Stratified Conformal Prediction for Neural Fluid Surrogates: Spatially Adaptive Uncertainty Quantification Framework

Afm Farhad [a,e], Tanvir Hossen Ekra [b,e], Md. Tanvir Azmain [c,e], Ashraf Mahmud Rayed[d,e],

Abstract

Deep learning surrogates, especially neural operators, provide rapid alternatives to traditional CFD solvers; nonetheless, they lack inherent physical guarantees, a significant drawback for safety-critical engineering applications. Conformal prediction offers distribution-free coverage guarantees; yet, traditional global calibration is inadequate for multi-regime fluid systems: a single threshold is overly conservative in laminar areas and insufficient in turbulent sections. We propose a physics-informed, regime-stratified conformal prediction approach for neural fluid surrogates that addresses the gap in three phases. First, we formulate a dimensionally consistent physics residual error based on the steady-state incompressible Navier-Stokes momentum equations, which serves as an unsupervised nonconformity score, validated by strong Spearman rank correlation with the actual prediction error (0.94 for cavity, 0.90 for cylinder). Second, we use regime-stratified calibration, dividing the parametric space according to Reynolds number and boundary-condition type to produce regime-specific thresholds that are up to 99.69% tighter than the global bound, while ensuring around 90% coverage across all regimes. Third, a modular spatially adaptive scaling stage converts uniform per-regime bounds into pixel-level uncertainty maps using any spatially varying signal; three candidates, vorticity magnitude, velocity magnitude and MC dropout standard deviation, are systematically evaluated with independently optimised hyperparameters. Against standalone MC dropout, the proposed framework produces intervals that are 3.3–16.2 times narrower at the same coverage target, with higher spatial error correlation. Among spatial signals within the framework, vorticity magnitude is adopted as the default: it is the only candidate that improves over the flat baseline across all regimes on both benchmarks at negligible computational cost relative to the 50-pass MC dropout alternative. A pressure-gradient extension demonstrates framework modularity across field variables, indicating that pressure prediction errors are more spatially concentrated than velocity errors, accounting for an additional efficiency improvement. All results are validated by a five-component statistical robustness suite, which includes bootstrap resampling, K-fold cross-validation, multi-seed stability, significance-level sweep, and calibration-size ablation. The framework is demonstrated on lid-driven cavity and cylinder flow benchmarks with a Fourier neural operator surrogate.

1. Introduction

Neural operator surrogates have evolved as a compelling substitute for traditional computational fluid dynamics (CFD) solvers, providing significant speed enhancements for parametric flow forecasting. However, their application in safety-critical engineering domains such as design optimisation, digital twins, and real-time control is constrained by a fundamental credibility gap: these models serve as statistical approximators that do not inherently adhere to physical conservation laws, and they lack a rigorous method for quantifying the reliability of their predictions in terms of time and location. Addressing this gap necessitates uncertainty quantification (UQ) methodologies that are both statistically rigorous and grounded in physical principles. Conformal prediction (CP) has drawn considerable interest as a distribution-free framework that converts any predictive scores into finite-sample valid prediction sets without assumptions on the model class or data distribution. When naively applied to multi-regime fluid systems, where the predictive challenge significantly differs between laminar diffusion-dominated flows and turbulent advection-dominated flows, standard global CP calibration yields a singular threshold that is overly conservative in straightforward regions and potentially inadequate in complex ones. This intrinsic constraint drives the current research. While there have

been notable advances in neural surrogate modelling (NSM) and UQ, there remain major limitations for neural fluid surrogates (NFS) applied in CFD. Bayesian, probabilistic, and conformal estimation of uncertainty have been presented for neural operators and flow reconstruction models, but typically without integrating regime-aware validity, physics-guided calibration, and spatially adaptive UQ within a single unified framework [1], [2], [3], [4]. This limitation is most pronounced for incompressible flows, where the prediction challenge is extremely heterogeneous across Reynolds number regimes and highly non-uniform across the spatial domain, particularly near shear layers, wakes, and strong pressure variation regions [3], [4]. Therefore, globally calibrated limits of uncertainty can be unnecessarily broad in simple flow regions and inadequately informed in localised complex structures. To address this gap, we introduce a physics-guided stratified conformal prediction model for neural fluid surrogates. The framework integrates three components: physics residual-based calibration, regime-stratified conformal prediction, and locally adaptive uncertainty scaling via vorticity and pressure-gradient fields.

The principal contributions of this work are the following:

1. **Physics-residual nonconformity scoring:** We develop a dimensionally consistent, spatially resolved measure of nonconformity based on the steady-state incompressible Navier-Stokes momentum equations, substituting heuristic or solely data-driven metrics. We evaluate this proxy using Spearman rank correlation analysis in respect to the actual L2 prediction error, confirming its appropriateness for conformal calibration.

2. **Regime-stratified conformal prediction (RSCP) with comprehensive statistical validation:** We formulate a physically grounded partitioning of the parametric calibration set based on Reynolds number regime and boundary-condition type, resulting in regime-specific quantiles that provide safety bounds up to two orders of magnitude tighter than a single global threshold while maintaining the finite-sample coverage assurance of conformal prediction. We further demonstrate the framework's robustness to the selection of regime boundaries via systematic binning tests. The comprehensive framework is substantiated by a five-component robustness suite that includes bootstrap confidence intervals, K-fold cross-validation, multi-seed stability, significance-level sweep, and calibration-size ablation to ensure reproducibility and finite-sample reliability at every stage.

3. **Modular spatially adaptive scaling:** We present a spatial scaling stage that transforms uniform regime-level constraints into pixel-specific uncertainty maps by utilising any spatially variable signal associated with local prediction difficulties. Three candidates, including vorticity magnitude, velocity magnitude, and MC dropout standard deviation, are rigorously evaluated against independently optimised hyperparameters. The vorticity magnitude is utilised as the default signal, as it was able to enhance performance beyond the flat baseline across all regimes on both benchmarks, with minimal inference cost compared to the 50-pass MC dropout alternative.

4. **Pressure-gradient adaptation:** The same spatial scaling architecture is applied to pressure-field predictions using pressure-gradient magnitude as the scaling signal, demonstrating framework modularity across field variables. A spatial error-concentration analysis reveals that pressure prediction errors occupy a smaller proportion of the fluid domain than velocity errors, accounting for the additional efficiency gain.

2. Literature Review

2.1. Structural UQ & Distribution-free predictive inference

A recent study at UQ has strongly recognised CP as one of the most promising frameworks for reliable and model-agnostic predictive inference. [5] [6] Its major advantages are that it lets you generate prediction sets that are valid for a finite number of samples and do not rely on strong assumptions. This makes it a distribution-free alternative to confidence measures that depend on heuristic approaches. [5] This basic concept has since been modified to include more dynamic data contexts in addition to standard interchangeable settings. Online conformal inference has proved, specifically, that prediction sets might be updated adaptively

when the distribution changes, which shows how important it is to manage local coverage when the data-generating process varies over time. [7] In both sequential and multivariate settings, there is also a clear trend toward structured uncertainty. Current work on joint prediction-region estimation for time series has moved away from the use of small one-step intervals and instead focused on coherent uncertainty sets over different future states. [8] Conformal approaches for multivariate responses and multivariate functions have additionally illustrated that predictions are capable of being established for complicated structure targets, rather than being limited to scalar outputs. [9] [10]. These studies collectively demonstrate that CP has transitioned from a generic statistical tool to a flexible framework for UQ in progressively structured learning challenges.

2.2. Physics informed ML and operator-based surrogate modelling

Along with improvement in distribution-free inference, scientific machine learning (SciML) has created a robust structure for adding physical laws directly into the model that are based on data. [11] [12] Physics informed neural networks established the significant paradigm of imposing governing differential equations throughout the training, improving data efficiency and limiting the hypothesis space by previous physical structure. [11] Later, this method was expanded from finding solutions to learning operators. Physics informed neural operators (PINOs), in particular, demonstrated the ability to integrate training data with PDE residual constraints across various resolutions to increase fidelity, enable super-resolution behavior, minimise dependability on totally labelled high-resolution data. [13] Recent progress in fluid mechanics has revealed that surrogate models for incompressible flow can be trained without using levels data by directly enforcing the Navier-Stokes equation and boundary conditions. Such a strategy is particularly useful for applications requiring multiple queries, where producing label CFD data is costly. [14] Additionally, physics constrained encoder-decoder techniques have shown that the surrogate prediction and UQ could be achieved even with no labelled target data and enhance the extrapolation of out-of-distribution inputs. [15] Application-orientated neural PDE surrogates and Fourier neural operators (PNO) add to the capability of operating learning approaches to enable accurate and rapid spatiotemporal prediction in challenging physical systems. [16] [17] In all, these analyses reveal that physical limitations in training, architectural, and operator design levels are progressively guiding modern surrogate models.

2.3. ML-based Flow Reconstruction, Reduced-Order Modelling, and Solver Integration

Machine learning has emerged as a widely adopted computational tool for fluid modelling, incorporating reconstruction, sensing, control, computational fluid dynamics acceleration, reduced-order modelling, and super-resolution [18]. Benchmarking studies of supervised machine learning algorithms in flow prediction demonstrate that the selection of architecture substantially influences reliability, computational expense, and reconstruction quality across several applications, including wake estimation and flow-field prediction [19]. Nonlinear reduced-order frameworks that integrate autoencoders with recurrent architectures have shown proficiency in reconstructing unsteady wakes of bluff bodies, even in unfamiliar geometries [20]. Furthermore, turbulence-aware superresolution methods indicate that reconstruction accuracy is enhanced when scale organisation and flow dynamics are integrated into the learning process [21], [22]. Convolutional methods utilised in wall-bounded turbulence highlight that localised physical structures can be discerned from sparse near-wall data [4]. These findings collectively suggest that effective fluid-learning models must incorporate local complexity, multiscale dynamics, and spatial structure instead of depending just on global flow statistics.

In the realm of solver integration, sensor-based deep model predictive control has demonstrated that effective control policies may be derived from limited real-world data utilising low-dimensional recurrent surrogates[18]. In turbulence modelling, investigations of data-driven, subgrid-scale parameterisation highlight that performance is significantly influenced by locality, stencil design, and the selection between pointwise and nonlocal architectures [23]. Recent advancements have reconfigured discretised PDE solvers in machine learning libraries, allowing convolution-like operators to emulate numerical discretisation on contemporary AI hardware [24]. These advancements underscore the increasing convergence of classical numerical techniques with data-driven models.

2.4. Physically constrained reliability and uncertainty-aware fluid learning

Awareness of uncertainty in fluid learning has grown and is now more and more essential as the neural models have been applied to advanced and critical flow problems. [25] [26] The general assessment of UQ in deep learning (DL) addresses the trade-off among ensemble methods, calibration accuracy, computing cost, dropout-based uncertainty, Bayesian approximations, and defining the difference between epistemic uncertainty and aleatoric uncertainty [25] Aerodynamic reorganisations based on sparse pressure are being applied in fluid-specific applications, incorporating Monte Carlo dropout and probabilistic regression to determine aerodynamic loads and flow fields with consideration of model uncertainty and data. The study also highlights the spatially heterogeneous character of uncertainty in fluid dynamics with low-order aerodynamic reconstructed flow, which suggests that predictive uncertainty is a function of local flow, sensor placement, and noise measurement features. [3] Fluid learning models are also dependent on physically constrained trustworthiness. Realisability-informed loss functions are being suggested in turbulence modelling in order to strengthen the stability and physical consistency of trained closure-based mappings, finding that accurate prediction is not enough when physical constraints are violated by the model outputs. [27] Further research of near-wall turbulence and pressure reconstruction also suggests that uncertainty may be increased by sparse measurements, localised coherent frameworks, and boundary effects [28] [29]. This set of works finds out that both physically meaningful structure and statistically meaningful measure of uncertainty are recognized as key requirements of fluid learning systems.

2.5. Consolidation of the literature

The literature collectively indicates a convergence of three significant developments. Initially, conformal and distribution-free UQ has made predictive inference more rigorous, and reliable uncertainty models for sequential, structured, and functional outputs are becoming more important. [5] [10] [7] Subsequently, physics-constrained surrogate learning, neural operators, and physics-informed neural network techniques that put physical equations directly into the model's inference and training have contributed to scientific machine learning (SML's) a stronger physical foundation. [11] [13] [14] Finally, uncertainty-aware prediction, super-resolution, flow reconstruction, and reduced-order modelling in contexts where the multiscale structure and local flow organization highly affect the accuracy of prediction have received more attention in fluid-based ML. [3] [30] [28] These usual trends suggest a more general pattern towards developing fluid surrogate models which are sensitive to local flow behavior, statistically informed, and physically constrained. However, there is still a vital methodological gap for CFD in neural fluid surrogates. Although current distribution-free and conformal techniques have excellent coverage guarantees, they are not frequently developed based on physically useful calibration signals suited to the prediction of flow fields. [5] [10] Conversely, by integrating governing equations, physics-informed and operator-based surrogate models improve prediction accuracy, but these models generally lack the ability to provide regime-aware conformal validity or spatially adaptive uncertainty bounds with a finite number of samples guarantees. [6] [14] Meanwhile reconstruction-orientated and fluid-specific uncertainty-aware research is also progressively revealing finding the substantial localisation of the prediction challenges, especially in close proximity of shear layers, wakes, pressure-sensitive structures and wall-bounded regions. [3] [4] However, such understanding still has not completely entered into a unified structure which can be considered spatially adaptive, regime-stratified, and physics-guided. The current study's main methodological inspiration arises from this unresolved gap.

3. Methodology

3.1. Dataset & Model

The datasets used in this study are derived from CFDONEval [31] datasets. Two classic fluid datasets are used here. One is Lid-Driven Cavity, and another one is Cylinder dataset. The proposed framework utilises a 2D Fourier Neural Operator (FNO) to learn the mapping between fluid states. The architecture initiates with an input lifting layer that transforms the input—consisting of the fluid state, spatial grid coordinates, geometry mask, and case-specific parameters—into a 32-dimensional latent space. The latent representation is subsequently transmitted through four iterative spectral layers. Each spectral layer employs a dual-path architecture consisting of a spectral path and a local residual path. A 2D Real Fast Fourier Transform (RFFT)

is executed in the spectral domain, where the lowest $k = 12$ modes are filtered and subsequently multiplied by trainable complex weight tensors. Concurrently, the residual path employs a 1×1 convolution to maintain local spatial information. The outputs from both paths are aggregated and subjected to a GELU activation function. The latent features are subsequently projected back to the physical domain via a projection stage that includes a hidden linear layer (width=128) and a concluding output layer. A masked mean squared error (M-MSE) loss is utilised to concentrate the model's learning exclusively on fluid physics. The network is trained using Backpropagation Through Time (BPTT) across the complete temporal sequence and optimised with the Adam algorithm (1×10^{-3}) utilising a step learning rate scheduler. All experiments were executed in a Google Colaboratory environment equipped with a single NVIDIA L4 GPU.

3.2. Problem Formulation

Let

$$\widehat{\mathbf{u}}_s = \mathcal{F}_\theta(\mathbf{u}_s^{t_1}, \mathbf{u}_s^{t_2}; \mathrm{Re}_s)$$

represents the prediction of pretrained Fourier Neural Network (FNO) where

$$\mathbf{u} = (u_x, v_y, p) \in \mathbb{R}^{64\times64\times3}$$

is the velocity-pressure field on a 64 X 64 grid. The dataset is divided into $\mathcal{D}_{\text{cal}}$ and test set $\mathcal{D}_{\text{test}}$ where $\mathcal{D}_{\text{cal}} \cap \mathcal{D}_{\text{test}} = \emptyset$. This partitioning guarantees that all uncertainty bounds are calibrated solely with the calibration data and assessed exclusively on the unseen test set, thereby fulfilling the exchangeability criteria inherent to the conformal prediction framework.

3.3. The Conformal Physics-Residual (CPR) Framework:

The pretrained Fourier Neural Operator ($\mathcal{F}_\theta$) delivers efficient point predictions of the flow field; nevertheless, it is fundamentally deterministic and does not possess a way to convey confidence in its outputs. In practical computational fluid dynamics, constraining this epistemic uncertainty is as essential as the prediction itself. To systematically quantify predictive uncertainty in the absence of the inaccessible ground truth $\mathbf{u}_s^*$ during inference, we build upon the Conformal Prediction with Physics Residual Error (CP-PRE) framework based on this framework. The fundamental idea of CP-PRE is to employ the breach of regulating physical laws as an unsupervised measure of non-conformity. Although our dataset comprises unsteady, time-dependent flow fields, previous studies have demonstrated that calculating complete spatiotemporal Navier-Stokes residuals' temporal derivatives ($\frac{\partial \mathbf{u}}{\partial t}$) from neural network predictions can result in considerable numerical noise. High-frequency artefacts can impair the monotonic correlation necessary for good conformal calibration. Thus, we implement a computationally efficient spatial momentum proxy that delineates the instantaneous convection-diffusion dynamics of the flow. When computing higher-order temporal and pressure gradients using surrogate predictions, numerical artefacts might reduce calibration efficiency. To maximise the signal-to-noise ratio of the non-conformity score, we choose this instantaneous spatial momentum proxy. Due to the presence of parametrically varying domain geometries within our datasets, a static global spatial discretisation proves to be inadequate. The physical grid spacings Δx_s and Δy_s are dynamically determined for each specific sample s from the coordinate grid. The Lid-Driven Cavity dataset employs a uniform Cartesian discretisation with consistent spacing throughout the entire domain. We determine the global step sizes by calculating the total extent of the domain and dividing it by the number of spatial intervals. $x_{s(\max)}$ and $x_{s(\min)}$ represent the spatial dimensions of the domain pertinent to the particular sample under consideration.

$$\Delta x_s = \frac{x_{s(\max)} - x_{s(\min)}}{W - 1} \tag{1}$$

$$\Delta y_s = \frac{y_{s(\max)} - y_{s(\min)}}{H - 1} \tag{2}$$

This formulation considers the parametric characteristics of our datasets; although the grid resolution (H, W) remains constant, the physical domain size—and consequently the differential spacing—fluctuates with each sample.The cylinder dataset utilises a non-uniform mesh for discretisation to accurately capture high-gradient

areas in the obstacle's boundary layer and the downstream wake. Thus, we define a local spacing tensor, wherein each entry denotes the physical distance corresponding to a central difference stencil:

$$\Delta x_{(i,j)} = |x_{i,(j+1)} - x_{i,(j-1)}| + \epsilon \tag{3}$$

$$\Delta y_{(i,j)} = |y_{(i+1),j} - y_{(i-1),j}| + \epsilon \tag{4}$$

where $\epsilon = 10^{-8}$ serves as a regularisation term to ensure numerical stability. This method guarantees that the Physics Residual Error (PRE) maintains physical consistency across areas with differing mesh densities. To guarantee that the Physics Residual Error (PRE) is assessed solely in regions where the complete numerical stencil is valid, we establish a Strict Mask (M_{strict}). For a pixel at (i, j) to be incorporated in the loss, it and all four of its adjacent neighbours must reside within the fluid domain ($M = 1$).This is mathematically executed as a logical "AND" operation across the shifted binary mask.

The Lid-Driven Cavity dataset features a basic square domain devoid of internal obstructions. Boundary management is executed through interior slicing. Due to the domain being confined by physical walls at the grid edges, we omit the boundary pixels $(0, N)$ from the residual computation. By assessing the physics residuals exclusively on the interior grid, we guarantee that the non-conformity score accurately reflects the model's capacity to adhere to momentum conservation laws within the active flow field, rather than its proficiency in reproducing fixed boundary conditions.

The main aim of the PRE is not to achieve flawless physical conservation, but to offer a computationally efficient, localised representation of physical discrepancies in the model's predictions.We calculate the Physics Residual Error (PRE) by estimating the partial derivatives of the flow field utilising a second-order central difference scheme on a discrete Eulerian grid.

1. **First-Order Spatial Gradients (∇)**

$$\frac{\partial f}{\partial x} \approx \frac{f_{(i+1),j} - f_{(i-1),j}}{2\Delta x} \tag{5}$$

$$\frac{\partial f}{\partial y} \approx \frac{f_{i,(j+1)} - f_{i,(j-1)}}{2\Delta y} \tag{6}$$

2. **2. The Laplacian Operator (∇^2)**

$$\begin{aligned}\mathcal{L}(f)_{i,j} = &\frac{f_{i+1,j} - 2f_{i,j} + f_{i-1,j}}{(\Delta x_s)^2} \\ &+ \frac{f_{i,j+1} - 2f_{i,j} + f_{i,j-1}}{(\Delta y_s)^2}\end{aligned} \tag{7}$$

The physics residuals evaluate the accuracy of the predicted fields in relation to the conservation of momentum. The discrete residuals for the x and y momentum components (r_u, r_v) are expressed as:

$$r_u = u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} + \frac{\partial p}{\partial x} - \nu\nabla^2 u \tag{8}$$

$$r_v = u\frac{\partial v}{\partial x} + v\frac{\partial v}{\partial y} + \frac{\partial p}{\partial y} - \nu\nabla^2 v \tag{9}$$

The localised Physics Residual Error is defined as the average magnitude of the steady-state incompressible Navier-Stokes momentum residuals at each valid interior grid point.

$$PRE(i, j) = \frac{|r_u| + |r_v|}{2} \tag{10}$$

3.4. Statistical Validation of the Physics-Informed Proxy

The operational validity of the proposed non-conformity measure depends on its capacity to function as a dependable, unsupervised proxy for the actual predictive error (L_2). Due to the inability to compute the true predictive error in the absence of ground-truth data during inference, we utilise the Physics-Residual Error (PRE) as a consistent measure of model uncertainty.

Spearman Rank Correlation Analysis: In order to substantiate this claim, we conduct an analysis of the correlation between the PRE score and the empirical L_2 error, employing the Spearman Rank Correlation Coefficient (ρ) as our statistical measure.In contrast to linear metrics, the Spearman coefficient evaluates the extent to which the relationship can be characterised by a monotonic function, which is essential for the rank-based calibration employed in Conformal Prediction.

3.5. Regime-Stratified Conformal Calibration (RSCP)

Standard CP frameworks employ a singular calibration set to obtain a global quantile $hatq$; however, this methodology is inadequate for parametric fluid dynamics because of the intrinsic spatial and parametric heteroscedasticity of the flow field.A laminar flow is regulated by linear diffusion, while high-Re flows are primarily influenced by non-linear advection and chaotic vortex shedding. Considering these as a singular distribution compels the CP framework to select a global quantile $\hat{q}_{global}$ that is:

- **Over-conservative for Laminar Flows:** Resulting in excessively broad uncertainty intervals that obscure the excellent precision of the surrogate in steady conditions.
- **Insufficient for Turbulent Flows:** Potentially underestimating the risk in high-gradient areas where the PRE scores are inherently several orders of magnitude greater.

Implementing Regime Stratified Conformal Prediction (RSCP) expressly reinstates physical consistency in the calibration process. Dividing the suite into distinct categories guarantees that each quantile $\hat{q}_b$ is obtained from physically interchangeable samples. This Physics-Stratified methodology enables the framework to deliver regime-aware efficiency, attaining a 90% coverage guarantee with intervals customised to the physical "tolerance" of each regime, thus optimising the utility of the uncertainty estimate for engineering decisions.

3.5.1 Cavity Dataset Regime Stratification

The cavity dataset is characterised by internal recirculation cells, with complexity increasing monotonically with the Reynolds number (Re). We divide the three regimes according to the Re thresholds.

- **Regime 1 (Steady Wake, $Re < 1000$):** Steady-state flow dominated by diffusion, wherein the surrogate attains optimal precision.
- **Regime 2 (Boundary Condition Driven):** This regime signifies separate cases where the boundary conditions are varied, keeping the Reynolds number constant.
- **Regime 3 (Unsteady Wake, $Re > 1000$):** Potentially unstable regimes in which high-frequency spatial gradients result in elevated PINC scores.

3.5.2 Cylinder Dataset Regime Stratification

The cylinder benchmark entails a more intricate interaction between fluid inertia and boundary limitations. To address this, we employ a mixed stratification method.

- **Regime 1 (Steady wake, $Re < 1000$):** Distinguished by consistent or mildly unstable divided flow behind to the cylinder.
- **Regime 2 (Boundary Condition Driven):** A unique regime in which the flow characteristics are determined by particular alterations in boundary conditions.
- **Regime 3 (Unsteady wake $Re > 1000$):** Indicates the shift towards chaotic vortex shedding, characterised by the surrogate's maximum L_2 inaccuracy.

3.5.3 Calibration and Inference Procedure

The Regime Stratified Conformal Prediction (RSCP) methodology is executed in two separate stages: an offline calibration phase to set physical standards and a real-time inference phase for validity evaluation.

Utilising a reserved calibration set $\mathcal{D}_{cal}$, we compute regime-specific thresholds $\{\hat{q}_0, \hat{q}_1, \hat{q}_2\}$. For each regime $\mathcal{B}_k$, the PRE scores (s_s) are calculated utilising computationally efficient Spatial Momentum Proxy described in section 3.3. The threshold is defined as the $(1-\alpha)$-th empirical quantile.

$$\hat{q}_k = \text{Quantile}\left(\{s_s\}_{s\in\mathcal{B}_k}, \frac{\lceil (n_k+1)(1-\alpha)\rceil}{n_k}\right) \tag{11}$$

where $\mathcal{B}_k$ denotes the subset of calibration samples associated with physical regime k (e.g., laminar or turbulent), and $n_k = |\mathcal{B}_k|$ signifies the quantity of samples within that regime.In this study, α represents the significance level, specified at 0.10, which outlines the maximum permissible likelihood of the model's physics residual surpassing the calibrated threshold.

For a new unseen parametric test sample from $\mathcal{D}_{test}$,the FNO model produces the predicted flow field $\hat{V}$.The sample is categorised into a distinct physical regime k according to its denormalised Reynolds number (Re) and relevant boundary condition characteristics.The actual grid spacing Δx_s and Δy_s is directly obtained from the input coordinate tensors to maintain dimensional consistency across the parametric suite.The PRE score (s_t) is being calculated and check validity if it satisfies the regime-specific threshold.

$$s_{test} \leq \hat{q}_k$$

3.6. Evaluation Metrics: Coverage and Efficiency

To systematically analyse the effectiveness of the RSCP framework, we establish two principal metrics : Empirical Coverage (Reliability) and Threshold Efficiency (Sharpness).
The primary guarantee of conformal prediction is marginal coverage. In our regime-stratified paradigm, coverage is defined as the empirical likelihood that unobserved test samples within a designated physical bin satisfy their assigned validity threshold.The empirical coverage C_k is defined as,

$$C_k = \frac{1}{N_{test,k}} \sum_{i=1}^{N_{test,k}} \mathbb{I}\left(s_{test,i} \leq \hat{q}_k\right) \tag{12}$$

where $\mathbb{I}$ is the indicator function to check whether it satisfies the regime specific threshold
Coverage determines reliability, while efficiency governs the practical applicability of the uncertainty estimate. A reduced threshold ($\hat{q}$) signifies a more rigorous constraint on physical validity. To measure the superiority of our method compared to global calibration, we define the Efficiency Multiplier (η_k) for each regime as follows:

$$\eta_k = \frac{\hat{q}_{\text{global}} - \hat{q}_k}{\hat{q}_{\text{global}}} \times 100\%$$

where $\hat{q}_{global}$ denotes the consolidated threshold computed throughout the entire dataset and $\hat{q}_k$ refers to regime wise threshold.

3.7. Empirical Validation and Statistical Robustness

Prior to delivering the definitive conformal prediction bounds, it is essential to thoroughly evaluate the stability, dependability, and data efficiency of the Regime-Stratified Conformal Prediction (RSCP) framework. To guarantee the framework's resilience to finite-sample variance and arbitrary data partitioning, the calibration engine underwent an extensive five-part statistical validation suite.

3.7.1 Multi-Seed Stability

To verify the framework's invariance to random calibration-test splits, the dataset underwent a 10-seed randomised shuffle test. This test verifies that the regime-specific calibration engine operates consistently, irrespective of the particular physical snapshots selected at random to the calibration pool.

3.7.2 K-Fold Cross-Validation

A standard K-fold cross-validation ($K = 5$) was performed on the complete dataset. The data was divided into five distinct folds, using four folds (80%) for calibration and the final fold (20%) for testing in an iterative manner. This test tries to measure the variance reduction obtained by our stratified methods by rigorously adjusting the calibration bounds throughout the whole parametric domain.

3.7.3 Bootstrap Resampling

A bootstrap approach with 1000 iterations and replacement was applied on the test set to determine 95% confidence intervals for the empirical coverage, so confirming the stability of the implemented thresholds.This approach isolates the variance of the inference phase, estimating the maximum statistical fluctuation of the framework when applied to unseen data.

3.7.4 Significance Level (α) Swap

A major theoretical requirement for Conformal Prediction is marginal validity, signifying that the empirical error rate must reliably correspond with the user-defined risk tolerance. To ascertain that our Physics-Informed Non-Conformity proxy consistently aligns with this theoretical risk, the nominal significance level was varied over 21 distinct levels ($\alpha \in [0.01, 0.30]$).

3.7.5 Calibration Size Ablation

Producing high-fidelity Computational Fluid Dynamics (CFD) data for offline calibration is resource-intensive. To ascertain the practical engineering constraints of our framework, An ablation research on sample size was performed, reducing the calibration pool from $N = 2000$ to $N = 100$, to find the minimal data threshold necessary for ensuring stable conformal bounds.

3.8. Motivation for Spatially Adaptive Calibration

The baseline framework effectively categorizes the data into discrete flow regimes; nevertheless, it is fundamentally flawed as it presumes that prediction uncertainty is uniformly distributed throughout each regime.This method overlooks the localized and highly variable characteristics of fluid dynamics by employing a uniform scalar bound ($\hat{q}_{bin}$) throughout the whole spatial domain.In practice, neural operators such as FNO, a deep learning framework designed to learn the mapping between distinct physical fields, demonstrate non-uniform distributions of error. Irrespective of the overall nature of the simulation, whether laminar or turbulent, the model generally exhibits strong performance in regions characterized by smooth and undisturbed fluid flow. However, it encounters significant challenges in areas marked by complex interactions, including shear layers, wake shedding, and boundary walls.As a result, the implementation of a uniform, regime-wide safety boundary necessitates an inefficient trade-off: the range becomes unnecessarily broad in stable areas, while concurrently posing a risk of inadequate coverage in chaotic sub-regions.Obtaining spatial efficiency while maintaining the statistical assurances of conformal prediction demands that the uncertainty bounds adjust to local flow complexity on a pixel-by-pixel basis. This section presents a modular scaling design that addresses the need for both velocity and pressure field predictions.

3.9. Spatially Adaptive Scaling Factor

3.9.1 Modular Scaling Architecture

The spatial scaling phase transforms uniform per-regime limits into pixel-specific uncertainty ranges. The architecture is modular: any scalar field $\sigma(\mathbf{x})$ that correlates with local prediction difficulty can operate as the scaling mechanism. All candidates follow to a uniform functional structure:

$$\sigma(\mathbf{x}) = 1 + \beta \left(\frac{S(\mathbf{x})}{S_{99}} \right) \tag{13}$$

Here, $S(\mathbf{x})$ denotes the spatial signal at location $\mathbf{x}$, S_{99} represents its 99th percentile calculated solely from the calibration set (to reduce the impact of isolated extreme values and prevent data leakage), and β is a scaling hyperparameter optimised independently for each candidate signal.

For velocity-field uncertainty, we evaluate three candidate signals:

1. **Vorticity magnitude**: $S(\mathbf{x}) = |\omega(\mathbf{x})|$, where $\omega = \nabla \times \mathbf{V}$ is the curl of the velocity field.
2. **Velocity magnitude**: $S(\mathbf{x}) = |\mathbf{V}(\mathbf{x})|$.
3. **MC Dropout standard deviation**: $S(\mathbf{x}) = \hat{\sigma}_{\mathrm{mc}}(\mathbf{x})$, the pointwise predictive standard deviation estimated from $T = 50$ stochastic forward passes.

The scaling parameter β is independently chosen for each candidate by minimising the expected interval width on the calibration set across the $\{0.0, 0.5, 1.0, 1.5, 2.0, 3.0, 5.0\}$:

$$\beta^* = \arg\min_{\beta} \frac{1}{B} \sum_{b=1}^{B} 2\hat{q}_b \cdot \bar{\sigma}_{\mathrm{cal},b} \tag{14}$$

where $\overline{\sigma}_{cal,b}$ is the mean scaling factor over calibration samples in bin b.This technique guarantees that no candidate adopts hyperparameter tuning from another; each is assigned its own optimised β via an identical protocol.In order to enhance the validation of the chosen β, the Spearman rank correlation is calculated between the adaptive interval width and the actual pointwise prediction error on the test set, serving as a post-hoc diagnostic measure.

3.9.2 Physical Motivation for Vorticity as the Default Signal

Among the three velocity-field candidates, vorticity magnitude is proposed as the primary scaling signal.In order to develop a spatially adaptive conformal interval, it is essential that the scaling factor $\sigma(x)$ is informed by a physical heuristic that exhibits a strong correlation with the localized prediction error of the neural network. In the analysis of the velocity field (u, v), we opted for the local vorticity magnitude, $|\omega(x)|$, as our primary metric, eschewing more straightforward measures like velocity magnitude ($|V|$) or kinetic energy. Vorticity is mathematically characterized as the curl of the velocity field ($\omega = \nabla \times V$), serving as a precise quantification of the local rotational and shearing dynamics within the fluid system. This serves as a direct mathematical representation of localized flow complexity. By establishing a connection between the conformal scaling factor and vorticity, we dictate the uncertainty bounds to broaden precisely in regions where fluid behavior exhibits significant non-linearity and poses challenges for model representation, while permitting the bounds to narrow in the more predictable, irrotational freestream.In an effort to substantiate this physical intuition, we conducted empirical tests to examine the correlation between localized model error and fluid vorticity. We calculated the Spearman rank correlation coefficient (Sp) to assess the relationship between the pixel-wise absolute velocity error and the local vorticity magnitude throughout the test set.In a two-dimensional incompressible flow, vorticity is defined as the z-component of the curl of the velocity field:

$$\omega_z = \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} \tag{15}$$

where u and v denote the velocity components in the x- and y-directions, respectively.This scalar quantity quantifies the local rotational dynamics of fluid components situated within the two-dimensional plane.For the purposes of to perform numerical computations based on the predicted velocity fields situated on a structured Cartesian grid, a second-order central finite difference scheme is employed at the interior nodes.

$$\tilde{\omega}_{(i,j)} = \left| \frac{v_{(i,\,j+1)} - v_{(i,\,j-1)}}{2\,\Delta x} - \frac{u_{(i+1,\,j)} - u_{(i-1,\,j)}}{2\,\Delta y} \right| \cdot \mathcal{M}_{i,j} \tag{16}$$

here, Δx and Δy represent the grid spacings along their respective axes. The term $\mathcal{M}_{i,j} \in \{0, 1\}$ denotes a binary fluid domain mask, which serves to confine the computational process to active fluid cells while

omitting solid boundaries and regions occupied by obstacles.We carried out a random sampling of thousands of spatial coordinates within the test dataset and calculated the Spearman rank correlation coefficient (Sp) between the true absolute velocity error at a specific pixel and the local vorticity magnitude at that identical pixel. The Spearman correlation found is **0.50** for the cavity dataset and **0.75** for the cylinder dataset. The low correlation of the cavity case was due to the lid-driven cavity constitutes a constrained domain significantly influenced by wall-bounded shear dynamics. The no-slip condition results in a pronounced peak in vorticity magnitude at all solid boundaries, including the straight walls, where the neural network consistently demonstrates high predictive accuracy. The model exhibits its most significant errors primarily concentrated at the corner singularities and secondary eddies. High vorticity regions include both straightforwardly predictable straight boundary layers and intricate separating corners, leading to a naturally diminished monotonic signal-to-noise ratio of the heuristic.

3.9.3 Calibration and Inference with Spatial Scaling

In the calibration phase, localized non-conformity scores are computed by di- viding the true absolute error by the local scaling factor

$$m_i = \frac{|V_{pred}(x_i) - V_{true}(x_i)|}{\sigma_{vel}(x_i)} \tag{17}$$

In order to address the discrepancies in difficulty throughout the grid, we compute adjusted scores by dividing the true absolute error by the corresponding local scaling factor.This adaptation serves to standardise the challenge of the problem, thereby ensuring that significant errors arising in physically complicated, high-vorticity areas are appropriately diminished in weight.We compute the 95th-percentile spatial score for each sample and subsequently establish the calibrated threshold $\hat{q}_{bin}$ as the $(1 - \alpha)$ empirical quantile of the aggregated scores.

$$\hat{S}_{vel}(x) = V_{pred}(x) \pm (\hat{q}_{bin} \cdot \sigma_{vel}(x)) \tag{18}$$

Finally, the adaptive prediction interval for the velocity at a designated spatial coordinate x is established by the multiplication of the regime's base quantile with the local physical scaling factor during the inference phase on a new test sample.

3.9.4 Evaluation Metrics

To evaluate the framework performance, we have employed four metrics.

- **Sample Coverage** This serves as the principal statistical validator. Within the proposed framework, a sample is regarded as "covered" when its spatial non-conformity score, which is defined as the 95th percentile of the absolute error across the interior grid, is less than or equal to the calibrated quantile $\hat{q}_{bin}$. From a mathematical perspective
- **Pixel-Wise Coverage:** Sample coverage functions at the aggregate image level, whereas pixel-wise coverage assesses the physical domain. The computation involves determining the raw percentage of individual spatial pixels within the test set for which the true absolute prediction error is effectively constrained by the local adaptive interval.
- **Mean Interval Width:** This denotes the mean physical dimensions of the uncertainty "box." For the baseline scenario, the width is defined as a constant scalar $(2 \cdot \hat{q}_{global})$. In the adaptive phases, the width exhibits variability across individual pixels, and we present the spatial mean accordingly:

$$\text{Mean Interval Width} = \frac{1}{H \cdot W} \sum_{x,y} 2 \cdot \hat{q}_{bin} \cdot \sigma(x, y) \tag{19}$$

- **Relative Spatial Efficiency (RSE):**

$$RSE = \frac{\text{Width}_{(\text{Base Framework})}}{\text{Width}_{(\text{Vorticity Adaptive})}} \tag{20}$$

3.9.5 Validation of Spatially Adaptive Framework

As like the base framework, we validate our primary vortiicty Adaptive Framework same five validation metrics.

- Multi-Seed Stability
- K-Fold Cross-Validation
- Bootstrap Resampling
- Significance Level (α) Swap
- Calibration Size Ablation

3.10. Pressure Field Adaptation via Gradient Scaling

The modular scaling design outlined in Section 3.9 seamlessly adapts to various output field variables. In velocity fields $(\hat{u}, \hat{v})$, vorticity magnitude is the preferred scaling heuristic. In scalar pressure fields ($hatp$), vorticity is not a suitable physical indicator, as it originates from the curl of the velocity field rather than from pressure dynamics.The apparent analogue for pressure is the magnitude of the pressure gradient $\|\nabla P\|$, which delineates areas of significant pressure variation where the surrogate is most prone to substantial prediction mistakes.In the framework of a discrete two-dimensional spatial grid, the magnitude of the pressure gradient at any internal coordinate (x, y) can be approximated through the application of second-order central difference methods.

$$\|\nabla P(x,y)\| = \sqrt{\left(\frac{\partial p}{\partial x}\right)^2 + \left(\frac{\partial p}{\partial y}\right)^2} \tag{21}$$

By employing this physical trigger, we establish the pressure-specific variance scaling factor, which is analogous to the vorticity scaling factor.

$$\sigma_{pres}(x) = 1 + \beta_P \left(\frac{\|\nabla P(x)\|}{\|\nabla P\|_{99,cal}}\right) \tag{22}$$

here, the quantity $\|\nabla P\|_{99,cal}$ represents the 99th percentile of the gradient magnitude and the β_P denotes a sensitivity hyperparameter calculated using calibration test.Normalization using the 99th percentile effectively mitigates the influence of isolated extreme values, thereby preserving the integrity of the scaling factor throughout the fluid domain and facilitating a balanced, dynamic interval.Subsequently, we calculate the physics-normalized non-conformity score for each pixel within the calibration set:

$$m_i = \frac{|P_{pred}(x_i) - P_{true}(x_i)|}{\sigma_{pres}(x_i)} \tag{23}$$

Then, for a new, unobserved test sample, the adaptive pressure prediction interval is established by projecting the calibrated threshold back through the local gradient field.

$$\hat{C}_{pres}(x) = P_{pred}(x) \pm (\hat{q}_{bin} \cdot \sigma_{vel}(x)) \tag{24}$$

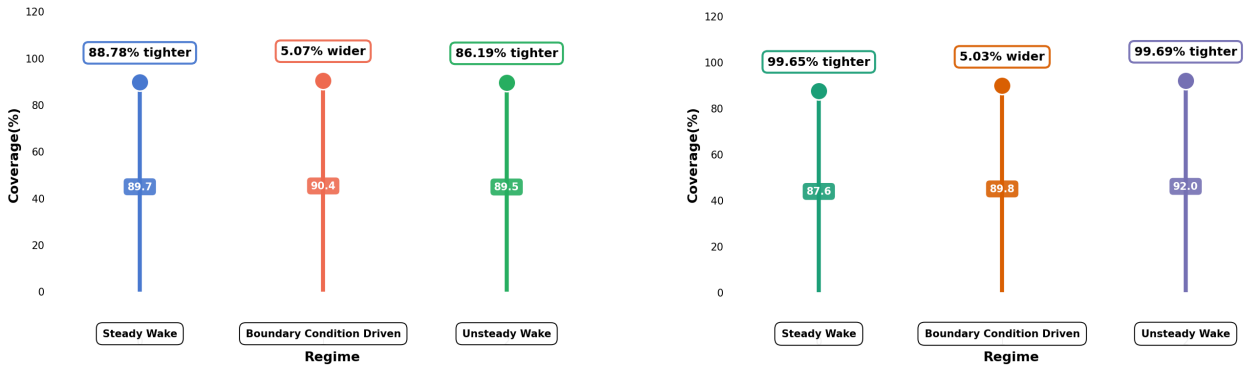
Ultimately, the framework is evaluated on the same four evaluation metrics as before: Sample Coverage, Pixel-Wise Coverage, Mean Interval Width, and Relative Spatial Efficiency (RSE).

4. Results & Discussion

4.1. Regime Stratified Conformal Prediction (RSCP) framework

4.1.1 Coverage & Efficiency

First, the Spearman rank correlation of physics residual error with prediction error was examined to verify the nonconformity of the physics proxy. It was found to be 0.94 for the cavity dataset and 0.90 for the cylinder dataset, suggesting strong correlation. Before dividing into regimes, the global threshold value was calculated. For the cavity dataset, it was found to be 3564.63. The physics residual calculations were conducted in physical space and were not normalized to the [0,1] range; consequently, the absolute values may vary based on the size of the domain and the scale of the variables involved. Nevertheless, this score was being used solely as a ranking-based non-conformity score within the conformal prediction framework. The regime-wise calculation adhered to the same framework; thus, the coverage guarantee of the conformal prediction framework remains valid irrespective of the residual scale. In the analysis of the CP-PRE framework applied to the cavity dataset, the original test set of the dataset comprising 4465 instances was partitioned into a calibration set and a test set to prevent data leakage. The calibration set contained 1000 instances, while the test set was composed of 3365 instances. Before dividing into regimes, the global threshold value was calculated. It was found to be 3564.63. The calibration set was categorized into three distinct regimes according to their Reynolds numbers and boundary conditions. The first zone was a steady wake zone, which has Reynolds numbers of less than 1000. The PRE threshold was calculated to be 400.26, representing an 88.78% tighter comparison to the global threshold. In the second zone, where boundary conditions vary while maintaining fixed Reynolds numbers, the PRE threshold was determined to be 3745.66, indicating a 5.07% increase in width. In the third regime, characterized by Reynolds numbers exceeding 1000, the calculated PRE threshold was found to be 492.02, representing an 86.19% reduction relative to the global threshold. The data underscores the necessity of a stratified framework for this regime. Employing a universal safety threshold across the dataset will include both steady wake and unsteady conditions, where a reduced threshold is essential. Conversely, it will be insufficiently addressed for BC-driven cases, which necessitate a higher threshold. It was found that, with the global threshold, only 68.1% of data was covered for the boundary condition cases. However, the efficacy of this framework is determined not solely upon this factor; it also necessitates a guarantee of coverage. With an (α) value set at 0.1, the regime-stratified framework demonstrated coverage rates of 89.7%, 90.4%, and 89.5% for the steady wake, BC-driven, and unsteady wake zones, respectively. The identical framework was applied to the cylinder dataset. The initial dataset, comprising 6762 instances, was partitioned into a calibration subset of 2000 instances and a testing subset of 4762 instances. In this instance, the global threshold was identified to be 3214.47. The regime threshold for a steady wake regime is only 11.09, which is 99.65% tighter than the global threshold. The threshold will be needlessly dragged to this enormous number if the global framework is used. The zone threshold for boundary condition-driven cases is 3376.30 , which is 5.03% wider than the global threshold. This suggests that the safety score for these complex cases needs to be enhanced. Lastly, the safety score only reaches 9.74, which is 99.69% tighter, for the unsteady wake with higher Reynolds numbers over 1000. The analysis reveals a conformal prediction coverage of 87.6% within the steady wake regime, which is slightly below the coverage limit, while a coverage of 89.8% is observed in the boundary condition-driven zone. The coverage increased to 92.0% concerning the unsteady wake phenomenon.



(a) (b)

Figure 1: Regime Stratifed Gain Over Global Quantile for (a) Cavity Dataset and (b) Cylinder Dataset

Table 1: Regime Stratified Conformal Prediction Coverage & Efficiency

Dataset	**Flow Regime**	N_{cal}	N_{test}	q_{glob}	**Global Coverage**	$\hat{q}_k$	**RSCP Coverage**	**RSCP Efficiency**
Cavity	Steady Wake ($Re < 1000$)	235	755	3564.63	100%	400.26	89.7%	88.78% tighter
	Boundary Condition Driven (Constant Re)	348	1240	3564.63	68.1%	3745.66	90.4%	5.07% wider
	Unsteady Wake ($Re > 1000$)	417	1470	3564.63	100%	492.02	89.5%	86.19% tighter
Cylinder	Steady Wake ($Re < 1000$)	285	695	3214.47	100%	11.09	87.6%	99.65% tighter
	Boundary Condition Driven (Constant Re)	1425	3377	3214.47	85.8%	3376.30	89.8%	5.03% wider
	Unsteady Wake ($Re > 1000$)	290	690	3214.47	100%	9.74	92%	99.69% tighter

*here Re represents Reynolds Number

4.1.2 Validation of the RSCP framework

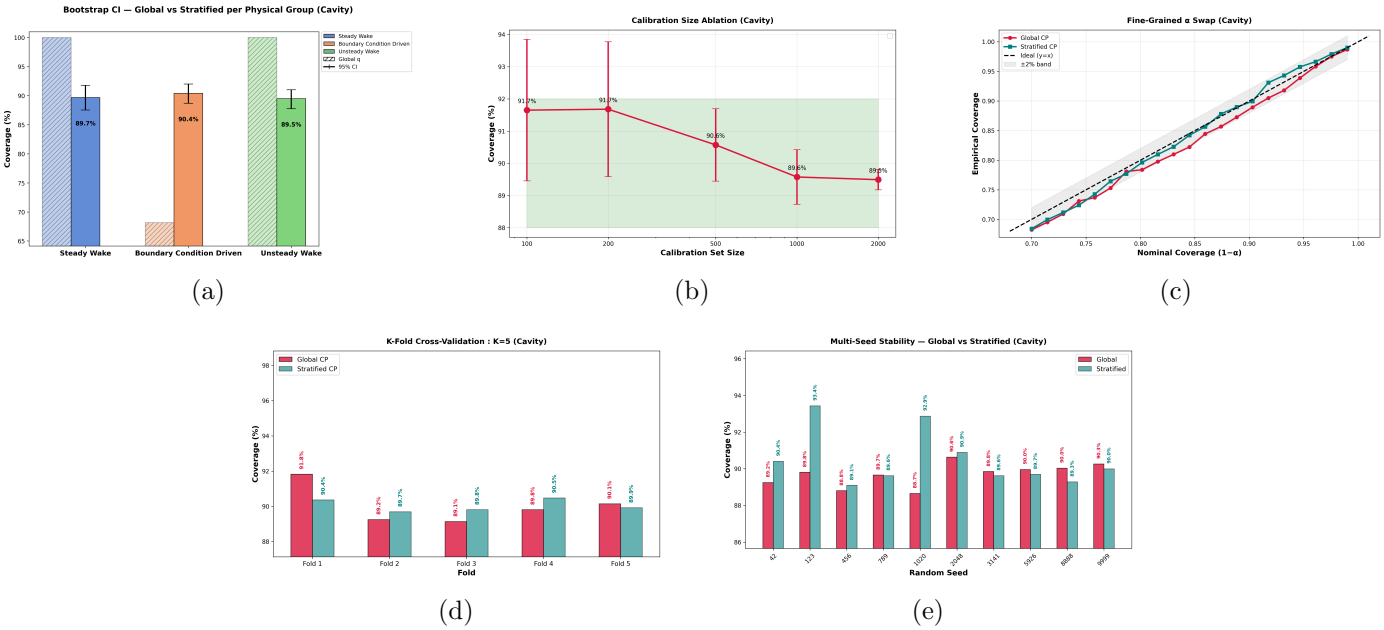


Figure 2: Regime Stratified Conformal Prediction Validation for Cavity Dataset : (a) Bootstrap Resampling, (b) Calibration Size Ablation, (c) Significance Level (α) Swap, (d) K-Fold Cross-Validation and (e) Multi-Seed Stability

To validate the framework, five comprehensive validation tests were run on both datasets. At first, for Bootstrap confidence intervals, a test was done for the datasets to check to quantify the uncertainty associated with the estimated performance coverage for each regime. Figures 2a and 3a show the result of this test. In the cavity case, the 95% bootstrap confidence interval for all three regimes ranged from 87.55% to 91.79%, exhibiting a maximum interval width of approximately 4.24%, indicating high precision and affirming the stability of the results across resampled datasets. Conversely, for the cylinder case, a 95% bootstrap confidence interval was established between steady zones of 85.18 and 90.07, exhibiting a maximum interval width of approximately 4.89%, thereby indicating accuracy and reliability across bootstrap resamples. For unsteady zones, the range is between 90.00% and 94.06%, while for boundary condition-driven cases, the interval width is much narrower at 2.1%. To validate the consistency of the observed coverage performance, ten independent experiments were executed utilizing different random seeds (42, 123, 456, 789, 1020, 2048, 3141, 5926, 8888, 9999). Figures 2e and 3e illustrates the outcomes of this random test. The cavity dataset reveals a minimum regime-wise coverage of 90.49%, along with a standard deviation of 1.42%, thereby affirming the framework's stability. The cylinder dataset reveals a minimum regime-wise coverage of 88.91%, which includes a standard deviation of 0.41%, thereby affirming the framework's reliability. To enhance the model's generalizability, 5-fold stratified cross-validation was conducted. The dataset was randomly divided into five equal folds, followed by the creation of a calibration set and a test split, after which the coverage for each regime was assessed, and results are shown in the figures 2d and 3d. The cavity dataset exhibits an average coverage of 90.06% with a standard deviation of 0.31%, while the cylinder dataset shows an average coverage of 89.96% with a standard deviation of 0.63%. exhibiting and demonstrating excellent stability. Initially, the alpha value for the conformal prediction framework is set at 0.1. To assess its robustness to varying alpha selections, a range of alpha values from 0.01 to 0.30 is utilized, with 21 alphas randomly selected to evaluate coverage performance, and figures 3c and 2c show the performance of the test. The highest deviation recorded for 21 distinct α values is 1.11% from the target in the cylinder dataset, with an average deviation of 0.49%. With a α value of 0.1, the calculated mean deviation is 0.21%. The cavity dataset indicates an average deviation from the target of 1.41%, with the maximum recorded deviation. 2.28 percent. With a α value of 0.1, the calculated average deviation is 1.05%. Ultimately, an ablation study is conducted to ascertain the requisite size of the calibration

dataset for ensuring coverage guarantee in the validation of the framework. The coverage percentages were evaluated in conjunction with the standard deviation. In the cavity dataset, a calibration set of 500 cases attained a mean coverage of 90.57% with a standard deviation of 1.13%, whereas a calibration set of 1000 cases yielded an average coverage of 89.58% with a standard deviation of 0.85%. In the cylinder dataset, a calibration set of 1000 cases yielded a mean coverage of 89.69% with a standard deviation of 0.75%, whereas a calibration set of 2000 cases resulted in an average coverage of 89.95% with a standard deviation of 0.39%.

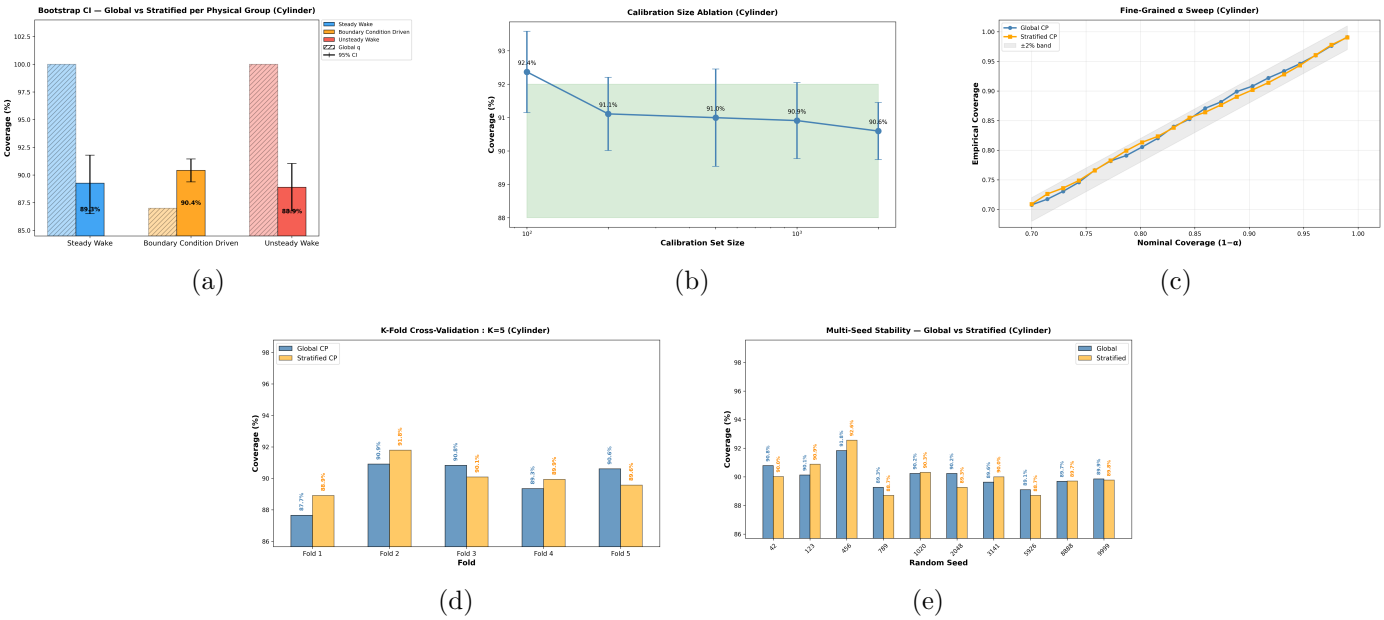


Figure 3: Regime Stratified Conformal Prediction Validation for Cylinder Dataset : (a) Bootstrap Resampling, (b) Calibration Size Ablation, (c) Significance Level (α) Swap, (d) K-Fold Cross-Validation and (e) Multi-Seed Stability

4.1.3 Robustness of the RSCP framework using different binning technique

To assess the robustness of the regime-stratified framework, it is essential to determine the flexibility of the regime selection boundaries. To address this, two distinct types of binning operations are conducted. Initially, as the Reynolds numbers dictated by boundary conditions are constant, we categorized them separately and partitioned the calibration and test sets according to distinct regime bounds, subsequently assessing the coverage and efficiency of conformal prediction. Subsequently, both categories of data are amalgamated and re-evaluated for conformal prediction coverage and efficacy. Initially, for the cavity dataset, excluding the boundary condition cases, the threshold distinguishing a steady wake from an unsteady wake was established at 4000. This figure appears elevated, attributable to the dataset's constrained range of 1000 to 4000. The purpose of this framework is not to determine the precise value that distinguishes steady from unsteady flow, but rather to demonstrate the necessity of establishing distinct thresholds instead of a singular global safety score. In the steady zone, the system attained 88.8% coverage with 74.11% tighter widths, while in the unsteady wake, it achieved 90.1% coverage with 86.19% tighter widths. The framework was evaluated by partitioning the calibration and test sets into two distinct regimes while mixing the split. In the first regime with Reynolds numbers below 2000, the application of the global threshold would cover only 80% of the test set, whereas the regime-stratified threshold attained 87.9% coverage with a narrower width of 86.2%. The second regime was determined to be equivalent to the threshold Reynolds number of 1000. The identity value resulted from the absence of cases within the Reynolds numbers of 1000 to 2000. Both the unsplit and split regimes affirm the robustness of the framework, indicating that the results are consistent and reliable across varying conditions and Reynolds numbers. The boundary condition cases for the cylinder dataset are maintained separately, similar to the cavity dataset. The calibration and test sets were partitioned based on Reynolds's 500 threshold. In the stable zone, the framework obtained 95.4% coverage with widths that are

99.56% tighter, while for the uns

Table 2: Regime Stratified Conformal Prediction Coverage & Efficiency for Different Binning

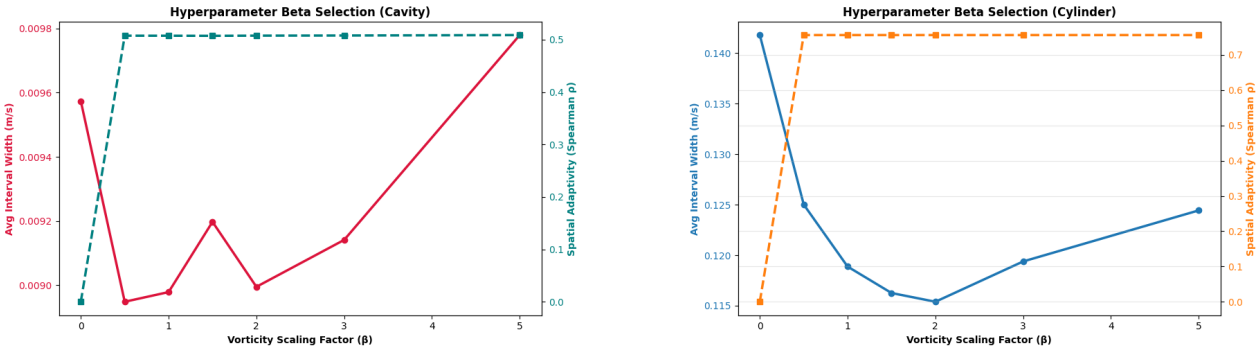
Dataset	**Flow Regime**	N_{cal}	N_{test}	q_{glob}	**Global Coverage**	$\hat{q}_k$	**RSCP Coverage**	**RSCP Efficiency**
Cavity (Unmixed Split)	Steady Wake ($Re < 4000$)	392	1353	3564.63	100%	922.72	88.8%	74.11% tighter
	Boundary Condition Driven (Constant Re)	348	1240	3564.63	68.1%	3745.66	90.4%	5.1% wider
	Unsteady Wake ($Re \geq 4000$)	260	872	3564.63	100%	492.02	90.1%	86.19% tighter
Cavity (Mixed Split)	Steady Wake ($Re < 2000$)	583	1995	3564.63	80.2%	3626.87	87.9%	1.74% wider
	Unsteady Wake ($Re \geq 2000$)	417	1470	3564.63	100%	492.02	89.5%	86.20% tighter
Cylinder (Unmixed Split)	Steady Wake ($Re < 500$)	77	217	3214.47	100%	14.04	95.4%	99.56% tighter
	Boundary Condition Driven (Constant Re)	1425	3377	3214.47	85.8%	3376.30	89.8%	5.03% wider
	Unsteady Wake ($Re \geq 500$)	498	1168	3214.47	100%	10.56	91.0%	99.67% tighter
Cylinder (Mixed Split)	Steady Wake ($Re < 2000$)	285	695	3214.47	100%	11.09	87.6%	99.65% tighter
	Unsteady Wake ($Re \geq 2000$)	1785	4067	3214.47	88.2%	3284.32	89.9%	2.17% wider

teady wake, it achieved 91.0% coverage with widths that are 99.67% tighter. Subsequent to the similarity, it was divided by amalgamating the split with a comparable Reynolds number threshold of 2000. In the initial regime, characterized by Reynolds numbers below 2000, the outcomes were observed to be analogous to the 1000 threshold. This phenomenon occurred due to the absence of cases between these two numbers, similar to the cavity dataset. Utilizing the global threshold, regardless of whether it is set at 2000 or 1000, unnecessarily drags the high safety score of 3257.25, whereas an 11.09 threshold can attain 87.6% coverage. The minimum coverage observed for both the dataset with the mixed split and the unmixed split was 87.6%, confirming the robustness of the regime-stratified framework.

4.2. Vorticity Adapted Regime Stratified Conformal Prediction Framework

4.2.1 β Selection

To calculate the spatially adaptive scaling factor $\sigma_{vel}(x)$, hyperparameter β selection is critical. The optimal scaling parameter (β) was ascertained via an efficiency ablation study conducted on the calibration test only. The objective of the selection was to minimize the expected interval width, thereby ensuring the narrowest predictive bounds. Subsequent to this selection, the framework's reliability was assessed through out-of-sample coverage to validate the $1-\alpha$ guarantee and spatial adaptivity (utilizing Spearman ρ correlation) to ascertain that the resultant heteroscedasticity adhered to the local physics of the flow



(a) (b)

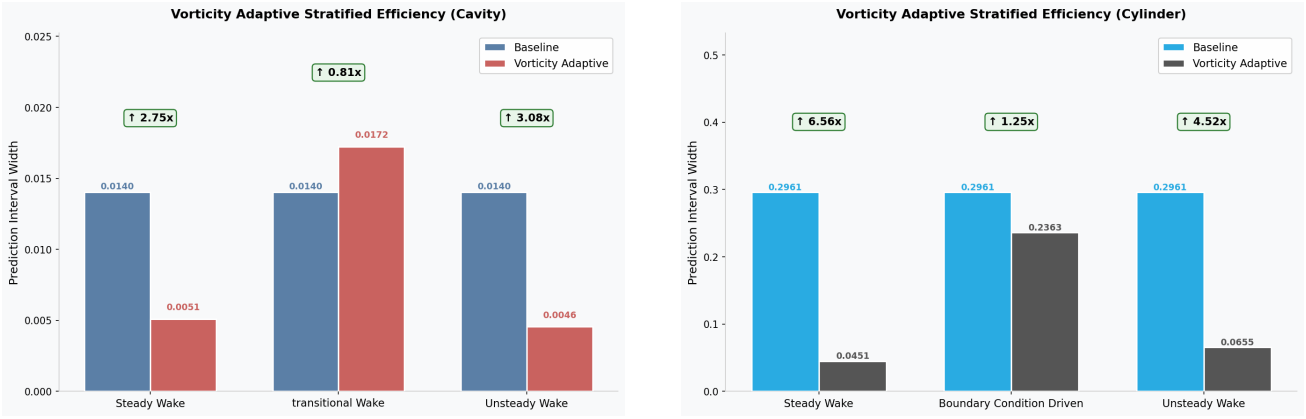
Figure 4: Hyperparameter beta selection using calibration dataset for (a) Cavity Dataset and (b) Cylinder Dataset

To select β from a range of numbers [0.0, 0.5, 1.0, 1.5, 2.0, 3.0, 5.0], we tried to find the average minimum safety bound. For cavity, it is found to be 0.5, which gives a minimum interval width of 0.00895 m/s. Figure 4a shows the result of the β selection cavity dataset. It is found that for that specific β Spearman correlation is 0.5077. It is understandable due to global topological change in vorticity across the dataset being moderate. Figure 4b illustrates the outcome of the β selection cylinder dataset. From the ablation test, the optimum β found for the cylinder dataset is 2, which gave the lowest safety width of 0.11539 m/s. When it was tested in the test set, the Spearman ρ correlation was found to be 0.7563. The discrepancy with the correlation compared to the cavity dataset was expected, as due to the nature of the cylinder dataset there will be higher vorticity change in the cylinder compared to the cavity dataset.It is worth mentioning that hyperparameter β was selected only using calibration and on the basis of minimum average interval width among regimes. Then, to check the reliability, they are examined on the test set. The figures showing here are the performance of that reliability test.

4.2.2 Performance of vorticity adaptive framework

After selecting β, the coverage and interval width are checked in the dataset. If β is set to zero, the Spearman correlation is found to be exact zero. This is the motivation of the vorticity adaptive framework. Though the base framework was able to improve the performance by setting a regime-stratified threshold instead of a global one, it will remain the same throughout the whole area of each specific steady or unsteady case. So with the base framework set before, getting zero correlation with that framework signifies that it is spatially constant and blind to local physics and incapable of correlating with the model's actual residuals. By enforcing a vorticity adaptive framework, it creates a spatially adaptive map where the safety interval width will expand in a high-gradient shear layer, whereas it will shrink in the relatively steady area. For the cylinder dataset, when β is selected zero, the average width found globally is 0.29609 regardless of any regime. When the vorticity adaptive framework was applied, for the steady zone, the safety avearge width goes down to 0.04511, which is 6.56 times less than the average global threshold. Figure 5b and table 3 show performance for the cylinder dataset. For unsteady wake cases, for the boundary condition-driven cases, the average minimum width also drops to 0.06548, which is 4.52 times less compared to the global threshold. In the case of boundary condition-driven data, as the boundary condition is changing, it is more difficult to reduce error and give better predictions. So, the mean safety interval width is necessary to be stretched. The framework result is consistent with this, giving the safety interval width of 0.23626, which is 5.23 times higher and 3.60 times higher compared to steady and unsteady cases, necessitating the need for adopting a vorticity adaptive framework.Now, apart from interval width adaption with different types of cases, coverage is also necessary to evaluate the performance of the framework.So sample-wise coverage is checked for each type of case with the new adaptive safety bounds and found 92.66%, 90.35%, and 90.72% for steady wake, boundary condition driven, and unsteady wakes, respectively.Now for the cavity dataset, same framework is followed.

At first, with the β value setting zero, the Spearman correlation is found to be zero as expected; with the vorticity adaptive method, it is found to be moderate, which is 0.50. It is due to the nature of the geometry of the lid-driven cavity case when the change in vorticity is moderate compared to cylinder cases.Figures 5a and table 3 represent the performance for cavity cases. Now for the steady cases, the average interval bound is 0.00510, which is 2.78 times less than the global threshold of 0.01403. For the unsteady wake cases, the mean safety width is 0.00456, which is also 3.078 times lower compared to the single global threshold. In terms of boundary-driven cases, it goes a bit higher compared to the global threshold but is still 3.38 times lower than steady and 3.78 times lower than unsteady cases, validating the need for vorticity adaptation, which is aware of the local physics instead of blindly giving one simple global threshold per cases.As stated before for the cylinder cases, the sample-wise coverage was also checked for each regime case. For steady-state cases, the coverage is found to be 89.67%, and for unsteady cases, it is found to be 89.18%. The boundary condition-driven cases covered 90.8%. with the adaptive safety width.



(a) (b)

Figure 5: Vorticity Adaptation Framework Gain in terms of Interval Width for (a) Cavity Dataset and (b) Cylinder Dataset

Table 3: Vorticity Adaptive Regime Stratified Conformal Prediction Coverage & Efficiency

Dataset	**Flow Regime**	N_{cal}	N_{test}	**Sample Coverage**	**Pixel Coverage**	**Base Width**	**Vorticity Width**	**Vorticity Framework Efficiency**
Cavity	Steady Wake ($Re < 1000$)	235	755	89.67%	96.52%	0.01403	0.00510	63.64% tighter
	Boundary Condition Driven (Constant Re)	348	1240	90.08%	96.63%	0.01403	0.01724	22.87% wider
	Unsteady Wake ($Re > 1000$)	417	1470	89.18%	97.32%	0.01403	0.00456	67.49% tighter
Cylinder	Steady Wake ($Re < 1000$)	285	695	92.66%	97.64%	0.29609	0.04511	84.77% tighter
	Boundary Condition Driven (Constant Re)	1425	3377	90.08%	94.88%	0.29609	0.23626	20.56% tighter
	Unsteady Wake ($Re > 1000$)	290	690	90.78%	96.41%	0.29609	0.06548	77.88% tighter

4.2.3 Validation of Vorticity Adaptive Framework

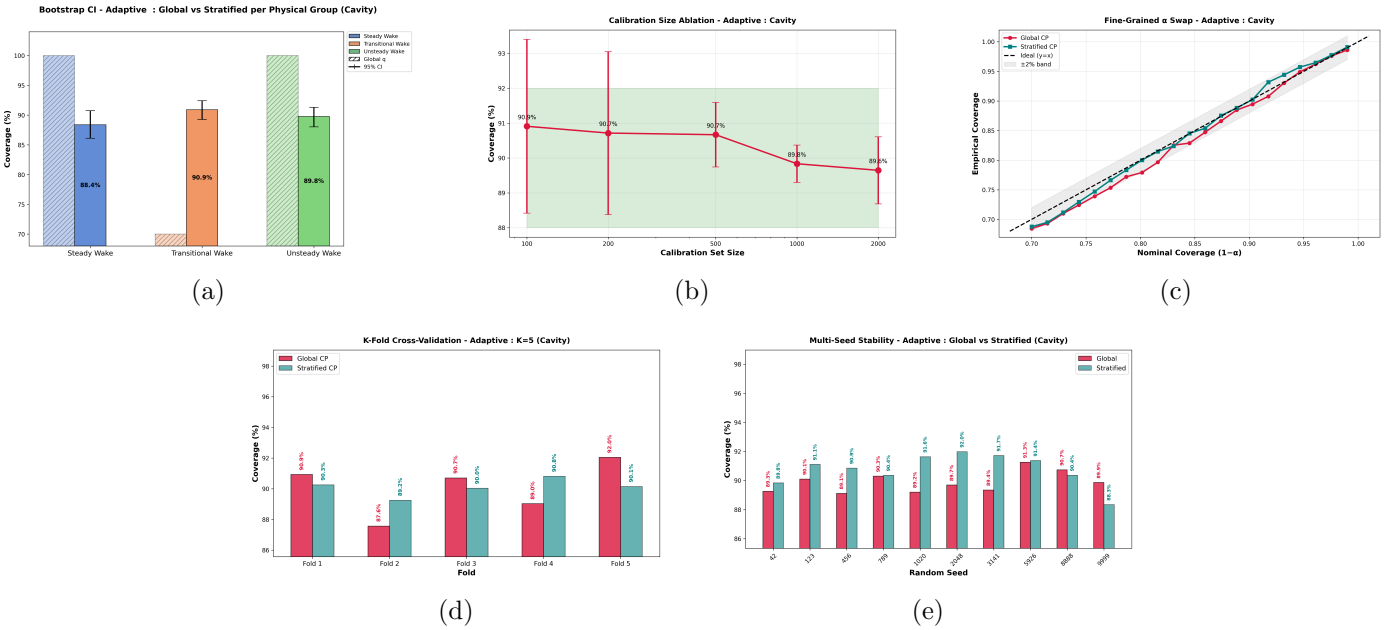


Figure 6: Vorticity Adaptive Regime Stratified Conformal Prediction Validation for Cavity Dataset : (a) Bootstrap Resampling, (b) Calibration Size Ablation, (c) Significance Level (α) Swap, (d) K-Fold Cross-Validation and (e) Multi-Seed Stability

Validation tests for vorticity-adaptive frameworks are conducted in both datasets, similar to the regime-stratified framework. The initial validation test, the bootstrap resampling test, for the cylinder dataset, indicated that the steady wake cases attained 92.73% coverage, with a confidence interval ranging from 90.65% to 94.53%. The interval width for boundary condition-driven cases is significantly narrower at 1.93, spanning from 89.96% to 91.89%. In unsteady scenarios, a coverage of 91.12% with a confidence interval ranging from 88.99% to 93.19% substantiates the coverage across various zones. In terms of the cavity dataset, the steady wake zone achieved coverage for 88.36% with an interval from 86.09% to 90.73%. But both the boundary condition-driven cases attained mean 90.90% and 89.77% coverage. Figure 7a and figure 6a show the result of this test. After that, 10 random seed tests were done for both datasets like before, and results are shown in the figures 7e and 6e. For the cylinder dataset, the mean coverage found for all three regimes is 90.39% with a 0.73% standard deviation, and for the cavity dataset, it is found to be 90.76% with a 1.03% standard deviation, validating the framework's reliability. Subsequently, to assess the robustness of the framework, K-fold cross-validation tests were conducted on both datasets. In a 5-fold test, the cylinder dataset attained a mean stratified coverage of 89.99% with a standard deviation of 0.24%. The mean stratified coverage of the cavity dataset was 90.10%, with a validation rate of 0.50%, confirming the system's robustness and reliability. Figures 6d and 7d show the performance of stratified coverage alongside global coverage. The framework, employing an alpha value of 0.1, is evaluated for various alpha values, and the results are presented in figures 6c and 7c. The maximum deviation observed for 21 distinct α values is 1.22% from the target in the cylinder dataset, with a mean deviation of 0.57%. For an α value of 0.1, the mean deviation found is 0.51%. The cavity dataset reveals an average deviation from the target of 1.16%, with the highest recorded deviation. 0.53%. For an α value of 0.1, the average deviation found is 0.5%. Finally, The calibration size ablation test was conducted as previously, with results presented in figures 6b and 7b. The cylinder dataset reveals that a calibration set size of 1000 achieves a coverage of 90.93% with a standard deviation of 1.0%, while a calibration set size of 3000 results in a coverage of 90.03% with a standard deviation of 0.50%. A calibration set size of 2000 yields a superior accuracy of 90.03% with a reduced standard deviation of 0.50%. The cavity dataset indicates that a calibration size of 1000 yields a mean coverage of 89.84% with a standard deviation of 0.54%. These tests validate the reliability and robustness of the vorticity adaptive framework across various

zones for both dataset types.

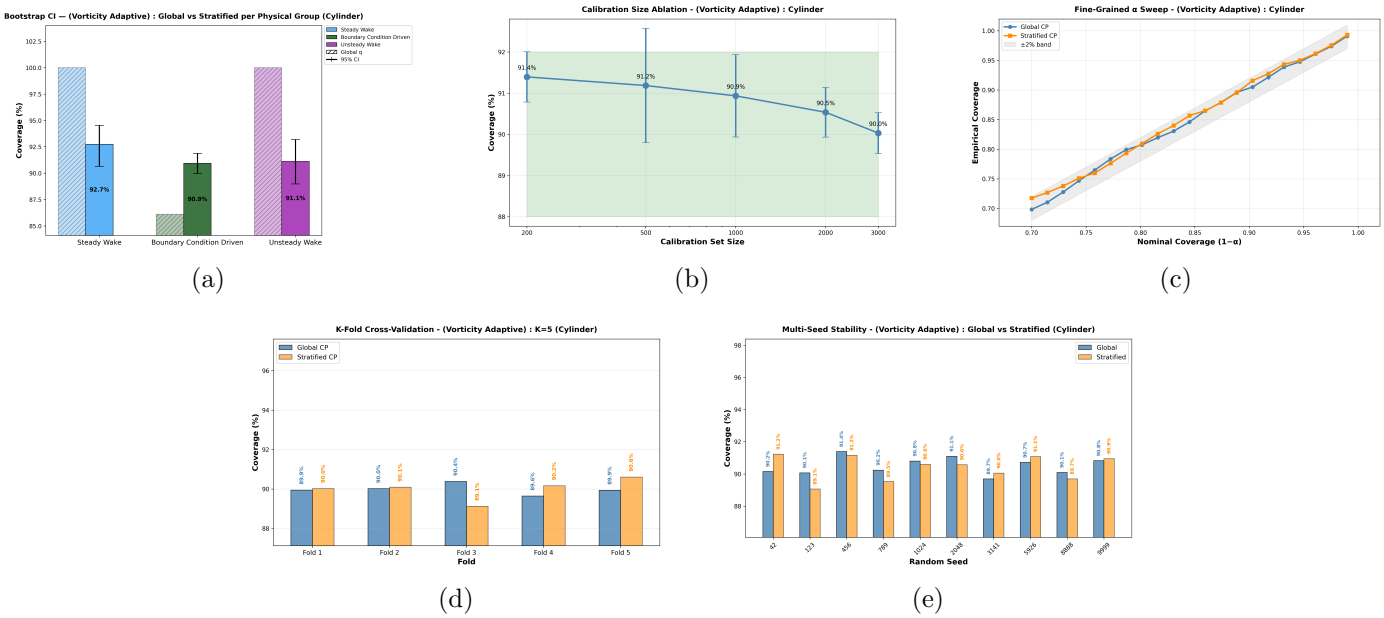


Figure 7: Vorticity Adaptive Regime Stratified Conformal Prediction Validation for Cylinder Dataset : (a) Bootstrap Resampling, (b) Calibration Size Ablation, (c) Significance Level (α) Swap, (d) K-Fold Cross-Validation and (e) Multi-Seed Stability

4.2.4 Robustness of Vorticity Adaptive Regime Stratified framework

We evaluated the robustness of the vorticity adaptive framework using varying bin counts in each regime, both by maintaining the separation of boundary condition-driven cases and by integrating them. We maintained the identical binning count as the basic framework. For the cavity dataset, at steady wake-less Reynolds number s below 4000, it provided a 53.67% tighter bound relative to the absence of vorticity. For datasets influenced by boundary conditions,

Table 4: Vorticity Adaptive Regime Stratified Conformal Prediction Coverage & Efficiency for Different Binning

Dataset	Flow Regime	N_{cal}	N_{test}	Sample Coverage	Base Width	Vorticity Width	Vorticity Framework Efficiency
Cavity (Unmixed Split)	Steady Wake ($Re < 4000$)	392	1353	89.9%	0.01403	0.0065	53.67% tighter
	Boundary Condition Driven (Constant Re)	348	1240	90.1	0.01403	0.0172	22.59% wider
	Unsteady Wake ($Re \geq 4000$)	260	872	89.2%	0.01403	0.0044	68.63% tighter
Cavity (Mixed Split)	Steady Wake ($Re < 2000$)	583	1995	89.6%	0.01403	0.0150	6.91% tighter
	Unsteady Wake ($Re \geq 2000$)	417	1470	89.2%	0.01403	0.0046	67.21% tighter
Cylinder (Unmixed Split)	Steady Wake ($Re < 500$)	77	217	94.93%	0.29609	0.02953	90.02% tighter
	Boundary Condition Driven (Constant Re)	1425	3377	90.94%	0.29609	0.24856	16.05% tighter
	Unsteady Wake ($Re \geq 500$)	484	1182	90.15%	0.29609	0.06365	78.50% tighter
Cylinder (Mixed Split)	Steady Wake ($Re < 2000$)	285	695	92.66%	0.29609	0.04511	84.79% tighter
	Unsteady Wake ($Re \geq 2000$)	1715	4067	90.44%	0.29609	0.23061	22.11% tighter

the width increased. As previously stated, it was anticipated that the safety width would increase due to the geometry of the lid-driven cavity, resulting in a 22.59% wider bound. For the unsteady wake, the boundary width decreased to 0.0044, in contrast to the global width of 0.01403, resulting in a 68.63% reduction. For the mixed split, in which boundary condition cases are also mixed, the safety width is recalculated. Within the range below 2000, where all boundary condition-driven scenarios lie, the safety margin rose to 0.0150 as anticipated, yielding a 6.91% broader margin; conversely, for cases with Reynolds numbers exceeding 2000, the margin contracted, resulting in 67.21% tighter bounds. Now, come to the cylinder dataset. For steady wake regions, the safety width is only 0.02953, whereas the global threshold was 0.29609. In terms of boundary-driven cases, the width decreased to 0.24856 and 16.05% tighter safety bounds. For unsteady wake cases, the safety width also shrunk, resulting in 78.05% tighter bounds. In terms of mixed split, the result was consistent, giving 84.79% and 22.11% tighter bounds for both steady and unsteady wakes.

4.3. Baseline Comparison

The prior sections demonstrated that the suggested three-stage framework , physics-residual nonconformity scores, regime-stratified calibration, and vorticity-based spatial adaptation yields more precise, coverage-controlled intervals with significant spatial error correlation. This section assesses the framework from two complementary perspectives: the effect of conformal calibration compared to independent uncertainty estimation (Section 4.3.1), and the modularity of the spatial adaptation phase across various scaling signals (Section 4.4). All baselines utilise an identical calibration-test split with the same random seed and, when relevant, independently optimised values chosen by the grid search method outlined in Section 3.9 .

4.3.1 Monte Carlo Dropout Evaluation

Monte Carlo Dropout is the most widely employed post-hoc uncertainty estimation technique for neural networks. To establish a baseline, dropout layers (p = 0.05) are incorporated following each spectral layer and the final hidden layer of the pretrained FNO during inference, without retraining.This post-hoc methodology does not retrain the surrogate using dropout as a regularizer, which might contribute to an underestimation of the quality of uncertainty estimations compared to a model trained with dropout from the outset.We adopt this setup deliberately in order to assess if MC Dropout's uncertainty signal can function as a useful spatial scaling function in our conformal framework rather than benchmarking it as a stand-alone UQ technique. Notably,[32] found that MC Dropout only achieved 44.05% empirical coverage against a 95% target on out-of-distribution Navier-Stokes predictions and 16.91% on magnetohydrodynamic systems even when dropout is incorporated during training. This shows that Bayesian UQ methods fail structurally to provide coverage guarantees without an additional calibration layer, regardless of training protocol.The lack of a calibration layer that may transform raw uncertainty estimates into coverage-controlled bounds at a user-specified confidence level is the structural flaw that causes both over-coverage and under-coverage.Table 5 compares standalone MC Dropout intervals against the proposed regime-stratified conformal framework at the 90% coverage target. MC Dropout surpasses the 90% target by achieving 100% sample coverage across all regimes.On both datasets, but at the expense of significantly inflated intervals. MC Dropout yields mean interval widths of 0.69–1.09 on the cylinder dataset, whereas the suggested framework yields mean interval widths of 0.045–0.236, indicating a reduction factor of 4.6 to 16.2. The reduction factor on the cavity dataset is between 3.3 and 6.0.The two approaches are further distinguished by the Spearman rank correlation between actual pointwise error and anticipated uncertainty. MC Dropout reaches $\rho = 0.512$ on the cylinder dataset, while vorticity obtains $\rho = 0.756$. The difference is greater on the cavity dataset ($\rho = 0.332$ compared to $\rho = 0.508$). Regardless of interval width, vorticity more precisely determines where errors concentrate.

4.3.2 Multi-Target Coverage Analysis

We subsequently investigate whether modifying the z-threshold independently can address the issue of over-coverage in the absence of conformal calibration. Table 6 assesses standalone MC Dropout over three coverage targets ($z \in \{1.282, 1.645, 1.960\}$, corresponding to 80%, 90%, and 95% under the Gaussian assumption) across all regimes for both datasets. In the cylinder dataset, all nine regime-target combinations achieve precisely

Table 5: Monte Carlo Dropout Evaluation Against Vorticity Adaptive Framework

Dataset	Regime	MC Width	Vorticity Width	MC Cov.	Ratio
Cylinder	Steady Wake	0.7294	0.0451	100%	16.2×
	BC-Driven	1.0895	0.2363	100%	4.6×
	Unsteady Wake	0.6881	0.0655	100%	10.5×
Cavity	Steady Wake	0.0276	0.0051	100%	5.4×
	BC-Driven	0.0563	0.0172	100%	3.3×
	Unsteady Wake	0.0274	0.0046	100%	6.0×

Table 6: MC Dropout Standalone Across Three Coverage Targets

Dataset	Target	z	Steady	BC-Driven	Unsteady	Coverage
Cylinder	80%	1.282	0.5683	0.8490	0.5364	100%
	90%	1.645	0.7292	1.0893	0.6883	100%
	95%	1.960	0.8689	1.2979	0.8201	100%
Cavity	80%	1.282	0.0215	0.0439	0.0213	99.6–100%
	90%	1.645	0.0276	0.0563	0.0274	100%
	95%	1.960	0.0328	0.0671	0.0326	100%

100% coverage. In the cavity dataset, the minimum recorded coverage is 99.60% (BC-Driven regime at the 80% objective), surpassing the most lenient specification by approximately 20 percentage points. Interval widths increase linearly with z, decreasing the cylinder's steady-wake width from 0.869 at z=1.960 to 0.568 at z=1.282, but coverage consistently maintains at 100%. The separation of width and coverage signifies that the surrogate's true error distribution is significantly more concentrated than the variation shown by MC Dropout, making z-threshold adjustment inadequate as a calibration method.

Furthermore, we determine that this over-coverage is not confined to the selected dropout rate by assessing standalone MC Dropout across$p \in \{0.01, 0.05, 0.10, 0.20\}$ at the 90% target on both datasets, utilising a subset of 500 test samples with $T = 10$ forward passes. Coverage grows monotonically with pp from 93.16% to 100% for the cavity and from 98.12% to 99.99% for the cylinder, with no tested rate reaching the designated 90% threshold, so demonstrating that the miscalibration is structural rather than a tuning effect.

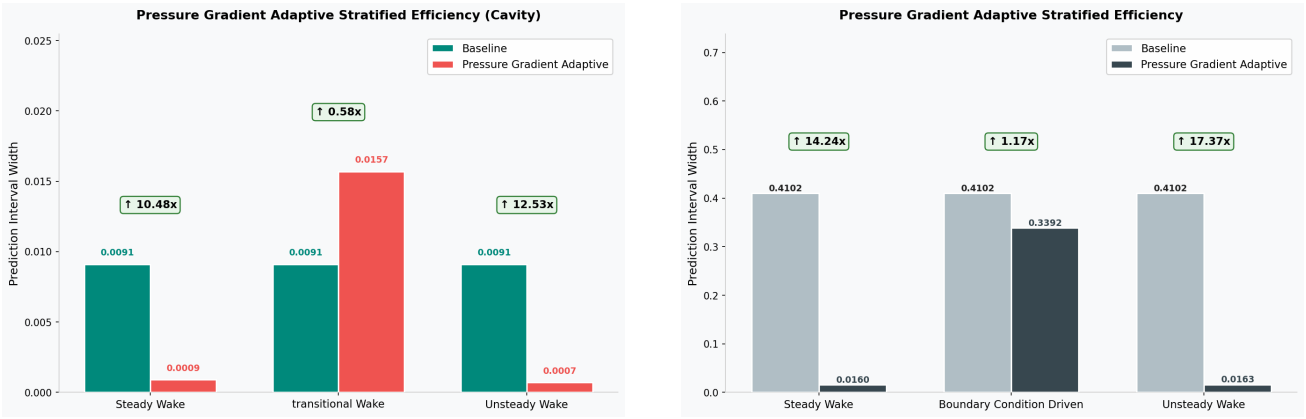
4.4. Spatial Scaling Signal Comparison

Within the same regime-stratified conformal framework, we assess which spatial signal yields the most precise calibrated intervals. Four configurations are evaluated, each assigned an independently optimised β through the calibration-only grid search outlined in Section 3.9: (i) Flat RSCP ($\beta = 0$, no spatial adaptation), (ii) velocity magnitude, (iii) MC Dropout standard deviation and (iv) vorticity magnitude. Table 7 reports per-regime mean interval widths and Spearman rank correlations between pixel-level interval width and prediction error. All spatially adaptive variants reduce interval width relative to the flat baseline, with the highest gains in the BC-Driven regime where flow heterogeneity is highest: vorticity reduces the cylinder BC-Driven width by 24.7% (0.314 to 0.236) and MC Dropout by 25.4% (0.314 to 0.234). On the cavity dataset, vorticity achieves a 9.4% reduction on the same regime. The narrow margins between the three adaptive signals particularly vorticity and MC Dropout on the cylinder reflect the design intent: conformal calibration absorbs much of the difference in raw signal quality, making the framework robust to the choice of spatial input. The essential difference is in consistency and computational expense. The velocity magnitude reduces to $\beta = 0$ in the cavity dataset, yielding no spatial information for a wall-bounded geometry when high velocity and high prediction error are inversely associated. MC Dropout achieves similar width and Spearman correlation on the cylinder (0.752 vs. 0.756) but requires $T = 50$ stochastic forward passes per sample and a selected $\beta = 5$, indicating that the inherent uncertainty signal required significant amplification to function as an effective normaliser. Vorticity is the sole candidate that surpasses the flat baseline across all six regime-dataset combinations, achieving the maximum Spearman correlation for both geometries (0.756 and 0.508), and requires one finite-difference computation at minimal inference cost.

Table 7: Spatial Scaling Signal Comparison Across Both Datasets

Dataset	Method	β	Steady	BC-Driven	Unsteady	Spearman
Cylinder	Flat RSCP	0.0	0.04546	0.31375	0.06612	—
	Velocity	1.5	0.04531	0.25932	0.06599	0.704
	MC Dropout	5.0	0.04551	**0.23410**	0.06625	0.752
	Vorticity (Ours)	2.0	**0.04511**	0.23626	**0.06548**	**0.756**
Cavity	Flat RSCP	0.0	0.00511	0.01902	0.00458	—
	Velocity	0.0	0.00511	0.01902	0.00458	0.394
	MC Dropout	1.0	0.00514	0.01795	0.00461	0.434
	Vorticity (Ours)	0.5	**0.00510**	**0.01724**	**0.00456**	**0.508**

4.5. Pressure Gradient Scaling Adaptation

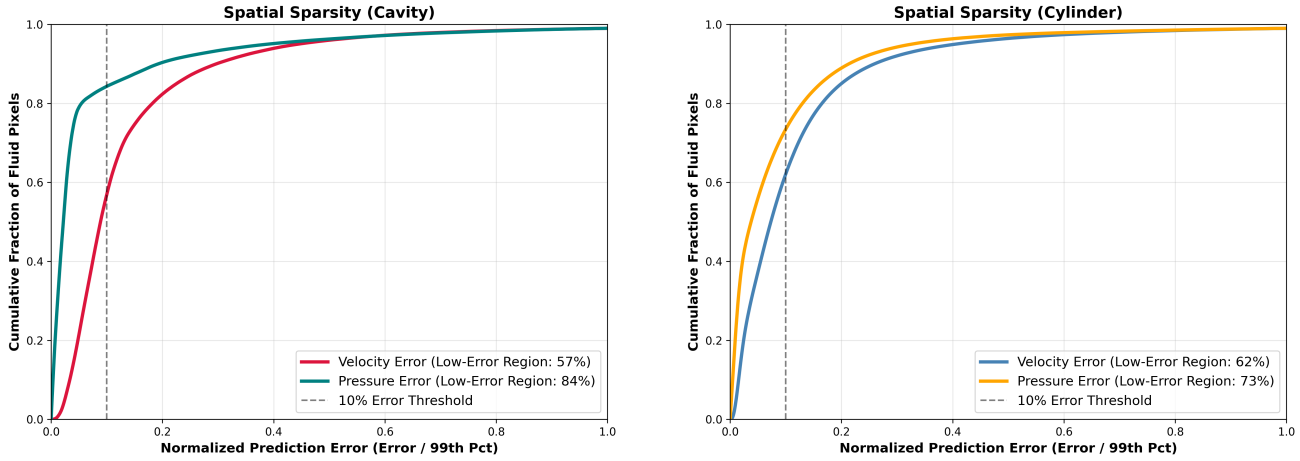


(a) (b)

Figure 8: Pressure Gradient Adaption Framework Gain in terms of Interval Width for (a) Cavity Dataset and (b) Cylinder Dataset

The previous sections evaluated the spatial scaling framework specifically for velocity-field uncertainty.This section demonstrates that the same can be applied to pressure-field uncertainty by using pressure gradient magnitude as the scaling heuristic, as described in Section 3.10. Like the vorticity framework, the β is selected by means of a calibration set only. Then, with that optimised value of β, the adaptive safety bound is calculated. To compare the efficiency, the global threshold one is calculated using the β value as zero. To validate the pressure gradient as a physics proxy, the Spearman correlation with the true error also counts for both datasets and is also calculated. For the cavity dataset, it was found to be 0.56, and for the cylinder dataset, it was found to be 0.69, validating the reliability of the pressure gradient as a physics heuristic. The global safety width found for the cavity dataset was 0.0091, but for the steady wake zones it was only 0.0009, which is 10.48 times less than the global one. For the unsteady zone it was merely 0.007, which is 12.53 times less than the global interval width. For boundary condition-driven cases, it was increased consistently with the previous sections' results and validated the need for regime-wise stratification. In terms of cylinder datasets, the same calculations were performed for different regimes. The global safety width, by selecting β as zero, was found to be 0.4102. For steady and unsteady wake zones, it was found to be only 0.0160 and 0.0163, respectively, resulting in 14.24 and 17.37 times reduction to global interval width. In terms of boundary-driven cases, it was found to be 0.3392, which is 1.17 times less than the global 0.4102.It is important to observe that the gain achieved in this pressure gradient adaptation is significantly greater than that in the vorticity adaptation framework. We attempted to ascertain the cause of this occurrence. We assessed the spatial distribution of errors to ascertain the underlying cause. To facilitate an equitable comparison among fields with inherently distinct units, scales, and boundary behaviours, the error fields are initially isolated and normalised. Prediction errors are carefully filtered with a fluid mask to eliminate solid boundaries and internal geometries, ensuring the analysis accurately represents only the active flow domain. The compressed error arrays for velocity (E_V) and pressure (E_P) are normalised by their corresponding 99th percentiles (P_{99}).

This mitigates the impact of isolated, anomalous error spikes on the distribution. The normalised error for a specific field is defined as $E_norm = E/P_{99}$, standardising all values on a comparable scale of 0.0 to 1.0 excluding the top 1% of outliers. We establish a low error region to characterise the spatial distribution of errors. This indicates the proportion of the fluid domain where the model's prediction error is minimal relative to the maximum pertinent error. A low-error threshold is set at $\tau = 0.10$ (10% of the 99th percentile). The spatial sparsity is determined as the ratio of fluid pixels for which $E_{norm} < \tau$ holds true.Figures 9a and 9b show the result of this calculation. For the cavity dataset, low-pressure-error regions were found to be 84% compared to 57% velocity-error regions. Thus maximum high-pressure errors are confined to a very small spatial region, compared to high-velocity errors, which spread over a much larger spatial region. In terms of the cylinder dataset, it is found to be consistently analogous to the cavity dataset, where there is a 62% low velocity error region compared to a high low-pressure region of 73%. This justifies the high regimewise pressure gradient gain compared to the vorticity adaptation framework.



(a) Cavity Dataset

(b) Cylinder Dataset

Figure 9: Spatial distribution of Pressure and Velocity Errors

5. Conclusion

This study addressed the significant issue of uncertainty quantification for neural operators in computational fluid dynamics. Although AI-driven surrogates provide remarkable inference speeds, their application in safety-critical engineering tasks is impeded by insufficient physical constraints. We found that the application of standard, global conformal prediction to multi-regime fluid systems does not yield physically significant calibration due to concealed numerical biases. We implemented Regime-Stratified Conformal Prediction (RSCP) to address this issue. Based on a solid, dimensionally consistent residual formulation that employs local grid-spacing extraction and stringent boundary masking to avoid wall-gradient singularities. We effectively utilised two datasets that, rather than employing global threshold regime-wise stratification, enhance performance by adjusting the safety margin as necessary. Comparison with standalone MC dropout demonstrated that uncalibrated Bayesian uncertainty estimates cannot meet a designated coverage target, irrespective of the dropout rate or z-threshold: independent intervals were 3.3 to 16.2 times broader than the constraints of the proposed framework while uniformly exceeding coverage at 100%. The comprehensive examination of three spatial scaling signals, vorticity magnitude, velocity magnitude, and MC dropout standard deviation within the framework, utilising independently optimised hyperparameters, confirmed that the design is actually modular. Vorticity was selected as the primary signal due to its superiority over the flat baseline in all six regime-dataset combinations, its achievement of the highest Spearman correlation in both geometries, and its requirement of only one finite-difference computation instead of 50 stochastic forward passes. A pressure-gradient extension further validated the flexibility of the approach across field variables. The current methodology employs a quasi-steady-state momentum proxy to guarantee computational viability. A potential route for future research is the expansion of the Physics Residual Error (PRE) to incorporate the

complete, unsteady Navier-Stokes equations. Integrating the comprehensive governing equations will convert the conformal score from a spatial momentum proxy into a detailed spatiotemporal physics metric. The proposed Regime-Stratified Conformal Prediction (RSCP) framework exhibits strong and physically accurate uncertainty calibration for 2D incompressible flows, and a natural extension of this study is its use with respect to three-dimensional (3D) fluid dynamics. This methodology is currently limited to 2D incompressible flows due to significant computational scaling costs and memory constraints associated with 3D neural operators. True turbulent flows, characterised by phenomena such as vortex stretching, the forward energy cascade, and intricate secondary flow structures—are intrinsically three-dimensional. The proposed approach provides a post-processing layer for pretrained neural fluid surrogates, eliminating the need to retrain the underlying model from a practical engineering perspective. The calibration process necessitates solely a reserved calibration dataset of simulation snapshots. This methodology could benefit applications, including predictive maintenance of fluid equipment, thermal management system design, and aerodynamic shape optimisation, where quantified uncertainty bounds are frequently essential in conjunction with point forecasts. The vorticity-adaptive scaling offers engineers a physically interpretable uncertainty map that describes high-risk spatial areas, including shear layers, separation points, and recirculation regions, potentially reducing the necessity for costly full-fidelity re-simulation. Thus, expanding the discrete Physics Residual Error (PRE) to encompass 3D Navier-Stokes momentum formulations constitutes a vital advancement. Ultimately, closing the gap to comprehensive physics-based, three-dimensional unsteady assessment constitutes the essential final phase for the trustworthy implementation of physics-informed artificial intelligence.

References

[1] S. Garg and S. Chakraborty, "Vb-deeponet: A bayesian operator learning framework for uncertainty quantification," *Engineering Applications of Artificial Intelligence*, vol. 118, Feb. 2023. DOI: 10.1016/j.engappai.2022.105685

[2] C. Moya, A. Mollaali, Z. Zhang, L. Lu, and G. Lin, "Conformalized-DeepONet: A distribution-free framework for uncertainty quantification in deep operator networks," *Physica D: Nonlinear Phenomena*, vol. 471, Jan. 2025. DOI: 10.1016/j.physd.2024.134418

[3] H. Mousavi and J. D. Eldredge, "Low-order flow reconstruction and uncertainty quantification in disturbed aerodynamics using sparse pressure measurements," *Journal of Fluid Mechanics*, vol. 1013, Jun. 2025. DOI: 10.1017/jfm.2025.10253

[4] L. Guastoni et al., "Convolutional-network models to predict wall-bounded turbulence from wall quantities," *Journal of Fluid Mechanics*, vol. 928, Dec. 2021. DOI: 10.1017/jfm.2021.812

[5] G. Shafer and V. Vovk, "A tutorial on conformal prediction," *Journal of Machine Learning Research*, vol. 9, no. 12, 2008.

[6] J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman, "Distribution-free predictive inference for regression," *Journal of the American Statistical Association*, vol. 113, no. 523, Jul. 2018. DOI: 10.1080/01621459.2017.1307116

[7] I. Gibbs and E. J. Candès, "Conformal inference for online prediction with arbitrary distribution shifts," *Journal of Machine Learning Research*, vol. 25, no. 162, 2024.

[8] E. English, E. Wong-Toi, M. Fontana, S. Mandt, P. Smyth, and C. Lippert, "JANET: Joint adaptive prediction-region estimation for time-series," *Machine Learning*, vol. 114, no. 8, Jun. 2025. DOI: 10.1007/s10994-025-06812-2

[9] F. Gan and Y. Liu, "Conformal prediction for multivariate responses with euclidean likelihood," *Journal of Multivariate Analysis*, vol. 210, Nov. 2025. DOI: 10.1016/j.jmva.2025.105494

[10] J. Diquigiovanni, M. Fontana, and S. Vantini, "Conformal prediction bands for multivariate functional data," *Journal of Multivariate Analysis*, vol. 189, May 2022. DOI: 10.1016/j.jmva.2021.104879

[11] M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," *Journal of Computational Physics*, vol. 378, Feb. 2019. DOI: 10.1016/j.jcp.2018.10.045

[12] G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang, "Physics-informed machine learning," *Nature Reviews Physics*, vol. 3, no. 6, Jun. 2021. DOI: 10.1038/s42254-021-00314-5

[13] Z. Li et al., "Physics-informed neural operator for learning partial differential equations," *ACM / IMS Journal of Data Science*, vol. 1, no. 3, May 2024. DOI: 10.1145/3648506

[14] L. Sun, H. Gao, S. Pan, and J.-X. Wang, "Surrogate modeling for fluid flows based on physics-constrained deep learning without simulation data," *Computer Methods in Applied Mechanics and Engineering*, vol. 361, Apr. 2020. DOI: 10.1016/j.cma.2019.112732

[15] Y. Zhu, N. Zabaras, P.-S. Koutsourelakis, and P. Perdikaris, "Physics-constrained deep learning for high-dimensional surrogate modeling and uncertainty quantification without labeled data," *Journal of Computational Physics*, vol. 394, Oct. 2019. DOI: 10.1016/j.jcp.2019.05.024

[16] Y. Poels, G. Derks, E. Westerhof, K. Minartz, S. Wiesen, and V. Menkovski, "Fast dynamic 1d simulation of divertor plasmas with neural pde surrogates," *Nuclear Fusion*, vol. 63, no. 12, Sep. 2023. DOI: 10.1088/1741-4326/acf70d

[17] V. Gopakumar, A. Jain, H. Kim, and A. Bhattacharjee, "Plasma surrogate modelling using fourier neural operators," *Nuclear Fusion*, vol. 64, no. 5, Apr. 2024. DOI: 10.1088/1741-4326/ad313a

[18] S. L. Brunton, B. R. Noack, and P. Koumoutsakos, "Machine learning for fluid mechanics," *Annual Review of Fluid Mechanics*, vol. 52, Jan. 2020. DOI: 10.1146/annurev-fluid-010719-060214

[19] K. Fukami, K. Fukagata, and K. Taira, "Assessment of supervised machine learning methods for fluid flows," *Theoretical and Computational Fluid Dynamics*, vol. 34, no. 4, Aug. 2020. DOI: 10.1007/s00162-020-00518-y

[20] K. Hasegawa, K. Fukami, T. Murata, and K. Fukagata, "Machine-learning-based reduced-order modeling for unsteady flows around bluff bodies of various shapes," *Theoretical and Computational Fluid Dynamics*, vol. 34, no. 4, Aug. 2020. DOI: 10.1007/s00162-020-00528-w

[21] H. Wu, Z. Wang, Y. Li, X. Zhang, L. Chen, and J. Wang, "Generalizable super-resolution turbulence reconstruction from minimal training data," *Journal of Fluid Mechanics*, vol. 1024, Dec. 2025. DOI: 10.1017/jfm.2025.10913

[22] J. Page, "Super-resolution of turbulence with dynamics in the loss," *Journal of Fluid Mechanics*, R3, Jan. 2025. DOI: 10.1017/jfm.2024.1202

[23] S. Pawar, O. San, A. Rasheed, and P. Vedula, "A priori analysis on deep learning of subgrid-scale parameterizations for kraichnan turbulence," *Theoretical and Computational Fluid Dynamics*, vol. 34, no. 4, Aug. 2020. DOI: 10.1007/s00162-019-00512-z

[24] B. Chen, C. E. Heaney, J. L. M. A. Gomes, O. K. Matar, and C. C. Pain, "Solving the discretised multiphase flow equations with interface capturing on structured grids using machine learning libraries," *Computer Methods in Applied Mechanics and Engineering*, vol. 426, Jun. 2024. DOI: 10.1016/j.cma.2024.116974

[25] M. Abdar et al., "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *Information Fusion*, vol. 76, Dec. 2021. DOI: 10.1016/j.inffus.2021.05.008

[26] E. Hüllermeier and W. Waegeman, "Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods," *Machine Learning*, vol. 110, no. 3, Mar. 2021. DOI: 10.1007/s10994-021-05946-3

[27] R. McConkey, N. Kalia, E. Yee, and F.-S. Lien, "Realisability-informed machine learning for turbulence anisotropy mappings," *Journal of Fluid Mechanics*, vol. 1019, Sep. 2025. DOI: 10.1017/jfm.2025.10618

[28] M. H. Parikh, X. Fan, and J.-X. Wang, "Conditional flow matching for generative modelling of near-wall turbulence with quantified uncertainty," *Journal of Fluid Mechanics*, vol. 1029, Feb. 2026. DOI: 10.1017/jfm.2026.11193

[29] H. Wang, F. Wu, Y. Liu, X. He, S. Feng, and S. Wang, "Machine-learning-based pressure reconstruction with moving boundaries," *Journal of Fluid Mechanics*, vol. 1008, Apr. 2025. DOI: 10.1017/jfm.2025.91

[30] K. Fukami, K. Fukagata, and K. Taira, "Super-resolution reconstruction of turbulent flows with machine learning," *Journal of Fluid Mechanics*, vol. 870, Jul. 2019. DOI: 10.1017/jfm.2019.238

[31] M. Liu et al., "CFDONEval: A comprehensive evaluation of operator-learning neural network models for computational fluid dynamics," in *Proceedings of the 34th International Joint Conference on Artificial Intelligence (IJCAI-25)*, International Joint Conferences on Artificial Intelligence, Aug. 2025. DOI: 10.24963/ijcai.2025/640

[32] V. Gopakumar et al., "Calibrated physics-informed uncertainty quantification," in *Proceedings of the 42nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 267, Vancouver, Canada: PMLR, 2025.