Flow-Based Conformal Predictive Distributions

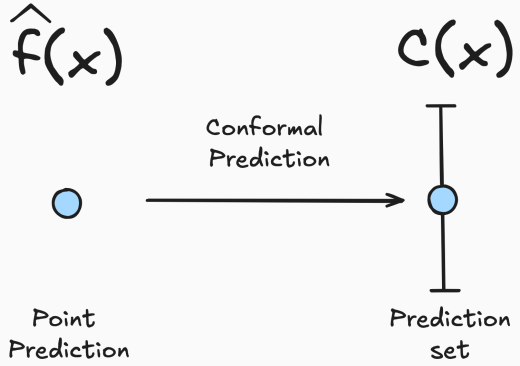
Trevor Harris

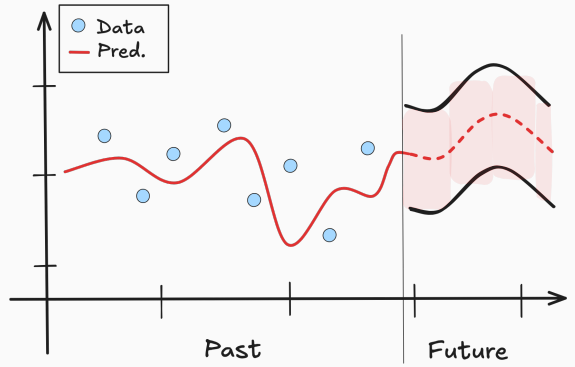
May 12, 2026

University of Connecticut
Department of Statistics

Carnegie Mellon University – Pittsburgh, PA

Conformal prediction and prediction sets



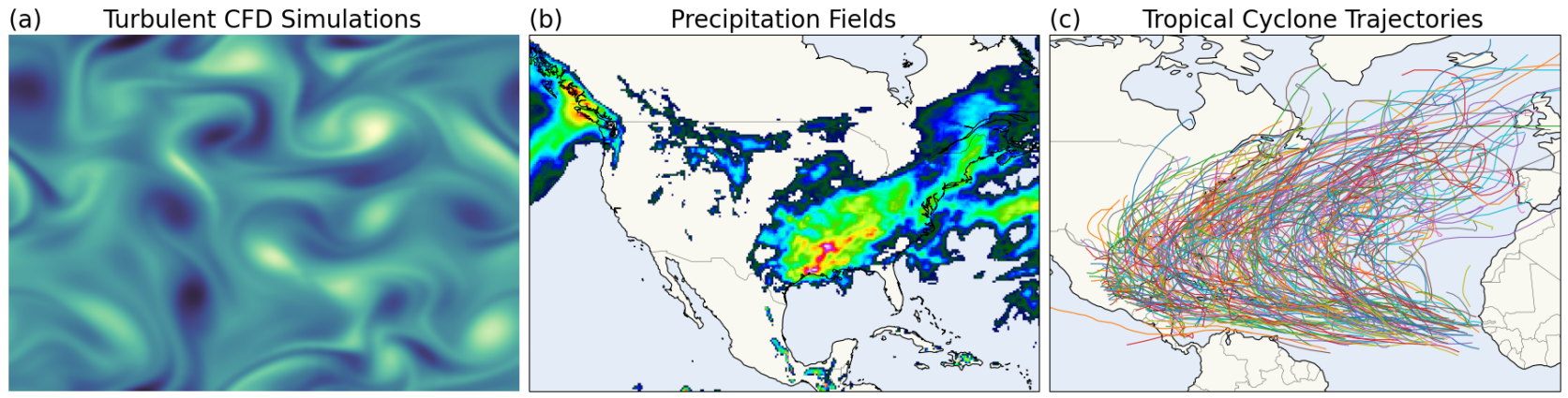


- **Conformal prediction** [Vovk et al. (2005)]: post-hoc, finite-sample valid prediction sets for *any* pretrained model $f_{\hat{\theta}} : \mathcal{X} \to \mathcal{Y}$
  1. Define a **nonconformity score** measuring prediction error, e.g.

$$S(x, y) = \|f_{\hat{\theta}}(x) - y\|_2 \qquad \text{req. exchangeable}$$

  2. Score calibration data: $S_i = S(x_i, y_i)$ for $(x_i, y_i) \in \mathcal{D}_{\text{cal}}$
  3. Threshold: $\tau_\alpha = S_{(k)}, \quad k = \lceil (1-\alpha)(n+1) \rceil$
  4. Prediction set: $C_\alpha(x) = \{y : S(x, y) \leq \tau_\alpha\}$
- $P(y_{n+1} \in C_\alpha(x_{n+1})) \geq 1 - \alpha$ — no asymptotics, no priors, no retraining
- **Problem:** $C_\alpha(x)$ is defined *implicitly* via the score

In structured spaces, prediction sets are inert



- **Predictive sets** $C_\alpha(x) = \{y : S(x, y) \leq \tau_\alpha\}$ — well-defined, but how do you apply it? Represent it?
- **Predictive distribution** — sample, forecast, simulate, or estimate risk
- Prior approaches: conformal predictive systems (*univariate only*) [Vovk et al. (2017)], generative models (*heavy, adds structure*) [Zheng and Zhu (2024)], OT maps (*scales poorly*) [Ndiaye (2025)]

$\Rightarrow$ **Goal:** Lift prediction sets into predictive distributions — high dim, scalably, without training.

Differentiable scores induce flows

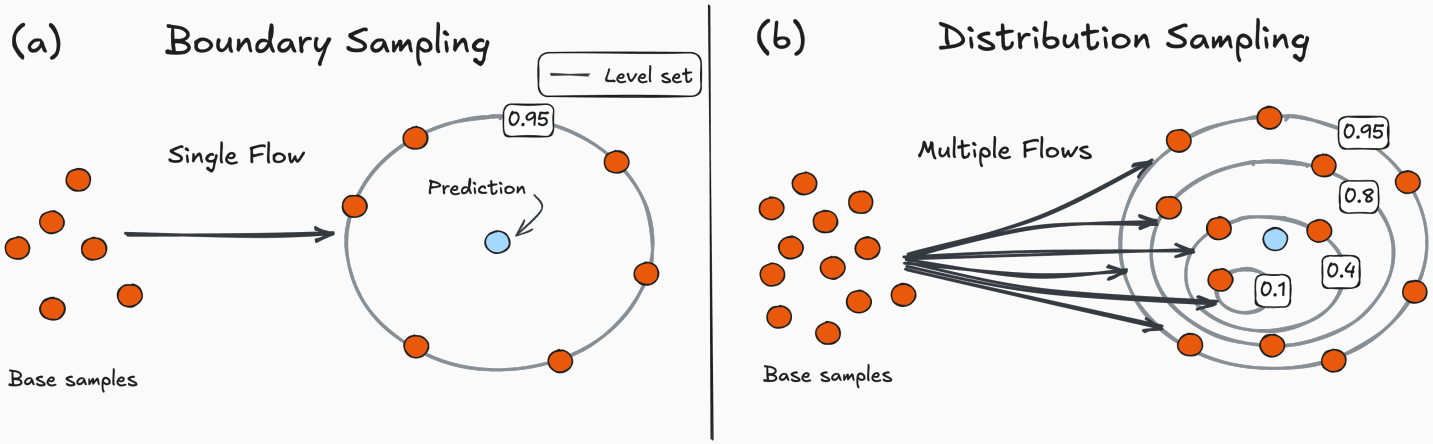


Figure 1: Score + confidence level determines a flow (a). Score + sequence of levels determines a sequence of flows (b).

- Any differentiable $S(x, y)$ gives a vector field $v(y) = \nabla_y S$ on $\mathcal{Y}$
- A vector field defines a **flow** Φ:

$$y'(t) = v(y(t)) \quad \Rightarrow \quad \Phi(t, y_0) := y(t) = y_0 + \int_0^t v(y(s))\, ds$$

- Design $v_\alpha(y)$ so $\partial C_\alpha(x) = \{y : S(x, y) = \tau_\alpha\}$ is an **attractor**
  - Flow at fixed $\alpha \Rightarrow$ boundary samples (a) — usable uncertainty objects
  - Mix flows over $\alpha \Rightarrow$ calibrated predictive distributions (b)

Nonconformity flows

- Fix $x \in \mathcal{X}$, $\alpha \in (0,1)$. Write $S(x,y) = S(y)$. Construct a flow via two coupled ODEs:

$$\underbrace{y'(t) = v(y(t))}_{\text{state evolution}} \qquad \underbrace{S'(y(t)) = -\lambda\big(S(y(t)) - \tau_\alpha\big)}_{\text{score controller: exponential decay to } \tau_\alpha}$$

- Chain rule links them: $S'(y(t)) = \nabla S^\top v$. Minimum-norm solution:

$$v_\alpha(y) = -\lambda\big(S(y) - \tau_\alpha\big)\frac{\nabla S(y)}{\|\nabla S(y)\|_2^2} \quad \Rightarrow \quad \Phi_\alpha(t, y_0) = y_0 + \int_0^t v_\alpha(y(s))\, ds$$

- Properties:
  - Analytic velocity: no training, estimation, no OT map
  - Only known quantities: S, ∇S (autodiff), τ_α (calibration)
  - Moves along ∇S — steepest path to level set; no tangential motion
  - $v_\alpha(y) = 0$ when $S(y) = \tau_\alpha$ — stops at boundary

Convergence to the conformal boundary

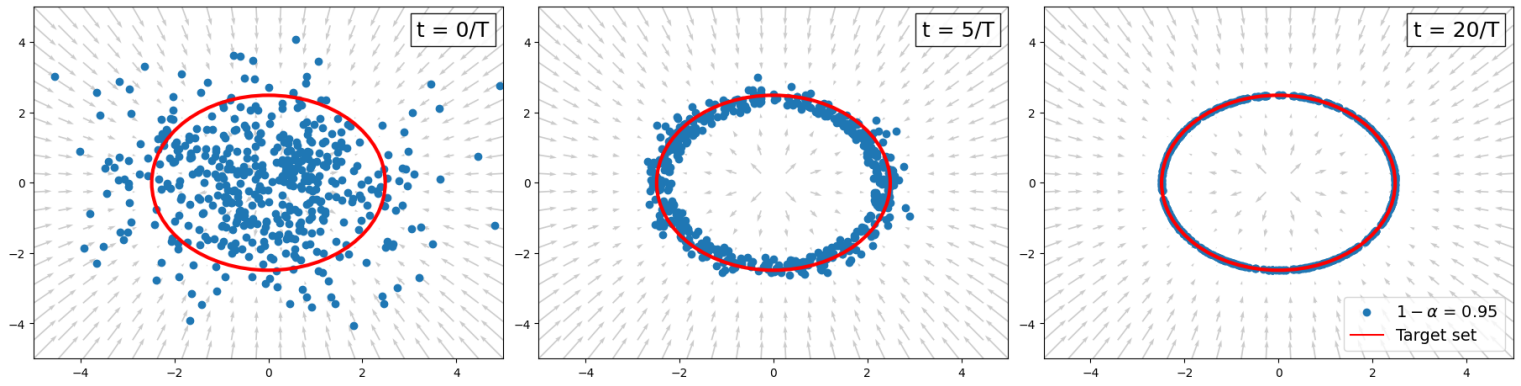


Figure 2: Random base samples y_0 flow to the $1 - \alpha = 0.95$ level set boundary.

- **Proposition (Convergence).** Def. score error $\varepsilon(t) := S(y(t)) - \tau_\alpha$.
  1. *Score convergence:* $\varepsilon'(t) = -\lambda\varepsilon(t) \;\Rightarrow\; \varepsilon(t) = \varepsilon(0)\, e^{-\lambda t}$
  2. *Pointwise convergence:* if $\|\nabla S(y)\|_2 \geq m > 0$ near $\partial C_\alpha(x)$, then

$$\|y(t) - y_\infty\|_2 \leq \frac{1}{m} |S(y_0) - \tau_\alpha| e^{-\lambda t}$$

- **Corollary (ε-hitting time).** Set $\lambda = \log(|S(y_0) - \tau_\alpha|/\varepsilon)$ to guarantee $|S(y(t)) - \tau_\alpha| \leq \varepsilon$ by $t = 1$. In practice: 5–20 Euler steps.

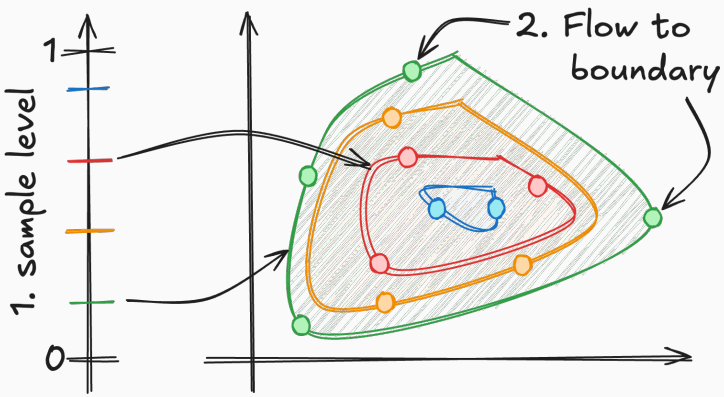
From boundaries to distributions

- We can sample $\partial C_\alpha(x)$ for any α — can we assemble these into a full predictive distribution? Yes, randomize over α.

Algorithm:

1. Sample $\alpha \sim \pi$ on $(0, 1)$
2. Sample $y_0 \sim \mu_x$, flow to $\partial C_\alpha(x)$

$$y = \Phi_\alpha(\infty, y_0)$$



- Conformal sets are nested: If $\alpha_1 \geq \alpha_2$ then $C_{\alpha_1}(x) \subseteq C_{\alpha_2}(x)$
  - Boundaries form quantile contours
  - Like level sets of a CDF, but in arbitrary dimension
- Infinitely many distributions are consistent with $\{C_\alpha(x)\}_\alpha$
- Define one with: μ_x (base measure), π (mixing measure), and Φ_α (flow)

Conformal Predictive Distributions (CPDs)

- **Definition (CPD):** Given mixing measure π on $(0,1)$ and boundary measures $\nu_{x,\alpha}$ on $\partial C_\alpha(x)$:

$$P_x^{\mathrm{CPD}}(A) = \int_0^1 \nu_{x,\alpha}(A)\, d\pi(\alpha)$$

- **Proposition (Calibration):** For any $\{\nu_{x,\alpha}\}$, on the empirical conformal sets:

$$P_x^{\mathrm{CPD}}\big(C_\alpha(x)\big) = \pi\big([\alpha,1)\big).$$

If $\pi = \mathrm{Unif}(0,1)$: $P_x^{\mathrm{CPD}}(C_\alpha(x)) = 1-\alpha$

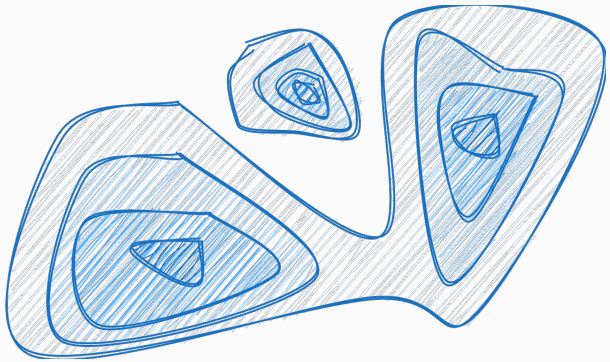


Figure 3: CPDs are mixtures over conformal level sets

- **CPD sampling:** (1) Sample $\alpha \sim \pi$, (2) Sample $y_0 \sim \mu_x$, (3) Flow $y = \Phi_\alpha(\infty, y_0)$
  - π controls calibration; $\nu_{x,\alpha} = \Phi_\alpha \# \mu_x$ controls predictive skill
  - Non-uniform π: oversample tails (π near 0) or center (π near 1)

Predictive distribution approximation

- CPDs are calibrated for *any* boundary measure, but how close is P_x^{CPD} to the true conditional $P_x^\star = P(\cdot \mid x)$?
- **Proposition (W_2 approximation bound).** Let P_x^{proj} be the dist. of projecting $Y \sim P_x^\star$ onto $\partial C_\alpha(x)$ over $\alpha \sim \mathrm{Unif}(0,1)$.

$$
\begin{aligned}
W_2(P_x^\star,\, P_x^{\mathrm{CPD}}) \;\le\; & W_2(P_x^\star,\, P_x^{\mathrm{proj}}) && \text{(i) score distortion} \\
& + \; (\mathbb{E}_\alpha L_\alpha^2)^{1/2}\, W_2(P_x^\star,\, \mu_x) && \text{(ii) base measure quality} \\
& + \; C\,(\mathbb{E}_\alpha[\bar{\kappa}(\alpha)^2 R_4(\alpha)])^{1/2} && \text{(iii) flow distortion}
\end{aligned}
$$

- Three components of forecast quality:
  1. **Score distortion**: small when level sets of S align with $P_x^\star$
  2. **Base measure**: small when $\mu_x \approx P_x^\star$ — use calibration responses
  3. **Flow distortion**: small when score has low curvature (straight paths)
- Prediction error low when $\mu_x \approx P_x^\star$ and $S(\cdot, y) \approx -\log p_x^\star(y)$

Flows converge across all settings

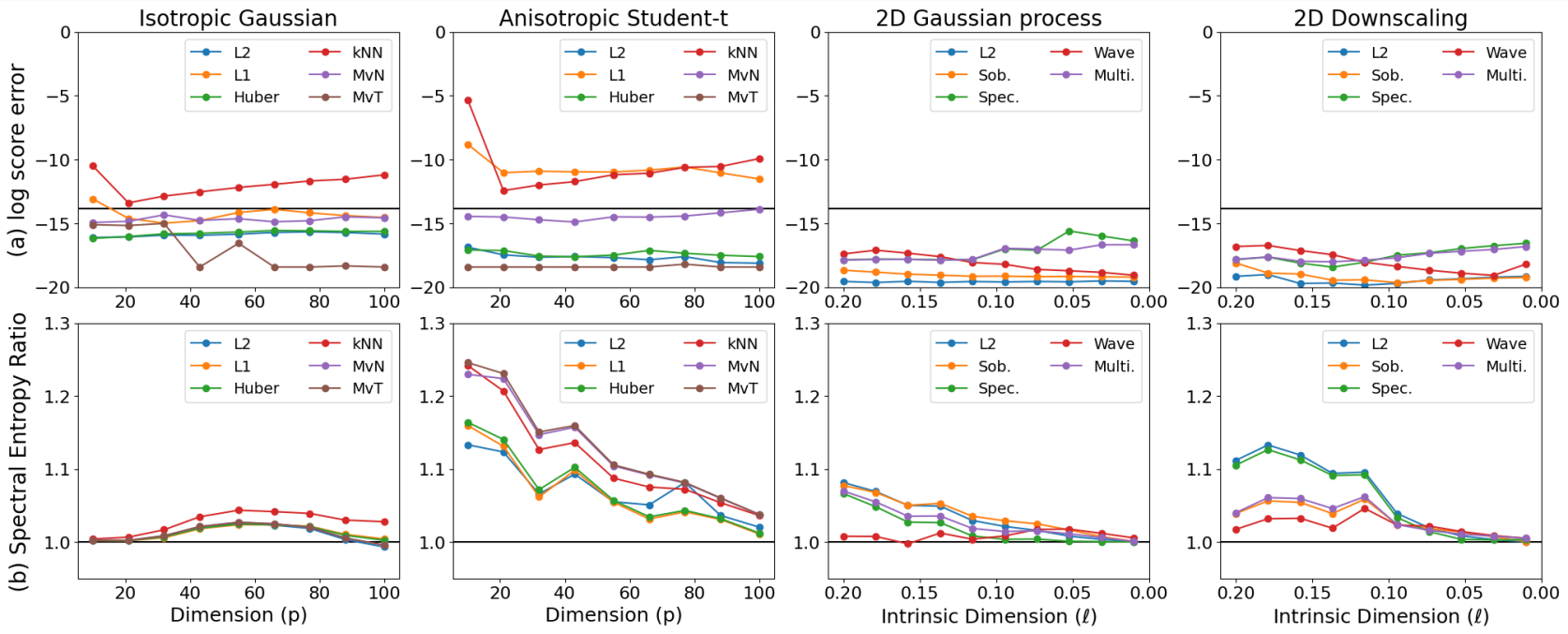


Figure 4: Flow convergence and diversity on 4 regression tasks. Black line = target

- **Top — Convergence:** $\searrow$ across all DGPs (iid, correlated, anisotropic) and scores (ℓ_2, Huber, Sobolev, spectral, . . .). Does not depend on dim.
- **Bottom— Diversity:** spectral entropy ratio ≥ 1 — boundary samples are as diverse as actual data; no subspace collapse even in high dim.

CPDs match or beat deep probabilistic baselines

Table 1: CPDs vs baselines across 5 structured regression tasks. **Bold** = best (↓).

	GP Regression			Elliptic PDE Inv.			Navier Stokes			Precip. Downscale			Climate Debias		
Method	ED	LSD	MMD	ED	LSD	MMD	ED	LSD	MMD	ED	LSD	MMD	ED	LSD	MMD
CPD-G	**0.313**	**0.069**	**0.023**	0.068	0.103	**0.006**	0.233	0.252	0.043	0.395	0.067	0.019	0.107	0.003	**0.024**
CPD-L	**0.313**	**0.069**	0.036	**0.067**	**0.099**	**0.006**	**0.229**	**0.166**	**0.039**	0.400	0.066	**0.016**	0.133	0.006	0.032
Drop.	0.400	0.083	0.187	0.115	0.263	0.158	0.295	0.245	0.142	0.418	0.072	0.075	0.127	0.004	0.211
D. Ens.	0.346	0.089	0.064	0.118	0.234	0.044	0.269	0.193	0.055	0.399	**0.059**	0.138	0.119	**0.002**	0.109
IQN	0.427	0.075	0.061	0.139	0.262	0.072	0.263	0.381	0.181	0.393	0.066	0.148	0.116	**0.002**	0.121
Flow	0.410	0.138	0.167	0.113	0.351	0.040	0.235	0.182	0.048	0.374	0.065	0.093	0.118	0.003	0.111
MDN	0.367	0.213	0.211	0.141	0.457	0.169	0.236	0.677	0.171	**0.361**	0.090	0.114	**0.105**	**0.002**	0.195
MVE	0.374	0.200	0.211	0.203	6.112	0.198	0.236	0.651	0.171	0.722	0.097	0.117	0.224	0.018	0.266

- **MMD:** CPDs best on all 5 tasks, often by 3–10× — fine-scale structure preserved
- **LSD/ED:** CPDs best (3/5) or competitive on all tasks
- Baselines approximate marginals but miss spatial structure and local detail in many cases.

CPDs produce realistic, diverse samples

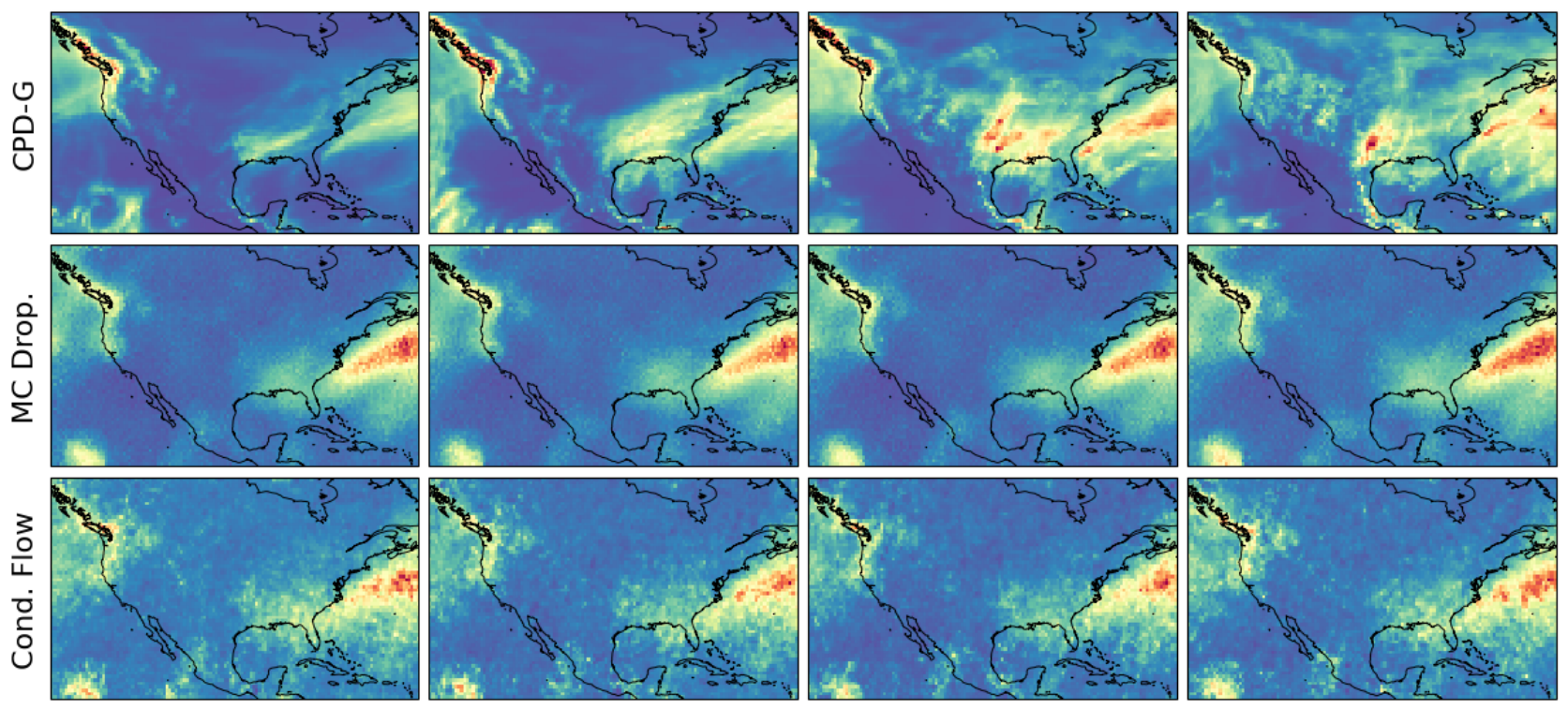
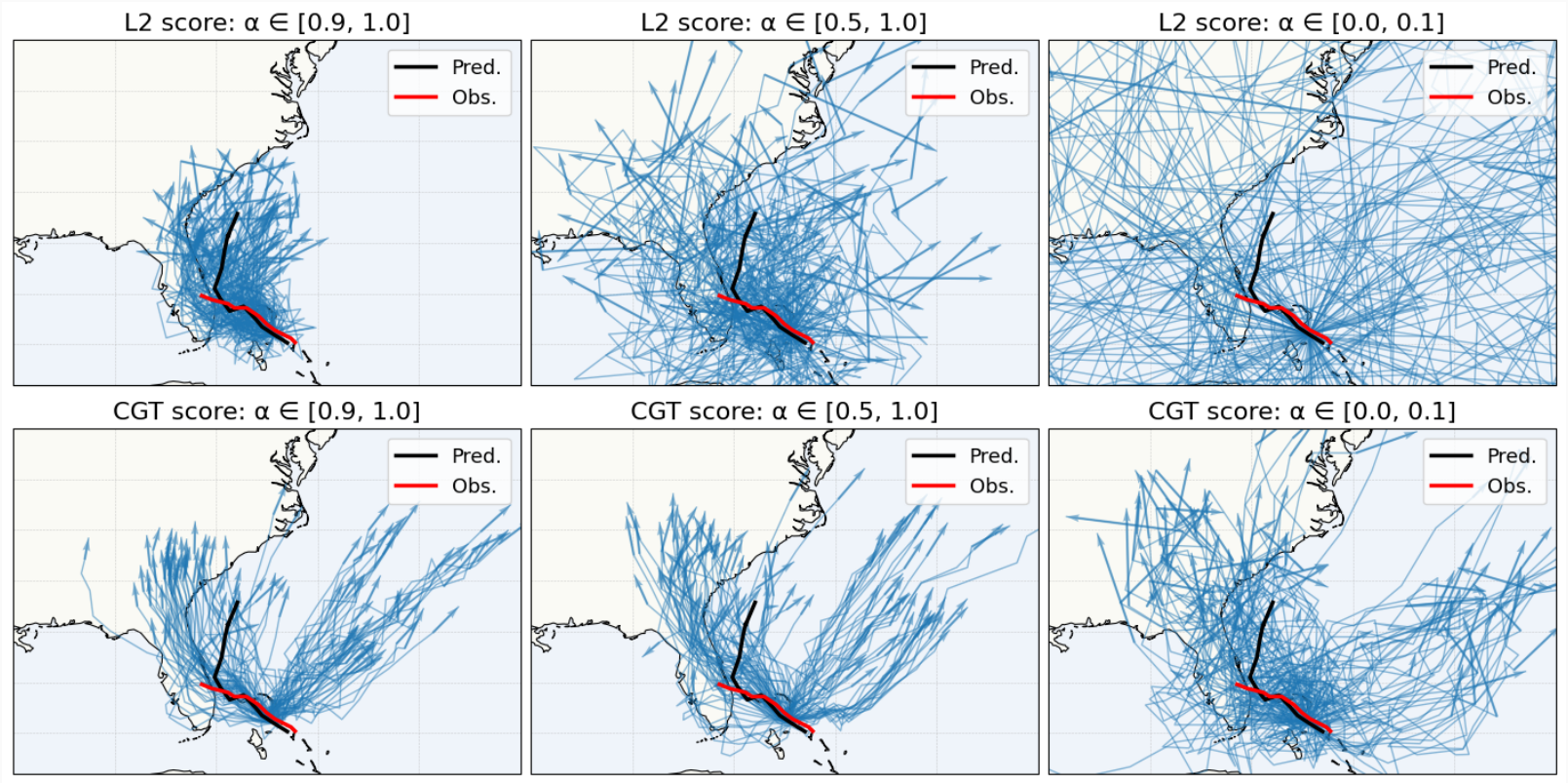


Figure 5: Sample precip. intensity from CPD-G, MC Dropout, and Flow Matching

- **CPD-G:** qualitatively different samples — realistic spatial variation
- **MC Dropout:** near-identical samples — mode collapse, excessive drizzle
- **Flow:** more variation than Dropout, oversmoothes fine structure

$\Rightarrow$ Low LSD/MMD for CPDs reflects structural diversity, not noise

Tropical cyclone forecasting — controllable generation



- **Columns:** varying π from high-confidence ($\alpha \in [0.9, 1]$) to extreme ($\alpha \in [0, 0.1]$) widens trajectory spread — controllable generation
- **Rows:** ℓ_2 score produces geometrically incoherent paths; CGT score (velocity, curvature, length) yields physically plausible trajectories
- Behavior emerges from the score, not the model. Scores matter!

Discussion

- **Problem:** conformal prediction sets are implicit and inert in high dim
  $\Rightarrow$ Cannot forecast, simulate, or estimate risk
- **Solution:** nonconformity flows
  - Any differentiable score $\Rightarrow$ deterministic flow to the conformal boundary
  - Training-free, scalable, dimension-free convergence
- **Result:** conformal predictive distributions
  - Calibrated by construction
  - Competitive with deep probabilistic baselines
  - Controllable: target any region of the distribution
- **Open questions:**
  - Full boundary exploration via tangential diffusion
  - Coupling with learned generative models
  - Non-exchangeable settings (temporal data)

ARXIV

GITHUB

Appendix

Velocity field derivation

- Chain rule expands the score controller into a linear constraint on v:

$$\nabla S(y)^\top v(y) = -\lambda(S(y) - \tau_\alpha)$$

- Choose the minimum-norm v satisfying this constraint:

$$v_\alpha(y) \in \arg\min_v \ \tfrac{1}{2}\|v\|_2^2 \quad \text{s.t.} \quad \nabla S(y)^\top v = -\lambda(S(y) - \tau_\alpha)$$

- Single-constraint QP $\Rightarrow$ Lagrangian:

$$\mathcal{L}(v, \eta) = \tfrac{1}{2}\|v\|_2^2 + \eta\big(\nabla S^\top v + \lambda(S(y) - \tau_\alpha)\big)$$

- First-order condition: $v = -\eta\,\nabla S(y)$. Enforce the constraint:

$$-\eta\|\nabla S(y)\|_2^2 = -\lambda(S(y) - \tau_\alpha) \quad \Rightarrow \quad \eta = \lambda\frac{S(y) - \tau_\alpha}{\|\nabla S(y)\|_2^2}$$

- Substitute back:

$$v_\alpha(y) = -\lambda(S(y) - \tau_\alpha)\frac{\nabla S(y)}{\|\nabla S(y)\|_2^2} \qquad \square$$

Regularity conditions

- **Basic assumptions** (all results):
  - (i) $S : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ with $\nabla_y S(\cdot, y)$ existing a.e.
  - (ii) $\mathcal{Y}$ admits a finite-dimensional differentiable representation
  - (iii) Calibration pairs (x_t, y_t) are exchangeable
- **Convergence** (Proposition 1):
  - (i) Score convergence: $\nabla S(y(t)) \neq 0$ whenever $S(y(t)) \neq \tau_\alpha$
  - (ii) Pointwise convergence: additionally, $\exists\, T < \infty$, $m > 0$ s.t. $\|\nabla S(y(t))\|_2 \geq m$ for $t \geq T$
- W_2 **approximation bound** (Proposition 3) — additionally:
  - (i) $S(x, \cdot) \in C^2$ with bounded gradient and Hessian along flow/projection paths: $0 < c \leq \|\nabla_z S\|_2 \leq C_S < \infty, \quad \|\nabla_z^2 S\|_{\mathrm{op}} \leq H_\alpha < \infty$
  - (ii) $\mathrm{proj}_{x,\alpha}$ onto $\partial C_\alpha(x)$ is unique a.s. and L_α-Lipschitz with $\mathbb{E}_\alpha L_\alpha^2 < \infty$
  - (iii) Curvature control: $\kappa_{\mathrm{proj}}(\alpha) := \mathrm{ess\,sup}_{y_0 \sim \mu_x} \|q_{x,\alpha}(y_0) - \mathrm{proj}_{x,\alpha}(y_0)\|_2 / \mathrm{dist}(y_0, \partial C_\alpha(x))^2 < \infty$
- **Calibration** (Proposition 2): no extra conditions — holds for *any* $\nu_{x,\alpha}$, π

Convergence theory (detail)

- **Score convergence.** Define $\varepsilon(t) := S(y(t)) - \tau_\alpha$. Under the nonconformity flow:

$$\varepsilon'(t) = -\lambda\varepsilon(t) \quad \Rightarrow \quad \varepsilon(t) = \varepsilon(0)\, e^{-\lambda t} \quad \Rightarrow \quad S(y(t)) \to \tau_\alpha$$

  Every accumulation point lies on $\partial C_\alpha(x)$; velocity vanishes at the boundary.

- **Pointwise convergence.** If $\|\nabla S(y)\|_2 \geq m > 0$ near $\partial C_\alpha(x)$, then each trajectory converges to a *unique* limit point $y_\infty \in \partial C_\alpha(x)$:

$$\|y(t) - y_\infty\|_2 \;\leq\; \frac{1}{m}\,|S(y_0) - \tau_\alpha|\, e^{-\lambda t}$$

  Exponentially fast, dimension-free, no oscillations or chaos.

- **ε-hitting time.** $T_\varepsilon(y_0) = \frac{1}{\lambda}\log\frac{|S(y_0)-\tau_\alpha|}{\varepsilon}$. Set $\lambda = \log(|S(y_0) - \tau_\alpha|/\varepsilon)$ for convergence in one unit of time.

The score determines the shape of uncertainty

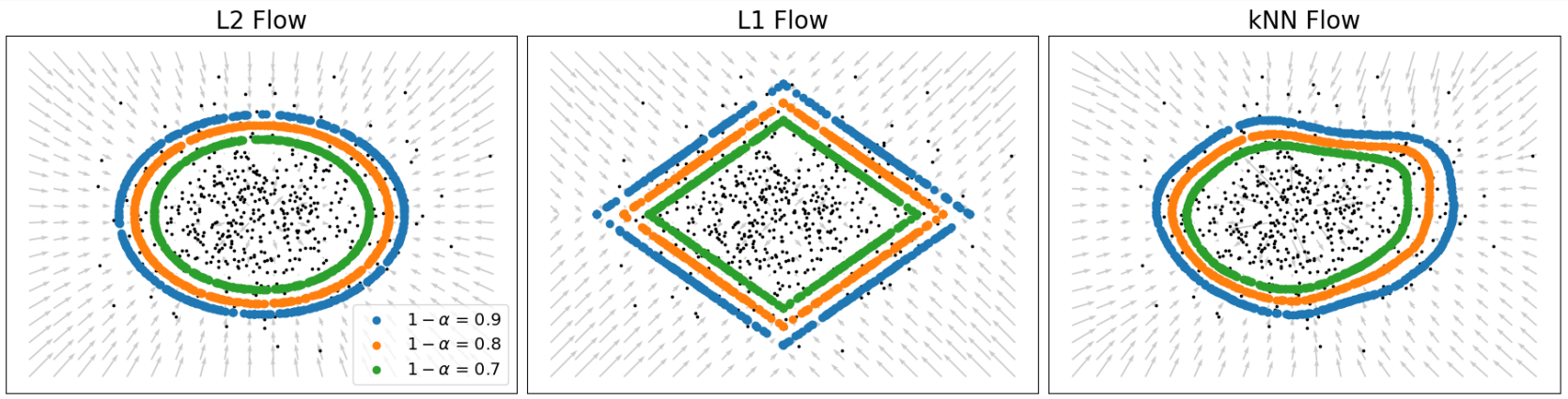


Figure 6: Same data. Different scores result in different flows and prediction sets.

- ℓ_2 vs. ℓ_1 vs. kNN ℓ_2 score.
- Across $\alpha \in \{0.1, 0.2, 0.3\}$ boundaries
- Non-differentiable, non-convex, multi-modal? No* problem.

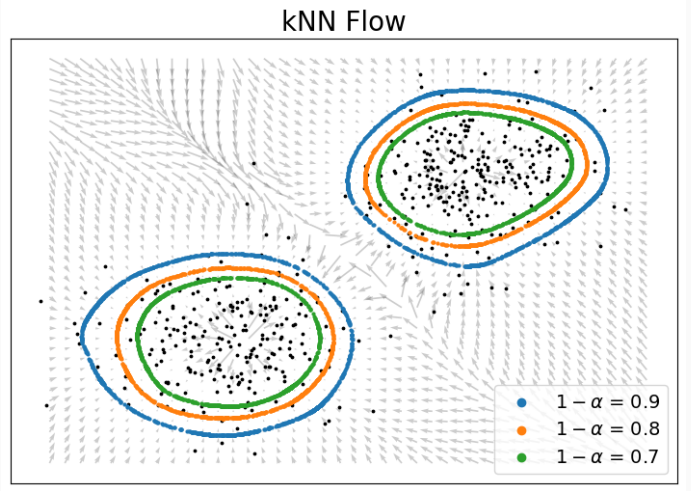


Figure 7: Bi-modal data.

Experimental details

- **Flow convergence tasks:**
  1. Isotropic Gaussian ($p = 10$–100)
  2. Anisotropic Student-t ($p = 10$–100)
  3. 2D GPs (64×64, $\ell = 0.2$–0.01)
  4. 2D GP $8\times$ upscaling
- **Scores (vectors):** ℓ_2, ℓ_1, Huber, kNN, Gaussian NLL, Student-t NLL
- **Scores (operators):** ℓ_2, Sobolev, Spectral, Wavelet, composite

- **CPD quality tasks:**
  1. 2D GP regression (64×64)
  2. Elliptic PDE Inv. (32×32)
  3. Navier-Stokes (64×64)
  4. Precip. downscale (64×128)
  5. Climate debias (64×128)
- **Models:** MLP (vectors), 4-layer 2D FNO (operators) [Li et al. (2020)]
- **Baselines:** MC Dropout, Deep Ens., IQN, Flow, MDN, MVE
- **Metrics:** ED, LSD, Patch MMD

Conditional Geometric Trajectory (CGT) score

- Hurricane setup: HURDAT2 (1850–2025), 6-hr resolution + wind/pressure. Predict next 12 steps from past 24. 1D CNN model.

- Six component terms from discrete derivatives of the trajectory:

$$\Delta y_t = y_{t+1} - y_t, \qquad \Delta^2 y_t = y_{t+2} - 2y_{t+1} + y_t$$

$$t_{\text{pos}} = \text{RMS}(\|y_t - \hat{y}_t\|_2) \qquad t_{\text{speed}} = \text{RMS}(\|\Delta y_t\|_2 - \|\Delta \hat{y}_t\|_2)$$
$$t_{\text{vel}} = \text{RMS}(\|\Delta y_t - \Delta \hat{y}_t\|_2) \qquad t_{\text{turn}} = \text{RMS}(\theta_t(y) - \theta_t(\hat{y}))$$
$$t_{\text{curv}} = \text{RMS}(\|\Delta^2 y_t - \Delta^2 \hat{y}_t\|_2) \qquad t_{\text{len}} = |L(y) - L(\hat{y})|$$

- Each term MAD-normalized: $\kappa_k = \text{median}(|t_k - \text{median}(t_k)|)$

- Final score — weighted RMS of normalized components: $s_{\text{CGT}} = \left(\sum_k w_k \, (t_k/\kappa_k)^2 / \sum_k w_k \right)^{1/2}$

Qualitative samples — Navier-Stokes

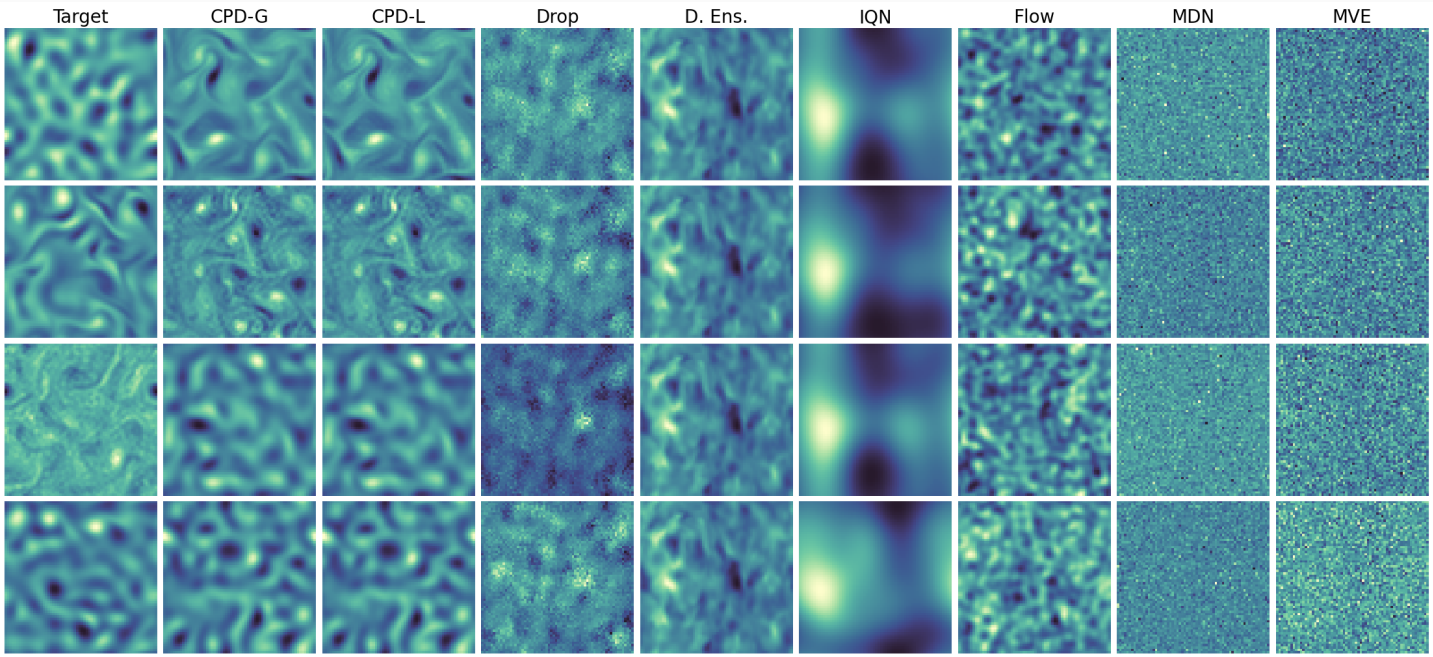


Figure 8: Column 1: test realizations. Remaining columns: samples from each UQ method.

- Baselines underestimate variability and miss fine-scale vortex structure
- CPDs preserve both large-scale flow patterns and small-scale turbulent features

Qualitative samples — Elliptic PDE inversion

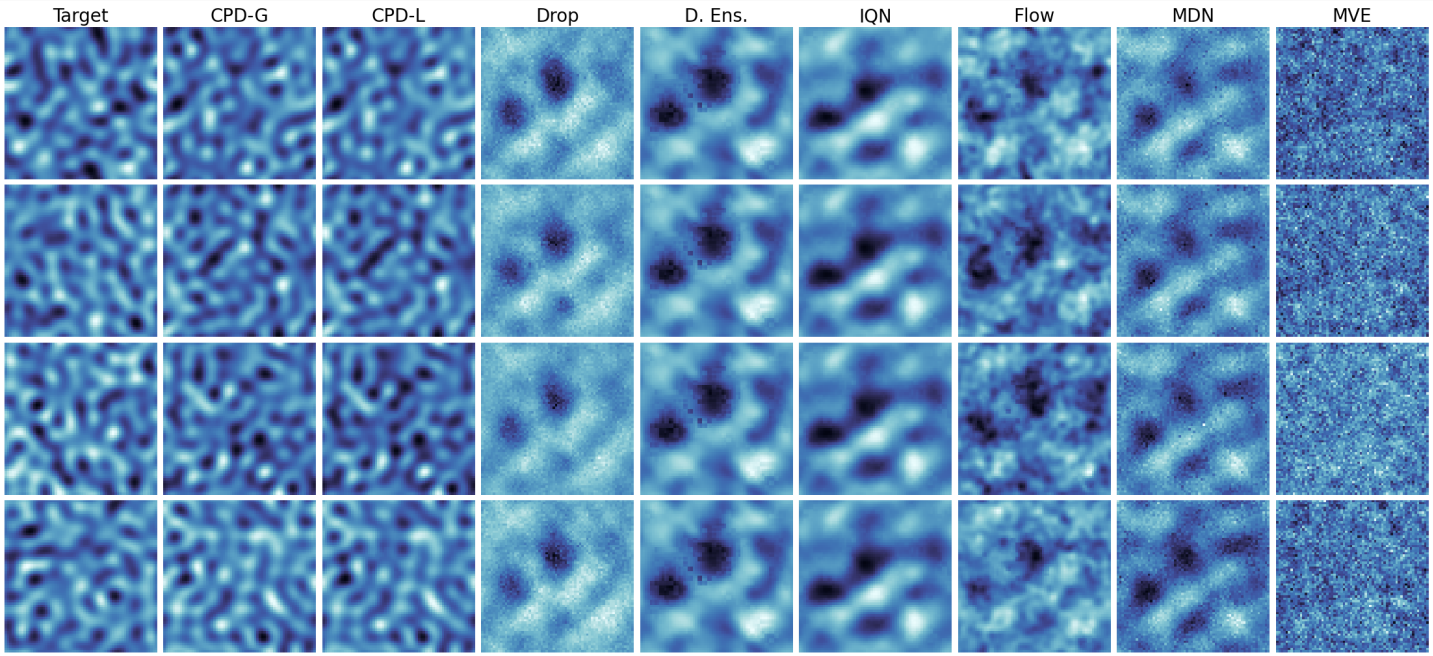


Figure 9: Column 1: test realizations. Remaining columns: samples from each UQ method.

- Target boundary conditions are Gaussian processes — baselines fail to capture the correct scale
- CPDs reproduce both the spatial correlation and amplitude of the forcing field

Computational scaling

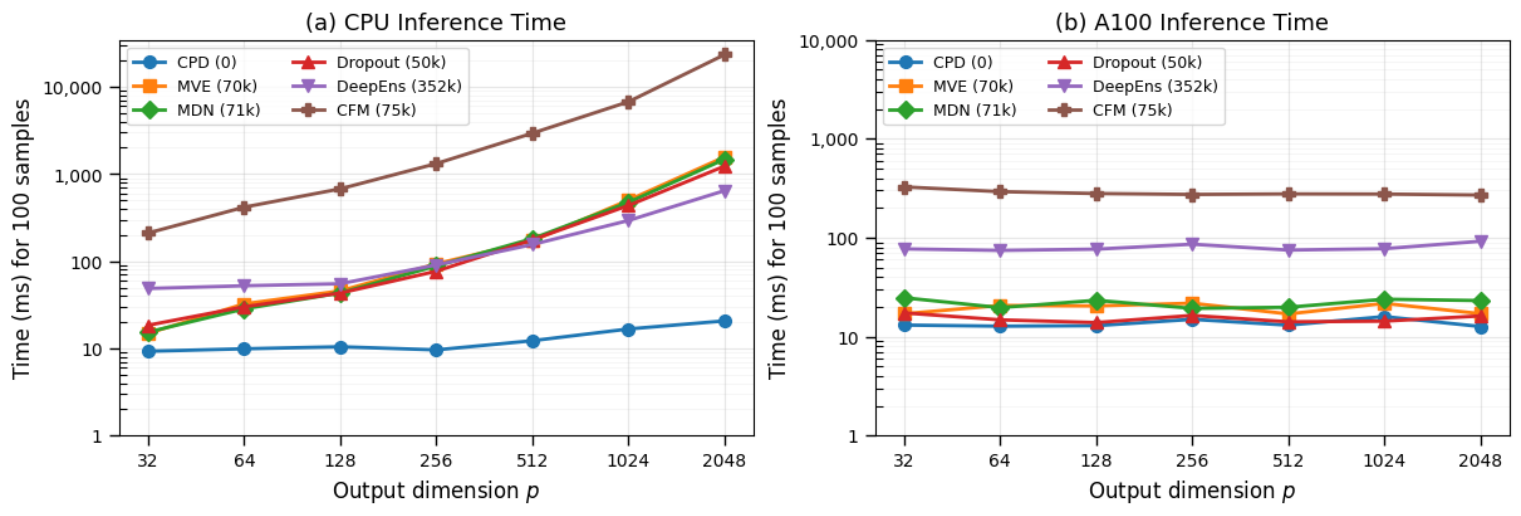


Figure 10: Inference time (100 samples) vs. output dimension p on 1D GP regression.

- **CPU (panel a):** CPD runtime nearly constant until $p \geq 512$; baselines scale dramatically worse — CPD requires no additional forward passes, only score gradients
- **GPU (panel b):** all methods parallelize to constant time on A100, but CPD has the lowest average runtime

Risk-controlling prediction bands

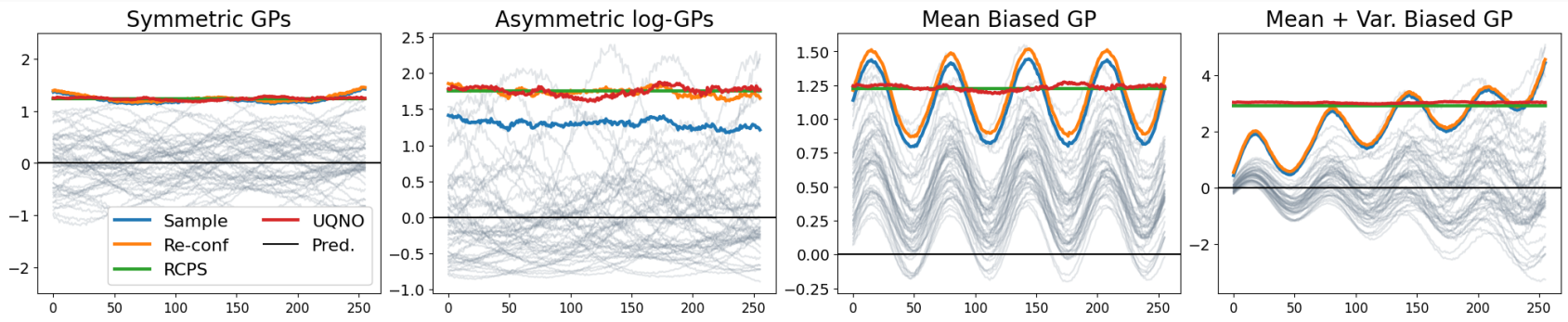


Figure 11: Sampled bands (Sample), reconformalized bands (Re-conf), and RCPS bands.

- Sample $\partial C_\alpha(x)$, form pointwise envelope, inflate by η until risk functional $\ell_\eta(x, y) = \sum_j \mathbf{1}\{y_j \notin \mathcal{B}_\eta(x)\} \leq \delta$ at level $1 - \alpha$
- Binary search over η (monotone)
- Re-conf adapts to asymmetry and bias

	Symm.	Asym.	$\Delta\mu$	$\Delta(\mu, \sigma)$
Sample	0.894	0.646	0.828	0.876
Re-conf	0.901	0.900	0.900	0.901
RCPS	0.900	0.901	0.901	0.902
UQNO	0.901	0.902	0.906	0.907
Sample	2.443	2.032	1.145	4.269
Re-conf	2.479	2.948	1.302	4.465
RCPS	2.460	3.487	2.451	5.715
UQNO	2.467	2.982	2.348	5.772

Boundary repulsion for exploration

- Minimum-norm flow converges **normally** to boundary
  - Samples may cluster in low-curvature regions
- **Fix:** after convergence, add a tangential repulsive velocity
  - Project kernel-based repulsion onto tangent space of $\partial C_\alpha(x)$
  - Samples spread along boundary without leaving it
- A few additional ODE steps; not needed for CPD sampling

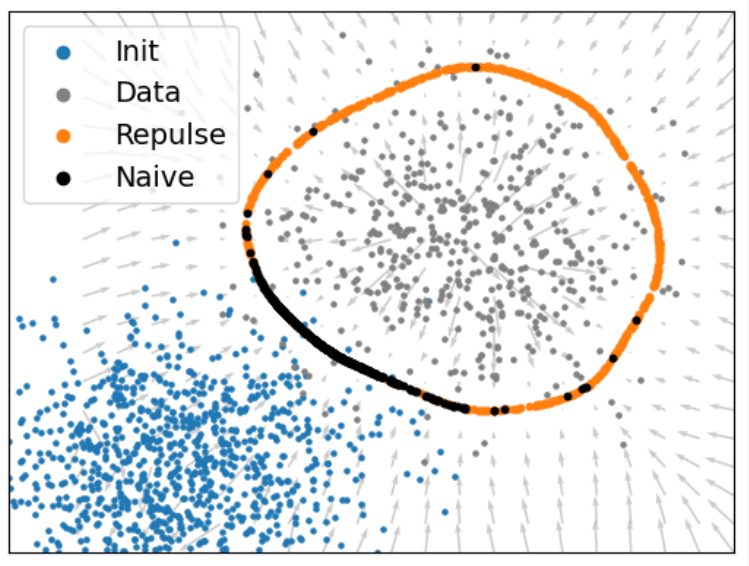


Figure 12: Repulsed boundary samples on bi-modal data.

- Boundary repulsion is a cheap heuristic. Future work: full tangent diffusion with target (uniform) mixing measure..