FLIPR: FLexible and Interpretable Prediction Regions for time series

Eshant English[1,2,3]*, **Christoph Lippert**[1,4]

[1]Hasso Plattner Institute for Digital Engineering, Germany
[2]University of Tokyo, Tokyo, Japan
[3]Centre for Advanced Intelligence Project, RIKEN, Tokyo, Japan
[4]Hasso Plattner Institute for Digital Health at Mount Sinai, USA

Constructing reliable and interpretable prediction regions remains a core challenge in time-series forecasting. While conformal prediction offers rigorous finite-sample coverage guarantees, most existing approaches focus on univariate intervals and fail to capture dependencies across multiple forecast horizons. We propose **FLIPR** (FLexible and Interpretable Prediction Regions), a flexible and interpretable conformal framework that constructs balanced joint prediction regions for multi-horizon forecasts. FLIPR for time series produces a K-th–order non-conformity score that jointly calibrates horizon-wise residuals using standardised mean and scale estimates, enabling explicit control of K-family-wise error while preserving interpretability. The resulting regions are rectangular yet adaptive, distributing coverage uniformly across horizons without requiring any additional learned model. Empirical results on synthetic and real-world datasets show that FLIPR achieves valid coverage with compact, well-calibrated prediction regions, outperforming existing conformal baselines in efficiency and interpretability.

1. Introduction

Conformal prediction (Angelopoulos and Bates, 2021; Vovk et al., 2005) provides a principled way to quantify uncertainty by constructing prediction regions $\Gamma_\epsilon(x)$ that contain the true outcome y with a prescribed probability $1-\epsilon$. In regression, this can be viewed as leveraging the empirical distribution of residuals on a held-out calibration set to make probabilistic statements about unseen data. *Inductive conformal prediction* (ICP) (Shafer and Vovk, 2008) formalizes this intuition while guaranteeing finite-sample marginal coverage.

For univariate outputs, ICP is straightforward: selecting the $(1-\epsilon)$ quantile of the calibration residuals $r_i = |y_i - \hat{f}(x_i)|$ directly yields a prediction interval $[\hat{f}(x) \pm q_{1-\epsilon}]$. Extending this idea to multivariate or multi-horizon settings, however, is less natural. One must first project the residual vector R onto a scalar non-conformity score—commonly via ℓ_1, ℓ_2, or ℓ_∞ norms—which implicitly defines the geometry of the resulting region. Such scalar mappings, while convenient, obscure dependencies among dimensions and limit interpretability.

For interpretability, it is often desirable to construct marginal prediction intervals for each forecast component h, such that their concatenation forms a joint region $\Gamma_\epsilon = \prod_{h=1}^{H} \Gamma_{\epsilon,h}$. This ensures that the entire forecast lies within the region with probability $1-\epsilon$. A common statistical remedy for joint coverage is to apply the Bonferroni correction (Dunn, 1961), replacing ϵ with ϵ/H. However, this becomes increasingly conservative as the horizon H grows and implicitly assumes that forecast errors are independent—an assumption rarely satisfied in time-series forecasting, where residuals are typically correlated across steps.

A simple alternative is to define the non-conformity score using the ℓ_∞ norm, $s_\infty(r) = \max_h |r_h|$, capturing the largest deviation across the horizon. Yet, this approach overemphasises the furthest forecast step—where model errors are usually largest—and produces intervals of uniform width

*contact email: eshant.english@hpi.de, eshantenglish@g.ecc.u-tokyo.ac.jp

Third Conference on Parsimony and Learning (CPAL 2026).

across time, ignoring heteroscedasticity in uncertainty. In summary, existing methods either sacrifice interpretability for efficiency or yield interpretable but overly conservative regions, limiting their practical value in decision-making contexts.

In many real-world applications, interpretability entails more than nominal coverage—it requires *actionable regions* that enable practitioners to reason about when and how predictions may fail. For example, if one forecasted component lies outside its interval $\Gamma_{\epsilon,h}$, it should be immediately clear that the joint forecast is unreliable, prompting recalibration or intervention. At the same time, strict joint validity may be unnecessarily conservative: one may wish to tolerate a few uncovered components without compromising reliability. Furthermore, fairness across horizons is desirable—no step should consistently underperform others in terms of coverage.

In this work, we introduce **FLIPR** (*FLexible and Interpretable Prediction Regions*) for time series, a conformal method that defines a non-conformity score based on standardised residuals and K-th–order deviations. This design balances interpretability and flexibility, allowing controlled K-family-wise error while maintaining balanced coverage across horizons. FLIPR enables the construction of prediction regions that are statistically valid, geometrically interpretable, and practically useful for sequential decision-making.

2. Background: Inductive Conformal Prediction

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ denote a dataset of n exchangeable samples drawn from an unknown distribution $\mathcal{P}(X, Y)$. In *Inductive Conformal Prediction* (ICP), the data are partitioned into a *training set* $\mathcal{D}_{\text{train}}$ and a *calibration set* $\mathcal{D}_{\text{cal}}$. A predictive model $\hat{f} : \mathcal{X} \to \mathcal{Y}$ is first fitted using $\mathcal{D}_{\text{train}}$.

For each calibration sample $(x_i, y_i) \in \mathcal{D}_{\text{cal}}$, we compute a *non-conformity score*, $\alpha_i = s(y_i, \hat{f}(x_i))$, where $s : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ quantifies how atypical the true outcome y_i is relative to its prediction $\hat{f}(x_i)$. A common choice in regression is the absolute residual: $\alpha_i = |y_i - \hat{f}(x_i)|$

Given a user-specified significance level $\epsilon \in (0, 1)$, we compute the empirical $(1 - \epsilon)$ quantile of the calibration scores:

$$q_{1-\epsilon} = \text{Quantile}_{1-\epsilon}\big(\{\alpha_i : (x_i, y_i) \in \mathcal{D}_{\text{cal}}\}\big).$$

This quantile serves as the *conformal threshold* controlling the prediction region width. For a new test input x^*, the conformal prediction region is then defined as

$$\Gamma_\epsilon(x^*) = \{\, y \in \mathcal{Y} : s(y, \hat{f}(x^*)) \leq q_{1-\epsilon} \,\}.$$

Under the assumption of exchangeability between calibration and test examples, ICP guarantees *finite-sample marginal coverage*:

$$\Pr_{(x,y)\sim\mathcal{P}} \big[y \in \Gamma_\epsilon(x)\big] \geq 1 - \epsilon.$$

This coverage guarantee holds regardless of the underlying model $\hat{f}$ or the data distribution, making ICP a powerful and distribution-free uncertainty quantification method.

3. FLIPR: FLexible and Interpretable Prediction Regions for time-series

Consider a collection of exchangeable time series $\{Z^{(i)}_{1:T+H}\}_{i=1}^N$, each of length $T + H$. For each series, we observe the past segment $Z^{(i)}_{1:T}$ and aim to predict the future segment $Z^{(i)}_{T+1:T+H}$. We denote

$$X^{(i)}_{1:T} = Z^{(i)}_{1:T}, \quad Y^{(i)}_{1:H} = Z^{(i)}_{T+1:T+H},$$

so that each training instance $(X^{(i)}_{1:T}, Y^{(i)}_{1:H})$ corresponds to a pair of observed history and its future trajectory. Under the assumption that these series are exchangeable realisations from an underlying stochastic process, the inductive conformal prediction framework can be directly applied to produce calibrated prediction regions for $Y_{1:H}$.

Property 1 (Rectangular Joint Prediction Region)**.** *Let $\Gamma_{\epsilon,h}(X_{1:T}) \subset \mathbb{R}$ denote the marginal prediction interval for each forecast step $h = 1, \ldots, H$, obtained at significance level ϵ_h. A* rectangular joint prediction region *(JPR) for the entire future trajectory $Y_{1:H} = (Y_1, \ldots, Y_H)$ is then defined as the Cartesian product of all marginal intervals:*

$$\Gamma_\epsilon(X_{1:T}) = \Gamma_{\epsilon,1}(X_{1:T}) \times \Gamma_{\epsilon,2}(X_{1:T}) \times \cdots \times \Gamma_{\epsilon,H}(X_{1:T}) \subset \mathbb{R}^H.$$

This rectangular construction implies that the joint region contains all trajectories whose individual components lie within their corresponding marginal intervals. While simple and interpretable, such regions ignore potential dependencies across horizons when created naively, leading to conservative coverage. (See Section 3.1 for an example of how simple projection leads to poor regions).

Now, for each input–output pair $(X^{(i)}_{1:T}, Y^{(i)}_{1:H})$ correspond to a time series segment, where $X^{(i)}_{1:T} = Z^{(i)}_{1:T}$ is the observed history and $Y^{(i)}_{1:H} = Z^{(i)}_{T+1:T+H}$ the H-step-ahead future trajectory. Given a trained forecasting model $\hat{f}$, we obtain predictions

$$\hat{Y}^{(i)}_{1:H} = \hat{f}(X^{(i)}_{1:T}) = (\hat{Y}^{(i)}_1, \ldots, \hat{Y}^{(i)}_H),$$

and define residual vectors $R^{(i)}_{1:H} = Y^{(i)}_{1:H} - \hat{Y}^{(i)}_{1:H}$.

3.1. Case Study: Fallacy of Projecting ℓ_2 Regions to Rectangular Regions

When conformity is defined via the ℓ_2 norm, the acceptance region corresponds to an ℓ_2 ball $\mathbb{B}_2(r) = \{u \in \mathbb{R}^2 : \|u\|_2 \leq r\}$, and the coverage probability is determined by the distribution of the residual vector $R = (R_1, R_2)$. For isotropic residuals, such as $R \sim \mathcal{N}(0, I_2)$, the joint density depends only on the radius $\rho = \|R\|_2$, so that $\Pr(R \in \mathbb{B}_2(r)) = F_{\chi_2}(r)$, where F_{χ_2} is the CDF of the chi distribution with two degrees of freedom.

If we approximate $\mathbb{B}_2(r)$ by a rectangular region $\mathbb{R}_{\text{out}}(r) = [-r, r]^2$, the resulting coverage is

$$\Pr(R \in \mathbb{R}_{\text{out}}(r)) = \Pr(|R_1| \leq r, |R_2| \leq r) = [F_{\mathcal{N}}(r) - F_{\mathcal{N}}(-r)]^2,$$

where $F_{\mathcal{N}}$ denotes the standard normal CDF. Because the joint distribution of (R_1, R_2) factorises under independence, the probability mass contained in the rectangle accumulates faster than that in the circular region. For example, for $r = 2.146$ (yielding 90% coverage under the ℓ_2 ball), we obtain $\Pr(R \in \mathbb{R}_{\text{out}}(r)) \approx 0.937$, confirming that the outer rectangular projection induces *over-coverage*.

Conversely, an inscribed rectangle $\mathbb{R}_{\text{in}}(r) = [-r/\sqrt{2}, r/\sqrt{2}]^2$ satisfies

$$\Pr(R \in \mathbb{R}_{\text{in}}(r)) = [F_{\mathcal{N}}(r/\sqrt{2}) - F_{\mathcal{N}}(-r/\sqrt{2})]^2 \approx 0.758,$$

which corresponds to *under-coverage*. Distributionally, these discrepancies arise because transforming a radially symmetric acceptance region into an axis-aligned rectangle implicitly reweights the density according to independent marginal tails rather than the joint radial probability. In practice, this means that projecting ℓ_2-based non-conformity scores to rectangular regions distorts the intended coverage: either inflating it when circumscribed or deflating it when inscribed; emphasising the need for conformity designs that preserve both interpretability and probabilistic calibration.

Property 2 (Flexibility with K-Family-Wise Error Control)**.** *Let $\Gamma_{\epsilon,h}(X_{1:T})$ denote the marginal prediction interval for each forecast horizon $h = 1, \ldots, H$. We define the number of violations across the prediction horizon as*

$$V = \sum_{h=1}^{H} \mathbf{1}\{Y_h \notin \Gamma_{\epsilon,h}(X_{1:T})\}.$$

*A joint prediction region is said to satisfy K-*family-wise error control *if*

$$\Pr[V > K] \leq \epsilon,$$

that is, the probability that more than K forecast components fall outside their respective marginal intervals does not exceed the prescribed significance level ϵ.

This relaxed form of simultaneous coverage generalises the standard family-wise error rate ($K = 1$) and allows a controlled number of violations, thereby providing a flexible trade-off between conservativeness and practical utility in multi-step forecasting.

Algorithm 1 FLIPR: Flexible and Interpretable Prediction Regions for time series

Require: Dataset $\mathcal{Z} = \{(X_{1:T}^{(i)}, Y_{1:H}^{(i)})\}_{i=1}^{N}$, significance ϵ, horizon H, error tolerance K
Ensure: Joint prediction region $\Gamma_\epsilon^{(K)}(X_{1:T}^*)$
1: Split $\mathcal{Z}$ into training $\mathcal{D}_{\text{tr}}$ and calibration $\mathcal{D}_{\text{cal}}$
2: Train predictor $\hat{f}$ on $\mathcal{D}_{\text{tr}}$; get $\hat{Y}_{1:H}^{(i)} = \hat{f}(X_{1:T}^{(i)})$
3: For $(X_{1:T}^{(i)}, Y_{1:H}^{(i)}) \in \mathcal{D}_{\text{cal}}$, compute residuals $R_h^{(i)} = Y_h^{(i)} - \hat{Y}_h^{(i)}$ for $h = 1, \ldots, H$
4: Estimate per-horizon mean/scale using $\mathcal{D}_{\text{train}}$: $\mu_h = \text{mean}(R_h^{(i)}),\ \sigma_h = \text{std}(R_h^{(i)})$
5: **for** $i = 1$ to n_{cal} **do**
6: Standardize $\tilde{R}_h^{(i)} = \left|R_h^{(i)} - \mu_h\right|/\sigma_h \quad (h = 1{:}H)$
7: Conformity score:

$$\alpha_i^{(K)} = \text{K-max}\{\tilde{R}_1^{(i)}, \ldots, \tilde{R}_H^{(i)}\}$$

8: **end for**
9: Threshold: $q_{1-\epsilon}^{(K)} = \text{Quantile}_{1-\epsilon}\big(\{\alpha_i^{(K)}\}_{i=1}^{n_{\text{cal}}}\big)$
10: For new $X_{1:T}^*$ with point forecast $\hat{Y}_{1:H}^* = \hat{f}(X_{1:T}^*)$, construct

$$\Gamma_\epsilon^{(K)}(X_{1:T}^*) = \prod_{h=1}^{H}\Big[\hat{Y}_h^* - \big(q_{1-\epsilon}^{(K)}\,\widehat{\sigma}_h + \widehat{\mu}_h\big),\ \hat{Y}_h^* + \big(q_{1-\epsilon}^{(K)}\,\widehat{\sigma}_h + \widehat{\mu}_h\big)\Big]$$

11: **return** $\Gamma_\epsilon^{(K)}(X_{1:T}^*)$

3.2. FLIPR for Univariate Time Series Trajectories

We define a weighted non-conformity score that captures heterogeneity across the forecast horizon:

$$\alpha_i^K = s_{\text{FLIPR}}(Y_{1:H}^{(i)}, \hat{f}(X_{1:T}^{(i)})) = \text{K-th largest}\left\{\frac{|R_1^{(i)} - \mu_1|}{\sigma_1}, \ldots, \frac{|R_H^{(i)} - \mu_H|}{\sigma_H}\right\}.$$

where μ_h and σ_h denote the empirical mean and standard deviation of residuals at step h in the calibration set. This formulation effectively *shifts* and *scales* residuals per horizon, ensuring that conformity reflects the relative deviation from the component-wise error magnitude.

For each calibration trajectory, the non-conformity score is computed as above, yielding $\{\alpha_i\}_{i=1}^{n_{\text{cal}}}$. We then determine the $(1-\epsilon)$ quantile

$$q_{1-\epsilon}^{(K)} = \text{Quantile}_{1-\epsilon}\big(\{\alpha_i^K\}_{i=1}^{n_{\text{cal}}}\big),$$

which ensures that at most K forecast components are allowed to violate the joint region with probability $\leq \epsilon$. The corresponding prediction region is

$$\Gamma_\epsilon^K(X_{1:T}^*) = \prod_{h=1}^{H}\Big[\hat{Y}_h^* - \big(q_{1-\epsilon}^K\widehat{\sigma}_h + \widehat{\mu}_h\big),\ \hat{Y}_h^* + \big(q_{1-\epsilon}^K\widehat{\sigma}_h + \widehat{\mu}_h\big)\Big].$$

Full details on the algorithm can be found in Algorithm 1. Additionally, the core idea for FLIPR can be extended to more non-conformity scores, for constructing regions that are asymmetric/one-sided (Section 3.4), have multi-dimensional time steps (Section 3.3), or need to control variances for specific components (Section 3.5).

Property 3 (Property of Balance)**.** *A joint prediction region (JPR) for the future path $Y_{1:H}$ is said to be* balanced *if the marginal coverage across all forecast horizons is approximately uniform. Formally, the JPR is balanced if*

$$\Pr\left[Y_h \in \Gamma_{\epsilon,h}(X_{1:T})\right] = 1 - \epsilon_h, \quad \textit{for some} \quad \epsilon_h, \quad \forall\, h = 1, \ldots, H,$$

such that each implied (simultaneous) prediction interval $\Gamma_{\epsilon,h}$ achieves the same coverage level across all horizons. This property ensures that no specific time step is systematically under- or over-covered, promoting homogeneity and interpretability across the sequence of predictions.

Since FLIPR rescales and recenters residuals independently per horizon, each component of the predictive sequence is calibrated on its own statistical scale. The resulting joint prediction region can be interpreted as a rectangular region whose sides adapt to both the local variability (σ_h) and systematic bias (μ_h) of the forecasting model. This yields *balanced coverage* across the horizon, where miscoverage is approximately uniform and intervals remain compact without sacrificing validity.

3.3. FLIPR for Multi-variate Time-steps

We now consider the general multivariate case where each forecasted step $Y_h \in \mathbb{R}^d$ is d-dimensional, resulting in a residual matrix $R^{(i)}_{1:H} \in \mathbb{R}^{H\times d}$ for each calibration trajectory i. To define a unified non-conformity score, we flatten the residual matrix into a vector of length $H \times d$ and compute a standardised deviation for each component. Specifically, we define

$$\alpha_i^K = s_{\text{FLIPR}}\big(Y^{(i)}_{1:H}, \hat{f}(X^{(i)}_{1:T})\big) = \text{K-max}\left\{ \frac{|R^{(i)}_{h,j} - \mu_{h,j}|}{\sigma_{h,j}} : 1 \le h \le H,\ 1 \le j \le d \right\},$$

where $\mu_{h,j}$ and $\sigma_{h,j}$ denote the empirical mean and standard deviation of residuals for component j at horizon h across the calibration set. This formulation *shifts* and *scales* residuals per dimension, ensuring that conformity reflects deviations relative to the local error magnitude within each horizon.

For all calibration trajectories $\{(X^{(i)}_{1:T}, Y^{(i)}_{1:H})\}_{i=1}^{n_{\text{cal}}}$, we obtain scores $\{\alpha_i^K\}_{i=1}^{n_{\text{cal}}}$ and compute the $(1-\epsilon)$ empirical quantile

$$q^{(K)}_{1-\epsilon} = \text{Quantile}_{1-\epsilon}\big(\{\alpha_i^K\}_{i=1}^{n_{\text{cal}}}\big).$$

The resulting joint prediction region for a new input $X^*_{1:T}$ is given by

$$\Gamma^{(K)}_\epsilon(X^*_{1:T}) = \prod_{h=1}^{H}\prod_{j=1}^{d}\Big[\hat{y}^*_{h,j} - \big(q^{(K)}_{1-\epsilon}\widehat{\sigma}_{h,j} + \widehat{\mu}_{h,j}\big),\ \hat{y}^*_{h,j} + \big(q^{(K)}_{1-\epsilon}\widehat{\sigma}_{h,j} + \widehat{\mu}_{h,j}\big)\Big].$$

3.4. FLIPR for One-Sided and Asymmetric Interval Regions

The FLIPR framework can be naturally extended to construct one-sided or asymmetric joint prediction regions by modifying the non-conformity score. Instead of using absolute residuals, we consider *signed* standardised residuals for each component, allowing directional coverage control.

For each calibration trajectory $(X^{(i)}_{1:T}, Y^{(i)}_{1:H})$ with residual matrix $R^{(i)}_{1:H} \in \mathbb{R}^{H\times d}$, we define

$$\alpha^K_{i,+} = \text{K-max}\left\{ \frac{R^{(i)}_{h,j} - \mu_{h,j}}{\sigma_{h,j}} : 1 \le h \le H,\, 1 \le j \le d \right\},$$

and analogously $\alpha^K_{i,-}$ with the sign flipped. Let $q^{(K,+)}_{1-\epsilon}$ and $q^{(K,-)}_{1-\epsilon}$ denote the $(1-\epsilon)$ empirical quantiles of $\{\alpha^K_{i,+}\}$ and $\{\alpha^K_{i,-}\}$, respectively.

The one-sided lower and upper FLIPR regions for a new input $X^*_{1:T}$ are then

$$\Gamma^{(K,+)}_{1-\epsilon}(X^*_{1:T}) = \prod_{h=1}^{H}\prod_{j=1}^{d}\Big(\hat{y}^*_{h,j} - q^{(K,+)}_{1-\epsilon}\widehat{\sigma}_{h,j} - \widehat{\mu}_{h,j}, \infty\Big),$$

$$\Gamma^{(K,-)}_{1-\epsilon}(X^*_{1:T}) = \prod_{h=1}^{H}\prod_{j=1}^{d}\Big(-\infty, \hat{y}^*_{h,j} + q^{(K,-)}_{1-\epsilon}\widehat{\sigma}_{h,j} + \widehat{\mu}_{h,j}\Big).$$

Asymmetric regions. When the predictive uncertainty is not symmetric, one can define distinct upper and lower error tolerances ϵ_+ and ϵ_- such that $\epsilon_+ + \epsilon_- = \epsilon$. The resulting asymmetric FLIPR region is

$$\Gamma^{(K,\pm)}_{1-\epsilon}(X^*_{1:T}) = \prod_{h=1}^{H}\prod_{j=1}^{d}\Big(\hat{y}^*_{h,j} - q^{(K)}_{1-\epsilon_-}\widehat{\sigma}_{h,j} - \widehat{\mu}_{h,j},\ \hat{y}^*_{h,j} + q^{(K)}_{1-\epsilon_+}\widehat{\sigma}_{h,j} + \widehat{\mu}_{h,j}\Big).$$

This formulation enables directional calibration and asymmetric coverage control, which is useful in applications where underestimation and overestimation risks differ.

3.5. FLIPR for Controlled Unbalancing via Weighted Residuals

The FLIPR framework can also accommodate *controlled unbalancing* across components of the forecast horizon or feature dimensions. In many applications, practitioners may wish to allocate uncertainty asymmetrically across components—allowing certain forecasted variables to fall outside their prediction intervals more frequently, while maintaining overall coverage control.

To achieve this, we introduce a set of non-negative *component weights* $w_{h,j} \in \mathbb{R}_+$ that modulate the influence of each residual $R_{h,j}$ in the non-conformity score. The weighted non-conformity score is then defined as

$$\alpha_i^K(w) = \text{K-max}\left\{ w_{h,j} \cdot \frac{|R_{h,j}^{(i)} - \mu_{h,j}|}{\sigma_{h,j}} : 1 \le h \le H,\, 1 \le j \le d \right\}.$$

A smaller weight $w_{h,j} < 1$ reduces the influence of component (h, j) in the joint non-conformity score, effectively tightening its corresponding marginal interval and increasing the likelihood of miscoverage for that component. Conversely, $w_{h,j} > 1$ expands its effective contribution, producing wider intervals and enhanced coverage for that dimension.

The resulting weighted prediction region for a new input $X_{1:T}^*$ becomes

$$\Gamma_\epsilon^{(K)}(X_{1:T}^*; w) = \prod_{h=1}^{H} \prod_{j=1}^{d} \left[\hat{y}_{h,j}^* - (q_{1-\epsilon}^{(K)} w_{h,j} \widehat{\sigma}_{h,j} + \widehat{\mu}_{h,j}),\ \hat{y}_{h,j}^* + (q_{1-\epsilon}^{(K)} w_{h,j} \widehat{\sigma}_{h,j} + \widehat{\mu}_{h,j}) \right].$$

By appropriately choosing or learning $w_{h,j}$, one can control the allocation of coverage across horizons or dimensions, enabling interpretable trade-offs between regional compactness and component-wise reliability.

4. Experiments

We evaluate **FLIPR** (Flexible and Interpretable Prediction Regions) across three experimental settings: (i) a controlled synthetic setup where we design the predictor to exhibit specific residual behaviours, (ii) particle-based simulation datasets, and (iii) a real-world dataset of COVID-19 case trajectories. This combination allows us to assess both the statistical validity and interpretability of the proposed prediction regions under increasing complexity.

For reference, we compare FLIPR against several established conformal methods. **JANET** (English et al., 2025) serves as a close baseline, differing mainly in that it lacks the residual-shifting term introduced in FLIPR's non-conformity score. **CF-RNN** (Stankeviciute et al., 2021) employs Bonferroni correction under conditional IID assumptions and typically produces conservative intervals at longer horizons. **VanillaCopula** and **CPTS** (Sun and Yu, 2024) are copula-based approaches that model cross-step dependencies but do not support adaptive K-family-wise control.

Both FLIPR and JANET are evaluated for three tolerance levels, $K \in \{1, 2, 3\}$, while other baselines correspond to $K = 1$. All methods target a nominal significance level of $\epsilon = 0.1$, and we report empirical coverage and average prediction-region width on held-out test sets. Each experiment is repeated over multiple random seeds, and we report the corresponding standard deviations to quantify variability. We defer to Section A for further details on the experiment setup.

4.1. Controlled Experiment Setting

To demonstrate the limitations of simple residual scaling and the benefits of FLIPR, we design a controlled synthetic experiment. We generate $N = 1000$ independent time series, each of length $T + H = 40 + 10 = 50$, according to

$$z_t = at + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2),$$

where the slope $a > 0$ is fixed and the intercept is zero. We then fit a linear predictor with an underestimated slope $\hat{a} < a$, leading to a systematic underfit and forecast errors that increase linearly with the prediction horizon:

$$e_{T+h} = z_{T+h} - \hat{a}(T+h) = (a - \hat{a})(T+h) + \epsilon_{T+h}.$$

Although ϵ_t has constant variance σ^2, the expected magnitude of the residuals grows with h due to bias accumulation.

In JANET's setting, conformity is computed by scaling the residuals with their empirical standard deviation. Since the mean residual increases with h, the conformity distribution is dominated by the final component e_{T+H}, effectively setting the threshold based on the farthest forecast step. This causes unbalanced coverage across the horizon—intervals become overly wide near the end and too narrow near the beginning. Table 2 reports empirical coverage and empirical width, where each

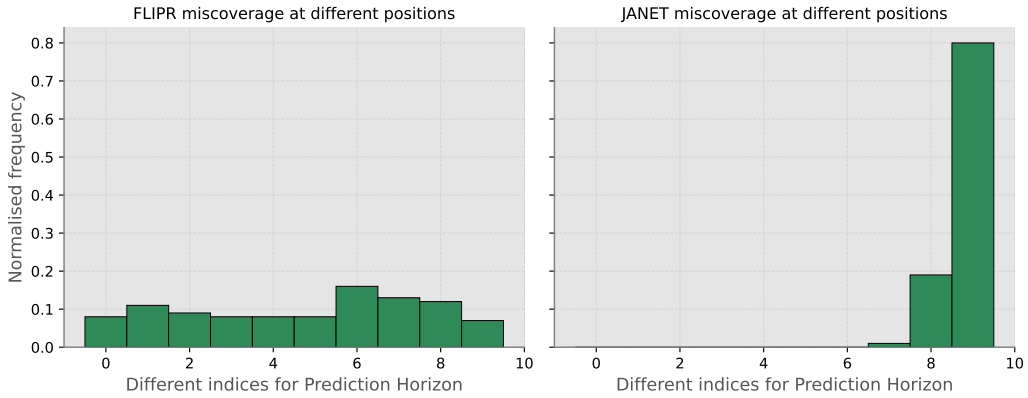

Figure 1: Distribution of errors across runs for the two methods. **Left:** FLIPR shows a more balanced error pattern, with deviations spread across multiple components. **Right:** JANET exhibits concentrated errors at specific positions, despite its built-in scaling factor, indicating that the residual correction is localised rather than uniformly distributed.

region's size is computed as a geometric width of all the predicted components as in (English et al., 2025), for all baselines. Both JANET and FLIPR achieve coverage close to the nominal level, while copula-based approaches severely under-cover, and CF-RNN exhibits strong over-coverage due to its conservative correction. Crucially, FLIPR maintains interpretability and allows adaptive K-FWER control ($K = 1, 2, 3$), offering a more balanced calibration.

Figure 1 visualises the distribution of miscoverage across forecast horizons for JANET and FLIPR. These two are the only methods that flexibly handle K while providing interpretable regions. FLIPR (left panel) spreads errors approximately uniformly across all forecast steps, whereas JANET (right panel) accumulates them predominantly at the final time step. This demonstrates FLIPR's advantage in adaptively weighting conformity across the horizon rather than relying on a single global scale.

Remark. In the idealised limit, one would expect JANET's errors to be entirely concentrated on the last forecast component, whereas FLIPR would achieve exact uniform miscoverage. Small deviations arise from estimation noise in the mean residual bias and the empirical scaling factor. Nonetheless, the experiment highlights how FLIPR achieves a more even and interpretable allocation of uncertainty across the forecasting horizon.

4.2. Particle Simulation Benchmarks

We begin with two synthetic datasets derived from particle interaction systems proposed by Kipf et al. (2018). Following the setup of Sun and Yu (2024), each trajectory consists of $T = 55$ observed time steps in $\mathbb{R}^2$, and the model predicts the next $H = 4$ steps. Independent Gaussian noise with

standard deviation $\sigma = 0.05$ and $\sigma = 0.01$ is added to generate two distinct experimental settings. Since the time steps are multidimensional, we apply the generalised non-conformity score that treats each dimension as a component. (see Section 3.3 for details)

Across all methods, the same base predictor is used (details in Section A). Table 1 reports results for the *particle5* experiment ($\sigma = 0.05$). FLIPR achieves coverage closest to the nominal level across all values of K, while maintaining compact regions. JANET performs comparably but tends to slightly over-cover, whereas CF-RNN also produces wider regions and systematic over-coverage due to its Bonferroni adjustment. Both copula-based variants (VanillaCopula and CPTS) show modest under-coverage, suggesting that explicit dependence modelling alone does not guarantee calibrated joint regions.

For the *particle1* experiment ($\sigma = 0.01$), FLIPR again demonstrates the most consistent calibration, especially for $K = 2$, where it achieves a good balance between flexibility and coverage. JANET remains close to the nominal level, though with marginally larger regions, while CF-RNN and Copula-based methods exhibit larger deviations. These results highlight that FLIPR's non-conformity score effectively handles correlated residuals without excessive conservativeness.

Table 1: Empirical coverage and average empirical width for all datasets at $\epsilon = 0.1$. FLIPR and JANET are evaluated for $K = 1, 2, 3$, while the other baselines correspond to $K = 1$. Methods with coverage closest to the nominal 0.9 are bolded. As shown, FLIPR performs the best for $K = 1$. All experiments are averaged over 10 random runs, and the standard deviations are reported accordingly.

	Method	**Particle1**		**Particle5**		**UK COVID-19**	
		Coverage	empirical width	Coverage	empirical width	Coverage	empirical width
$K = 1$	CopulaCPTS	0.92 ± 0.06	0.1476 ± 0.0185	0.88 ± 0.06	0.3433 ± 0.0305	0.88 ± 0.06	921.6 ± 164.5
	VanillaCopula	0.89 ± 0.06	0.1398 ± 0.0185	0.87 ± 0.06	0.3214 ± 0.0249	0.88 ± 0.05	818.5 ± 99.5
	CF-RNN	0.97 ± 0.02	0.1651 ± 0.0155	0.94 ± 0.04	0.3739 ± 0.0307	0.97 ± 0.03	1415.9 ± 227.8
	JANET_1	0.94 ± 0.03	0.1437 ± 0.0167	0.93 ± 0.04	0.3452 ± 0.0261	0.92 ± 0.05	894.3 ± 83.7
	FLIPR_1 (ours)	$\mathbf{0.89 \pm 0.12}$	$\mathbf{0.1388 \pm 0.0209}$	$\mathbf{0.89 \pm 0.06}$	$\mathbf{0.3316 \pm 0.0254}$	$\mathbf{0.91 \pm 0.05}$	$\mathbf{883.9 \pm 97.0}$
$K = 2$	JANET_2	0.94 ± 0.02	0.1198 ± 0.0157	0.93 ± 0.05	0.2749 ± 0.0261	$\mathbf{0.92 \pm 0.05}$	$\mathbf{768.7 \pm 98.2}$
	FLIPR_2 (ours)	$\mathbf{0.90 \pm 0.10}$	$\mathbf{0.1158 \pm 0.0203}$	$\mathbf{0.90 \pm 0.05}$	$\mathbf{0.2624 \pm 0.0185}$	0.93 ± 0.05	761.6 ± 94.1
$K = 3$	JANET_3	0.94 ± 0.03	0.1030 ± 0.0148	$\mathbf{0.92 \pm 0.05}$	$\mathbf{0.2156 \pm 0.0169}$	0.92 ± 0.04	647.2 ± 114.1
	FLIPR_3 (ours)	$\mathbf{0.90 \pm 0.09}$	$\mathbf{0.1013 \pm 0.0178}$	0.87 ± 0.10	0.2036 ± 0.0194	$\mathbf{0.92 \pm 0.05}$	$\mathbf{639.9 \pm 117.7}$

4.3. UK COVID-19 Case Forecasting

We next evaluate the methods on the UK COVID-19 dataset (Stankeviciute et al., 2021; Sun and Yu, 2024), consisting of daily infection counts across 380 regions. The forecasting task is to predict $H = 10$ future days based on the previous $T = 100$ observations. Although the independence assumption is not strictly satisfied across regions, approximate validity is expected from the inductive conformal construction.

Results in Table 1 show that FLIPR maintains near-nominal coverage with narrower and more stable regions than competing methods. JANET performs similarly well, with slight overcoverage. CF-RNN remains conservative, showing persistent over-coverage, while copula-based methods slightly under-cover at longer horizons, consistent with their reliance on approximate dependence corrections.

5. Related Work

Recent conformal approaches have extended uncertainty quantification to structured and sequential prediction tasks. BJ-RNN (Alaa and Van Der Schaar, 2020) applies the Jackknife+ method (Barber et al., 2021) to recurrent models, but its theoretical guarantee of $(1 - 2\epsilon)$ coverage is looser than the standard $(1 - \epsilon)$ bound. CF-RNN (Stankeviciute et al., 2021) assumes conditionally i.i.d. residuals and employs Bonferroni correction (Dunn, 1961), which becomes increasingly conservative as the forecast horizon grows. Cleaveland et al. (2024) proposed a linear-programming approach for multi-step error modelling, but it requires an additional calibration set and extensive parameter

Table 2: Empirical coverage and average prediction region empirical widths (mean $\pm$ std) for the *Controlled* (*linear-trend*) dataset at $\epsilon = 0.1$. FLIPR and JANET are shown for $K = 1, 2, 3$; other baselines correspond to $K = 1$. Coverage is reported as a fraction. Methods with nominal coverage closest to (0.90) are highlighted in bold with a tie resolved through empirical widths. These results show FLIPR performs the best, closely followed by JANET. However, JANET has a much bigger empirical width.

	Method	**Coverage**	**empirical widths (mean $\pm$ std)**
$K = 1$	CopulaCPTS	0.81 ± 0.05	164.9168 ± 0.1534
	VanillaCopula	0.76 ± 0.05	164.0531 ± 0.1534
	CF-RNN	0.86 ± 0.04	164.7690 ± 0.2960
	JANET_1	0.92 ± 0.03	179.4928 ± 1.5585
	FLIPR_1 (ours)	**0.92 ± 0.02**	**164.9722 ± 0.2190**
$K = 2$	JANET_2	0.85 ± 0.05	173.9479 ± 1.0036
	FLIPR_2 (ours)	**0.94 ± 0.01**	**163.6723 ± 0.1453**
$K = 3$	JANET_3	0.87 ± 0.04	171.2284 ± 0.8433
	FLIPR_3 (ours)	**0.92 ± 0.01**	**162.9582 ± 0.1456**

tuning. CopulaCPTS (Sun and Yu, 2024) builds on earlier copula-based dependency modelling (Messoudi et al., 2021) to account for temporal correlations, though at the cost of data inefficiency and gradient-based training overhead. More recently, methods such as CAFHT (Zhou et al., 2024) and JANET (English et al., 2025) introduced rectangular prediction regions that adapt to context, improving interpretability but remaining geometrically restrictive. JAPAN (English and Lippert, 2025) builds upon English and Lippert (2024) to form efficient regions based on density estimation of normalising flows, but the regions are not interpretable.

In contrast, **FLIPR** provides a nonparametric alternative that balances interpretability and flexibility. By using K-th–order non-conformity scores with horizon-wise standardisation, FLIPR achieves controlled K-family-wise error without assuming independence or requiring learned density models, yielding compact and balanced joint prediction regions across multiple horizons.

6. Conclusion and Discussion

We presented **FLIPR**, a conformal prediction framework designed to produce *flexible*, *interpretable*, and *balanced* prediction regions for time-series forecasting. FLIPR extends standard inductive conformal prediction by incorporating K-family-wise error control and distributing conformity adaptively across the forecast horizon. This enables fine-grained uncertainty quantification where coverage remains consistent across components while preserving interpretability and efficiency.

Through controlled synthetic experiments and real-world datasets, we demonstrated the limitations of purely rescaled non-conformity scores that ignore systematic bias and do not account for residual shifts. Our results show that such scaling-only methods concentrate errors at the end of the forecast horizon, whereas FLIPR produces balanced and stable prediction regions that reflect temporal uncertainty more faithfully.

While FLIPR achieves adaptivity across forecast components, it does not yet adapt prediction intervals based on the *conditioning history*. A natural extension is to integrate a learned calibration model, similar to JANET (English et al., 2025), or to employ generative models such as *normalising flows* (Kobyzev et al., 2021; Papamakarios et al., 2021) to estimate the conditional distribution of residuals. These models could provide sample-dependent shift and scale parameters, allowing non-conformity scores to adapt not only along the horizon but also with respect to the past context. Exploring such history-aware calibration mechanisms represents a promising direction for future work.

Acknowledgements

This research was supported by the HPI Research School for Digital Health. We gratefully acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 459422098.

References

Alaa, A. and Van Der Schaar, M. (2020). Frequentist uncertainty in recurrent neural networks via blockwise influence functions. In *International Conference on Machine Learning*, pages 175–190. PMLR.

Angelopoulos, A. N. and Bates, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*.

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. (2021). Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486–507.

Cleaveland, M., Lee, I., Pappas, G. J., and Lindemann, L. (2024). Conformal prediction regions for time series using linear complementarity programming. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 20984–20992.

Dunn, O. J. (1961). Multiple Comparisons among Means. *Journal of the American Statistical Association*, 56(293):52–64.

English, E. and Lippert, C. (2024). Conformalised conditional normalising flows for joint prediction regions in time series. In *NeurIPS 2024 Workshop on Bayesian Decision-making and Uncertainty*.

English, E. and Lippert, C. (2025). Japan: Joint adaptive prediction areas with normalising-flows. *arXiv preprint arXiv:2505.23196*.

English, E., Wong-Toi, E., Fontana, M., Mandt, S., Smyth, P., and Lippert, C. (2025). Janet: Joint adaptive prediction-region estimation for time-series. *Machine Learning*, 114(8):177.

Kipf, T., Fetaya, E., Wang, K.-C., Welling, M., and Zemel, R. (2018). Neural relational inference for interacting systems. In *International conference on machine learning*, pages 2688–2697. Pmlr.

Kobyzev, I., Prince, S. J., and Brubaker, M. A. (2021). Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3964–3979.

Messoudi, S., Destercke, S., and Rousseau, S. (2021). Copula-based conformal prediction for multi-target regression. *Pattern Recognition*, 120:108101.

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B. (2021). Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64.

Shafer, G. and Vovk, V. (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3).

Stankeviciute, K., M. Alaa, A., and van der Schaar, M. (2021). Conformal Time-series Forecasting. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P. S., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 6216–6228. Curran Associates, Inc.

Sun, S. H. and Yu, R. (2024). Copula conformal prediction for multi-step time series prediction. In *The Twelfth International Conference on Learning Representations*.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic learning in a random world*, volume 29. Springer.

Zhou, Y., Lindemann, L., and Sesia, M. (2024). Conformalized adaptive forecasting of heterogeneous trajectories. In *International Conference on Machine Learning*, pages 62002–62056. PMLR.

A. Experiment Setup

All experiments were conducted using NVIDIA A100 GPUs with 20 cores. Including failed runs, hyperparameter tuning, and iterative experimental development, the total compute time across all models is estimated to be approximately 5 GPU-days.

A.1. Controlled Experimental Setting

In the controlled experiment, we explicitly design the predictor to exhibit systematic bias across the forecasting horizon. Each synthetic trajectory is generated from a linear trend (slope coefficient $= 2$) with additive Gaussian noise $\sigma = 1$, consisting of 50 time steps, out of which the last 10 are used as the forecasting horizon. The fitted predictor intentionally underestimates the slope, resulting in increasing residual magnitudes over time. We fit a linear trend with slope coefficient $= 0.5$ as the base predictor and fit a simple linear regression model to estimate per-step scaling factors. This setting allows us to isolate and analyse the effect of residual structure on the conformal threshold. The shifting and scaling parameters for the non-conformity scores are obtained through the residuals on the training set across all the settings.

A.2. Particle Simulation Experiments

Following the setup of Sun and Yu (2024), we use a single-layer sequence-to-sequence LSTM network (EncDec) in which both the encoder and decoder contain one LSTM layer with embedding size 24, followed by a linear output layer. Each model is trained for 200 epochs with a batch size of 128.

For every dataset, we generate 5,000 synthetic trajectories and split them into 100 sequences for calibration and 80 sequences for the test set. Methods that required another split during the calibration phase use the calibration set's 50-50 split for the same.

A.3. UK COVID-19 Experiments

We adopt the same network architectures as in the particle simulations and apply them to the *UK COVID-19* dataset from Stankeviciute et al. (2021). Models are trained for 150 epochs with an embedding size of 512 and a batch size of 128. Out of the 380 regional time series, we use 200 sequences for training, 100 for calibration, and the remaining 80 for testing.