SPLIT CONFORMAL PREDICTION IN THE FUNCTION SPACE VIA NEURAL OPERATOR LEARNING

David Millard & Ali Baheri
Department of Mechanical Engineering
Rochester Institute of Technology
Rochester, NY 14623, USA
{djm3622,akbeme}@rit.edu

Lars Lindemann
Automatic Control Laboratory
ETH Zurich
Zurich, 8092, Switzerland
{llindemann}@ethz.ch

ABSTRACT

Uncertainty quantification for neural operators remains an open problem in the infinite-dimensional setting due to the lack of finite-sample coverage guarantees over functional outputs. While conformal prediction offers finite-sample guarantees in finite-dimensional spaces, it does not directly extend to function-valued outputs. Existing approaches require strong distributional assumptions or yield conservative coverage.This work extends split conformal prediction to function spaces via a discretize–then–lift construction. We first establish finite-sample coverage guarantees in a finite-dimensional space using a discretization map in the output function space. These guarantees are then lifted to the function space under a bilipschitz discretization assumption linking the discrete and continuous norms. To characterize the effect of resolution, we decompose the conformal radius into discretization, calibration, and misspecification components. This decomposition motivates a regression-based correction to transfer calibration across resolutions. Additionally, we propose two diagnostic metrics (conformal ensemble score and internal agreement) to quantify forecast degradation in autoregressive settings. Empirical results show that our method maintains calibrated coverage with less variation under resolution shifts while achieving improved coverage in super-resolution tasks.

1 INTRODUCTION

Conformal prediction (CP) provides a framework to obtain prediction sets with distribution-free coverage guarantees in finite-sample settings. CP guarantees that ground-truth outcomes lie within predicted sets with a probability $1-\alpha$. While CP has been extensively studied for finite-dimensional prediction tasks involving scalar or vector outputs, comparatively little work has investigated how its finite-sample guarantees extend to function-valued predictions. This gap poses unresolved computational and theoretical challenges Fontana et al. (2023). Existing approaches to functional uncertainty quantification (UQ), such as Gaussian processes Schulz et al. (2018), Bayesian neural networks Goan & Fookes (2020), or quantile neural operators Ma et al. (2024), rely on restrictive distributional assumptions or provide probably approximately correct (PAC) bounds Haussler & Warmuth (2018). While PAC-style bounds can offer desirable conditional guarantees, they often require assumptions on model class complexity or generalization behavior. CP, by contrast, guarantees marginal coverage for any predictor.

Addressing the challenges of CP in function spaces is especially relevant for neural operators. This class of models approximates mappings between infinite-dimensional input–output pairs. Architectures such as Fourier neural operators (FNOs) Li et al. (2021) and DeepONets Lu et al. (2021) offer the potential for discretization-invariant predictions, although real-world performance depends on both the resolution (the number of points sampled) and grid scheme (the method we use to sample those points) Bahmani et al. (2025). Recent work have extended neural operators to generative settings Kiranyaz et al. (2021); Liu & Tang (2025); Ranganath et al. (2016), enabling stochastic sampling of solutions. However, their suitability for CP remains unexplored. More broadly, little prior work has both (i) formally extended CP to the infinite-dimensional function spaces and (ii) studied this in relation to neural operators.

To bridge these gaps, we develop a conformal prediction framework tailored to function-valued outputs and neural operators. Our approach combines theoretical extensions with practical tools for uncertainty quantification in infinite-dimensional settings. Our main contributions are as follows:

1. An extension of split conformal prediction to function spaces. This method reduces the variation, w.r.t the standard L^2 norm, of the conformal radius (the $1-\alpha$ quantile on the residual norm) dependent on grid geometry.
2. A heuristic model of the resolution-dependent conformal radius. This method enables for the improvement in downstream task such as super-resolution.
3. Two diagnostic metrics, conformal ensemble score and internal agreement. These metrics increase the reliability of ensemble forecasting by building on the components of conformal prediction.

2 Related Work

Uncertainty Quantification in Operator Learning. A key challenge in operator learning is producing reliable uncertainty estimates. Prior methods often relied on distributional assumptions like approximate Gaussian processes Akhare et al. (2023); Zou et al. (2025), pointwise variance estimates Guo et al. (2024), or used loss-based formulations that lacked formal coverage guarantees Lara Benitez et al. (2024). While a significant advance by Ma et al. Ma et al. (2024) provides probably approximately correct (PAC) guarantees for simultaneous, pointwise coverage on a discretized grid, this does not ensure coverage for the function in the continuous domain. Similarly, prior work hierarchically applies split conformal prediction at varying scales and computes the union of their bounds Baheri & Shahbazi (2025). Recent work has also explored conformal prediction for function-valued models, including approaches that construct adaptive prediction sets or geometric uncertainty sets for operator and surrogate models Harris & Liu (2025); Gray et al. (2025). In contrast, our approach uses split conformal prediction to construct a single uncertainty set over the entire function space. This yields a domain-wide prediction set whose coverage guarantee is obtained by lifting the discrete conformal guarantee under a bilipschitz discretization assumption.

Uncertainty Quantification in Ensemble Forecasting. Ensemble forecasting is a standard technique for UQ, especially in weather modeling, using initial condition perturbation or deep ensembles Tran et al. (2020); Scoccimarro (1998). This is often an empirical approach, lacking guarantees on the resulting distribution. To approximate the posterior of such models, methods like Monte Carlo (MC) Dropout Folgoc et al. (2021) and MC-Sampling Shapiro (2003) are commonly used but do not inherently provide a guaranteed coverage probability. The quality of these probabilistic forecasts is typically evaluated using metrics like the continuous ranked probability score (CRPS) Pic et al. (2023); Bülte et al. (2025). However, a good CRPS does not translate to a formal guarantee for any single forecast. While post-hoc techniques like temperature scaling can improve upon this lack of guarantees Kull et al. (2019), the literature has not sufficiently investigated forecasting quality via CP Qian et al. (2023); Durasov et al. (2021); Rahaman et al. (2021). To this end, we integrate ensemble methods with conformal calibration to create a diagnostic tool that indicates when an autoregressive ensemble forecast has degraded beyond a predefined reliability threshold of $1-\alpha$.

3 Preliminaries and Problem Formulation

Conformal Prediction. CP provides a distribution-free framework for constructing prediction sets with finite-sample coverage guarantees. Given i.i.d. training data $\{(x_i, y_i)\}_{i=1}^n \sim \mathcal{D}$ and a predictive model $\hat{f}\colon \mathcal{X} \to \mathcal{Y}$, the goal is to construct a prediction set $\Gamma_\alpha(x)$ such that:

$$\mathbb{P}_{\{(x_i,y_i)\}_{i=1}^n,\,(x,y)\sim\mathcal{D}}\left[y \in \Gamma_\alpha(x)\right] \geq 1-\alpha. \tag{1}$$

Split CP begins by partitioning the data into two disjoint sets: one for model training $\mathcal{D}_{\text{train}}$ and another for calibration $\mathcal{D}_{\text{cal}}$. A model $\hat{f}$ is trained on $\mathcal{D}_{\text{train}}$, and nonconformity scores are computed on $\mathcal{D}_{\text{cal}} = \{(x_i, y_i)\}_{i=1}^m$ via $s_i = \mathcal{A}(x_i, y_i) = \|\hat{f}(x_i) - y_i\|$. The quantile τ_α is then defined as the $\lceil (1-\alpha)(n+1) \rceil$-th value of the ordered scores from the calibration set, and the prediction set is given by:

$$\Gamma_\alpha(x) = \left\{ y : s(\hat{f}(x), y) \leq \tau_\alpha \right\}, \tag{2}$$

where $s(\hat{f}(x), y)$ denotes the nonconformity score between the prediction $\hat{f}(x)$ and candidate output y. In our case, this is the relative weighted L^2 error. This procedure guarantees that for a new test point, the nonconformity score $s(\hat{f}(x), y)$ will not exceed this threshold with probability at least $1-\alpha$: $\mathbb{P}_{\{(x_i,y_i)\}_{i=1}^n,(x,y)\sim\mathcal{D}}(s(\hat{f}(x), y) \leq \tau_\alpha) \geq 1-\alpha$.

Neural Operators. Neural operators generalize classical neural networks to learn mappings between infinite-dimensional function spaces. Formally, given a (possibly nonlinear) operator $\mathcal{G}\colon \mathcal{X} \to \mathcal{Y}$ acting on Banach spaces $\mathcal{X}, \mathcal{Y}$, the goal is to learn a parametric approximation $\mathcal{G}_\theta$ such that $\mathcal{G}_\theta : f \mapsto u \approx \mathcal{G}(f)$, where $f \in \mathcal{X}$ is typically a coefficient or source term in a PDE, and $u \in \mathcal{Y}$ is the corresponding solution. Unlike standard deep learning models, which learn pointwise mappings, operator learning learns entire function-to-function maps, enabling inference on new input functions and resolutions. A generic neural operator maps $f \mapsto u$ via a composition of a lifting map, multiple integral transformations, and a final projection back to the input domain. We adopt the standard Fourier Neural Operator (FNO) architecture in this work.

Problem Formulation. Given a new input function $f \in \mathcal{X}$, our primary objective is to construct a set-valued predictor $\Gamma_\alpha(f) \subset \mathcal{Y}$ that contains the true, unknown solution $u = \mathcal{G}(f)$ with a user-specified probability $1-\alpha$. We seek to satisfy the classical marginal coverage guarantee of conformal prediction $\mathbb{P}\left(u \in \Gamma_\alpha(f)\right) \geq 1-\alpha$, where the probability is taken over the draw of the i.i.d. training and calibration data, as well as the unseen test pair (f, u). Achieving this goal in the infinite-dimensional setting of neural operators introduces three core challenges that this paper addresses:

1. How can a scalar nonconformity score be defined to meaningfully capture the distance between a predicted function, $\hat{u}$, and the ground truth, u? How can this score be designed to distinguish the neural operator's intrinsic prediction error from the error arising from discretizing the underlying continuous function domain?
2. How does the choice of grid scheme affect the calibrated uncertainty bounds? Furthermore, how can we decompose these bounds to characterize their constituent sources of error?
3. How can we provide meaningful UQ for autoregressive forecasting tasks where error accumulation violates standard conformal assumptions? Can we adapt metrics from ensemble learning to develop more informative diagnostics that quantify forecast degradation over time?

4 Theoretical Foundation

To formally extend CP to the function-space we follow a two step method. We first establish finite-sample coverage guarantees in a finite-dimensional, discretized space via the standard split CP framework. Then these guarantees are lifted to the function-space by considering the asymptotic convergence as the discretization is refined.

4.1 CP in Discretized Function Spaces

Let $\{(f_i, u_i)\}_{i=1}^{n+1}$ be an i.i.d. sample from a data-generating distribution on an input-output function space $\mathcal{X} \times \mathcal{Y}$. We use a discretization operator $P_d : \mathcal{Y} \to \mathbb{R}^d$ that maps each function $u_i \in \mathcal{Y}$ to its evaluations on a fixed grid over the domain $\Omega \subset \mathbb{R}^d$. Given a neural operator $\mathcal{G}_\theta$ trained on set of points collected with P_d, we define a nonconformity score for each sample (f_i, u_i) in a separate calibration set as the distance between the discretized prediction and the discretized ground truth: $s_i = d(P_d(\mathcal{G}_\theta(f_i)), P_d(u_i))$, where $d(\cdot,\cdot)$ can be any vector norm. Using split CP, we compute the threshold τ_α. This provides a standard, distribution-free coverage guarantee in the discretized space: $\mathbb{P}(s_{n+1} \leq \tau_\alpha) \geq 1-\alpha$ where $\mathbb{P}(\cdot)$ captures the randomness in $\{(f_i, u_i)\}_{i=1}^{n+1}$.

4.2 From Discrete to Continuous Guarantees

The critical step is to ensure this guarantee translates back to the infinite-dimensional function space $\mathcal{Y}$. This first requires a formal link between the discrete and continuous geometries. To achieve this, we define the distance metric $d(\cdot,\cdot)$ using a quadrature-weighted L^2 norm. First, consider a continuous function $u \in \mathcal{Y}$. We discretize this function using P_d on a structured Cartesian grid over $\Omega \subset \mathbb{R}^d$ to obtain a discrete vector, $h = P_d(u)$, with coordinate vectors

$x^{(k)} = \{x_{j_k}^{(k)}\}_{j_k=0}^{N_k}$ along each dimension $k = 1, \ldots, d$. For each multi-index $\mathbf{j} = (j_1, \ldots, j_d) \in \mathcal{J}$, where $\mathcal{J} = \prod_{k=1}^{d} \{0, \ldots, N_k - 1\}$, we define the quadrature weight (i.e., the cell's volume) as $w_{\mathbf{j}} = \prod_{k=1}^{d} \Delta x_{j_k}^{(k)}$, where $\Delta x_{j_k}^{(k)} = x_{j_k+1}^{(k)} - x_{j_k}^{(k)}$ denotes the grid spacing in the k-th coordinate direction. Therefore the norm is defined as $\|h\|_{w,2}^2 = \sum_{\mathbf{j} \in \mathcal{J}} w_{\mathbf{j}}\, h(\mathbf{x}_{\mathbf{j}})^2$, where $\mathbf{x}_{\mathbf{j}} = (x_{j_1}^{(1)}, \ldots, x_{j_d}^{(d)})$ denotes the grid point corresponding to multi-index $\mathbf{j}$ over the d-dimensional domain. By construction, this is a Riemann sum for the integral of u^2 over the domain Ω. Consequently, it provably converges to the continuous $L^2(\Omega)$ norm as the discretization is refined (see Theorem 4 in the appendix for a formal proof). Next, to connect the discrete and continuous spaces, the projection P_d must preserve the geometry of the function space. That is, the distance between two functions in the continuous space must be comparable to the distance between their discretized representations. Next, we assume that the target function $u \in \mathcal{Y}$ belongs to a Sobolev space $H^s(\Omega)$ with $s > d/2$ Le & Dik (2024); Katende (2025). Since solutions generated by neural operators to partial differential equations (PDEs) are typically smooth, this assumption is justified Furuya et al. (2024). This motivates the following:

Assumption 1 (Bilipschitz Discretization)**.** *The discretization map P_d is bilipschitz, meaning there exist constants $0 < c_1 \leq 1 \leq c_2$ such that for all $u, v \in \mathcal{Y}$:*

$$c_1 \|u - v\|_{\mathcal{Y}} \leq \|P_d(u) - P_d(v)\|_{w,2,d} \leq c_2 \|u - v\|_{\mathcal{Y}} \tag{3}$$

Here, the constants c_1 and c_2 quantify this approximation error, which for smooth functions on a grid with mesh size η is known to scale as $1 \pm \mathcal{O}(\eta^k)$ for some $k > 0$. And again, since the PDE solutions we are interested in are smooth, we expect $c_1 \to 1$ for a sufficiently fine grid, making our discrete norm a near-isometric proxy for the continuous one Hicken & Zingg (2013).

Remark 1. *In practice, such inequalities hold when P_d is implemented as (i) a sufficiently fine grid-based sampling of $u \in \mathcal{Y}$, or (ii) a truncated spectral expansion that retains enough terms to capture the relevant energy in $\|\cdot\|_{\mathcal{Y}}$. Moreover, when $d \gg N$, the constants c_1 and c_2 are approximatly 1, implying that P_d becomes a near-isometric embedding of the subspace of interest.*

Under this assumption, we derive our main theoretical result.

Theorem 1 (Functional Conformal Coverage)**.** *Let τ_α be the threshold calibrated on the discrete scores $s_i = \|P_d(\hat{u}_i) - P_d(u_i)\|_{w,2,d}$ where $\hat{u}_i$ is the predicted function. Under Assumption 1, the functional prediction set: $\Gamma_\alpha^{func}(f) = \{v \in \mathcal{Y} : \|\mathcal{G}_\theta(f) - v\|_{\mathcal{Y}} \leq \tau_\alpha / c_1\}$,satisfies the coverage guarantee:*

$$\mathbb{P}\left(u_{n+1} \in \Gamma_\alpha^{func}(f_{n+1})\right) \geq 1 - \alpha, \tag{4}$$

where (f_{n+1}, u_{n+1}) is an i.i.d. test point drawn from the data-generating distribution.

Proof Sketch. The proof proceeds by connecting the established discrete-space guarantee to the continuous function space via Assumption 1. From split CP, we have the guarantee $\mathbb{P}(s_{n+1} \leq \tau_\alpha) \geq 1 - \alpha$ in the discrete space, where $s_{n+1} = \|P_d(\hat{u}_{n+1}) - P_d(u_{n+1})\|_{w,2,d}$. The left-hand side of the inequality in Assumption 1 states that $c_1 \|\hat{u}_{n+1} - u_{n+1}\|_{\mathcal{Y}} \leq s_{n+1}$. Combining these, the event $s_{n+1} \leq \tau_\alpha$ implies the event $c_1 \|\hat{u}_{n+1} - u_{n+1}\|_{\mathcal{Y}} \leq \tau_\alpha$, which is equivalent to $\|\hat{u}_{n+1} - u_{n+1}\|_{\mathcal{Y}} \leq \tau_\alpha / c_1$. Since the first event implies the second, the probability of the second event must be at least as large as the first: $\mathbb{P}(\|\hat{u}_{n+1} - u_{n+1}\|_{\mathcal{Y}} \leq \tau_\alpha / c_1) \geq \mathbb{P}(s_{n+1} \leq \tau_\alpha) \geq 1 - \alpha$. This is precisely the coverage guarantee for the functional prediction set $\Gamma_\alpha^{\text{func}}(f)$. The full proof is provided in the appendix. □

4.3 A Heuristic Model for the Conformal Radius

Similar to the classic bias-variance decomposition of prediction error Brofos et al. (2019), we view the conformal radius $\tau_\alpha(d)$–where d denotes the number of evaluation points used by the discretization operator P_d–as arising from three dominant sources: discretization of the underlying function, finite-sample calibration, and model misspecification. To capture these effects, we propose a first-order Hastie et al. (2009) heuristic model that decomposes $\tau_\alpha(d)$ as:

$$\tau_\alpha(d) \approx \underbrace{\varepsilon_{\text{disc}}(d)}_{\text{discretization}} + \underbrace{\varepsilon_{\text{cal}}}_{\text{calibration}} + \underbrace{\varepsilon_{\text{misspec}}(d)}_{\text{misspecification}}. \tag{5}$$

Here, $\varepsilon_{\text{disc}}(d) = \|u - P_d u\|_{w,2}$ is the discretization error, decaying as $\mathcal{O}(d^{-p})$ Lanthaler et al. (2024), $\varepsilon_{\text{cal}} = \mathcal{O}(1/\sqrt{n})$ is the finite-sample calibration error Ghosh et al. (2023), and $\varepsilon_{\text{misspec}}(d)$ is the model's generalization error at resolution d. Due to the dependence of $\varepsilon_{\text{misspec}}$ on the predictor, developing a theoretical bound on its error is challenging and inefficient. Instead, we analyze the distribution of τ_α values across resolutions, per predictors. Empirically, we find the distributions of τ_α are approximately log-linear evaluated at resolutions beyond the training resolution when using a FNO, making it particularly useful for super-resolution tasks without requiring retraining or recalibration (see Figures 5 and 6 in the Appendix). To estimate the super-resolution conformal radius τ_α, we fit a regression of the form $\tau(R) = \exp(s \cdot R + b)$, where $\log \tau(R_i) = s \cdot R_i + b$. Although no formal coverage guarantee exists for the extrapolated value τ_α, we find that it yields substantial improvements in coverage accuracy, detailed in Table 5.

4.4 Time-Series Forecasting

Until now, we considered spatial mappings of the form $f \mapsto u$, where split conformal prediction guarantees hold under exchangeability. We now extend this perspective to temporal mappings, where a neural operator predicts the evolution over time t. Formally, let u_t denote the true system state at time t and $\hat{f}_t$ the operator's forecasted input at time t. Because each forecasted state $\hat{f}_t$ depends on previous predictions, the sequence $\{(\hat{f}_t, u_t)\}_{t=1}^T$ is no longer exchangeable. This violates the core assumption of split conformal prediction, causing a coverage gap as prediction errors compound over time. Rather than "correcting" the conformal radius to maintain a guarantee under drift Cleaveland et al. (2024), we reinterpret the coverage gap itself as a diagnostic signal.

Theorem 2 (Drift-Aware Functional Coverage)**.** *Let τ_α be the conformal threshold computed on discrete nonconformity scores $s_i = \|P_d(\hat{u}_i) - P_d(u_i)\|_{w,2,d}$ under a calibration distribution $\mathbb{P}_{\text{cal}}$, where $P_d : \mathcal{Y} \to \mathbb{R}^d$ is bilipschitz as in Assumption 1. At forecast steps $t = 1, \ldots, T$, let the data distribution drift to $\mathbb{P}_t$. Define the functional prediction set: $\Gamma_\alpha^{\textit{func}}(f_t) = \left\{v \in \mathcal{Y} : \|\hat{u}_t - v\|_{\mathcal{Y}} \leq \tau_\alpha / c_1\right\}$, where c_1 is the constant from Assumption 1. Then, for each t,*

$$\mathbb{P}_t\left(u_t \in \Gamma_\alpha^{\textit{func}}(f_t)\right) \geq 1 - \alpha - d_{\text{TV}}(\mathbb{P}_{\text{cal}}, \mathbb{P}_t), \tag{6}$$

where d_{TV} is the total variation distance between the calibration and forecast-time distributions.

By applying a fixed, conformally-calibrated threshold at each forecast step, a violation of the conformal bound serves as an interpretable signal that the accumulated model drift, quantified by d_{TV}, has become significant enough to degrade coverage below the desired level.

5 Methodology and Implementation

Functional Calibration. Following supervised training, we perform conformal calibration on a held-out set of input–output function pairs $\{(f_i, u_i)\}_{i=1}^n$. Each ground-truth function $u_i^{(d)} = P_d(u_i)$ and prediction $\hat{u}_i^{(d)} = P_d(\mathcal{G}_\theta(f_i))$ are discretized via the operator P_d. The nonconformity score for the i-th calibration sample is the relative weighted L^2 error $s_i = \frac{\|\hat{u}_i - u_i\|_{w,2}}{\|\hat{u}_i\|_{w,2}}$, where $\|\cdot\|_{w,2}$ is the quadrature-weighted norm. We use the relative form so the score is scale-invariant across functions of different magnitudes and therefore more meaningful. In implementation we include a small constant $\epsilon > 0$ in the denominator to avoid numerical instability when $\|\hat{u}_i\|_{w,2}$ is close to zero. Given we are working with fine resolutions we assume $c_1 = 1$, following Remark 1. Using this score, we compute the threshold τ_α, and the prediction set for any new input function f is:

$$\Gamma_\alpha(f) = \left\{u : \frac{\|\hat{u} - u\|_{w,2}}{\|\hat{u}\|_{w,2}} \leq \tau_\alpha\right\}. \tag{7}$$

Monte Carlo Bounding. Generative methods characterize uncertainty by producing multiple realizations of the solution. Popular examples include MC-dropout, variational autoencoders, and diffusion models, which generate an ensemble of n candidate realizations $\hat{u}^{(j)}(x) \sim S_\theta(x)$. The spread of these samples is then used to quantify variability. To construct pointwise prediction intervals, we compute lower and upper bounds at each spatial location j by taking the minimum and maximum over the ensemble, forming a conservative bounding envelope $\hat{u}^{\min}(x) = \min_{1 \leq j \leq n} \hat{u}^{(j)}(x)$,

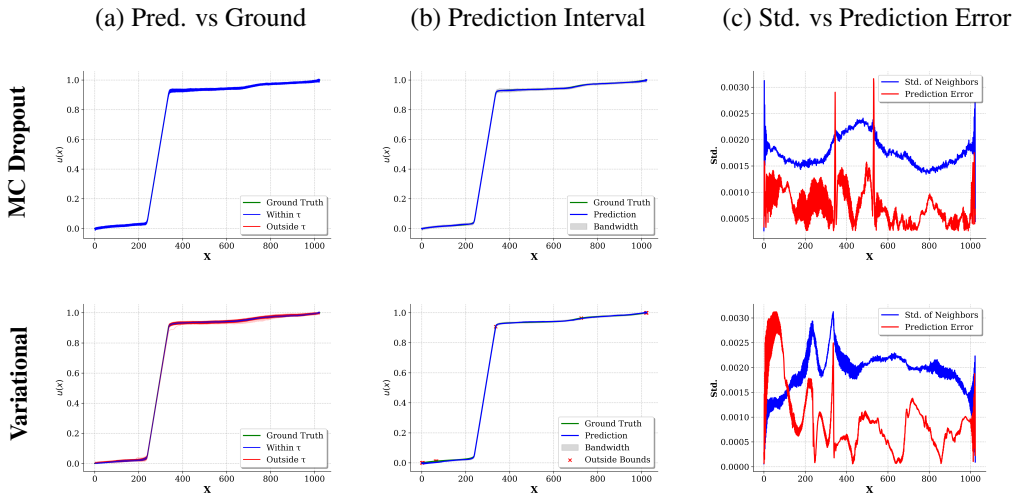


Figure 1: Darcy 1D: MC Bounds, with τ_α calibration at level $\alpha = 0.1$. Shown is a sample drawn and evaluated on a FNO with MC-Dropout and a variational-FNO. **(a)** Samples drawn for one specific instance, with samples within τ_α shown in blue and those outside τ_α in red. **(b)** Bounds inferred from the minimum and maximum of the samples within τ_α. **(c)** Bandwidth versus true error.

and $\hat{u}^{\max}(x) = \max_{1\le j\le n} \hat{u}^{(j)}(x)$. Under this formulation, we obtain conservative estimators of the functional upper and lower bounds directly at inference time. An alternative approach would involve using the sample standard deviation to form prediction bands; however, we find that this often leads to undercoverage. To incorporate the conformal threshold, we modify the previous procedure by conditioning the interval construction on the calibrated threshold τ_α, such that $\hat{u}^{\min}(x) = \min_{1\le j\le n} \hat{u}^{(j)}(x) | \hat{u}(x) \le \tau_\alpha$, and $\hat{u}^{\max}(x) = \max_{1\le j\le n} \hat{u}^{(j)}(x) | \hat{u}(x) \le \tau_\alpha$.

Quantile Bounds Adjustment. Alternatively, models may learn a direct heuristic of the error field, thereby producing explicit estimates of the prediction interval. In this work, we consider the triplet formulation $(\hat{u}^{\text{lo}}, \hat{u}^{\text{mid}}, \hat{u}^{\text{hi}})$, where the upper and lower bounds are learned via quantile regression using the pinball loss Huang et al. (2013). To align the resulting interval $[\hat{u}^{\text{lo}}, \hat{u}^{\text{hi}}]$ with the target coverage level, we perform a post hoc adjustment. For a given bound $\hat{u}$ (either lower or upper), we define its offset from the central prediction as $\delta = \hat{u} - \hat{u}^{\text{mid}}$, and compute the initial error:

$$r_{\text{initial}} = \frac{\|\hat{u} - \hat{u}^{\text{mid}}\|}{\|\hat{u}^{\text{mid}}\|} = \frac{\|\delta\|}{\|\hat{u}^{\text{mid}}\|}. \tag{8}$$

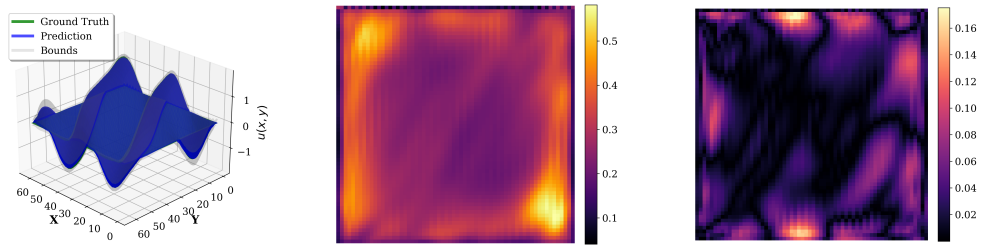
Next, we derive a scaling factor s as the ratio of the target relative error τ_α to the initial error: $\mathbf{s} = \tau_\alpha / r_{\text{initial}}$. Finally, adjusted bound $\hat{u}^{\text{adj}}$ is then found by scaling the original offset:

$$\hat{u}^{\text{adj}} = \hat{u}^{\text{mid}} + s \cdot \delta = \hat{u}^{\text{mid}} + \left(\frac{\tau_\alpha}{r_{\text{initial}}}\right) \cdot (\hat{u} - \hat{u}^{\text{mid}}). \tag{9}$$

This procedure, applied to both $\hat{u}^{\text{lo}}$ and $\hat{u}^{\text{hi}}$, yields a recalibrated interval $[\hat{u}^{\min}, \hat{u}^{\max}]$ whose bounds are exactly τ_α distant from the center in relative terms.

Table 1: Coverage Summary for Darcy 1D. Both scalars are set for a significance level $\alpha = 0.1$. Our method calibrates the overly conservative baselines to achieve our desired significance level.

	Technique	Scalar	Func.	Point.
Uncalibrated	MC Dropout	–	–	1.000
	Variational	–	–	1.000
Calibrated	MC Dropout	0.0044	0.9003	**0.9200**
	Variational	0.0035	**0.9000**	0.8807



(a) Prediction vs Ground Truth (b) Std. of Neighbors (c) Prediction Error

Figure 2: Poisson 2D: Adjusted Quantiles, with τ_α calibration at level $\alpha = 0.1$. **(a)** Mean prediction against the ground truth, with upper and lower bounds scaled to the τ_α distance. **(b)** Bandwidth of the scaled bounds. **(c)** Prediction error between the mean prediction and the ground truth.

Stochastic Forecast Bounding. For time-series models that generate stochastic trajectories, we first obtain the conformal radius τ_α, calibrated on the one initial one step distribution. We then apply this radius across all forecast steps to monitor for performance degradation. At each time step t, the model produces an ensemble of forecasted solutions, $\{\hat{u}_t^{(j)}\}_{j=1}^n$. The forecast's quality is then evaluated using a the following metrics. We first define the calibration error as the normalized distance between the ensemble mean, $\bar{u}_t$, and the ground-truth solution, u_t. This term is directly comparable to τ_α and provides a binary assessment of whether the forecast remains within its initial bounds. Next, we compute the ensemble spread to quantify forecast sharpness by measuring the average normalized distance of individual ensemble members from their mean:

$$\text{ES}_t = \frac{1}{n}\sum_{j=1}^{n} \frac{\|\hat{u}_t^{(j)} - \bar{u}_t\|}{\|\bar{u}_t\|}, \quad \text{for} \quad \bar{u}_t := \frac{1}{n}\sum_{j=1}^{n} \hat{u}_t^{(j)}. \tag{10}$$

These two components are summed to form the conformal ensemble score (CES), a metric analogous to the continuous ranked probability score (CRPS) Pic et al. (2023). Finally, to measure the ensemble's internal consensus, we define the conformal prediction set $\Gamma_\alpha^{(t)}$ as:

$$\Gamma_\alpha^{(t)} = \left\{ \hat{u}_t^{(j)} : \frac{\|\hat{u}_t^{(j)} - \bar{u}_t\|}{\|\bar{u}_t\|} \le \tau_\alpha \right\}. \tag{11}$$

The proportion of members within $\Gamma_\alpha^{(t)}$ serves as our internal agreement (IA) metric. A drop in IA over time reflects a loss of forecast confidence and growing internal divergence. While no formal coverage guarantee holds at individual time steps $t > 1$ due to non-exchangeability in the time-series sequence, this approach provides an estimate of the temporal degradation.

Evaluation We assess calibration at three levels. First, pointwise coverage measures the proportion of spatial points within the conformal bounds, $C_e = \frac{1}{N}\sum_{i=1}^{N} \mathbf{I}[\hat{u}_i^{\min}(x) \le u_i(x) \le \hat{u}_i^{\max}(x)]$. Second, functional coverage evaluates whether entire functions satisfy the calibrated radius, $C_f = \frac{1}{K}\sum_{j=1}^{K} \mathbf{I}\left[\frac{\|\hat{u}_j - u_j\|_{w,2}}{\|\hat{u}_j\|_{w,2}} \le \tau_\alpha\right]$. Finally, for long-horizon forecasts, we track diagnostic metrics (CES and IA) and qualitatively interpret their signals.

Table 2: Coverage Summary for Poisson 2D. Both scalars are set for a significance level $\alpha = 0.1$. The functional and pointwise coverage reaches slightly above our target coverage of 90%.

	Method	Scalar	Func.	Point.
Uncalibrated	Pinball Only	–	–	0.6911
Calibrated	Risk Controlling	0.1480	–	**0.9101**
	Our Bounds	0.1486	**0.9010**	0.9108

Table 3: Evolution of key metrics for the autoregressive forecast, evaluated at a significance level of $\alpha = 0.1$ with a conformal threshold of $\tau = 0.023$. The mean distance serves as the calibration error, the within τ_α column gives a binary check of whether this error exceeds the threshold, and the ensemble spread measures forecast sharpness. Our CES score combines calibration error and sharpness to evaluate overall quality, while our IA metric quantifies the forecast's consensus.

Metric	$t = 1$	$t = 2$	$t = 3$	$t = 4$	$t = 5$	$t = 6$	$t = 7$	$t = 8$
Within τ_α	Yes	Yes	Yes	Yes	Yes	**No**	**No**	**No**
Mean Distance	0.013	0.012	0.014	0.017	0.020	0.026	0.032	0.039
Ensemble Spread	0.013	0.013	0.015	0.019	0.025	0.032	0.040	0.047
IA (Ours)	98.0%	98.0%	92.2%	69.0%	36.3%	20.0%	13.8%	10.7%
CES (Ours)	0.026	0.025	0.029	0.036	0.046	0.058	0.072	0.086
CRPS	0.008	0.007	0.009	0.011	0.014	0.017	0.021	0.025

6 Numerical Results & Discussion

To validate our framework, we evaluate the resulting prediction sets on progressively more challenging PDE benchmarks: Monte Carlo bounding for Darcy flow, post hoc quantile adjustment for the Poisson equation, and bounded ensemble forecasting for autoregressive Navier–Stokes dynamics. In addition, we conduct two ablation studies to assess the framework's robustness to grid discretization and resolution misspecification.

6.1 Darcy 1D: Monte Carlo Bounds

We first evaluate ability of MC methods to produce calibrated bounds for generative UQ in 1D Darcy flow using variational and MC-dropout neural operators. Shown in Table 1, the uncalibrated approach yields overly consecrative coverage, whereas our calibration procedure successfully produces uncertainty sets that meet the target coverage. Additionally, we highlight a trade-off between methods: MC-Dropout achieves robust coverage with a larger conformal radius, while the variational method yields a tighter bound at the cost of slight under-coverage, a finding visually confirmed in Figure 1. Notably, the standard MC approach lacks functional coverage, as there is no global scalar metric to compare entire functions; therefore it can only provide pointwise uncertainty estimates. While calibration is effective in this setting, the computational overhead of sampling-based methods presents a significant bottleneck for deployment in high-resolution, complex PDEs.

6.2 Poisson 2D: Adjusted Quantiles

Next, we evaluate a more computationally efficient approach on the 2D Poisson equation using a neural operator that directly outputs both the solution and an error heuristic. As shown in Table 2, our method achieves pointwise coverage levels comparable to the risk-controlling PAC bounds formulated by Ma et al. (2024). A key strength of our approach is its flexibility: since calibration is performed post hoc using a functional norm, the method accommodates both pointwise and functional uncertainty quantification, which the risk controlling bounds cannot do. Figure 2 illustrates that the learned error heuristic closely tracks the true error, though the resulting prediction intervals tend to be wider in many regions—a characteristic of models trained with the pinball loss. Although, the reliance on the pinball loss introduces a potential failure mode if it fails to accurately capture the true error distribution, a limitation noted in prior work.

6.3 Navier-Stokes 2D: Bounded Ensemble Forecasting

As the underlying dynamics become increasingly complex, learned error heuristics may fail, restricting us to quantifying distributional drift instead. For our final case, we consider stochastic autoregressive forecasting of the 2D Navier–Stokes vorticity field. As shown in Table 3, forecast quality progressively degrades over time, evidenced by increasing CES and CRPS values and a sharp decline in IA across the forecast horizon. A key advantage of our framework is its ability to relate forecast degradation to the significance level. The `Within` τ_α column provides a binary indicator of calibration validity, flagging violations of the predefined safety threshold. Furthermore,

Table 4: Comparison of calibrated thresholds τ_α across grid geometries for $\alpha = 0.1$. The weighted norm yields more consistent τ_α values with less variation under grid scheme.

Grid Type	Relative Norm	Weighted Norm
Uniform	0.02810	0.02810
Clustered Center	0.02866	0.02821
Clustered Boundary	0.03280	0.03079
Std. of Thresholds	0.00257	**0.00148**
Coeff. of Variation	8.6%	**5.1%**

because all components are expressed in the same relative units as τ_α, the magnitudes of continuous diagnostic values can be directly interpreted with respect to the calibrated conformal radius. Finally, by jointly analyzing IA and CES, we observe that even when the ensemble mean remains within calibrated bounds, individual members may drift beyond a reasonable significance level, indicating loss of forecast reliability.

6.4 Ablation Studies

Conditional Coverage. Assessing the effect of grid geometry on calibration is essential for demonstrating the necessity of our weighted norm. Using the 2D Poisson problem, we generate data on three distinct grid types: uniform, center-clustered, and boundary-clustered. For each grid, we compute τ_α using both the standard relative L^2 norm and the weighted relative L^2 norm. As shown in Table 4, the weighted norm yields τ_α values with lower relative variation across grid types. In practice, neural operators often develop biases toward the discretization used during training. Our method mitigates this bias by incorporating the grid geometry, allowing the nonconformity score to more faithfully approximate the continuous-space error.

Super-Resolution Coverage. Next, we evaluate the super-resolution task introduced in the Theoretical Foundation section. Table 5 shows that directly applying a conformal threshold τ_α calibrated at a low resolution results in substantial undercoverage at higher resolutions, with the Poisson case falling nearly 70% below the target level. After adjustment, coverage for the Darcy case is fully restored to the desired 90%, while the Poisson case, though still imperfect, improves markedly from 19.9% to 74.9%. These results indicate that the proposed adjustment substantially improves reliability, but does not fully recover target coverage under large resolution shifts, suggesting that the heuristic fails to capture the full complexity of the underlying error behavior. Furthermore, estimating the regression slope requires careful selection of calibration points, and thus relies on prior knowledge of the resolution-dependent behavior of $\tau_\alpha(d)$.

7 Conclusion

This work extends split CP to function spaces using a quadrature-weighted norm, enabling calibrated uncertainty quantification for neural operators across varying grid resolutions. The proposed method improves stability under discretization shifts, achieves near target coverage in super-resolution tasks via a heuristic model of τ_α, and enables reliability diagnostics in autoregressive forecasting through CES and IA. Empirical results demonstrate improved functional and pointwise coverage. Extensions to strongly non-smooth or discontinuous PDE solutions remain necessary work.

Table 5: Coverage summary for Darcy 1D and Poisson 2D super-resolutions, calibrated at significance level $\alpha = 0.1$.

	Instance	Scalar	Coverage
Unadjusted	Darcy	0.00771	0.878
	Poisson	0.09659	0.199
Adjusted	Darcy	0.00811	**0.909**
	Poisson	0.10767	**0.749**

REFERENCES

Deepak Akhare, Tengfei Luo, and Jian-Xun Wang. Diffhybrid-UQ: Uncertainty quantification for differentiable hybrid neural modeling, 2023. URL https://arxiv.org/abs/2401.00161.

Ali Baheri and Marzieh Amiri Shahbazi. Conformal prediction across scales: Finite-sample coverage with hierarchical efficiency. *Results in Applied Mathematics*, 26:100589, 2025. ISSN 2590-0374. doi: https://doi.org/10.1016/j.rinam.2025.100589. URL https://www.sciencedirect.com/science/article/pii/S2590037425000536.

Bahador Bahmani, Somdatta Goswami, Ioannis G. Kevrekidis, and Michael D. Shields. A resolution independent neural operator. *Computer Methods in Applied Mechanics and Engineering*, 444: 118113, 2025. ISSN 0045-7825. doi: https://doi.org/10.1016/j.cma.2025.118113. URL https://www.sciencedirect.com/science/article/pii/S0045782525003858.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816 – 845, 2023. doi: 10.1214/23-AOS2276. URL https://doi.org/10.1214/23-AOS2276.

James Brofos, Rui Shu, and Roy R Lederman. A bias-variance decomposition for bayesian deep learning. In *NeurIPS 2019 Workshop on Bayesian Deep Learning*, 2019.

Christopher Bülte, Nina Horat, Julian Quinting, and Sebastian Lerch. Uncertainty quantification for data-driven weather models. *Artificial Intelligence for the Earth Systems*, 2025.

Matthew Cleaveland, Insup Lee, George J Pappas, and Lars Lindemann. Conformal prediction regions for time series using linear complementarity programming. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 20984–20992, 2024.

Nikita Durasov, Timur Bagautdinov, Pierre Baque, and Pascal Fua. Masksembles for uncertainty estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13539–13548, 2021.

Loic Le Folgoc, Vasileios Baltatzis, Sujal Desai, Anand Devaraj, Sam Ellis, Octavio E Martinez Manzanera, Arjun Nair, Huaqi Qiu, Julia Schnabel, and Ben Glocker. Is mc dropout bayesian? *arXiv preprint arXiv:2110.04286*, 2021.

Matteo Fontana, Gianluca Zeni, and Simone Vantini. Conformal prediction: A unified review of theory and new challenges. *Bernoulli*, 29(1):1 – 23, 2023. doi: 10.3150/21-BEJ1447. URL https://doi.org/10.3150/21-BEJ1447.

Takashi Furuya, Koichi Taniguchi, and Satoshi Okuda. Quantitative approximation for neural operators in nonlinear parabolic equations. *arXiv preprint arXiv:2410.02151*, 2024.

Subhankar Ghosh, Yuanjie Shi, Taha Belkhouja, Yan Yan, Jana Doppa, and Brian Jones. Probabilistically robust conformal prediction. In Robin J. Evans and Ilya Shpitser (eds.), *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 681–690. PMLR, 31 Jul–04 Aug 2023. URL https://proceedings.mlr.press/v216/ghosh23a.html.

Ethan Goan and Clinton Fookes. *Bayesian Neural Networks: An Introduction and Survey*, pp. 45–87. Springer International Publishing, Cham, 2020. ISBN 978-3-030-42553-1. doi: 10.1007/978-3-030-42553-1_3. URL https://doi.org/10.1007/978-3-030-42553-1_3.

Ander Gray, Vignesh Gopakumar, Sylvain Rousseau, and Sébastien Destercke. Guaranteed prediction sets for functional surrogate models. *arXiv preprint arXiv:2501.18426*, 2025.

Ling Guo, Hao Wu, Yan Wang, Wenwen Zhou, and Tao Zhou. Ib-uq: Information bottleneck based uncertainty quantification for neural function regression and neural operator learning. *Journal of Computational Physics*, 510:113089, 2024. ISSN 0021-9991. doi: https://doi.org/10.1016/j.jcp.2024.113089. URL https://www.sciencedirect.com/science/article/pii/S0021999124003383.

Trevor Harris and Yan Liu. Locally adaptive conformal inference for operator models. *arXiv preprint arXiv:2507.20975*, 2025.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

David Haussler and Manfred Warmuth. The probably approximately correct (pac) and other learning models. *The Mathematics of Generalization*, pp. 17–36, 2018.

Jason E Hicken and David W Zingg. Summation-by-parts operators and high-order quadrature. *Journal of Computational and Applied Mathematics*, 237(1):111–125, 2013.

Xiaolin Huang, Lei Shi, and Johan AK Suykens. Support vector machine classifier with pinball loss. *IEEE transactions on pattern analysis and machine intelligence*, 36(5):984–997, 2013.

Ronald Katende. Stability analysis of physics-informed neural networks via variational coercivity, perturbation bounds, and concentration estimates. *arXiv preprint arXiv:2506.13554*, 2025.

Serkan Kiranyaz, Junaid Malik, Habib Ben Abdallah, Turker Ince, Alexandros Iosifidis, and Moncef Gabbouj. Self-organized operational neural networks with generative neurons. *Neural Networks*, 140:294–308, 2021. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2021.02.028. URL https://www.sciencedirect.com/science/article/pii/S0893608021000782.

Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. *Advances in neural information processing systems*, 32, 2019.

Samuel Lanthaler, Andrew M. Stuart, and Margaret Trautner. Discretization error of Fourier neural operators, 2024. URL https://arxiv.org/abs/2405.02221.

Jose Antonio Lara Benitez, Takashi Furuya, Florian Faucher, Anastasis Kratsios, Xavier Tricoche, and Maarten V. de Hoop. Out-of-distributional risk bounds for neural operators with applications to the helmholtz equation. *Journal of Computational Physics*, 513:113168, September 2024. ISSN 0021-9991. doi: 10.1016/j.jcp.2024.113168. URL http://dx.doi.org/10.1016/j.jcp.2024.113168.

Vu-Anh Le and Mehmet Dik. A mathematical analysis of neural operator behaviors. *arXiv preprint arXiv:2410.21481*, 2024.

Zongyi Li, Nikola Borislavov Kovachki, Kamyar Azizzadenesheli, Burigede liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=c8P9NQVtmnO.

Xiaoyi Liu and Hao Tang. Difffno: Diffusion fourier neural operator. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 150–160, June 2025.

Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via deeponet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3(3):218–229, March 2021. ISSN 2522-5839. doi: 10.1038/s42256-021-00302-5. URL http://dx.doi.org/10.1038/s42256-021-00302-5.

Ziqi Ma, David Pitt, Kamyar Azizzadenesheli, and Anima Anandkumar. Calibrated uncertainty quantification for operator learning via conformal prediction. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=cGpegxy12T.

Romain Pic, Clément Dombry, Philippe Naveau, and Maxime Taillardat. Distributional regression and its evaluation with the crps: Bounds and convergence of the minimax risk. *International Journal of Forecasting*, 39(4):1564–1572, October 2023. ISSN 0169-2070. doi: 10.1016/j.ijforecast.2022.11.001. URL http://dx.doi.org/10.1016/j.ijforecast.2022.11.001.

Weizhu Qian, Dalin Zhang, Yan Zhao, Kai Zheng, and James JQ Yu. Uncertainty quantification for traffic forecasting: A unified approach. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, pp. 992–1004. IEEE, 2023.

Rahul Rahaman et al. Uncertainty quantification and deep ensembles. *Advances in neural information processing systems*, 34:20063–20075, 2021.

Rajesh Ranganath, Dustin Tran, Jaan Altosaar, and David Blei. Operator variational inference. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/d947bf06a885db0d477d707121934ff8-Paper.pdf.

Eric Schulz, Maarten Speekenbrink, and Andreas Krause. A tutorial on gaussian process regression: Modelling, exploring, and exploiting functions. *Journal of Mathematical Psychology*, 85:1–16, 2018. ISSN 0022-2496. doi: https://doi.org/10.1016/j.jmp.2018.03.001. URL https://www.sciencedirect.com/science/article/pii/S0022249617302158.

Roman Scoccimarro. Transients from initial conditions: a perturbative analysis. *Monthly Notices of the Royal Astronomical Society*, 299(4):1097–1118, 1998.

Alexander Shapiro. Monte carlo sampling methods. *Handbooks in operations research and management science*, 10:353–425, 2003.

Vinh Ngoc Tran, M Chase Dwelle, Khachik Sargsyan, Valeriy Y Ivanov, and Jongho Kim. A novel modeling framework for computationally efficient and accurate real-time ensemble flood forecasting with uncertainty quantification. *Water Resources Research*, 56(3):e2019WR025727, 2020.

Zongren Zou, Xuhui Meng, and George Em Karniadakis. Uncertainty quantification for noisy inputs–outputs in physics-informed neural networks and neural operators. *Computer Methods in Applied Mechanics and Engineering*, 433:117479, 2025. ISSN 0045-7825. doi: https://doi.org/10.1016/j.cma.2024.117479. URL https://www.sciencedirect.com/science/article/pii/S0045782524007333.

A CODE AVAILABILITY

Our repository is available under the Apache License version 2.0 at `https://github.com/SAILRIT/CFNO`. Our checkpoints and output can be obtained upon request from the corresponding author.

B DETAILED PROOF OF THEOREM 1

Proof. Let $\{(f_i, u_i)\}_{i=1}^{n+1}$ be an i.i.d. sample from the data-generating distribution on $\mathcal{X} \times \mathcal{Y}$. Let $\hat{u}_i = \mathcal{G}_\theta(f_i)$ be the model prediction for input f_i. Our goal is to construct a prediction set for u_{n+1} in the function space $\mathcal{Y}$ with a coverage guarantee. The proof proceeds in five steps.

Step 1: Calibration in Discretized Space. We define a nonconformity score on the discretized space $\mathbb{R}^d$ using the projection operator $P : \mathcal{Y} \to \mathbb{R}^d$ and the quadrature-weighted norm $||\cdot||_{w,2,d}$. For each of the first n samples in the calibration set, we compute the score:

$$s_i = ||P(\hat{u}_i) - P(u_i)||_{w,2,d}. \tag{12}$$

We then compute the conformal threshold τ_α as the $\lceil (1-\alpha)(n+1) \rceil / (n+1)$-th empirical quantile of the scores $\{s_1, \ldots, s_n\}$. By the standard theory of split conformal prediction, the score for the test point s_{n+1} satisfies:

$$\mathbb{P}(s_{n+1} \leq \tau_\alpha) \geq 1 - \alpha. \tag{13}$$

This provides a finite-sample, distribution-free guarantee in the discretized space.

Step 2: Relating Discrete and Continuous Nonconformity. The core of our functional guarantee relies on relating the discrete score s_i to a true nonconformity score in the continuous function space, $s_i^{\text{cont}} = ||\hat{u}_i - u_i||_{\mathcal{Y}}$. This relationship is formalized by the bilipschitz assumption (Assumption 1), which states that there exist constants $c_1, c_2 > 0$ such that for any $u, v \in \mathcal{Y}$:

$$c_1 ||u - v||_{\mathcal{Y}} \leq ||P(u) - P(v)||_{w,2,d} \leq c_2 ||u - v||_{\mathcal{Y}}. \tag{14}$$

Applying this to our nonconformity scores, we have:

$$c_1 s_i^{\text{cont}} \leq s_i \leq c_2 s_i^{\text{cont}} \quad \forall i = 1, \ldots, n+1. \tag{15}$$

Step 3: Deriving the Functional Coverage Guarantee. We can now translate the coverage guarantee from the discrete space to the continuous space. The event $s_{n+1} \leq \tau_\alpha$ is the basis of our $1 - \alpha$ guarantee. From the bilipschitz inequality, this event implies a condition on the continuous score: if $s_{n+1} \leq \tau_\alpha$, then $c_1 s_{n+1}^{\text{cont}} \leq s_{n+1} \leq \tau_\alpha$, which implies $s_{n+1}^{\text{cont}} \leq \tau_\alpha / c_1$. Because the event $\{s_{n+1} \leq \tau_\alpha\}$ implies the event $\{s_{n+1}^{\text{cont}} \leq \tau_\alpha / c_1\}$, the probability of the latter must be at least as large as the probability of the former:

$$\mathbb{P}(s_{n+1}^{\text{cont}} \leq \tau_\alpha / c_1) \geq \mathbb{P}(s_{n+1} \leq \tau_\alpha) \geq 1 - \alpha. \tag{16}$$

Step 4: Defining the Functional Prediction Set. This result allows us to define a prediction set directly in the function space $\mathcal{Y}$. We define the set $\Gamma_\alpha^{\text{func}}(f_{n+1})$ as a ball in $\mathcal{Y}$ with radius τ_α / c_1:

$$\Gamma_\alpha^{\text{func}}(f_{n+1}) = \{v \in \mathcal{Y} : ||\hat{u}_{n+1} - v||_{\mathcal{Y}} \leq \tau_\alpha / c_1\}. \tag{17}$$

From Step 3, the probability that the true function u_{n+1} lies in this set is:

$$\mathbb{P}(u_{n+1} \in \Gamma_\alpha^{\text{func}}(f_{n+1})) = \mathbb{P}(s_{n+1}^{\text{cont}} \leq \tau_\alpha / c_1) \geq 1 - \alpha. \tag{18}$$

This provides a valid finite-sample coverage guarantee for a prediction set in the function space $\mathcal{Y}$.

Step 5: The Asymptotic Argument. The "asymptotic" nature of the guarantee refers to the fact that the radius of the guaranteed set, τ_α / c_1, converges to the calibrated discrete radius τ_α as the discretization becomes arbitrarily fine. As the discretization is refined (i.e., $d \to \infty$), the quadrature-weighted norm converges to the continuous L^2 norm, and the projection P becomes a near-isometry for the class of functions of interest. In this limit, the distortion constant $c_1 \to 1$. Thus, for sufficiently fine discretizations, the guaranteed function-space ball has a radius that approaches the empirically calibrated threshold τ_α. □

Remark 2 (Asymptotic Convergence of the Prediction Set Radius.)**.** *The practical utility of the guarantee depends on the constant c_1, which quantifies the geometric distortion introduced by the discretization map P. The "asymptotic" aspect of our framework relates to the behavior of this constant. As the discretization is refined (i.e., $d \to \infty$), the quadrature-weighted norm converges to the continuous L^2 norm (as shown in Theorem 4), and the projection P becomes a near-isometry for the class of functions of interest. In this limit, the distortion constant $c_1 \to 1$. Consequently, for sufficiently fine discretizations, the radius of the guaranteed function-space ball, τ_α/c_1, approaches the empirically calibrated and directly computable threshold τ_α. This ensures that our functional prediction set is not only theoretically valid but also practically useful, as its size is determined by a well-behaved, empirical quantity.*

C Detailed Proof of Theorem 2

Proof. Let $\{(f_i, u_i)\}_{i=1}^n$ be an i.i.d. calibration set drawn from $\mathbb{P}_{\text{cal}}$, and let $\hat{u}_i = \mathcal{G}_\theta(f_i)$ denote model predictions. We aim to characterize the functional coverage when the same conformal threshold τ_α is applied to data at forecast time t from a potentially different distribution $\mathbb{P}_t$. The proof follows three steps.

Step 1: Coverage under Distribution Shift. By the standard split CP guarantee, the discrete scores $s_i = \|P_d(\hat{u}_i) - P_d(u_i)\|_{w,2,d}$ satisfy $\mathbb{P}_{\text{cal}}\big(s \le \tau_\alpha\big) \ge 1 - \alpha$. When this threshold is applied to samples drawn from $\mathbb{P}_t$, the drift bound of Barber et al. (2023) implies

$$\mathbb{P}_t\big(s \le \tau_\alpha\big) \ge 1 - \alpha - d_{\text{TV}}(\mathbb{P}_{\text{cal}}, \mathbb{P}_t), \tag{19}$$

where d_{TV} is the total variation distance between $\mathbb{P}_{\text{cal}}$ and $\mathbb{P}_t$.

Step 2: Lifting to the Function Space. From Assumption 1, we have $c_1\|\hat{u}_t - u_t\|_{\mathcal{Y}} \le s$. Thus, the event $s \le \tau_\alpha$ implies

$$\|\hat{u}_t - u_t\|_{\mathcal{Y}} \le \tau_\alpha/c_1. \tag{20}$$

Step 3: Functional Prediction Set Coverage. Define the functional prediction set $\Gamma_\alpha^{\text{func}}(f_t) = \{v \in \mathcal{Y} : \|\hat{u}_t - v\|_{\mathcal{Y}} \le \tau_\alpha/c_1\}$. By Step 2, $s \le \tau_\alpha$ implies $u_t \in \Gamma_\alpha^{\text{func}}(f_t)$. Combining this implication with the drift-adjusted discrete guarantee from Step 1

$$\mathbb{P}_t\left(u_t \in \Gamma_\alpha^{\text{func}}(f_t)\right) \ge 1 - \alpha - d_{\text{TV}}(\mathbb{P}_{\text{cal}}, \mathbb{P}_t), \tag{21}$$

which establishes the stated result. □

D Additional Theoretical Results

Now we present three additional theoretical results, which further illuminate the connection between discrete and functional coverage. The first result Lemma 1 formalizes a stability property of the projection operator, ensuring that small discretization errors cannot drastically change the geometry of the output space. The second result Theorem 3 incorporates this stability into a PDE setting, where the solution operator is assumed to be stable with respect to perturbations in boundary conditions or source terms. The third result Theorem shows the convergence of the weighted norm to the continuous L^2 norm.

Lemma 1 (Projection Stability)**.** *Let $\mathcal{Y}$ be a normed function space equipped with $\|\cdot\|_{\mathcal{Y}}$, and let $P : \mathcal{Y} \to \mathbb{R}^d$ be a discretization (or projection) map. Suppose that there exist constants $c_1, c_2 > 0$ such that for all $u, v \in \mathcal{Y}$,*

$$c_1\|u - v\|_{\mathcal{Y}} \le \|P(u) - P(v)\|_{\mathbb{R}^d} \le c_2\|u - v\|_{\mathcal{Y}}. \tag{22}$$

Then P is bilipschitz on $\mathcal{Y}$, up to the constants c_1 and c_2. Consequently, any ball in $\mathcal{Y}$ is mapped to a comparable ball in $\mathbb{R}^d$, and vice versa.

Proof. The assumed double inequality implies that P preserves distances in $\mathcal{Y}$ up to multiplicative factors c_1 and c_2. Specifically,

$$c_1\|u - v\|y \le \|P(u) - P(v)\|_{\mathbb{R}^d} \le c_2\|u - v\|_y \tag{23}$$

Hence, if $\|u-v\|y \le \epsilon$, it follows that $\|P(u)-P(v)\|_{\mathbb{R}^d} \le c_2\epsilon$. Conversely, $\|P(u)-P(v)\|_{\mathbb{R}^d} \le \delta$ implies $\|u - v\|y \le \frac{\delta}{c_1}$. Thus, neighborhoods in $\mathcal{Y}$ map to neighborhoods in $\mathbb{R}^d$ with at most a constant change in radius, establishing bilipschitz continuity of P. □

Theorem 3 (Conformal Coverage for PDE Solutions Under Operator Stability). *Let $\mathcal{A} : \mathcal{X} \to \mathcal{Y}$ be a forward operator that maps an input function (e.g., initial/boundary data, source terms) $f \in \mathcal{X}$ to the corresponding PDE solution $u = \mathcal{A}(f) \in \mathcal{Y}$. Assume $\mathcal{A}$ is Lipschitz stable with constant $L > 0$; that is, for any $f_1, f_2 \in \mathcal{X}$,*

$$\|\mathcal{A}(f_1) - \mathcal{A}(f_2)\|_{\mathcal{Y}} \leq L \|f_1 - f_2\|_{\mathcal{X}}$$

Suppose further that the discretization map $P : \mathcal{Y} \to \mathbb{R}^d$ satisfies the bilipschitz property in the Lemma above, with constants $c_1 \leq c_2$. Let $\mathcal{G}_\theta$ be a neural-operator-based approximation of $\mathcal{A}$, and let $\widehat{\Gamma}_\alpha(\cdot) \subseteq \mathbb{R}^d$ be a (discretized) conformal set calibrated to achieve

$$\mathbb{P}\left[P(u_{n+1}) \notin \widehat{\Gamma}_\alpha(f_{n+1})\right] \leq \alpha$$

under i.i.d. draws $\{(f_i, u_i)\} \sim \mathcal{D}$. Define the functional conformal set

$$\Gamma_\alpha(f) = \left\{v \in \mathcal{Y} : P(v) \in \widehat{\Gamma}_\alpha(f)\right\}$$

Then, for sufficiently fine discretization (i.e., adequately large d), there exists $\delta \geq 0$ such that

$$\mathbb{P}\left[u_{n+1} \notin \Gamma_\alpha(f_{n+1})\right] \leq \alpha + \delta$$

where $u_{n+1} = \mathcal{A}(f_{n+1})$. Moreover, $\delta \to 0$ as the mesh or basis in P is refined and the training sample size n grows.

Proof. he key is to formalize the role of the PDE operator $\mathcal{A}$ in ensuring that the conditions for Theorem 1 are met.

Step 1: Defining the Relevant Function Subspace. The set of all possible PDE solutions forms a specific subset of the larger function space $\mathcal{Y}$. We define this subset as the range of the solution operator, $\mathcal{Y}_\mathcal{A} = \{\mathcal{A}(f) \mid f \in \mathcal{X}\}$. The Lipschitz stability of $\mathcal{A}$ implies that solutions in $\mathcal{Y}_\mathcal{A}$ possess a certain regularity (e.g., they lie within a ball in a Sobolev space). This regularity is critical because it makes the bilipschitz assumption on the discretization map P plausible.

Step 2: Applying the Bilipschitz Assumption on the Subspace. We assume that the discretization map $P : \mathcal{Y} \to \mathbb{R}^d$ is bilipschitz specifically on the subspace of solutions $\mathcal{Y}_\mathcal{A}$. That is, there exist constants $c_1, c_2 > 0$ such that for any two solutions $u, v \in \mathcal{Y}_\mathcal{A}$:

$$c_1||u - v||_{\mathcal{Y}} \leq ||P(u) - P(v)||_{w,2,d} \leq c_2||u - v||_{\mathcal{Y}}. \tag{24}$$

This is a more targeted assumption than requiring the property to hold over all of $\mathcal{Y}$. The stability of $\mathcal{A}$ makes this assumption reasonable for standard discretization methods (e.g., sufficiently fine grids for smooth solutions).

Step 3: Invoking Theorem 3. The calibration data consists of pairs (f_i, u_i) where $u_i = \mathcal{A}(f_i)$, so all true solutions u_i belong to $\mathcal{Y}_\mathcal{A}$. The neural operator $\mathcal{G}_\theta$ is trained to approximate $\mathcal{A}$, so its predictions $\hat{u}_i = \mathcal{G}_\theta(f_i)$ are also expected to lie in or near this subspace. Since the bilipschitz condition holds for the relevant functions (the true solutions and their approximations), all conditions for Theorem 3 are met. We calibrate a threshold τ_α in the discrete space as per Step 1 of the proof of Theorem 3. Then, following Steps 2-4 of that proof, we can construct a functional prediction set:

$$\Gamma_\alpha^{\text{func}}(f_{n+1}) = \{v \in \mathcal{Y} : ||\hat{u}_{n+1} - v||_{\mathcal{Y}} \leq \tau_\alpha / c_1\}, \tag{25}$$

which is guaranteed to have coverage of at least $1 - \alpha$:

$$\mathbb{P}(u_{n+1} \in \Gamma_\alpha^{\text{func}}(f_{n+1})) \geq 1 - \alpha, \tag{26}$$

where $u_{n+1} = \mathcal{A}(f_{n+1})$. The asymptotic nature relates to $c_1 \to 1$ as the discretization is refined. This completes the proof. □

Remark 3. *The Lipschitz stability assumption on $\mathcal{A}$ arises naturally in many PDE problems with wellposedness guarantees, where small perturbations in boundary conditions or source terms lead to proportionally bounded changes in the solution field. The theorem then assures that, for stable PDE operators and suitably refined discretization, the functional conformal predictor $\Gamma_\alpha(f_{n+1})$ offers robust finite-sample coverage in the infinite-dimensional setting.*

Theorem 4 (Convergence to the Continuous L^2 Norm)**.** *Let f be a Riemann integrable function on a bounded domain $\Omega \subset \mathbb{R}^2$. Let P_d be a sequence of partitions of Ω into d cells, indexed by $d \in \mathbb{N}$, such that the norm of the partition, $\|P_d\| = \max_{i=1,\dots,d}(diam(C_i))$, tends to zero as $d \to \infty$. Let $\|f\|^2_{w,2,d}$ be the squared quadrature-weighted norm computed on the partition P_d. Then, the sequence of discrete norms converges to the squared continuous L^2 norm:*

$$\lim_{d\to\infty} \|f\|^2_{w,2,d} = \|f\|^2_{L^2(\Omega)} \tag{27}$$

Proof. We establish the equivalence in four steps, following the definition of the Riemann integral.

Step 1: Formal Definitions. The squared continuous L^2 norm of a function $f : \Omega \to \mathbb{R}$ is given by the definite integral:

$$\|f\|^2_{L^2(\Omega)} = \int_\Omega f(x)^2 \, dx \tag{28}$$

A partition P_d of the domain Ω consists of a set of d non-overlapping cells $\{C_1, C_2, \dots, C_d\}$ such that $\cup_{i=1}^d C_i = \Omega$. For each cell C_i, we denote its area by $w_i = \text{Area}(C_i)$ and choose a sample point $x_i^* \in C_i$. The squared quadrature-weighted discrete norm is defined on this partition as:

$$\|f\|^2_{w,2,d} = \sum_{i=1}^d w_i f(x_i^*)^2 \tag{29}$$

Step 2: Identification as a Riemann Sum. Let the function $g(x) = f(x)^2$. Since f is Riemann integrable, g is also Riemann integrable on Ω. The expression for the squared quadrature-weighted norm,

$$\sum_{i=1}^d f(x_i^*)^2 w_i = \sum_{i=1}^d g(x_i^*) w_i \tag{30}$$

is precisely the definition of a Riemann sum for the function $g(x)$ over the domain Ω with respect to the partition P_d and the sample points $\{x_i^*\}$.

Step 3: Convergence via the Definition of the Riemann Integral. The definite integral of a function g over a domain Ω is defined as the limit of its Riemann sums as the norm of the partition (i.e., the maximum diameter of any cell in the partition) approaches zero. Formally,

$$\int_\Omega g(x) \, dx = \lim_{\|P_d\|\to 0} \sum_{i=1}^d g(x_i^*) w_i \tag{31}$$

This limit exists and is independent of the choice of sample points x_i^* because g is Riemann integrable. As we refine our grid such that $d \to \infty$ and $\|P_d\| \to 0$, our discrete norm calculation becomes an increasingly accurate approximation of this integral.

Step 4: Conclusion of Equivalence. By substituting $g(x) = f(x)^2$ and applying the definition of the Riemann integral from Step 3, we directly connect the limit of the discrete norm to the continuous norm:

$$\lim_{d\to\infty} \|f\|^2_{w,2,d} = \lim_{\|P_d\|\to 0} \sum_{i=1}^d f(x_i^*)^2 w_i = \int_\Omega f(x)^2 \, dx = \|f\|^2_{L^2(\Omega)} \tag{32}$$

This convergence proves that the quadrature-weighted discrete norm is fundamentally equivalent to the continuous L^2 norm in the limit. This property justifies its use as a reliable and geometrically sound nonconformity score for function-space conformal prediction. □

E PDEs & Data Generation

E.1 Darcy Flow

For the first case we consider a one-dimensional variant of Darcy flow governed by the steady-state elliptic problem

$$-\frac{d}{dx}\left(k(x)\frac{du}{dx}\right) = 0 \quad \text{on } x \in [0,1] \tag{33}$$

subject to Dirichlet boundary conditions $u(0) = 0$ and $u(1) = 1$. The coefficient $k(x)$ represents a permeability field that varies in space. In order to emulate heterogeneous media, we draw $k(x)$ from a random field with mild smoothness, ensuring that it remains strictly positive. The essential objective is to approximate the mapping $k(\cdot) \mapsto u(\cdot)$ and then provide set-valued predictions $\Gamma_\alpha(\cdot)$ with a guaranteed coverage level $1 - \alpha$.

For data generation, we discretize the domain $[0, 1]$ into 1024 uniform points and construct each random permeability $k(x)$ by summing a few random Fourier modes. This yields fields with different oscillatory patterns, constrained so that $k(x)$ remains between 0.01 and 10. For each sample, we solve the linear system that arises from the finite-difference approximation of the Darcy equation, thereby obtaining a ground-truth solution $u(x)$. We generated 100,000 such permeability-solution pairs and subdivided them into training, calibration, and test sets in the ratio $80\%, 10\%, 10\%$.

E.2 Poisson Equation

In our second case we consider a two-dimensional Poisson problem, in which the governing equation is

$$-\nabla \cdot (\nabla u(x, y)) = f(x, y), \quad (x, y) \in [0, 1]^2, \tag{34}$$

subject to Dirichlet boundary conditions $u(x, y) = 0$ on $\partial[0, 1]^2$. The function $f(x, y)$ represents a spatially varying source term that influences the solution $u(x, y)$. We generate a family of such problems by sampling f from a random field with bounded support, then numerically solving for the corresponding solution u. Our goal is to learn the mapping $f(\cdot, \cdot) \mapsto u(\cdot, \cdot)$ and and again provided set-valued predictions $\Gamma_\alpha(\cdot)$.

To generate synthetic examples, we discretize the domain $[0, 1]^2$ using three types of grids: (i) a standard uniform $N \times N$ grid with evenly spaced coordinates, (ii) a non-uniform grid obtained by applying a cubic transformation to a uniform reference grid to cluster points toward the center, and (iii) a grid generated via a sine transformation to concentrate points near the domain boundaries. These coordinate mappings introduce structured resolution variation while preserving geometric continuity for the finite difference solver. We construct random forcing fields f by summing a small number of random Fourier modes. For each forcing sample, we solve the equation $\Delta u = f$ with boundary condition $u = 0$ using either a finite-difference or Jacobi iterative method, yielding the ground-truth solution u. This process yields a dataset $\{(f_i, u_i)\}$, which we partition into 5,000 training, 1,000 calibration, and 1,000 test instances.

E.3 Unsteady Navier–Stokes Dynamics

Our final case study considers the two-dimensional incompressible Navier–Stokes equations in vorticity form:

$$\partial_t \omega + \mathbf{u} \cdot \nabla \omega = \nu \Delta \omega + f, \quad (x, y) \in [0, 1]^2, \ t \in [0, T], \tag{35}$$

where $\omega(x, y, t)$ denotes the vorticity, $\nu > 0$ is the kinematic viscosity, and $f(x, y)$ is a stationary external forcing term. The velocity field $\mathbf{u}$ is recovered from ω via the stream function ψ, solving $\Delta \psi = \omega$, followed by $\mathbf{u} = (-\partial_y \psi, \partial_x \psi)$. We impose periodic boundary conditions in both spatial dimensions.

We generate the initial vorticity field $\omega_0(x, y)$ by sampling from a two-dimensional Gaussian random field with spectral decay. The external forcing $f(x, y)$ is chosen as a fixed sinusoidal function. We simulate the evolution of $\omega(x, y, t)$ using a pseudo-spectral method with dealiasing and implicit treatment of the diffusion term. The solver records the vorticity field at uniformly spaced time steps, producing a spatio-temporal trajectory $\omega(\cdot, \cdot, t)$ over a time horizon $T = 50$ with 200 snapshots. We then generate 1,200 initial conditions and their corresponding time evolutions on a 64×64 spatial grid. Next we split this into 1,000 training, 100 calibration, and 100 test instances. Because we don't provide a qualitative evaluation of this case in the paper, we refer the reader to Figure 3 for a visualization.

F Discussion on Resolution Adjustment

As shown in Figure 5, only two dominant patterns are observed in the Poisson data—excluding the lower resolutions, which are dominated by $\varepsilon_{\text{disc}}$. In the Darcy case, a similar trend emerges, albeit

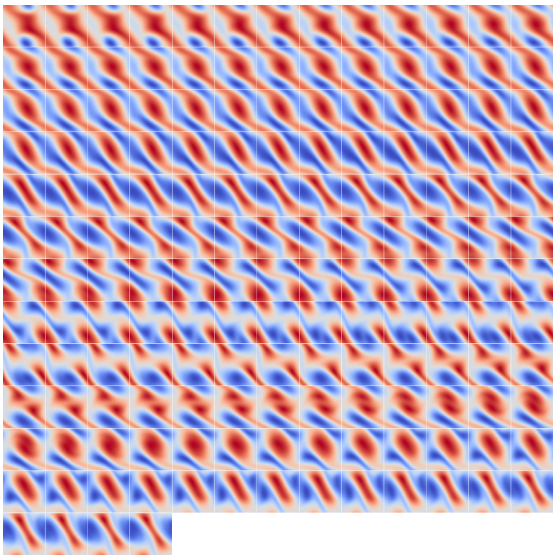


Figure 3: Unsteady Navier–Stokes Dynamics. Visualized is a single predicted trajectory from the evaluation set.

with a slightly weaker spike. Thus, our regression approach proves effective for calibration transport across resolutions, especially for super-resolution tasks. In some cases, particularly for resolutions

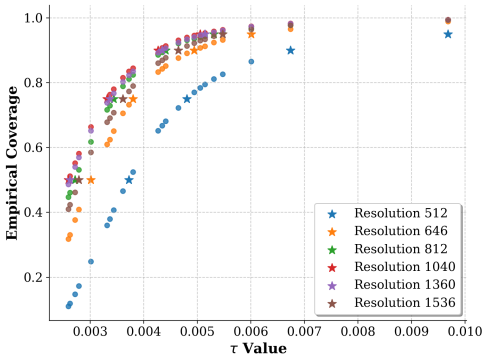


Figure 4: Coverage versus τ_α, grouped by resolution. We mark each resolution's own calibrated τ_α/coverage with a colored star; circles represent τ_α values evaluated at non-native resolutions.

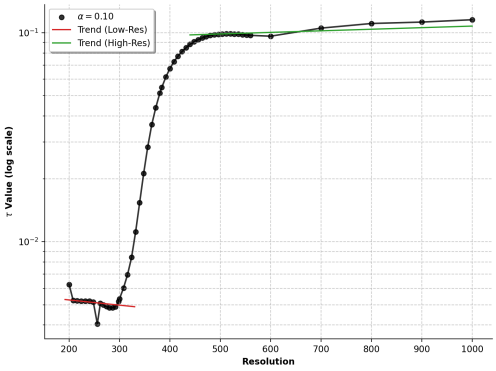


Figure 5: We plot the empirical distribution of τ_α values across varying resolutions for the Poisson case study. We propose two small regressions to transport τ_α within the most consistent reigns.

close to the calibration resolution, this transportation adjustment is not as necessary. Figure 4 shows the inter-evaluation of τ_α values across a small subset of resolutions within the Darcy flow case. As we increase the resolution locally, the drop in significance—while still present—remains relatively minor compared to larger resolution shifts. For example, consider the τ_α value produced at resolution 1,040 for a significance level of $\alpha = 0.1$. As the resolution increases to 1,536, the coverage drops only slightly to 85%. In contrast, when the resolution decreases by a similar amount to 512, the coverage drops more substantially to 65%.s

G Experimental Setup

All experiments were conducted on SPORC (Scheduled Processing On Research Computing), Rochester Institute of Technology's High Performance Computing (HPC) cluster. For all training,

evaluation, and calibration tasks, we used a single NVIDIA A100 GPU and two CPU cores. This configuration was sufficient for all neural operator models and associated conformal calibration routines. No distributed training or multi-GPU setups were required. Job submissions were managed through Slurm, and all experiments used fixed random seeds for reproducibility. We provide a summary of each methods computational budget in Table 6.

Table 6: Computational Summary. Reported losses are dependent on training procedure (e.g., variational methods include the KL loss.) Each result is produced from an average of 5 runs.

Model	Epochs	Dataset Size	# Parameters	Loss	Training Time
Darcy 1D					
MC Dropout	700	20,000	454K	0.000006	0:00:34
Variational	700	20,000	480K	0.000065	0:00:34
Poisson 2D					
Triplet	500	5,000	19.7M	0.005017	0:01:02
Quantile	500	5,000	19.7M	0.005783	0:01:01
Navier-Stokes 2D					
Deterministic	100	1,200	226.9M	0.000004	3:18:28
Variational	100	1,200	227.0M	0.000116	2:18:28

H SUPER RESOLUTION

Figures 6 and 7 present qualitative visualizations of individual test instances evaluated far outside the training and calibration resolution. In both cases, the displayed instance corresponds to the resolution level to which the conformal calibration threshold τ_α was transported using the empirical regression procedure described in Section 6.4. These examples illustrate the structure of the predicted field, the associated conformal bounds, and the alignment between model uncertainty and true error, despite a significant resolution shift relative to the calibration set. The ability to apply calibrated uncertainty estimates under such resolution mismatch is critical for high-fidelity scientific applications and demonstrates the practical effectiveness of our resolution-agnostic calibration framework.

(a) Prediction vs Ground Truth (b) Prediction Interval (c) Std. vs Prediction Error

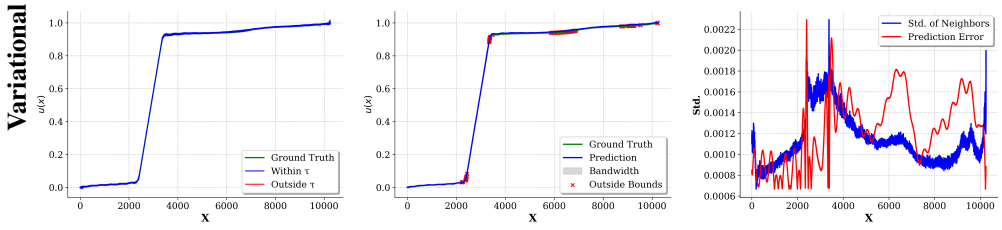
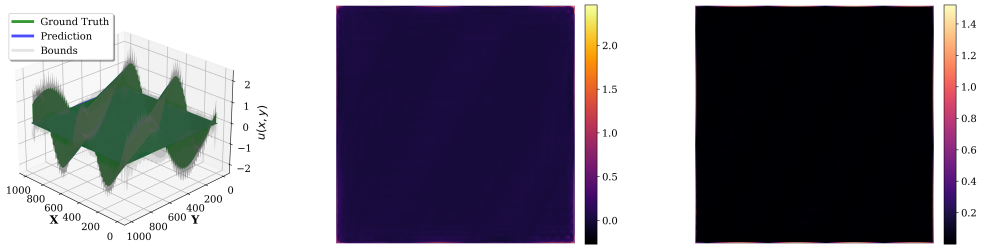


Figure 6: Visualization of a Darcy flow test instance evaluated at a resolution significantly higher than that of the training and calibration data. This specific instance corresponds to the resolution level used for transporting the conformal calibration scalar τ_α via the regression procedure. Subfigures show (a) the predicted solution alongside the ground truth, (b) the conformal prediction interval derived from sampled trajectories, and (c) the local standard deviation compared to the pointwise prediction error.



(a) Prediction vs Ground Truth (b) Std. of Neighbors (c) Prediction Error

Figure 7: Visualization of a Poisson equation test instance at a super-resolved grid. This instance was used as the evaluation target for applying the transported conformal threshold τ_α. Shown are (a) the predicted solution versus the ground truth, (b) the local ensemble spread, and (c) the absolute prediction error. These visualizations demonstrate the model's behavior under resolution shift beyond the training range.

I GRID GEOMETRIES

Let $[0,1]^2$ denote the spatial domain, and let n_x, n_y denote the number of discretization points along each axis. We construct grid coordinates by first defining a parametric variable $t \in [-1,1]$ sampled uniformly with n_x and n_y points in the x- and y-directions, respectively. The final grid coordinates (x_i, y_j) are obtained via one of the following continuous mappings $T : [-1,1] \to [0,1]$ applied elementwise:

$$\text{Uniform:} \quad T_{\text{uniform}}(t) = \tfrac{1}{2}(t+1), \tag{36}$$

$$\text{Clustered Center:} \quad T_{\text{center}}(t) = \tfrac{1}{2}(t^3+1), \tag{37}$$

$$\text{Clustered Boundary:} \quad T_{\text{boundary}}(t) = \tfrac{1}{2}(\sin(\tfrac{\pi}{2}t)+1). \tag{38}$$

These mappings define the spatial concentration of grid points. The uniform mapping yields evenly spaced coordinates across the domain. The cubic mapping clusters points toward the domain center, due to its vanishing first derivative at $t = \pm 1$ and maximal slope at $t = 0$. Conversely, the sinusoidal mapping clusters points near the boundaries, as its derivative vanishes at $t = 0$ and grows toward $t = \pm 1$. In each case, the transformation ensures that grid coordinates remain in $[0,1]$ and preserve the geometric continuity of the domain. These geometries enable controlled evaluation of calibration robustness under structured discretization shift. We visualize each method in Figure 8.

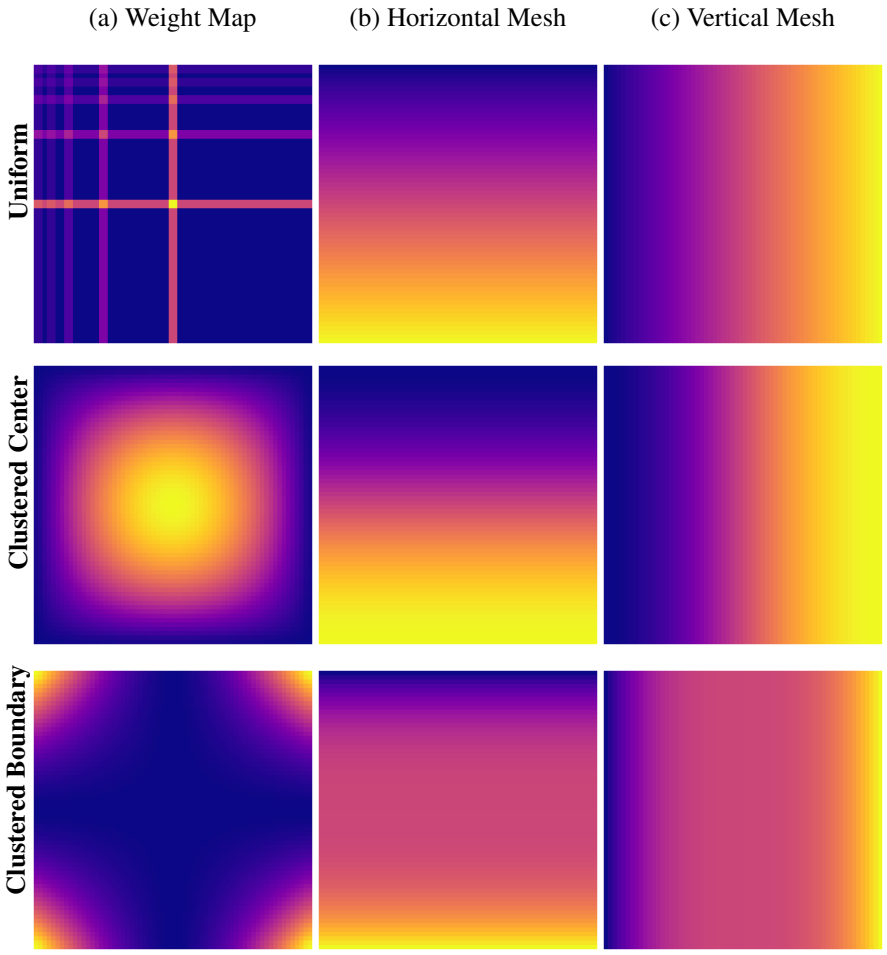


Figure 8: Visualization of the three grid geometries used in our experiments. Each row corresponds to a different grid mapping: uniform (top), clustered toward the center (middle), and clustered toward the boundary (bottom). Columns show the numerical quadrature weight map (left), horizontal coordinate mesh x (center), and vertical coordinate mesh y (right). The visible banding in the uniform case is a result of floating-point rounding errors and does not meaningfully affect the calibration procedure.