for interpretation: Sharp (minimum-width but unreliable), Deploy (balanced), and Research (conservative). The vertical dotted line marks the nominal 90% coverage target.

The resulting operating points clarify how deployment decisions change with the OOD threshold τ. This trade-off is summarized graphically in Fig.4, which shows the coverage-width envelopes, and in Fig.5, which reports the accepted-set risk-coverage curves. The corresponding numerical operating points are reported in Table 4. Lenient thresholds accept most inputs and preserve relatively narrow intervals, but overall coverage remains limited under geometric shift. Stricter thresholds reduce acceptance and enlarge the overall uncertainty bounds, but they also improve two-tier coverage by sending weakly supported cases to the conservative tier.

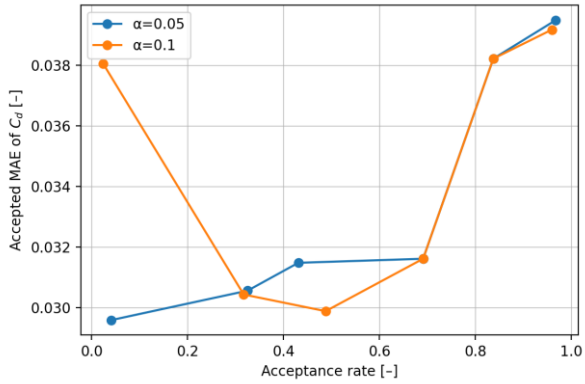


Fig.5. Risk-coverage curves for OOD-gated selective prediction under LOAO.
Note: Accepted-set mean absolute error (MAE) is plotted against acceptance rate for the representative two-tier policy under leave-one-angle-out (LOAO) testing at α = 0.05 and α = 0.10. Stricter OOD thresholds reject more weakly supported cases and reduce accepted-set error by routing them to the conservative tier.

The θ =180° diagnostics are particularly informative. Under strict thresholds, this fold is frequently rejected and handled conservatively, indicating that the OOD gate is functioning as intended by identifying a hydraulically unsupported regime. In practice, this means that the framework does not attempt to "repair" extrapolation through overly confident point prediction, but instead responds by enlarging uncertainty bounds when geometric support is insufficient. A schematic interpretation of this deployment logic is provided in Fig.7.

Table 4. Representative two-tier operating points under LOAO with 95% Wilson confidence intervals for empirical coverage.

α	Operating point	Method	τ	Accept rate	Accepted PI coverage [95% CI]	Accepted PI width	Two-tier coverage [95% CI]	Two-tier width
0.05	Sharp (min width)	Baseline_GlobalSplit	10.0	0.967	0.555 [0.465, 0.641]	0.079	0.569 [0.481, 0.653]	0.114
0.05	Deploy (≥0.80 cov)	Baseline_GlobalSplit	1.5	0.431	0.585 [0.451, 0.707]	0.065	0.805 [0.726, 0.865]	0.238
0.05	Research (≥0.95 cov)	CVPooled_Global	2.0	0.691	1.000 [0.957, 1.000]	0.230	1.000 [0.970, 1.000]	0.515
0.10	Sharp (min width)	Kernel_m=0.75	10.0	0.959	0.458 [0.370, 0.547]	0.059	0.480 [0.393, 0.567]	0.083
0.10	Deploy (≥0.80 cov)	Baseline_GlobalSplit	1.5	0.317	0.513 [0.362, 0.661]	0.054	0.805 [0.726, 0.865]	0.208
0.10	Research (≥0.95 cov)	CVPooled_Global	3.0	0.691	0.988 [0.936, 0.998]	0.194	0.992 [0.955, 0.999]	0.507

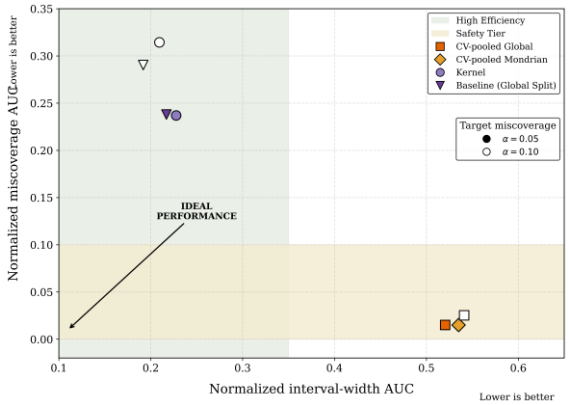


Fig.6. Normalized AUC comparison of selected two-tier uncertainty-quantification strategies under LOAO.
Note: Points show normalized AUC values for interval width and miscoverage; lower values indicate more favorable overall trade-offs. Filled markers denote α=0.05, whereas hollow markers denote α=0.10. Shaded regions are schematic guides to interpretation and do not represent statistically estimated boundaries.

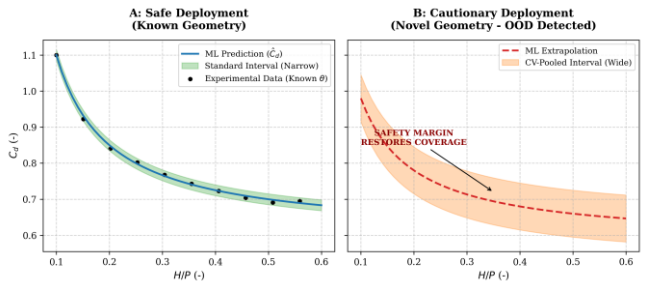


Fig.7. Schematic illustration of selective deployment for discharge-coefficient prediction under known and novel weir geometries.
Note: Panel A illustrates safe deployment for a known geometry, where a narrower standard prediction interval accompanies the ML prediction. Panel B illustrates cautionary deployment for a novel geometry detected as out-of-distribution, where a wider CV-pooled conformal interval is used to provide more conservative uncertainty quantification. The Fig.7 emphasizes the practical distinction between routine deployment for in-support geometries and cautionary deployment for weakly supported geometries.

3.6 Operating-point selection and AUC-based comparison

The coverage-width envelopes in Fig.4 and the risk-coverage curves in Fig.5 support three representative operating regimes: a Sharp regime with high acceptance and narrow intervals but insufficient reliability under shift; a Deploy regime that offers a more balanced compromise between acceptance, coverage, and interval width; and a Research regime that prioritizes reliability through conservative calibration at the cost of wider intervals. The corresponding representative operating points are reported in Table 4.

To compare methods across the full threshold range, Fig.6 and Table 5 summarize normalized AUC values for two-tier miscoverage, two-tier width, and accepted-set MAE. Lower values indicate more favourable overall trade-offs. These summaries should be interpreted as comparative descriptors rather than as formal guarantees. Their main value is to provide a compact decision aid for selecting uncertainty-quantification strategies under geometric domain shift.

Taken together, the results show that no single strategy dominates all deployment objectives. Methods that preserve narrow intervals tend to under-cover under LOAO, whereas methods that restore reliability do so at the cost of width. The most practically useful operating points are therefore intermediate ones, where interval sharpness, empirical coverage, and acceptance remain reasonably balanced for deployment beyond the calibration envelope.

4 DISCUSSION

4.1 Generalization limits and hydraulic interpretation

The results show that the apparent strength of data-driven discharge-coefficient prediction under random splitting does not transfer directly to deployment on a novel weir geometry. When all vertex angles are mixed across training and testing sets, the models operate in an interpolation regime and achieve high apparent accuracy. Under leave-one-angle-out (LOAO) evaluation, however, the task changes fundamentally: the model must extrapolate from

previously observed planforms to a withheld geometry. The marked degradation observed under LOAO therefore indicates that geometry-specific flow behavior is not fully captured by interpolation-based training.

Table 5. AUC-based summaries and best-at-coverage operating points for two-tier uncertainty quantification under LOAO.

Panel A: AUC rankings (lower is better).

α	Method	AUC two-tier miscov (norm)	AUC two-tier width (norm)	AUC MAE(norm)
0.050	CVPooled_Global	0.0171	0.5244	0.0319
0.050	CVPooled_Moddrian_2x2	0.0171	0.5390	0.0319
0.050	Kernel_m=2	0.2389	0.2316	0.0326
0.050	Baseline_GlobalSplit	0.2400	0.2210	0.0326
0.100	CVPooled_Global	0.0232	0.5371	0.0329
0.100	Adaptive_kNN_k=20	0.2779	0.2309	0.0334
0.100	Baseline_GlobalSplit	0.2885	0.1877	0.0334
0.100	Mondrian_3x2	0.2934	0.2735	0.0334

Panel B: Best-at-coverage selections.

α	coverage target	Method	τ	accept_rate	two-tier cov	two-tier width
0.050	0.800	Baseline_GlobalSplit	1.5000	0.4309	0.8049	0.2376
0.050	0.950	CVPooled_Global	2.0000	0.6911	1.0000	0.5154
0.100	0.800	Baseline_GlobalSplit	1.5000	0.3171	0.8049	0.2076
0.100	0.950	CVPooled_Global	3.0000	0.6911	0.9919	0.5068

Note: Panel A reports normalized area-under-the-curve (AUC) summaries for two-tier miscoverage, two-tier mean PI width, and accepted-set MAE across the evaluated OOD-threshold grid; lower values indicate better performance. Panel B reports the smallest-width operating points achieving the target two-tier coverage level for each α. These summaries are intended as comparative descriptors of the coverage-sharpness and risk-acceptance trade-offs across methods.

This limitation is hydraulically meaningful rather than purely statistical. The strongest degradation occurs near the θ = 180° regime, which corresponds to a transition from strongly contracted triangular planform behavior toward the linear-weir limit. In the more acute configurations, discharge is strongly influenced by lateral contraction, streamline convergence, and enhanced interaction between nappe structures. As θ increases toward 180°, these effects weaken and the flow approaches a simpler overflow configuration. A model trained on the more contracted regimes therefore tends to carry over geometric influences that are no longer dominant in the linear-limit case, which explains the large point-prediction errors observed in the θ = 180° fold.

From this perspective, LOAO should not be viewed only as a machine-learning validation strategy, but also as a physically meaningful stress test of whether the predictor remains reliable across changes in flow regime. The present results suggest that, for triangular planform weirs, a substantial part of the prediction error under geometric shift is a direct consequence of the underlying hydraulic transition rather than of random model instability.

In hydraulic terms, this transition is associated with weakening lateral contraction, reduced streamline convergence, and a redistribution of local energy-loss mechanisms as the planform approaches the linear-weir limit.

4.2 Uncertainty quantification under geometric shift

The uncertainty-quantification results lead to a similar conclusion. Under random splitting, conformal prediction intervals behave reasonably well because calibration and test points originate from similar geometry mixtures. Under LOAO, the residual distribution shifts with the geometry, and intervals calibrated on observed angles can become strongly under-covering when transferred to a withheld angle. In practical terms, this means that narrow uncertainty intervals obtained under interpolation may provide a misleading impression of reliability when the model is applied outside the calibration envelope.

The severe under-coverage of the baseline global split conformal method under LOAO therefore has a direct engineering interpretation. It indicates that residual variability is tied to geometry-dependent flow behavior and cannot be assumed constant across vertex-angle regimes. In hydraulic design or discharge estimation, such optimistic intervals would understate uncertainty precisely in the situations where extrapolation risk is highest. Conservative residual pooling improves empirical coverage by widening the intervals and thereby introducing an explicit reliability buffer. This does not eliminate geometric shift, but it makes the uncertainty bounds more consistent with the level of epistemic uncertainty associated with unsupported geometries.

Conditional variants such as Mondrian and kernel-conditional conformal provide only limited recovery of coverage in the present benchmark. This suggests that local conditioning in the available feature space is not sufficient, by itself, to overcome a geometry-driven residual shift when the withheld regime differs qualitatively from those seen during calibration.

4.3 Towards safe deployment: OOD-aware two-tier prediction

A central practical contribution of the study is the use of out-of-distribution (OOD) detection to convert statistical shift awareness into a deployment-oriented decision rule. By combining k-nearest-neighbor distance-based OOD detection with a two-tier conformal policy, the framework distinguishes between geometries that are well supported by the calibration data and geometries that are weakly supported or novel. Inputs judged to be in-support can retain relatively sharp prediction intervals, whereas flagged inputs are routed to a more conservative interval tier.

This mechanism is particularly important for the θ = 180° regime. Under stricter thresholds, this geometry is frequently rejected by the OOD gate and handled conservatively, which shows that the system is not “fixing” the extrapolation problem through improved point prediction, but rather managing it by explicitly enlarging the uncertainty bound. In engineering terms, this is preferable to returning a narrow interval that appears precise but does not reflect the true uncertainty of the unsupported configuration.

The operating-point framework provides a useful way to express this trade-off. A “Sharp” regime offers high acceptance and narrow

intervals, but insufficient protection under geometric shift. A more balanced "Deploy" regime sacrifices some acceptance in exchange for more reliable coverage. A highly conservative "Research" regime provides the strongest protection, but at the cost of substantially wider intervals. This interpretation is directly relevant to weir design and discharge measurement, because the acceptable balance between sharpness and reliability depends on whether the prediction is used for routine estimation, exploratory analysis, or safety-relevant hydraulic decision-making.

4.4 Broader implications for hydraulic modelling

The broader implication of this study is that geometric shift should be treated as a first-order issue in data-driven hydraulic prediction. Strong interpolation performance within a pooled dataset is not sufficient evidence that a model can be safely applied to a modified geometry. For triangular planform weirs, the present benchmark shows that prediction failures and interval failures under LOAO are not isolated statistical anomalies; they are signatures of regime-dependent hydraulic structure that is incompletely represented in the training data.

Accordingly, safer deployment requires more than accurate point prediction. It requires evaluation protocols that reflect realistic extrapolation conditions, uncertainty bounds that remain meaningful under shift, and explicit mechanisms for identifying unsupported inputs. In this sense, the proposed framework should be understood not as a replacement for hydraulic understanding, but as a way to embed that caution directly into data-driven prediction practice.

5 LIMITATIONS AND FUTURE EXPERIMENTAL NEEDS

The present study has several limitations that should be considered when interpreting the results. First, the dataset is relatively small (n = 123) and was reconstructed from a single experimental campaign. Although this design is useful for isolating geometry-driven effects and avoids inter-laboratory inconsistencies, it also limits the generalizability of the benchmark and does not provide an independent external validation dataset. The reported results should therefore be interpreted as a controlled demonstration of geometric domain shift rather than as a fully comprehensive assessment across facilities, measurement protocols, and operating environments.

Second, the available dataset does not explicitly include several hydraulic descriptors that may influence the discharge coefficient, including aeration or ventilation state, detailed approach-flow conditions, and direct measurement uncertainty. These omitted factors may contribute to residual variability and could become especially important when transferring the framework to broader experimental settings. In this sense, the present benchmark captures the dominant role of planform geometry, but not the full range of hydraulic and experimental variability that may arise in practice.

Third, the leave-one-angle-out (LOAO) protocol provides a physically meaningful test of geometric extrapolation, but it remains limited to six discrete vertex-angle regimes derived from one source dataset. As a result, the transition toward the $\theta = 180°$ regime can be identified, but not densely resolved. Additional experiments in the transitional range would help determine more precisely how the flow evolves from strongly contracted labyrinth behavior toward the linear-weir limit, and how rapidly this transition affects both point-prediction error and uncertainty calibration.

A further limitation is that the empirical coverage values reported for the uncertainty-quantification methods are estimated from a limited number of test cases and therefore remain subject to sampling variability. For this reason, coverage estimates should be interpreted together with their uncertainty bounds rather than as exact performance constants. Likewise, repeated-run summaries under random splitting should be viewed as stable comparative estimates of interpolation performance, not as exhaustive Monte Carlo characterizations.

These limitations point directly to future experimental priorities. The most important next step is the development of larger multi-source datasets spanning a wider range of vertex angles, operating heads, and laboratory conditions. Such datasets would make it possible to distinguish more clearly between geometry-driven shift and facility-specific variability. A second priority is richer metadata reporting, particularly for aeration state, approach-flow conditions, and measurement uncertainty, so that future uncertainty-quantification frameworks can better separate physical variability from calibration uncertainty. A third priority is targeted experimentation in the transitional range approaching $\theta = 180°$, where the present results indicate a regime change with important consequences for both prediction error and interval reliability.

Beyond data collection, future work should examine whether recent conformal-prediction methods developed for covariate shift can be adapted to hydraulic deployment settings in which geometry, operating head, or experimental support differ between calibration and application domains (Yang et al., 2024; Jonkers et al., 2024; Qin et al., 2025). More broadly, it will be important to assess whether shift-aware calibration strategies, hybrid physics-informed models, or limited geometry-specific recalibration can reduce extrapolation risk while preserving the practical advantages of data-driven prediction.

5 CONCLUSION

This study shows that evaluating discharge-coefficient predictors for triangular planform weirs by random train-test splitting can substantially overestimate performance when the target application involves a change in geometry. Using a reconstructed run-level dataset spanning six vertex-angle regimes, the leave-one-angle-out (LOAO) protocol revealed a marked loss of generalization relative to random splitting, confirming that interpolation accuracy within a mixed-geometry dataset should not be interpreted as evidence of reliable performance on a novel planform.

The strongest degradation was observed near the $\theta = 180°$ regime, which is consistent with a hydraulic transition from strongly contracted triangular planform behavior toward the linear-weir limit. In this sense, the largest prediction errors are not merely statistical anomalies, but indicators of a change in the underlying flow structure that is not represented in the training data.

For uncertainty quantification, the results show that conformal prediction intervals calibrated under interpolation can fail under geometric domain shift. Global split conformal intervals were severely under-covering under LOAO, indicating that calibration residuals were not transferable across vertex-angle regimes in the present setting. More conservative residual pooling substantially improved empirical coverage, but only at the cost of wider intervals. The resulting coverage-sharpness trade-off is therefore a central practical constraint for deployment on unsupported geometries.

A more deployment-oriented alternative is to combine prediction intervals with out-of-distribution detection and a two-tier policy. In this framework, geometries that are well supported by calibration data can be handled with relatively sharp intervals, whereas weakly supported or novel geometries are routed to a conservative interval tier. This provides a practical mechanism for avoiding overconfident discharge-coefficient predictions when extrapolating beyond the calibration envelope.

The broader implication of the study is that geometric shift should be treated as a first-order issue in data-driven hydraulic prediction. For triangular planform weirs, predictive failure under LOAO reflects not only statistical shift but also a physically meaningful transition in contraction, streamline organization, and overflow behavior across planform geometries. Future work should therefore prioritize denser experiments in the transitional range approaching $\theta = 180°$, richer reporting of hydraulic metadata and measurement uncertainty, and evaluation on larger multi-source datasets. Such developments would help extend the present

benchmark toward a more deployable framework for uncertainty-aware discharge prediction under unsupported weir geometries.

NOTATION

Latin symbols

Symbol	Definition (units)
B	Channel width (m).
P	Crest height (weir height above bed) (m).
H	Upstream head over crest (m).
L	Crest length (m).
y	Approach depth, $y = P + H$ (m).
V	Mean approach velocity, $V = Q/(B\ y)$ (m/s).
Q	Discharge (m³/s).
g	Gravitational acceleration (m/s²).
Cd	Discharge coefficient (dimensionless, "--"), defined from Eq. (2)
Fr	Froude number, $Fr = V/\sqrt{g\ y}$ (dimensionless).
$PI(x)$	Prediction interval for Cd at input x.
x	Predictor (feature) vector used as model input.
$r(x)$	OOD score from kNN distances in standardized feature space.

Greek symbols

Symbol	Definition (units)
θ	Vertex angle (deg).
α	Miscoverage level for prediction intervals (e.g., 0.05, 0.10).
τ	OOD acceptance threshold (accept if $r(x) \leq \tau$).

Subscripts / superscripts / derived groups

Notation	Definition (units)
$\hat{C}_d(x)$	Point prediction of C_d at input x.
e_i	Absolute calibration residual. C_{d,i}-\hat{C}_d(x_i)
$q_{1-\alpha}$	Conformal quantile ("radius") used in symmetric split-conformal. intervals.
$b(x)$	Bin/group label used in Mondrian (group-conditional) conformal.
L/P	Dimensionless crest-length ratio.
P/B	Dimensionless crest-height ratio.
$B\sin(\theta)/L$	Dimensionless geometry group.
$y/(B\sin^2(\theta/2))$	Stable dimensionless depth group used to avoid singularity at $\theta = 180^{\circ}$
$\theta/180$	Normalized vertex angle (dimensionless).

Abbreviations

Abbreviation	Meaning
ML	Machine learning.
UQ	Uncertainty quantification.
LOAO	Leave-one-angle-out.
PI	Prediction interval (PIs = prediction intervals).
OOD	Out-of-distribution.
kNN	k-nearest neighbors.
GBR	Gradient boosting regressor.
RF	Random forest.
MAE	Mean absolute error.
RMSE	Root mean square error.
R^2	Coefficient of determination.
CV	Cross-validation (as in "CV-pooled").
AUC	Area under the curve.

REFERENCES

Angelopoulos, A.N., Bates, S., 2023. Conformal prediction: A gentle introduction. Foundations and Trends in Machine Learning 16(4), 494-591. https://doi.org/10.1561/2200000101

Asgharzadeh-Bonab, A., Bijanvand, S., Parsaie, A., Afaridegan, E., 2025. Machine learning-based estimation of discharge coefficient for semicircular labyrinth weirs. Scientific Reports 15(1), 33002. https://doi.org/10.1038/s41598-025-18230-4

Barber, R.F., Candès, E.J., Ramdas, A., Tibshirani, R.J., 2021. Predictive inference with the jackknife+. The Annals of Statistics 49(1), 486-507. https://doi.org/10.1214/20-AOS1965

Barber, R.F., Candès, E.J., Ramdas, A., Tibshirani, R.J., 2023. Conformal prediction beyond exchangeability. The Annals of Statistics 51(2), 816-845. https://doi.org/10.1214/23-AOS2276

Belaabed, F., Arabet, L., Goudjil, K., Ouamane, A., 2025. Optimized machine learning models for accurate prediction of the discharge coefficient in hydraulic weirs. Journal of

Hydrology and Hydromechanics 73(3), 295–309. https://doi.org/10.2478/johh-2025-0023

Breiman, L., 2001. Random forests. Machine Learning 45, 5-32. https://doi.org/10.1023/A:1010933404324

Buckingham, E., 1914. On physically similar systems; illustrations of the use of dimensional equations. Physical Review 4(4), 345-376. https://doi.org/10.1103/PhysRev.4.345

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. The Annals of Statistics 29(5), 1189-1232. https://doi.org/10.1214/aos/1013203451

Gibbs, I., Candès, E., 2021. Adaptive conformal inference under distribution shift. Advances in Neural Information Processing Systems, 34, 1660-1672. https://doi.org/10.48550/arXiv.2106.00170

Haghbin, M., Sharafati, A., 2022. A review of studies on estimating the discharge coefficient of flow control structures based on the soft computing models. Flow Measurement and Instrumentation 83, 102119. https://doi.org/10.1016/j.flowmeasinst.2021.102119

Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12(1), 55-67. https://doi.org/10.1080/00401706.1970.10488634

Jonkers, J., Van Wallendael, G., Duchateau, L., Van Hoecke, S., 2024. Conformal Predictive Systems Under Covariate Shift. In: Vantini, S., Fontana, M., Solari, A., Boström, H., Carlsson, L. (Eds.): Proceedings of the Thirteenth Symposium on Conformal and Probabilistic Prediction with Applications. Proceedings of Machine Learning Research 230, 406–423.

Kumar, S., Ahmad, Z., Mansoor, T., 2011. A new approach to improve the discharging capacity of sharp-crested triangular planform weirs. Flow Measurement and Instrumentation 22(3), 175-180. https://doi.org/10.1016/j.flowmeasinst.2011.01.006

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R., Wasserman, L., 2018. Distribution-free predictive inference for regression. Journal of the American Statistical Association 113(523), 1094-1111. https://doi.org/10.1080/01621459.2017.1307116

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J.V., Lakshminarayanan, B., Snoek, J., 2019. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. Advances in Neural Information Processing Systems 32, 13969-13980

Qin, J., Liu, Y., Li, M., Huang, C.-Y., 2025. Distribution-Free Prediction Intervals Under Covariate Shift, With an Application to Causal Inference. Journal of the American Statistical Association 120(549), 559–571. https://doi.org/10.1080/01621459.2024.2356886

Romano, Y., Patterson, E., Candès, E., 2019. Conformalized quantile regression. Advances in Neural Information Processing Systems, 32.

Seyedian, S.M., Haghiabi, A.H., Parsaie, A., 2023. Reliable prediction of the discharge coefficient of triangular labyrinth weir based on soft computing techniques. Flow Measurement and Instrumentation 92, 102403. https://doi.org/10.1016/j.flowmeasinst.2023.102403

Seyedian, S.M., Kisi, O., 2024. Uncertainty analysis of discharge coefficient predicted for rectangular side weir using machine learning methods. Journal of Hydrology and Hydromechanics 72(1), 113-130. https://doi.org/10.2478/johh-2023-0043

Sun, Y., Ming, Y., Zhu, X., Li, Y., 2022. Out-of-distribution detection with deep nearest neighbors. In: Proceedings of the 39th International Conference on Machine Learning (ICML). PMLR 162, 20827-20840.

Tibshirani, R.J., Barber, R.F., Candès, E.J., Ramdas, A., 2019. Conformal prediction under covariate shift. Advances in Neural Information Processing Systems, 32.

Yang, Y., Kuchibhotla, A.K., Tchetgen Tchetgen, E., 2024. Doubly robust calibration of prediction sets under covariate shift. Journal of the Royal Statistical Society Series B: Statistical Methodology 86(4), 943–965. https://doi.org/10.1093/jrsssb/qkae009

Received 2 February 2026
Accepted 15 May 2026

APPENDIX

Appendix A. Technical details of conformal calibration and AUC-based ranking

The formal derivations of the conformal residual definition, finite-sample quantile construction, OOD-gated two-tier routing rule, and AUC-based summaries are moved to this appendix in order to keep the main text focused on engineering interpretation. This presentation follows the conformal and distribution-shift literature used in the manuscript, including Angelopoulos and Bates (2023), Barber et al. (2021, 2023), Lei et al. (2018), Romano et al. (2019), Gibbs and Candès (2021), Ovadia et al. (2019), and Sun et al. (2022).

A.1. Nonconformity score and conformal quantile

Let $\hat{\boldsymbol{f}}(\boldsymbol{x})$ denote the point prediction produced by the regression model for input *x*. For a calibration sample $\{(\mathbf{x_i}, \mathbf{y_i})\}_{\mathbf{i=1}}^{\mathbf{n_{cal}}}$, the absolute nonconformity score is defined as

$$r_i = \left|y_i - \hat{f}(x_i)\right| \tag{A1}$$

For a target miscoverage level α, the finite-sample conformal index is

$$k = \lceil (n_{cal} + 1)(1 - \alpha) \rceil \tag{A2}$$

The corresponding conformal radius is then

$$q_{1-\alpha} = r_{(k)} \tag{A3}$$

and the resulting symmetric split-conformal prediction interval is:

$$\hat{C}_\alpha(x) = \left[\hat{f}(x) - q_{1-\alpha}\,, \hat{f}(x) + q_{1-\alpha}\right] \tag{A4}$$

A.2. Mondrian and conditional conformal construction

For Mondrian conformal prediction, let *g(x)* denote a discrete bin assignment. For each bin *b*, a bin-specific conformal radius $\boldsymbol{q}_{\boldsymbol{1-\alpha}}^{(\boldsymbol{b})}$ is computed, yielding

$$\hat{C}_\alpha^{(b)}(x) = \left[\hat{f}(x) - q_{1-\alpha}^{(b)}\,, \hat{f}(x) + q_{1-\alpha}^{(b)}\right] \tag{A5}$$

A.3. OOD-gated two-tier policy

Let *s(x)* denote the OOD score and τ the acceptance threshold. A point is treated as in-support if

$$s(x) \le \tau \tag{A6}$$

Under the two-tier policy, the deployed prediction interval is

$$\hat{C}_\alpha^{2-tier}(x) = \begin{cases} \hat{C}_\alpha^{base}(x), & s(x) \le \tau, \\ \hat{C}_\alpha^{cons}(x), & s(x) > \tau. \end{cases} \tag{A7}$$

A.4. AUC-based summaries

For any threshold-dependent metric *m(τ)*, the discrete AUC over the evaluated OOD-threshold grid is computed as

$$AUC(m) = \sum_{j=1}^{j-1} \frac{m(\tau_j) + m(\tau_{j+1})}{2} \left|a\left(\tau_{j+1}\right) - a\left(\tau_j\right)\right| \tag{A8}$$

where $a\left(\tau_j\right)$ is the acceptance rate at threshold τ_j.

The three criteria used in the manuscript are defined as

$$m_{mis}(\tau) = 1 - Coverage_{2-tier}(\tau) \tag{A9}$$

$$m_{wid}(\tau) = Width_{2-tier}(\tau) \tag{A10}$$

$$m_{risk}(\tau) = MAE_{accepted}(\tau) \tag{A11}$$

To compare methods on a common scale, the corresponding AUC values are normalized to the unit interval:

$$AUC_{norm}^{(m)} = \frac{AUC^{(m)} - min_l AUC_l^{(m)}}{max_l AUC_l^{(m)} - min_l AUC_l^{(m)}} \tag{A12}$$

A.5. Interpretation in the LOAO setting

In the random-split setting, calibration and test data are drawn from similar geometry mixtures, so the global conformal baseline is expected to behave relatively well. In contrast, LOAO evaluation withholds an entire vertex-angle regime, which creates a geometry-dependent residual shift. For this reason, the role of the appendix-level derivations is not to claim exact validity under shift, but to define the interval constructions and ranking criteria used for empirical comparison in the main text. The main manuscript therefore focuses on engineering interpretation, while the present appendix records the formal definitions needed for reproducibility.