Calibrated Predictive Distributions from Sample-Based Generators

Wen-Ting Wang (egpivo@gmail.com)[a], ShengLi Tzeng (sltzeng@nchu.edu.tw)[a], Yu-Ting Fan (tinafan8181@gmail.com)[b], Hsin-Cheng Huang (hchuang@stat.sinica.edu.tw)[c,d,*]

[a]*Department of Applied Mathematics and Graduate Institute of Statistics, National Chung Hsing University, Taiwan*
[b]*Texas Instruments, Taiwan*
[c]*Institute of Statistical Science, Academia Sinica, Taiwan*
[d]*Institute of Statistics and Data Science, National Tsing Hua University, Taiwan*

Abstract

Conditional generative models, including diffusion models and ensemble forecasters, often produce predictive samples without a tractable likelihood representation. Such sample-based predictive distributions can be systematically biased and poorly calibrated. We propose bias-corrected conformal probability integral transform (PIT) calibration, a split-sample post-processing framework that outputs a calibrated predictive distribution rather than a single fixed-level prediction interval. The method first estimates an affine location-scale correction on a held-out bias split, then calibrates randomized PIT values using a conformal calibrator. The resulting predictive law is represented as a weighted empirical distribution on the generator order statistics, enabling the direct computation of threshold-coherent exceedance probabilities, arbitrary quantiles, highest-density intervals, expected tail losses, and calibrated resamples. In contrast, standard conformal prediction primarily provides fixed-level prediction sets or threshold decisions and does not directly estimate predictive probabilities or high-density regions. We establish finite-sample calibration in probability under exchangeability and show how an optional split-conformal wrapper based on a PIT-centrality score gives nested prediction intervals with finite-sample marginal coverage at user-specified levels. Simulation studies with controlled misspecification and a WeatherBench-2 precipitation-forecasting application demonstrate substantial improvements in probabilistic calibration and downstream distributional summaries relative to uncalibrated sample-based forecasts and interval-only conformal baselines.

Keywords: Bias correction; conformal prediction; continuous ranked probability score; Cramér-von Mises distance; probability integral transform; sample-based forecasting.

*Corresponding author

1. Introduction

Modern conditional generative models are increasingly used for predictive inference. Given covariates x, a generator produces samples $Y^{(1)}, \ldots, Y^{(m)} \sim \widehat{P}(\cdot|x)$ that are intended to approximate the conditional distribution of Y given x. In many applications, the model is accessed only through such samples rather than through an explicit likelihood or a closed-form density. This setting arises, for example, in diffusion-based inverse problems, in simulation-based science where forward simulators produce realizations, and in probabilistic forecasting pipelines that rely on ensemble-style outputs. In principle, such models deliver full predictive distributions rather than only point predictions. In practice, however, limited training data, model misspecification, and optimization artifacts can introduce substantial systematic bias and probability miscalibration into $\widehat{P}(\cdot|x)$, which in turn undermines downstream uncertainty quantification.

Conformal prediction provides distribution-free predictive inference with finite-sample marginal validity under exchangeability (Vovk et al., 2005; Shafer and Vovk, 2008; Lei et al., 2018). Classical conformal regression typically begins with a point predictor and a scalar nonconformity score, such as an absolute residual, and returns a prediction interval at a specified miscoverage level α. In sample-only generative settings, one common workaround is to compress the generated sample cloud to a point summary and then apply a standard conformal method. This approach discards distributional information, including skewness, heteroskedasticity, and multimodality, that may be crucial for downstream probability calculations.

The distinction between a valid prediction set and a calibrated predictive distribution is consequential for downstream analysis. Many scientific and operational users do not only ask whether a future response lies in a fixed 90% interval. They ask for threshold probabilities such as $P(Y > c|x)$, tail summaries such as $\mathbb{E}\{(Y-c)_+|x\}$, Bayes actions under asymmetric losses, or simulated inputs for a subsequent decision model. These are functionals of the predictive law. A fixed-level conformal interval can support a valid set-valued statement, and a one-sided conformal construction can certify a threshold decision at a chosen error level, but neither of these outputs determines a predictive cumulative distribution function (CDF) or a calibrated estimate of an arbitrary exceedance probability.

Recent work has begun to use conditional random samples from a generator directly for conformal inference. Notably, Wang et al. (2023) develop probabilistic conformal prediction using conditional random samples and construct prediction sets with marginal coverage guarantees. Such methods are well-suited when the target is a set-valued prediction region

at a single level. Our objective differs in that we aim to produce a calibrated predictive distribution rather than merely a prediction set at a fixed level. A predictive cumulative distribution function (CDF) supports calibrated quantiles, equal-tail summaries, and highest-density intervals at all levels, enables calibrated resampling for downstream Monte Carlo decision-making, and can be evaluated using calibration diagnostics such as the Cramér-von Mises distance and proper scoring rules such as the continuous ranked probability score (CRPS) (Gneiting and Raftery, 2007).

Conformal predictive systems (CPSs) and conformal calibrators provide a general framework for constructing predictive distributions that are calibrated in probability (Vovk et al., 2017, 2020). Our method shares this objective but is designed for a sample-only setting, in which the base predictive distribution is unavailable in closed form and is represented only by draws from an implicit generator. Consequently, for a finite generator sample of size m, the empirical predictive CDF is both discrete and random. We address these features by developing a randomized probability integral transform (PIT) tailored to sample-based predictors, together with an efficient representation that assigns nonuniform weights to the generator's order statistics. This representation enables rapid evaluation of calibrated quantiles and direct resampling from the calibrated predictive distribution, without retraining the underlying generator.

Distributional conformal prediction (DCP) calibrates an estimated conditional CDF $\widehat{F}_x(y)$ using PIT ranks and studies approximate conditional validity under additional regularity and consistency conditions (Chernozhukov et al., 2021). Our setting differs in that $\widehat{F}_x(y)$ is typically not directly evaluable and is accessible only through generated samples. Consequently, we focus on finite-sample marginal calibration in probability in the sense of CPS. Extensions toward more localized forms of validity, such as localized conformal methods or Mondrian variants, can in principle be incorporated, but they are not our primary focus (Guan, 2023; Bostrom et al., 2021). This is consistent with known impossibility results for exact distribution-free conditional coverage (Barber et al., 2021).

A practical distinction from both CPS-style calibration and DCP-style PIT adjustment is that we introduce an explicit bias-correction operation, estimated on a held-out split, prior to conformal calibration. In many applications of generative models, systematic location and scale distortions are the dominant sources of error. We address this with an affine correction in two forms: a simple global adjustment based on residual location and scale, and a more flexible x-dependent correction in which the residual mean and log-scale ratio are modeled as smooth functions of the covariates using generalized additive model (GAM) regressions on the bias split.

The proposed procedure, bias-corrected conformal PIT calibration (CPIT),[1] has two distributional outputs. First, it produces a calibrated predictive CDF represented as a weighted empirical distribution on adjusted generator order statistics. This is the main object of the paper. It can be queried for event probabilities, arbitrary quantiles, HDIs, tail expectations, proper scores, and calibrated resamples. Second, it provides a smoothed weighted CDF when full-support summaries are desired. When exact coverage is needed at specified levels, the CPIT CDF can also be used inside a PIT-centrality split-conformal wrapper that returns nested fixed-level intervals. For comparison, we adapt three interval-only split-conformal baselines to the sample-based forecasting setting, using empirical quantiles, scaled residuals, and empirical CRPS scores, allowing us to distinguish fixed-level coverage calibration from calibration of the full predictive distribution.

Our contributions are as follows.

1. We develop CPIT, a split-sample post-processing framework for sample-based conditional generators. The method combines affine location-scale bias correction with conformal calibration of randomized PIT values, and applies to black-box predictive samples without requiring a tractable likelihood. The proposed method outputs a calibrated predictive distribution, not only a fixed-level prediction interval.

2. We establish finite-sample calibration in probability for the randomized calibrated PIT values under exchangeability. We also provide an optional PIT-centrality split-conformal wrapper that gives nested fixed-level prediction intervals with finite-sample marginal coverage, and give stability and approximation results for the smoothed weighted CPIT distribution.

3. We demonstrate, through controlled simulations and a WeatherBench 2 precipitation-forecasting application, that CPIT improves distributional calibration and supports downstream probabilistic analyses, including threshold-coherent exceedance probabilities, expected exceedance severity, tail-risk summaries, and highest-density intervals.

The rest of the paper is organized as follows. Section 2 introduces the sample-based forecasting setup and the CPIT procedure, including bias correction, PIT calibration, the weighted empirical predictive distribution, and its smoothed version. Section 3 presents fixed-level conformal prediction methods for sample-based forecasts, including an exact PIT-centrality wrapper for CPIT and the interval-only conformal baselines used in the empirical

[1] An open-source implementation is available on GitHub (https://github.com/egpivo/bc-cpit) and PyPI (https://pypi.org/project/bc-cpit).

comparisons. Section 4 gives the calibration, fixed-level coverage, and stability theory. Section 5 presents simulation studies that separate global bias, covariate-dependent bias, and distributional shape misspecification. Section 6 applies CPIT to WeatherBench 2 precipitation forecasts and evaluates distributional calibration, interval summaries, exceedance probabilities, and threshold-coherence. Section 7 concludes with limitations and directions for future work.

2. Proposed Method

2.1. Statistical setup

Let (X_i, Y_i), $i = 1, \ldots, n$, denote observations, where $X_i \in \mathcal{X}$ contains the covariates used for prediction and $Y_i \in \mathbb{R}$ is the scalar response. We assume access to a fitted sample-only predictive model. For any covariate value x, this model returns an m-vector of predictive draws,

$$\widehat{\boldsymbol{Y}}(x) = \big(\widehat{Y}^{(1)}(x), \ldots, \widehat{Y}^{(m)}(x)\big) \sim \widehat{Q}_x^{(m)}.$$

Here $\widehat{Q}_x^{(m)}$ denotes the joint law of the m generated values induced by the fitted model at x. The generator may be biased or miscalibrated, and $\widehat{Q}_x^{(m)}$ need not match the true conditional law of $Y \mid X = x$. In particular, we do not require the generated values to be conditionally independent.

For the observed case i, we write

$$\widehat{Y}_i^{(b)} = \widehat{Y}^{(b)}(X_i), \qquad b = 1, \ldots, m.$$

Thus the basic forecast-response object is

$$\mathcal{O}_i = \big(X_i, Y_i, \widehat{Y}_i^{(1)}, \ldots, \widehat{Y}_i^{(m)}\big). \tag{1}$$

If the fitted generator is randomized, its Monte Carlo randomness is included in $\mathcal{O}_i$. The conformal guarantees below are conditional on the fitted generator and require exchangeability of the corresponding forecast-response objects across the calibration and test cases.

We use four disjoint splits, $\mathcal{I}_{tr}$, $\mathcal{I}_{\text{bias}}$, $\mathcal{I}_{cal}$, and $\mathcal{I}_{test}$, for training, bias correction, PIT calibration, and testing. When exact fixed-level interval coverage is desired in addition to the CPIT predictive CDF, Section 3 describes a final split-conformal wrapper that uses an additional interval-calibration split $\mathcal{I}_{\text{int}}$, or a held-out portion of the available calibration data. This optional interval split is not used to estimate the CPIT calibration map. All finite-sample statements below are conditional on the fitted generator and bias-correction rule.

2.2. Location-scale bias correction

Before applying the affine correction, we allow an optional monotone transformation of the response. Let $T(\cdot)$ be a fixed, strictly increasing transformation from the support of Y to an interval on the real line, with inverse $T^{-1}(\cdot)$. Examples include the identity transformation, a log transformation with an offset, or a Box-Cox transformation. The purpose of $T(\cdot)$ is to place the outcome on a scale where location and scale corrections are more appropriate, for example, by reducing skewness.

2.2.1. Global affine correction

Let

$$\widehat{\mu}(x) = \frac{1}{m}\sum_{j=1}^{m} T\big(\widehat{Y}^{(j)}(x)\big), \qquad \widehat{\sigma}^2(x) = \frac{1}{m-1}\sum_{j=1}^{m} \big\{T\big(\widehat{Y}^{(j)}(x)\big) - \widehat{\mu}(x)\big\}^2$$

denote the sample mean and variance of the transformed generator samples. The global correction applies a constant location shift and a positive scale factor on the transformed scale,

$$T\big(\widehat{Y}^{(j)}_{\text{adj}}(x)\big) = a_0 + b_0\, T\big(\widehat{Y}^{(j)}(x)\big), \qquad b_0 > 0.$$

On the bias split $\mathcal{I}_{\text{bias}}$, we estimate (a_0, b_0) by the heteroskedastic Gaussian quasi-likelihood on the transformed scale,

$$(\widehat{a}_0, \widehat{b}_0) \in \underset{a_0\in\mathbb{R},\, b_0>0}{\operatorname{arg\,min}} \sum_{i\in\mathcal{I}_{\text{bias}}} \left[2\log b_0 + \frac{\{T(Y_i) - a_0 - b_0\,\widehat{\mu}(X_i)\}^2}{b_0^2\,\widehat{\sigma}^2(X_i)}\right]. \tag{2}$$

For fixed b_0, the minimizer in a_0 is

$$\widehat{a}_0(b_0) = \frac{\sum_{i\in\mathcal{I}_{\text{bias}}} \big\{T(Y_i) - b_0\,\widehat{\mu}(X_i)\big\}/\widehat{\sigma}^2(X_i)}{\sum_{i\in\mathcal{I}_{\text{bias}}} 1/\widehat{\sigma}^2(X_i)},$$

so the criterion can be minimized by profiling over the single positive parameter b_0. The resulting corrected samples are therefore

$$\widehat{Y}^{(j)}_{\text{adj}}(x) = T^{-1}\big(\widehat{a}_0 + \widehat{b}_0\, T\big(\widehat{Y}^{(j)}(x)\big)\big), \qquad j = 1, \ldots, m.$$

2.2.2. Covariate-dependent correction

When the generator bias varies with the covariates, we allow the center and spread of the transformed sample distribution to depend on x. Let $\mu_{\text{adj}}(x)$ denote the adjusted transformed mean and let $h(x)$ denote the log-variance multiplier. We define

$$T\big(\widehat{Y}^{(j)}_{\text{adj}}(x)\big) = \mu_{\text{adj}}(x) + \exp\{h(x)/2\}\big\{T\big(\widehat{Y}^{(j)}(x)\big) - \widehat{\mu}(x)\big\}, \qquad j = 1, \ldots, m. \tag{3}$$

For each fixed x, this remains an affine transformation of the generated samples, with $\mu_{\mathrm{adj}}(x)$ and $\exp\{h(x)/2\}$ controlling their transformed center and spread, respectively.

Let $\boldsymbol{\phi}(x) = (\phi_1(x), \ldots, \phi_p(x))'$ be a low-dimensional feature map containing covariates or summaries of the generated sample. We use separate additive predictors for the adjusted center and the log-variance multiplier:

$$\mu_{\mathrm{adj}}(x) = \alpha_0 + \beta_0\, \widehat{\mu}(x) + \sum_{\ell=1}^{p} f_{m,\ell}\{\phi_\ell(x)\}, \qquad h(x) = \gamma_0 + \sum_{\ell=1}^{p} f_{h,\ell}\{\phi_\ell(x)\}. \tag{4}$$

Here β_0 calibrates the relation between the observed response and the generator center, whereas $\exp(\gamma_0/2)$ is the baseline spread multiplier. The smooth terms describe covariate-dependent departures from these baseline relationships and are centered over the bias-correction split for identifiability.

We estimate the mean and scale functions jointly by minimizing the penalized normal quasi-likelihood

$$\begin{aligned}(\widehat{\mu}_{\mathrm{adj}}, \widehat{h}) = \underset{\mu_{\mathrm{adj}}, h}{\arg\min} \Bigg\{ & \sum_{i \in \mathcal{I}_{\mathrm{bias}}} \left[h(X_i) + \frac{\{T(Y_i) - \mu_{\mathrm{adj}}(X_i)\}^2}{\widehat{\sigma}^2(X_i) \exp\{h(X_i)\}} \right] \\ & + \sum_{\ell=1}^{p} \lambda_{m,\ell} J_{m,\ell}(f_{m,\ell}) + \sum_{\ell=1}^{p} \lambda_{h,\ell} J_{h,\ell}(f_{h,\ell}) \Bigg\},\end{aligned} \tag{5}$$

where $J_{m,\ell}$ and $J_{h,\ell}$ are spline roughness penalties induced by (4). This criterion corresponds to the working model

$$T(Y_i) \mid X_i \approx N\big(\mu_{\mathrm{adj}}(X_i), \widehat{\sigma}^2(X_i) \exp\{h(X_i)\}\big)\,.$$

In practice, (5) can be optimized by alternating two standard GAM updates.

- Given a current estimate $h^{(t)}(\cdot)$, update $\mu_{\mathrm{adj}}(\cdot)$ by fitting a weighted Gaussian GAM with weights

$$w_i^{(t)} = \left[\widehat{\sigma}^2(X_i) \exp\{h^{(t)}(X_i)\}\right]^{-1}.$$

  Equivalently,

$$\mu_{\mathrm{adj}}^{(t+1)}(\cdot) = \underset{\mu_{\mathrm{adj}}(\cdot)}{\arg\min} \sum_{i \in \mathcal{I}_{\mathrm{bias}}} w_i^{(t)} \{T(Y_i) - \mu_{\mathrm{adj}}(X_i)\}^2 + \sum_{\ell=1}^{p} \lambda_{m,\ell} J_{m,\ell}(f_{m,\ell}).$$

- Given the updated mean fit $\mu_{\mathrm{adj}}^{(t+1)}(\cdot)$, form

$$q_i^{(t+1)} = \frac{\{T(Y_i) - \mu_{\mathrm{adj}}^{(t+1)}(X_i)\}^2}{\widehat{\sigma}^2(X_i)}, \qquad i \in \mathcal{I}_{\mathrm{bias}}.$$

Update $h(\cdot)$ by solving

$$h^{(t+1)}(\cdot) = \arg\min_{h(\cdot)} \sum_{i \in \mathcal{I}_{\text{bias}}} \left[q_i^{(t+1)} \exp\{-h(X_i)\} + h(X_i) \right] + \sum_{\ell=1}^{p} \lambda_{h,\ell} J_{h,\ell}(f_{h,\ell}).$$

This is equivalent to fitting a Gamma GAM with log link and response $q_i^{(t+1)}$, for which

$$\mathbb{E}\big(q_i^{(t+1)} \big| X_i\big) = \exp\{h(X_i)\}.$$

We initialize the iteration at the global affine fit,

$$\mu_{\text{adj}}^{(0)}(x) = \widehat{a}_0 + \widehat{b}_0\, \widehat{\mu}(x), \qquad h^{(0)}(x) = 2 \log \widehat{b}_0,$$

and alternate the mean and scale updates until the criterion in (5) stabilizes. With the resulting estimates $\widehat{\mu}_{\text{adj}}$ and $\widehat{h}$, the covariate-dependent corrected samples are

$$\widehat{Y}_{\text{adj}}^{(j)}(x) = T^{-1}\Big(\widehat{\mu}_{\text{adj}}(x) + \exp\{\widehat{h}(x)/2\} \left\{ T\big(\widehat{Y}^{(j)}(x)\big) - \widehat{\mu}(x) \right\} \Big), \qquad j = 1, \ldots, m. \tag{6}$$

With either the global or covariate-dependent transformed-scale correction, the bias-corrected empirical base CDF is

$$\widehat{F}_x^{\text{adj}}(y) = \frac{1}{m} \sum_{j=1}^{m} \mathbf{1}\big\{ \widehat{Y}_{\text{adj}}^{(j)}(x) \le y \big\}. \tag{7}$$

This distribution serves as the base predictive CDF to be calibrated in the next step.

2.3. Conformal calibrator and calibrated predictive CDF

For each calibration pair (X_i, Y_i) with $i \in \mathcal{I}_{\text{cal}}$, we draw m samples from the generator and apply the bias correction from Section 2.2. Let $\hat{Y}_{i,\text{adj}}^{(1)}, \ldots, \hat{Y}_{i,\text{adj}}^{(m)}$ denote the resulting adjusted samples at X_i. Define

$$N_i = \#\{j : \hat{Y}_{i,\text{adj}}^{(j)} < Y_i\}, \qquad N_i^* = \#\{j : \hat{Y}_{i,\text{adj}}^{(j)} = Y_i\}.$$

The rank-cell randomized PIT is

$$u_i = \frac{N_i + V_i(N_i^* + 1)}{m + 1}, \qquad V_i \sim \text{Unif}(0, 1), \tag{8}$$

where the V_i are independent across i and independent of the data. When there are no ties, $N_i^* = 0$ and $u_i = (N_i + V_i)/(m + 1)$. Thus u_i is uniformly randomized within the rank cell indexed by N_i. Under an ideal continuous generator for which $\{Y_i, \hat{Y}_{i,\text{adj}}^{(1)}, \ldots, \hat{Y}_{i,\text{adj}}^{(m)}\}$ are exchangeable, N_i is uniform on $\{0, \ldots, m\}$, and hence $u_i \sim \text{Unif}(0, 1)$ exactly.

Algorithm 1 CPIT (Bias-Corrected Conformal PIT Calibration)

Require: Training split, bias split $\mathcal{I}_{\text{bias}}$, calibration split $\mathcal{I}_{\text{cal}}$, generator sample size m

1: Fit generator G on the training split
2: Estimate either $(\widehat{a}_0, \widehat{b}_0)$ by (2) or $\widehat{\mu}_{\text{adj}}(\cdot), \widehat{h}(\cdot)$ by (6) on $\mathcal{I}_{\text{bias}}$
3: **for** $i \in \mathcal{I}_{\text{cal}}$ **do**
4: Draw $\hat{y}^{(1:m)}(X_i) \sim \widehat{Q}^{(m)}_{X_i}$ and apply the bias correction
5: Compute u_i using (8)
6: **end for**
7: Construct $\widehat{C}$ by (9)
8: For a new x, draw $\hat{Y}^{(1:m)}(x)$, apply the bias correction, and output the calibrated predictive CDF $\widetilde{F}_x$ with rank-cell weights $w_1, \ldots, w_m$ by (11)-(12)

Using the calibration PIT values $\{u_i : i \in \mathcal{I}_{\text{cal}}\}$, define the conformal calibration map

$$\widehat{C}(t) = \begin{cases} 0, & t = 0, \\ \dfrac{1 + \sum_{i \in \mathcal{I}_{\text{cal}}} \mathbf{1}\{u_i \le t\}}{|\mathcal{I}_{\text{cal}}| + 1}, & 0 < t \le 1. \end{cases} \tag{9}$$

The randomized PIT in (8) is defined on an $(m+1)$-cell rank grid, whereas the adjusted empirical distribution is supported on m generated values. To align these two representations, define the finite-support calibration map

$$\widehat{C}_m(t) = \frac{\widehat{C}(mt/(m+1))}{\widehat{C}(m/(m+1))}, \qquad 0 \le t \le 1. \tag{10}$$

Because $\widehat{C}_m$ is nondecreasing and satisfies $\widehat{C}_m(0) = 0$ and $\widehat{C}_m(1) = 1$, it is a proper calibration map on $[0, 1]$.

For a new covariate value x, the CPIT predictive CDF is

$$\widetilde{F}_x(y) = \widehat{C}_m\big(\widehat{F}^{\text{adj}}_x(y)\big). \tag{11}$$

The rescaling by $m/(m+1)$ maps the empirical-CDF grid j/m to the rank-cell grid $j/(m+1)$, while the denominator normalizes the resulting finite-support distribution. When $\widehat{C}$ is the identity map, $\widehat{C}_m$ is also the identity, and (11) reduces exactly to the adjusted empirical CDF $\widehat{F}^{\text{adj}}_x$ in (7). The finite-sample result in Theorem 1 concerns the randomized conformal transform $\widehat{C}(u)$. Equation (11) is its deterministic finite-support projection and is the predictive CDF used for downstream distributional summaries.

Algorithm 1 summarizes the complete procedure.

2.4. Quantiles, highest-density intervals, and calibrated resampling

A central advantage of CPIT is that it outputs a calibrated predictive distribution rather than only a prediction interval at a single nominal level. This section describes how to com-

pute quantiles, equal-tail summaries, HDIs, calibrated resamples, and a smoothed calibrated CDF from $\widetilde{F}_x$.

Let $\widehat{\boldsymbol{Y}}_{\mathrm{adj}}(x) = \{\widehat{Y}^{(1)}_{\mathrm{adj}}(x), \ldots, \widehat{Y}^{(m)}_{\mathrm{adj}}(x)\}$ be the bias-corrected generator samples at covariate value x. We write

$$\widehat{Y}_{\mathrm{adj},(1)}(x) \le \cdots \le \widehat{Y}_{\mathrm{adj},(m)}(x)$$

for their order statistics, with ties counted according to their multiplicity. The normalized rank-cell weights are

$$w_j = \widehat{C}_m\left(\frac{j}{m}\right) - \widehat{C}_m\left(\frac{j-1}{m}\right), \qquad j = 1, \ldots, m. \tag{12}$$

Because $\widehat{C}$ is nondecreasing, the weights are nonnegative and sum to one. Thus the calibrated predictive distribution can be written as the weighted empirical measure

$$\widetilde{P}_x = \sum_{j=1}^{m} w_j \, \delta_{\widehat{Y}_{\mathrm{adj},(j)}(x)}, \tag{13}$$

with CDF

$$\widetilde{F}_x(y) = \sum_{j=1}^{m} w_j \, \mathbf{1}\{\widehat{Y}_{\mathrm{adj},(j)}(x) \le y\}, \tag{14}$$

which is the same as that in (11). The same telescoping argument remains valid when adjusted draws are tied, because all copies of a tied value are consecutive in the ordered sample. If the conformal calibration map is the identity, then $w_j = 1/m$ for all j. CPIT therefore reduces to the ordinary empirical distribution of the adjusted generator draws.

The corresponding quantile is

$$\widetilde{Q}_x(\alpha) = \inf\{y : \widetilde{F}_x(y) \ge \alpha\}, \qquad \alpha \in (0,1).$$

An equal-tail $(1-\alpha)$ empirical CPIT interval is therefore

$$\widetilde{I}_\alpha(x) = \big[\widetilde{Q}_x(\alpha/2),\, \widetilde{Q}_x(1-\alpha/2)\big]. \tag{15}$$

The intervals in (15) are summaries of the calibrated predictive CDF rather than separate split-conformal prediction sets. Section 3 shows how to add a final PIT-centrality conformal layer when exact fixed-level coverage is desired.

The equal-tail CPIT intervals in (15) are only one type of interval that can be extracted from the calibrated predictive distribution. Because CPIT returns a full weighted predictive law, we can also construct a highest-density-region (HDR) interval, defined here as the shortest interval whose calibrated predictive mass is at least $1-\alpha$. With $w_0 = 0$, define

$$(k_\alpha, \ell_\alpha) \in \operatorname*{arg\,min}_{1 \le k \le \ell \le m} \Big\{ \widehat{Y}_{\mathrm{adj},(\ell)}(x) - \widehat{Y}_{\mathrm{adj},(k)}(x) : w_k + \cdots + w_\ell \ge 1 - \alpha \Big\}, \tag{16}$$

and the empirical CPIT HDR interval is

$$\widetilde{H}_\alpha(x) = \big[\widehat{Y}_{\mathrm{adj},(k_\alpha)}(x), \widehat{Y}_{\mathrm{adj},(\ell_\alpha)}(x)\big]. \tag{17}$$

After sorting, (16) can be computed by a one-pass two-pointer search over the cumulative weights. This construction highlights a key distinction between CPIT and interval-only conformal methods. Once a calibrated predictive law is available, intervals can be chosen by optimizing over calibrated probability mass, rather than being fixed in advance by a single nonconformity score.

The same weights yield calibrated resampling. To draw from $\widetilde{P}_x$, sample $J \in \{1, \ldots, m\}$ with probabilities $w_1, \ldots, w_m$ and return $\widehat{Y}_{\mathrm{adj},(j)}(x)$. More generally,

$$\mathbb{E}_{\widetilde{P}_x}\{f(Y)\} = \sum_{j=1}^{m} w_j \, f\big(\widehat{Y}_{\mathrm{adj},(j)}(x)\big)$$

for any measurable function f. This weighted representation is useful in downstream Monte Carlo tasks such as estimating exceedance probabilities, computing risk measures, or propagating predictive uncertainty through a decision rule.

2.5. Smoothed weighted CPIT distribution

The weighted empirical distribution in (14) is supported only on the adjusted generator draws. To obtain a continuous predictive law with full support, we smooth this distribution by Gaussian convolution. For a bandwidth $\tau_x > 0$, define

$$\widetilde{F}_x^{\tau}(y) = \sum_{j=1}^{m} w_j \Phi\left(\frac{y - \widehat{Y}_{\mathrm{adj},(j)}(x)}{\tau_x}\right), \tag{18}$$

where Φ denotes the standard normal CDF. Let $\widetilde{P}_x^{\tau}$ denote the probability distribution associated with (18). Because the weights are nonnegative and sum to one, $\widetilde{F}_x^{\tau}$ is a proper continuous CDF. Moreover, as $\tau_x \to 0$, $\widetilde{F}_x^{\tau}(y)$ converges to the weighted empirical CDF at each of its continuity points.

For the Gaussian kernel in (18), the normal-reference rule for kernel distribution-function estimation gives

$$\tau_x = 4^{1/3} \widehat{s}_x m^{-1/3},$$

where $\widehat{s}_x$ is the unweighted sample standard deviation of the adjusted generator draws $\widehat{Y}_{\mathrm{adj}}^{(1)}(x), \ldots, \widehat{Y}_{\mathrm{adj}}^{(m)}(x)$ (Lopez-de Ullibarri, 2015). Because this rule is derived for an ordinary unweighted kernel CDF estimator, we use it as a simple scale-adaptive default rather than as an optimal bandwidth for the calibration-dependent CPIT weights. More data-adaptive

alternatives include plug-in bandwidth selection (Altman and Leger, 1995) and CDF-specific cross-validation (Bowman et al., 1998).

The corresponding smoothed quantile function is

$$\widetilde{Q}_x^\tau(q) = \inf\left\{y : \widetilde{F}_x^\tau(y) \geq q\right\}, \qquad 0 < q < 1.$$

Hence, the smoothed equal-tail $(1-\alpha)$ CPIT interval is

$$\widetilde{I}_\alpha^\tau(x) = \left[\widetilde{Q}_x^\tau(\alpha/2), \widetilde{Q}_x^\tau(1-\alpha/2)\right]. \tag{19}$$

The density associated with (18) is the Gaussian mixture

$$\widetilde{f}_x^\tau(y) = \sum_{j=1}^m w_j \tau_x^{-1} \phi\left(\frac{y - \widehat{Y}_{\text{adj},(j)}(x)}{\tau_x}\right),$$

where ϕ is the standard normal density. A smoothed highest-density region is

$$\widetilde{H}_\alpha^\tau(x) = \left\{y : \widetilde{f}_x^\tau(y) \geq \lambda_\alpha(x)\right\},$$

where

$$\lambda_\alpha(x) = \sup\left\{\lambda \geq 0 : \widetilde{P}_x^\tau\left(\widetilde{f}_x^\tau(Y) \geq \lambda\right) \geq 1-\alpha\right\}.$$

For a multimodal predictive distribution, $\widetilde{H}_\alpha^\tau(x)$ may be a union of disjoint intervals. The smoothed distribution can be sampled by first drawing J with probabilities $w_1, \ldots, w_m$ and then drawing from $N\left(\widehat{Y}_{\text{adj},(J)}(x), \tau_x^2\right)$.

3. Fixed-Level Conformal Prediction

This section collects fixed-level split-conformal constructions for sample-based predictive distributions. The first construction is an optional wrapper for CPIT. It uses the calibrated CDF only through a scalar PIT-centrality score and therefore gives the usual finite-sample marginal coverage guarantee at any chosen significance level. The remaining constructions are interval-only conformal baselines used in the empirical comparisons. They are useful fixed-level competitors, but their output is a prediction set at a selected nominal level rather than a calibrated predictive CDF.

3.1. Generic split-conformal construction

Let $\mathcal{I}_{conf}$ be a conformal calibration split that is not used to fit the prediction rule or score function. For a fixed nonconformity score $s(x, y)$, compute

$$S_i = s(X_i, Y_i), \qquad i \in \mathcal{I}_{conf},$$

and let $S_{(1)} \le \cdots \le S_{(N_{conf})}$ denote their order statistics, where $N_{conf} = |\mathcal{I}_{conf}|$. For a target miscoverage level α, set

$$k_\alpha = \lceil (N_{conf}+1)(1-\alpha) \rceil, \qquad q_{\alpha,s} = \begin{cases} S_{(k_\alpha)}, & k_\alpha \le N_{conf}, \\ +\infty, & k_\alpha > N_{conf}. \end{cases}$$

The split-conformal prediction set is

$$\mathcal{C}_{\alpha,s}(x) = \{y : s(x,y) \le q_{\alpha,s}\}.$$

Under exchangeability of the conformal calibration cases and the test case, this set has marginal coverage at least $1-\alpha$, conditional on all data used to construct the score. This generic construction will be used in two ways. For the CPIT-centrality wrapper below, $\mathcal{I}_{conf} = \mathcal{I}_{int}$, which is separate from the PIT-calibration split. For the interval-only baselines, $\mathcal{I}_{conf} = \mathcal{I}_{cal}$ is the usual conformal calibration split for the corresponding fixed-level score.

3.2. PIT-centrality conformal intervals from CPIT

The equal-tail CPIT interval in (15) is a distributional summary of $\widetilde{F}_x$. If exact finite-sample coverage is required at a chosen level, we can add a final split-conformal layer after the CPIT CDF has been constructed. Let $\mathcal{I}_{int}$ be an interval-calibration split not used to estimate the affine correction or the CPIT calibration map, and define the PIT-centrality score

$$S_{\mathrm{pit}}(x,y) = \left|2\widetilde{F}_x(y) - 1\right|. \tag{20}$$

Let q_α^{pit} be the split-conformal quantile of $S_{\mathrm{pit}}(X_i, Y_i)$, $i \in \mathcal{I}_{int}$, computed as in Section 3.1. The resulting conformalized CPIT set is

$$\mathcal{C}_\alpha^{\mathrm{pit}}(x) = \left\{y : S_{\mathrm{pit}}(x,y) \le q_\alpha^{\mathrm{pit}}\right\}. \tag{21}$$

For a continuous and strictly increasing $\widetilde{F}_x$, this set is the interval

$$\mathcal{C}_\alpha^{\mathrm{pit}}(x) = \left[\widetilde{Q}_x\big((1-q_\alpha^{\mathrm{pit}})/2\big), \widetilde{Q}_x\big((1+q_\alpha^{\mathrm{pit}})/2\big)\right]. \tag{22}$$

For a general right-continuous CDF, (21) is the exact conformal set and should be used directly. A closed interval containing this set can instead be formed from the lower endpoint $\inf\{y : \widetilde{F}_x(y) \ge (1-q_\alpha^{\mathrm{pit}})/2\}$ and the upper endpoint $\sup\{y : \widetilde{F}_x(y) \le (1+q_\alpha^{\mathrm{pit}})/2\}$. The same calibration scores can be queried at several values of α, producing a nested family because q_α^{pit} is monotone in $1-\alpha$. Theorem 2 gives the finite-sample coverage guarantee for the exact score sublevel set at each chosen level. Thus CPIT supplies the calibrated predictive CDF, while the PIT-centrality wrapper supplies fixed-level marginal coverage when it is needed.

3.3. Interval-only conformal baselines

For the interval-only baselines, the split $\mathcal{I}_{\text{cal}} = \mathcal{I}_{\text{conf}}$ is used directly as the split-conformal calibration set. For each $i \in \mathcal{I}_{\text{cal}}$, obtain the adjusted predictive samples $\widehat{Y}^{(1)}_{\text{adj}}(X_i), \ldots, \widehat{Y}^{(m)}_{\text{adj}}(X_i)$. The same notation covers the Raw, Global, and GAM baselines by taking the bias correction to be, respectively, the identity, the global correction, or the covariate-dependent correction.

Recall the adjusted empirical CDF $\widehat{F}^{\text{adj}}_x$ from (7), and define its quantile function by

$$\widehat{Q}^{\text{adj}}_x(u) = \inf\left\{y : \widehat{F}^{\text{adj}}_x(y) \geq u\right\}, \qquad 0 < u < 1.$$

Quantile residual score

For a target miscoverage level α, the quantile residual score is

$$s_{\text{QR},\alpha}(x, y) = \max\left\{\widehat{U}_{\text{adj},\alpha/2}(x) - y,\ y - \widehat{Q}_{\text{adj},(1-\alpha)/2}(x)\right\}. \tag{23}$$

This is the conformalized quantile regression score (Romano et al., 2019), applied here to empirical quantiles of the bias-adjusted generator samples. Because the score is signed, the conformal correction may be negative. Thus the calibrated interval can shrink an overly conservative base interval as well as expand an undercovering one.

Let q^{QR}_α be the split-conformal quantile computed from the calibration scores $s_{\text{QR},\alpha}(X_i, Y_i)$. The resulting prediction interval is

$$\mathcal{C}^{\text{QR}}_\alpha(x) = \left[\widehat{Q}_{\text{adj},\alpha/2}(x) - q^{\text{QR}}_\alpha,\ \widehat{Q}_{\text{adj},(1-\alpha)/2}(x) + q^{\text{QR}}_\alpha\right],$$

with the convention that the interval is empty if the lower endpoint exceeds the upper endpoint.

Scaled residual score

The scaled residual score uses the empirical center and spread of the bias-adjusted sample cloud:

$$s_{\text{SR}}(x, y) = \frac{|y - \widehat{\mu}_{\text{adj}}(x)|}{\widehat{\sigma}_{\text{adj}}(x)}. \tag{24}$$

This score adapts the residual magnitude to the local dispersion of the generator samples. Let q^{SR}_α be the split-conformal quantile of the scaled residual scores. Inverting $s_{\text{SR}}(x, y) \leq q^{\text{SR}}_\alpha$ gives

$$\mathcal{C}^{\text{SR}}_\alpha(x) = \left[\widehat{\mu}_{\text{adj}}(x) - q^{\text{SR}}_\alpha \widehat{\sigma}_{\text{adj}}(x),\ \widehat{\mu}_{\text{adj}}(x) + q^{\text{SR}}_\alpha \widehat{\sigma}_{\text{adj}}(x)\right].$$

Empirical CRPS score

The empirical CRPS score evaluates the full bias-adjusted sample cloud rather than only its center and spread (Gneiting and Raftery, 2007):

$$s_{\mathrm{CRPS}}(x,y) = \frac{1}{m}\sum_{b=1}^{m} |\widehat{Y}_{\mathrm{adj}}^{(b)}(x) - y| - \frac{1}{2m^2}\sum_{b=1}^{m}\sum_{b'=1}^{m} |\widehat{Y}_{\mathrm{adj}}^{(b)}(x) - \widehat{Y}_{\mathrm{adj}}^{(b')}(x)|. \tag{25}$$

For fixed samples $\widehat{Y}_{\mathrm{adj}}^{(1:m)}(x)$, the function $s_{\mathrm{CRPS}}(x,y)$ is convex and piecewise linear in y. Therefore, the split-conformal prediction set

$$\mathcal{C}_\alpha^{\mathrm{CRPS}}(x) = \{y : \ s_{\mathrm{CRPS}}(x,y) \le q_\alpha^{\mathrm{CRPS}}\}$$

is an interval, where q_α^{CRPS} is the split-conformal quantile of the empirical CRPS scores.

4. Theory

This section establishes the theoretical guarantees for CPIT and the optional fixed-level construction in Section 3. We first prove a finite-sample rank-calibration result for the conformally transformed randomized PIT. We then establish marginal coverage and nesting for the PIT-centrality split-conformal wrapper. Finally, we study the weighted CPIT law, showing that Gaussian smoothing yields a full-support distribution and that estimation error in the one-dimensional calibration map propagates stably to the predictive CDF.

Let $\mathcal{D}_{tr}$ and $\mathcal{D}_{bias}$ collect the data and auxiliary randomness used to fit the generator and the bias-correction rule. These quantities are treated as fixed in the conditional statements below. We write $N = |\mathcal{I}_{cal}|$.

4.1. Finite-sample PIT calibration

For a calibrated continuous predictive distribution, the randomized PIT is uniform on $[0,1]$. Because CPIT starts from an empirical distribution supported on generated values, its basic finite-sample statement is instead a conformal rank result. The next theorem concerns the transformed value $\widehat{C}(u_{n+1})$, not the raw randomized PIT u_{n+1}.

Theorem 1 (Finite-sample CPIT rank calibration)**.** *For every $i \in \mathcal{I}_{\mathrm{cal}} \cup \{n+1\}$, let u_i be defined by (8) using the auxiliary variable V_i. Suppose that, conditional on $\mathcal{D}_{\mathrm{tr}}$ and $\mathcal{D}_{\mathrm{bias}}$,*

$$\{(\mathcal{O}_i, V_i) : i \in \mathcal{I}_{\mathrm{cal}} \cup \{n+1\}\}$$

is exchangeable, where $\mathcal{O}_i$ is defined in (1). Then the calibration map $\widehat{C}$ in (9) satisfies

$$\mathbb{P}\Big\{\widehat{C}(u_{n+1}) \le t \,\Big|\, \mathcal{D}_{\mathrm{tr}}, \mathcal{D}_{\mathrm{bias}}\Big\} \le t, \qquad t \in [0,1]. \tag{26}$$

Moreover, conditional on $\mathcal{D}_{\mathrm{tr}}$ and $\mathcal{D}_{\mathrm{bias}}$, $\widehat{C}(u_{n+1})$ is discrete uniform on $\{1/(N+1), \ldots, 1\}$.

Theorem 1 is the finite-sample calibration-in-probability statement for CPIT. The conformal transform of the future PIT is super-uniform, and in fact exactly uniform on the conformal rank grid. The theorem does not claim that the untransformed PIT u_{n+1} is uniform under misspecification.

4.2. Finite-sample fixed-level coverage from PIT centrality

When coverage at a prespecified level is required, the CPIT CDF can be followed by a separate split-conformal layer. This layer uses the CDF only through the scalar PIT-centrality score, so its validity follows from the usual exchangeable rank argument. The CPIT CDF remains the object used for probability, risk, and other distributional summaries.

Theorem 2 (PIT-centrality conformal sets)**.** *Let $\widetilde{F}_x$ be an empirical or smoothed CPIT predictive CDF constructed without using the interval-calibration split $\mathcal{I}_{\text{int}}$. Let $\mathcal{D}_{CPIT}$ denote all data and auxiliary randomness used to construct the CPIT rule, excluding $\mathcal{I}_{\text{int}}$ and the test case. Suppose that, conditional on $\mathcal{D}_{CPIT}$,*

$$\{\mathcal{O}_i : i \in \mathcal{I}_{int} \cup \{n+1\}\}$$

is exchangeable. For $i \in \mathcal{I}_{\text{int}}$, define the exact score sublevel set by (21). Then

$$\mathbb{P}\big\{Y_{n+1} \in \mathcal{C}_\alpha^{pit}(X_{n+1}) \,\big|\, \mathcal{D}_{CPIT}\big\} \geq 1-\alpha. \tag{27}$$

If, in addition, the $N_{\text{int}}+1$ calibration and test scores are almost surely distinct conditional on $\mathcal{D}_{CPIT}$, then

$$\mathbb{P}\big\{Y_{n+1} \in \mathcal{C}_\alpha^{pit}(X_{n+1}) \,\big|\, \mathcal{D}_{CPIT}\big\} \leq 1-\alpha+\frac{1}{|\mathcal{I}_{int}|+1}.$$

Moreover, if $\alpha_1 < \alpha_2$, then $\mathcal{C}_{\alpha_1}^{pit}(x) \supseteq \mathcal{C}_{\alpha_2}^{pit}(x)$ for every x.

For a continuous and strictly increasing $\widetilde{F}_x$, the exact score set is the interval in (22). For a discontinuous CDF, the score sublevel set in (21) is the object covered by the theorem; replacing it by a larger closed interval preserves the lower coverage bound but can invalidate the upper bound. Reusing the same calibration scores across several values of α gives nested sets with levelwise marginal guarantees, not a simultaneous coverage statement for the entire random family.

4.3. Smoothed weighted CPIT CDF

The weighted CPIT law $\widetilde{P}_x$ is discrete because it is supported on the m adjusted generator order statistics. Gaussian smoothing converts this law into a continuous full-support

distribution. We quantify the perturbation using the 1-Wasserstein distance. For probability measures P and Q on $\mathbb{R}$ with finite first moments,

$$W_1(P,Q) = \inf_{\gamma \in \Gamma(P,Q)} \int |u-v| \, d\gamma(u,v),$$

where $\Gamma(P,Q)$ is the set of couplings of P and Q. By the Kantorovich–Rubinstein duality,

$$W_1(P,Q) = \sup_{\|h\|_{\mathrm{Lip}} \le 1} \left| \mathbb{E}_P\{h(Y)\} - \mathbb{E}_Q\{h(Y)\} \right|.$$

Thus W_1 directly controls the error of Lipschitz downstream summaries (see, e.g., Villani (2009)).

Proposition 1. *For any x and $\tau_x > 0$, the smoothed CPIT CDF $\widetilde{F}_x^{\tau}$ in (18) is the CDF of an absolutely continuous distribution with density*

$$\widetilde{f}_x^{\tau}(y) = \sum_{j=1}^{m} w_j \, \tau_x^{-1} \phi\left(\frac{y - \widehat{Y}_{\mathrm{adj},(j)}(x)}{\tau_x} \right). \tag{28}$$

The distribution has support $\mathbb{R}$, and $\widetilde{F}_x^{\tau}$ is strictly increasing. Let $\widetilde{P}_x$ be the weighted empirical distribution in (13), and let $\widetilde{P}_x^{\tau}$ be the distribution with CDF $\widetilde{F}_x^{\tau}$. Then

$$W_1(\widetilde{P}_x^{\tau}, \widetilde{P}_x) \le \tau_x \sqrt{2/\pi}. \tag{29}$$

Consequently, for every L-Lipschitz function h,

$$\left| \mathbb{E}_{\widetilde{P}_x^{\tau}}\{h(Y)\} - \mathbb{E}_{\widetilde{P}_x}\{h(Y)\} \right| \le L \tau_x \sqrt{2/\pi}. \tag{30}$$

Proposition 1 is a stability statement rather than an additional conformal guarantee. It shows that smoothing regularizes the predictive law while perturbing any Lipschitz summary by at most a quantity proportional to the bandwidth.

4.4. Stability with respect to the calibration map

The CPIT predictive distribution depends on the fitted calibration map only through the normalized rank-cell weights in (12). The next theorem shows that uniform estimation of this one-dimensional map yields uniform control of the resulting predictive CDF. The normalization by $\widehat{C}\{m/(m+1)\}$ conditions the first m rank-cell masses on the portion of the grid represented by the m generated order statistics. The remaining rank cell, corresponding to a response above all generated values, has no separate atom in the finite-support predictive law.

Theorem 3 (Stability of the weighted CPIT CDF)**.** *Suppose that, conditional on* $\mathcal{D}_{\text{tr}}$ *and* $\mathcal{D}_{bias}$*, the calibration PIT values are independent with common CDF*

$$C^\star(t) = \mathbb{P}\{u_i \le t \mid \mathcal{D}_{\text{tr}}, \mathcal{D}_{bias}\}, \qquad t \in [0,1].$$

Assume $C^\star\{m/(m+1)\} > 0$*, and define the oracle normalized rank-cell weights by*

$$w_j^\star = \frac{C^\star\{j/(m+1)\} - C^\star\{(j-1)/(m+1)\}}{C^\star\{m/(m+1)\}}, \qquad j = 1, \dots, m.$$

Let $F^\tau_{x,C^\star}$ *be the smoothed CDF formed from these oracle weights, using the same ordered adjusted samples and bandwidths as* $\widetilde{F}^\tau_x$*. For* $\varepsilon > 0$*, define*

$$\delta_{N,\varepsilon} = \varepsilon + \frac{1}{N+1}.$$

If $\delta_{N,\varepsilon} < C^\star\{m/(m+1)\}$*, then for every collection* $\mathcal{X}_0$ *of covariate values,*

$$\mathbb{P}\left\{ \sup_{x\in\mathcal{X}_0} \sup_{y\in\mathbb{R}} \left|\widetilde{F}^\tau_x(y) - F^\tau_{x,C^\star}(y)\right| > \frac{2\delta_{N,\varepsilon}}{C^\star\{m/(m+1)\} - \delta_{N,\varepsilon}} \,\middle|\, \mathcal{D}_{\text{tr}}, \mathcal{D}_{bias} \right\} \le 2\exp(-2N\varepsilon^2). \tag{31}$$

The same bound holds for the corresponding unsmoothed weighted empirical CDFs.

The theorem compares the fitted CPIT CDF with the oracle finite-support CDF obtained from the population PIT calibration map $C^\star$. The term $(N+1)^{-1}$ is the conformal rank correction in $\widehat{C}$. The bound is uniform in the covariate values, adjusted sample locations, and bandwidths because these quantities enter both CDFs identically, with only their common weights differing. Together, Theorems 1 and 3 separate two roles of the calibration sample. Conformal ranking gives finite-sample calibration in probability, while a one-dimensional empirical-process bound controls estimation of the predictive CDF.

To distinguish this calibration-map error from other sources of approximation, let

$$F^0_x(y) = \mathbb{P}(Y \le y \mid X = x)$$

be the true conditional CDF, let H_x be the population CDF of the bias-adjusted generator, and define the population oracle recalibration

$$F^\star_x(y) = C^\star\{H_x(y)\}.$$

For each x,

$$\sup_y |\widetilde{F}^\tau_x(y) - F^0_x(y)| \le \sup_y |\widetilde{F}^\tau_x(y) - F^\star_x(y)| + \sup_y |F^\star_x(y) - F^0_x(y)|.$$

The first term is controlled by Theorem 3. The second term is an oracle approximation error. It vanishes only when the true conditional law can be represented as a global PIT recalibration of the bias-adjusted generator,

$$F_x^0 = C^\star \circ H_x.$$

Thus CPIT provides finite-sample rank calibration and a stable predictive law, whereas closeness to the full conditional distribution additionally depends on the adequacy of the bias-adjusted generator and the use of a global calibration map.

5. Simulation Studies

5.1. Setup and metrics

We use three controlled simulations to separate the main ways in which a sample-only generator can fail. In each Monte Carlo replicate, the generator is specified analytically, so no training split is needed, and the source of misspecification is known. The data are split into bias, calibration, and test sets with

$$n_{\text{bias}} = 1000, \qquad n_{\text{cal}} = N = 1000, \qquad n_{\text{test}} = 5000.$$

Unless otherwise stated, we use $m = 100$ generator samples per covariate value and repeat each configuration over 100 replications. All post-processing methods receive only the generator samples and the observed responses.

The three designs are as follows.

1. Global location-scale distortion: $X \sim \text{Unif}(0,1)$, $Y \mid X = x \sim N\{\mu(x), 0.15^2\}$, and $\mu(x) = \sin(2\pi x)$. The generator is globally biased:

$$\widehat{Y} \mid X = x \sim N\left(\frac{\mu(x) - 0.25}{1.3}, \left(\frac{0.15}{1.3}\right)^2\right).$$

   This design is favorable to a global affine correction.

2. Covariate-dependent location-scale distortion: $X \sim \text{Unif}(0,1)$, $Y = \sin(2\pi X) + \sigma(X)\varepsilon$, $\sigma(x) = 0.05 + 0.25x$, and $\varepsilon \sim N(0,1)$. The generator has an x-dependent mean and scale error:

$$\widehat{Y} \mid X = x \sim N\Big(0.85\sin(2\pi x) + 0.15x - 0.05, \{0.08 + 0.10(1-x)\}^2\Big).$$

   This design requires covariate-adaptive correction.

3. Distributional shape misspecification: $X \sim \mathrm{Unif}(0,1)$, $Y = \sin(2\pi X) + \sigma(X)(\eta - 1)$, $\sigma(x) = 0.10 + 0.20x$, and $\eta \sim \mathrm{Exp}(1)$. The generator matches the first two conditional moments but imposes a Gaussian predictive shape:

$$\widehat{Y} \mid X = x \sim N\big(\sin(2\pi x), \sigma^2(x)\big).$$

Thus affine correction alone cannot repair the skewed conditional law.

We compare the raw generator, global and GAM affine bias corrections, CPIT applied after each bias correction, and the fixed-level split-conformal baselines in Section 3. We assess distributional calibration using randomized PIT histograms, the Cramér–von Mises (CvM) distance of the PIT distribution from uniformity, and quantile calibration error (QErr). Specifically, QErr is the mean absolute difference between the empirical coverage of each predictive quantile and its nominal level, averaged over $\{0.05, 0.10, \ldots, 0.95\}$. We also report mean CRPS, central interval coverage and length, and local diagnostics within five bins of X. Detailed metric definitions, interval results for $\alpha \in \{0.02, 0.05, 0.10, 0.20\}$, local bin diagnostics, and the Simulation 3 upper-tail QErr results are provided in the supplementary material.

5.2. Results

Tables 1 and 2 summarize the global distributional diagnostics and the nominal 90% interval results, respectively. Figure 1 shows the corresponding randomized PIT histograms, averaged over the 100 Monte Carlo replications.

In Simulation 1, the global affine model is correctly specified. BC(Global) reduces CvM from 474.5615 to 0.5884, QErr from 0.2745 to 0.0098, and mean CRPS from 0.1915 to 0.0854. CPIT retains essentially the same CRPS and yields nearly uniform PIT histograms. Its CvM values, 0.9462 and 0.9433, are somewhat larger than the 0.5884 attained by the correctly specified global affine correction. As shown in Table 2, CPIT also moves central 90% coverage closer to the nominal level: coverage increases from 0.883 and 0.885 under BC(Global) and BC(GAM) to 0.890 and 0.891 under the corresponding CPIT corrections, with only modest increases in average length.

Simulation 2 highlights the need for covariate-adaptive correction. BC(Global) improves the raw generator but leaves appreciable miscalibration, with CvM 3.8295, QErr 0.0243, and central 90% coverage 0.801. BC(GAM) lowers CvM and QErr to 0.6950 and 0.0101, respectively, and raises coverage to 0.885. Both CPIT variants yield strong global calibration, with CvM below 0.96 and QErr 0.0111. CPIT(GAM) additionally preserves the lower mean CRPS of 0.1002 and moves coverage to 0.891, while maintaining a relatively small average length. The supplementary binwise diagnostics reveal a distinction that is not apparent

Table 1: Global distributional diagnostics for the three simulations over 100 replications. Entries are means, with Monte Carlo standard errors in parentheses. Smaller values indicate better performance for CvM, QErr, and mean CRPS.

Simulation	Method	CvM	QErr	Mean CRPS
1	Raw	474.5615 (1.3530)	0.2745 (0.0004)	0.1915 (0.0002)
	BC(Global)	0.5884 (0.0701)	0.0098 (0.0005)	0.0854 (0.0001)
	BC(GAM)	0.6671 (0.0729)	0.0100 (0.0005)	0.0856 (0.0001)
	CPIT(Global)	0.9462 (0.0828)	0.0114 (0.0005)	0.0854 (0.0001)
	CPIT(GAM)	0.9433 (0.0832)	0.0111 (0.0005)	0.0857 (0.0001)
2	Raw	47.3461 (0.2167)	0.0885 (0.0002)	0.1436 (0.0002)
	BC(Global)	3.8295 (0.1877)	0.0243 (0.0007)	0.1138 (0.0002)
	BC(GAM)	0.6950 (0.0754)	0.0101 (0.0005)	0.1002 (0.0001)
	CPIT(Global)	0.8816 (0.0743)	0.0111 (0.0004)	0.1135 (0.0002)
	CPIT(GAM)	0.9550 (0.0810)	0.0111 (0.0005)	0.1002 (0.0001)
3	Raw	35.6368 (0.2793)	0.0757 (0.0003)	0.1066 (0.0002)
	BC(Global)	35.0022 (0.5079)	0.0746 (0.0006)	0.1066 (0.0002)
	BC(GAM)	31.4727 (0.6454)	0.0706 (0.0007)	0.1069 (0.0002)
	CPIT(Global)	1.0241 (0.1086)	0.0120 (0.0005)	0.1023 (0.0002)
	CPIT(GAM)	1.0251 (0.1130)	0.0120 (0.0006)	0.1028 (0.0002)

from the global summaries: CPIT(Global) remains uneven across X, whereas BC(GAM) and CPIT(GAM) achieve nearly uniform coverage across the five covariate bins. Thus, favorable marginal diagnostics can mask residual covariate-dependent miscalibration.

Simulation 3 isolates shape misspecification that affine location-scale correction cannot remove. Bias correction alone has little effect: CvM remains above 31, QErr remains near 0.07, and mean CRPS is essentially unchanged. By contrast, CPIT reduces CvM to about 1.02 and QErr to 0.0120, while also lowering mean CRPS to 0.1023 under global correction and 0.1028 under GAM correction. Supplementary Table S7 reports upper-tail calibration over quantile levels $\tau \in \{0.96, 0.97, 0.98, 0.99\}$. The corresponding upper-tail QErr decreases from 0.0328 for the raw generator and 0.0331 after BC(Global) to 0.0134 for CPIT(Global) and 0.0162 for CPIT(GAM). The central CPIT intervals in Table 2 have 90% coverages 0.888 and 0.887, with average lengths 0.5867 and 0.5956, respectively. The smoothed HDR variants reported in the supplementary material have coverages 0.916 and 0.915. Together, these results illustrate CPIT's role as a distributional recalibrator rather than merely a location-scale adjustment.

Table 2 also compares CPIT with the interval-only QR, SR, and eCRPS split-conformal baselines. As expected, these methods attain coverage close to the nominal 90% level in all three simulations. CPIT's central intervals have coverage between 0.887 and 0.891 and are competitive in length in Simulations 1 and 2, particularly after GAM correction in Simulation 2; the score-based baselines are in general slightly shorter in Simulation 3. The purpose of this comparison is therefore not to suggest that nominal fixed-level coverage is

Table 2: Nominal 90% interval coverage and average length for the three simulations. Each entry reports mean coverage followed by mean length, separated by a slash. Standard errors, additional nominal levels, and HDR variants are reported in the supplementary material.

Method	Simulation 1	Simulation 2	Simulation 3
Raw	0.444 / 0.3683	0.641 / 0.4150	0.924 / 0.6384
BC(Global)	0.883 / 0.4794	0.801 / 0.5463	0.923 / 0.6386
BC(GAM)	0.885 / 0.4838	0.885 / 0.5668	0.917 / 0.6264
CPIT(Global)	0.890 / 0.4901	0.887 / 0.7735	0.888 / 0.5867
CPIT(GAM)	0.891 / 0.4930	0.891 / 0.5783	0.887 / 0.5956
QR(Raw)	0.900 / 0.9677	0.902 / 0.8571	0.901 / 0.5584
QR(Global)	0.901 / 0.5052	0.902 / 0.7592	0.901 / 0.5612
QR(GAM)	0.901 / 0.5076	0.902 / 0.5908	0.901 / 0.5763
SR(Raw)	0.899 / 0.9705	0.902 / 0.9839	0.900 / 0.5302
SR(Global)	0.901 / 0.5007	0.902 / 0.8164	0.900 / 0.5338
SR(GAM)	0.901 / 0.5033	0.902 / 0.5902	0.900 / 0.5516
eCRPS(Raw)	0.899 / 0.9635	0.902 / 0.8035	0.899 / 0.5361
eCRPS(Global)	0.900 / 0.4961	0.901 / 0.6816	0.899 / 0.5371
eCRPS(GAM)	0.900 / 0.4975	0.902 / 0.6002	0.899 / 0.5422

difficult to obtain, but to identify what CPIT adds beyond it. Unlike the interval-only baselines, CPIT returns a calibrated predictive CDF, so its quantiles, intervals, threshold probabilities, tail summaries, and resamples all arise from a single coherent predictive law.

6. WeatherBench 2 Precipitation Applications

We evaluate CPIT using WeatherBench 2 forecasts of 24-hour accumulated precipitation from the 50-member ECMWF Integrated Forecasting System ensemble (IFS-ENS), with ERA5 reanalysis fields used for verification (Rasp et al., 2024). The forecasts are initialized twice daily, at 00 and 12 UTC, and evaluated at a 24-hour lead over 2018-2022. Forecasts and verifications are represented on the 1.5° equiangular grid comprising 240 longitude points and 121 latitude points, including both poles. The data are available through the WeatherBench 2 data archive.[2] After removing three initialization times for which either the IFS-ENS forecast or the corresponding ERA5 verification is unavailable, $n = 3{,}649$ forecast-verification cases remain. We randomly partition these cases into a bias-correction set ($n_{\text{bias}} = 1{,}204$), a PIT-calibration set ($n_{\text{cal}} = N = 1{,}204$), and a test set ($n_{\text{test}} = 1{,}241$).

For each forecast-verification case, we map the full spatial field to a scalar target. The sample-only predictive distribution for that target is the empirical distribution of the $m = 50$ IFS-ENS ensemble members after the same spatial mapping. Let $Z_i(s)$ denote the ERA5 verifying precipitation at grid cell s, and let $\widehat{Z}_i^{(b)}(s)$, $b = 1, \ldots, m$, denote the corresponding

[2] https://weatherbench2.readthedocs.io/en/latest/data-guide.html

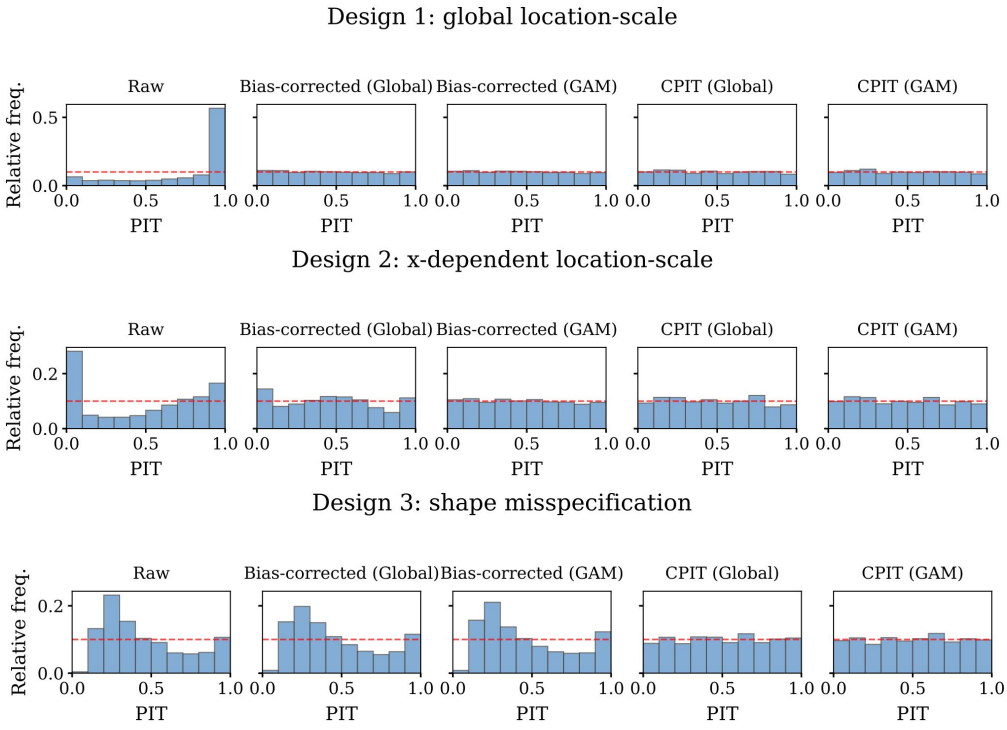

Figure 1: Randomized PIT histograms for the three simulation designs, averaged over 100 replications. Within each row, the panels show the raw generator, global and GAM bias correction, and CPIT after each bias correction. The dashed line indicates the uniform reference.

IFS-ENS ensemble precipitation forecasts. For target functional T_ℓ, define

$$Y_{i\ell} = T_\ell(Z_i), \qquad \widehat{Y}_{i\ell}^{(b)} = T_\ell(\widehat{Z}_i^{(b)}), \qquad b = 1, \ldots, m.$$

Thus $Y_{i\ell}$ is the scalar verifying response, while $\widehat{Y}_{i\ell}^{(1)}, \ldots, \widehat{Y}_{i\ell}^{(m)}$ are the sample-only predictive draws supplied by IFS-ENS for the same target and verification time. The covariate $X_{i\ell}$ collects the information available for post-processing at forecast initialization, including the date, the ensemble forecast, and auxiliary atmospheric fields. For the GAM affine correction, this information is summarized by a low-dimensional feature vector.

We consider two precipitation experiments. The Europe experiment uses the region 35°-75°N and 12.5°W-42.5°E. It evaluates both the area-weighted regional mean, T_{EU}^{mean}, and the regional upper-tail target, $T_{EU}^{0.95}$. The Taiwan experiment focuses on four individual WeatherBench-2 grid cells covering western and eastern Taiwan. Their domains are shown in Figure 2.

For a region R, let a_s denote the area weight of grid cell s. The area-weighted mean target is

$$T_R^{mean}(Z_i) = \frac{\sum_{s \in R} a_s Z_i(s)}{\sum_{s \in R} a_s}.$$

We also consider the weighted empirical upper-tail summary $T_R^q(Z_i)$, defined as the weighted

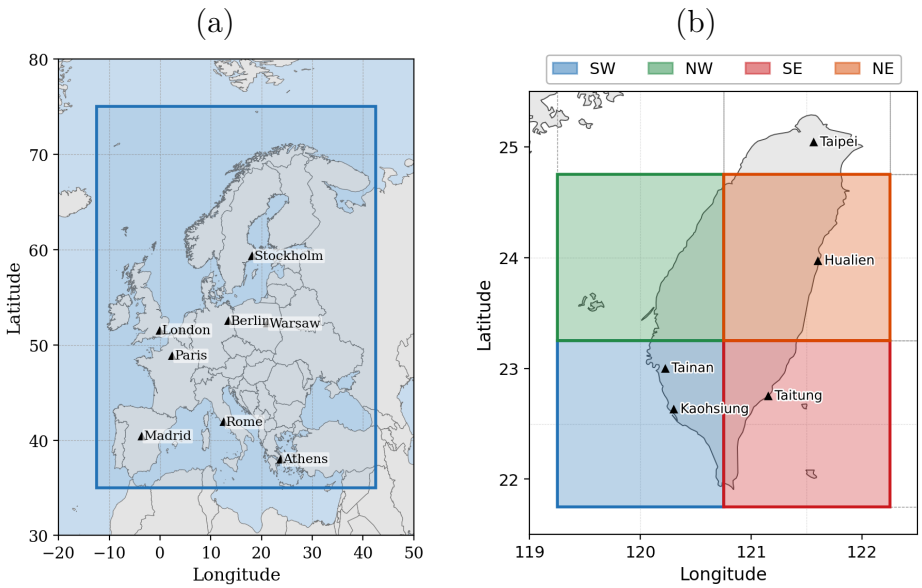

Figure 2: (a) Europe WeatherBench-2 region used for area-weighted aggregation: 35°-75°N and −12.5°-42.5°E. (b) Four WB2 1.5° grid cells used in the Taiwan experiment. The cells are SW (22.5°N, 120°E), NW (24°N, 120°E), SE (22.5°N, 121.5°E), and NE (24°N, 121.5°E).

q-quantile of $\{Z_i(s) : s \in R\}$ with weights $\{a_s : s \in R\}$. For the Taiwan four-cell experiment, each target is a singleton cell. If s_ℓ is one of the four selected Taiwan cells, then $T_\ell(Z_i) = Z_i(s_\ell)$.

For both Europe and Taiwan, precipitation is retained in meters for CPIT calibration and evaluation, while the affine correction is fitted on the log-millimeter scale

$$g(y) = \log(y + 0.1),$$

where y is precipitation in millimeters. On the transformed scale, define

$$\widehat{G}_{i\ell}^{(b)} = g\big(\widehat{Y}_{i\ell}^{(b)}\big), \qquad \widehat{\mu}_{i\ell} = \frac{1}{m}\sum_{b=1}^{m} \widehat{G}_{i\ell}^{(b)},$$

and

$$\widehat{\sigma}_{i\ell}^2 = \frac{1}{m-1}\sum_{b=1}^{m} \big(\widehat{G}_{i\ell}^{(b)} - \widehat{\mu}_{i\ell}\big)^2.$$

For Europe, the GAM affine correction uses the three-feature vector

$$\phi_{i\ell}^{\mathrm{EU}} = \big(d_i, \widehat{\mu}_{i\ell}, \log(\widehat{\sigma}_{i\ell} + \varepsilon)\big)',$$

whereas for Taiwan it uses the six-feature vector

$$\phi_{i\ell}^{\mathrm{TW}} = \big(d_i, \widehat{\mu}_{i\ell}, \log(\widehat{\sigma}_{i\ell} + \varepsilon), u_{850,i\ell}, v_{850,i\ell}, \mathrm{TCWV}_{i\ell}\big)'.$$

Here $\varepsilon > 0$ is a small numerical constant. In both analyses, d_i denotes day of year and is modeled with a cyclic P-spline to capture annual periodicity. Each smooth term uses five spline basis functions, and the smoothing penalty is selected by GCV. The variables $u_{850,i\ell}$ and $v_{850,i\ell}$ are the local zonal and meridional wind components at 850 hPa, and $\mathrm{T}CWV_{i\ell}$ denotes total column water vapor.

We assess threshold-event performance using Brier skill scores. In the formulas below, c is expressed on the original millimeter scale. For threshold c, let

$$\widehat{\pi}_{i\ell}(c) = 1 - \widetilde{F}^{\tau}_{i\ell}(c)$$

denote the CPIT exceedance probability on the original precipitation scale, where $\widetilde{F}^{\tau}_{i\ell}$ is the smoothed weighted CDF in (18). The Brier score is

$$\mathrm{BS}_{\ell}(c) = \frac{1}{n_{test}} \sum_{i \in \mathcal{I}_{test}} \left\{ \widehat{\pi}_{i\ell}(c) - \mathbf{1}(Y_{i\ell} > c) \right\}^2,$$

and the Brier skill score is

$$\mathrm{BSS}_{\ell}(c) = 1 - \frac{\mathrm{BS}_{\ell}(c)}{\mathrm{BS}_{\ell,\mathrm{ref}}(c)}.$$

Here $\mathrm{BS}_{\ell,\mathrm{ref}}(c)$ is the Brier score of the climatological reference forecast, using the empirical exceedance rate from the combined bias and calibration splits. For Taiwan, we consider $c \in \{10, 20, 40, 80\}$. The 80 mm event occurs only about five times in the 1241 Taiwan test cases, so BSS at this threshold is statistically unstable and can become negative after only a small number of false alarms.

6.1. Europe regional targets

The European experiment evaluates CPIT for two regional precipitation functionals with distinct meteorological interpretations. The domain shown in Figure 2(a), spanning 35°-75°N and 12.5°W-42.5°E, covers the North Atlantic-European storm track, the Mediterranean basin, and northern Europe. All spatial summaries use area weights, so that each grid cell contributes according to its physical area rather than receiving equal weight on the latitude-longitude grid. The regional mean summarizes the domain-wide 24-hour precipitation burden. The p95 target, defined as the area-weighted 95th percentile of 24-hour precipitation across grid cells, summarizes the spatial upper tail and is therefore more sensitive to localized precipitation maxima associated with frontal, convective, or orographic processes.

Table 3 reports complementary diagnostics of distributional reliability, sharpness, and threshold-event skill. We use randomized PIT CvM and QErr to assess full-distribution

Table 3: WeatherBench-2 Europe results for various methods.

Target	Method	CvM	QErr	90% Cov	Mean CRPS	BSS(> 2.5)	BSS(> 3.0)	BSS(> 3.5)	BSS(> 4.0)
Mean	Raw ensemble	2.8318	0.0344	0.929	0.05	0.892	0.904	0.895	0.886
Mean	BC (Global)	2.9386	0.0374	0.932	0.05	–	–	–	–
Mean	BC (GAM)	0.3303	0.0164	0.867	0.04	–	–	–	–
Mean	CPIT (Raw)	0.8739	0.0250	0.892	0.04	0.895	0.904	0.903	0.891
Mean	CPIT (Global)	0.7056	0.0225	0.879	0.04	0.895	0.908	0.911	0.890
Mean	CPIT (GAM)	0.9078	0.0260	0.882	0.04	0.900	0.914	0.918	0.885
p95	Raw ensemble	4.7708	0.0524	0.910	0.31	1.000	1.000	0.932	0.482
p95	BC (Global)	1.6337	0.0259	0.911	0.30	–	–	–	–
p95	BC (GAM)	0.1232	0.0090	0.856	0.29	–	–	–	–
p95	CPIT (Raw)	0.7052	0.0204	0.898	0.30	1.000	1.000	0.952	0.417
p95	CPIT (Global)	0.5433	0.0195	0.886	0.30	1.000	1.000	0.965	0.425
p95	CPIT (GAM)	0.4660	0.0188	0.870	0.29	1.000	1.000	0.911	0.529

and quantile calibration, empirical coverage of central 90% intervals to assess interval validity, mean CRPS to summarize proper-score performance, and Brier skill scores relative to empirical climatology for exceedance probabilities.

For the Europe mean target, the raw ensemble is slightly conservative at the central 90% level, with coverage 0.929, but its CvM statistic is 2.8318. The global affine correction does not improve this behavior, whereas BC(GAM) reduces CvM to 0.3303 and QErr to 0.0164 at the cost of undercoverage, 0.867. The unsmoothed CPIT variants reduce CvM to between 0.7056 and 0.9078 and QErr to between 0.0225 and 0.0260. Their central coverages, 0.879-0.892, are slightly below nominal, but smoothing raises them to 0.919-0.927 in Table 4. CPIT(Global) has the smallest CvM and QErr among the CPIT variants, while CPIT(GAM) has the strongest Brier skill at the 2.5, 3.0, and 3.5 mm thresholds.

The p95 target shows stronger distributional error in the raw ensemble. Its 90% coverage is 0.910, yet its CvM statistic is 4.7708. BC(GAM) produces very small PIT and quantile errors, CvM 0.1232 and QErr 0.0090, but undercovers at 0.856. CPIT reduces CvM to 0.7052, 0.5433, and 0.4660 for the raw, global, and GAM variants, respectively, while their unsmoothed central coverages are 0.898, 0.886, and 0.870. The smoothed variants raise these coverages to 0.919, 0.920, and 0.904. CPIT(GAM) has the smallest p95 CvM and QErr and the largest BSS at the highest displayed threshold. The lower p95 thresholds are nearly saturated in this split, so their BSS values should be interpreted cautiously.

Table 4 shows the same validity-sharpness tradeoff for both targets. Unsmoothed central and HDR intervals are generally shorter than the raw intervals but tend to undercover. Kernel smoothing increases coverage with a moderate increase in length. For example, the p95 CPIT(Global) interval changes from 0.886/1.76 without smoothing to 0.920/1.99 with smoothing, whereas the p95 CPIT(GAM) interval changes from 0.870/1.64 to 0.904/1.76. Thus the updated results support using the unsmoothed law for compact empirical summaries and the smoothed law when interval calibration is the primary goal.

Table 4: WeatherBench-2 Europe 90% interval coverage and average length for various methods. Each entry is "coverage / length", with length in millimeters. HDR denotes the shortest weighted interval from the calibrated predictive distribution. Smooth indicates using the kernel-smoothed CPIT distribution.

Method	Mean	p95
Raw	0.929 / 0.32	0.910 / 1.94
BC (GAM)	0.867 / 0.23	0.856 / 1.56
CPIT (Raw)	0.892 / 0.28	0.898 / 1.87
CPIT (Raw, smooth)	0.925 / 0.31	0.919 / 2.01
CPIT (Raw, HDR)	0.874 / 0.26	0.862 / 1.66
CPIT (Raw, HDR, smooth)	0.927 / 0.31	0.922 / 1.98
CPIT (Global)	0.879 / 0.26	0.886 / 1.76
CPIT (Global, smooth)	0.927 / 0.31	0.920 / 1.99
CPIT (Global, HDR)	0.864 / 0.25	0.861 / 1.62
CPIT (Global, HDR, smooth)	0.928 / 0.31	0.922 / 1.96
CPIT (GAM)	0.882 / 0.24	0.870 / 1.64
CPIT (GAM, smooth)	0.919 / 0.26	0.904 / 1.76
CPIT (GAM, HDR)	0.851 / 0.22	0.836 / 1.47
CPIT (GAM, HDR, smooth)	0.915 / 0.26	0.911 / 1.74
QR (GAM)	0.905 / 0.26	0.886 / 1.68
SR (GAM)	0.904 / 0.25	0.885 / 1.62
eCRPS (GAM)	0.896 / 0.24	0.870 / 1.60

Table 5: Taiwan four-cell experiment: cell locations and observed exceedance frequencies in the test split.

Cell	Location	$P(Y > 10\text{m}m)$	$P(Y > 20\text{m}m)$	$P(Y > 40\text{m}m)$	$P(Y > 80\text{m}m)$
SW	(22.5°N, 120°E)	0.128	0.061	0.022	0.005
NW	(24°N, 120°E)	0.115	0.053	0.015	0.003
SE	(22.5°N, 121.5°E)	0.147	0.058	0.019	0.006
NE	(24°N, 121.5°E)	0.276	0.091	0.023	0.005

6.2. Taiwan four-cell experiment

Figure 2(b) shows the four WeatherBench-2 grid cells used in the Taiwan experiment. This setting is more local and meteorologically more demanding than the Europe regional-average experiment. Each target is a single 1.5° grid cell, so spatial averaging does not smooth out displacement error, land-sea contrast, or unresolved orographic effects. The four cells provide coarse proxies for southwestern Taiwan (SW), northwestern Taiwan (NW), southeastern Taiwan (SE), and northeastern Taiwan (NE).

Table 5 reports the observed exceedance frequencies in the test split. The NE cell is the wettest at the 10 mm threshold, with $P(Y > 10\text{m}m) = 0.276$. The 80 mm event is rare in all four cells, with frequencies between 0.003 and 0.006, corresponding to only about 4-7 events per cell in $n_{test} = 1241$ cases. The 80 mm BSS values should therefore be read as sensitivity diagnostics rather than stable estimates of operational tail skill.

Table 6 reports distributional calibration and probabilistic forecasting scores. The raw ensemble is substantially underdispersed in SW, NW, and NE, with central 90% coverages 0.763, 0.737, and 0.734. Their CvM statistics are 12.6724, 7.4014, and 5.3007, respectively. The SE cell is less severely miscalibrated but still undercovers at 0.820.

The global affine correction is unreliable in this local setting. It improves SW, SE, and

Table 6: WeatherBench-2 Taiwan four-cell results for various methods.

Cell	Method	CvM	QErr	90% Cov	Mean CRPS	BSS(> 10)	BSS(> 20)	BSS(> 40)	BSS(> 80)
SW	Raw ensemble	12.6724	0.0976	0.763	0.95	0.707	0.710	0.555	0.017
SW	BC (Global)	1.0747	0.0334	0.800	0.94	–	–	–	–
SW	BC (GAM)	0.2163	0.0084	0.871	0.94	–	–	–	–
SW	CPIT (Raw)	0.0900	0.0114	0.862	0.95	0.705	0.703	0.527	0.124
SW	CPIT (Global)	0.1482	0.0101	0.857	0.94	0.709	0.702	0.536	0.106
SW	CPIT (GAM)	0.1679	0.0137	0.870	0.94	0.706	0.720	0.540	0.034
NW	Raw ensemble	7.4014	0.0761	0.737	0.85	0.695	0.715	0.601	0.715
NW	BC (Global)	10.2603	0.0861	0.712	0.85	–	–	–	–
NW	BC (GAM)	0.3066	0.0159	0.876	0.82	–	–	–	–
NW	CPIT (Raw)	0.7067	0.0249	0.877	0.84	0.707	0.706	0.601	0.694
NW	CPIT (Global)	0.6024	0.0244	0.853	0.85	0.709	0.704	0.590	0.680
NW	CPIT (GAM)	0.2306	0.0130	0.890	0.82	0.713	0.727	0.605	0.756
SE	Raw ensemble	0.4966	0.0253	0.820	1.14	0.687	0.583	0.533	-0.370
SE	BC (Global)	0.2791	0.0216	0.823	1.13	–	–	–	–
SE	BC (GAM)	0.2145	0.0102	0.879	1.10	–	–	–	–
SE	CPIT (Raw)	0.1770	0.0128	0.873	1.13	0.685	0.588	0.538	-0.390
SE	CPIT (Global)	0.1749	0.0111	0.874	1.14	0.684	0.588	0.543	-0.561
SE	CPIT (GAM)	0.1139	0.0092	0.889	1.10	0.697	0.617	0.543	-0.915
NE	Raw ensemble	5.3007	0.0654	0.734	1.75	0.513	0.491	0.498	0.620
NE	BC (Global)	4.5055	0.0639	0.731	1.78	–	–	–	–
NE	BC (GAM)	0.1711	0.0179	0.839	1.66	–	–	–	–
NE	CPIT (Raw)	0.3065	0.0214	0.827	1.75	0.528	0.482	0.467	0.623
NE	CPIT (Global)	0.2593	0.0181	0.856	1.75	0.537	0.478	0.414	0.612
NE	CPIT (GAM)	0.1044	0.0120	0.895	1.66	0.580	0.492	0.503	0.324

NE to varying degrees but worsens NW, where CvM increases from 7.4014 to 10.2603. By contrast, BC(GAM) has average CvM 0.2271, compared with 6.4678 for the raw ensemble, and average QErr 0.0131, compared with 0.0661. Its average 90% coverage is 0.8663, so a flexible location-scale correction alone still does not fully calibrate predictive uncertainty.

Averaged over the four cells, the CvM statistics for CPIT(Raw), CPIT(Global), and CPIT(GAM) are 0.3201, 0.2962, and 0.1542, respectively. CPIT(GAM) therefore reduces average CvM by about 98% relative to the raw ensemble. Its average QErr is 0.0120. The average unsmoothed central coverage improves from 0.7635 for the raw ensemble to 0.8860 for CPIT(GAM), with NW and NE improving to 0.890 and 0.895. The smoothed CPIT(GAM) intervals in Table 7 raise average coverage further to 0.9275.

For the more frequent 10 and 20 mm events, CPIT(GAM) improves average BSS from 0.6505 to 0.6740 and from 0.6248 to 0.6390, respectively. At 40 mm the average BSS is essentially unchanged. At 80 mm, a few false alarms or misses produce large cell-to-cell changes, including strongly negative values in SE, so this threshold is not used to rank methods.

The raw intervals have average coverage 0.7635 and average length 5.06 mm. For CPIT(GAM), the unsmoothed central intervals have average coverage 0.8860 and length 6.85 mm, and smoothing increases these to 0.9275 and 7.43 mm. The unsmoothed HDR intervals are shorter, with average length 5.68 mm, but undercover at 0.8520. Smoothed HDR intervals provide a more balanced compromise, with average coverage 0.9168 and length 7.01

Table 7: WeatherBench-2 Taiwan 90% interval coverage and average length for the log-millimeter analysis. Each entry is coverage / length, with length in millimeters. HDR denotes the shortest weighted interval from the calibrated predictive distribution. Smooth intervals use the kernel-smoothed CPIT distribution.

Method	SW	NW	SE	NE
Raw	0.763 / 4.23	0.737 / 3.78	0.820 / 5.57	0.734 / 6.65
BC (GAM)	0.871 / 5.28	0.876 / 4.57	0.879 / 6.34	0.839 / 8.57
CPIT (Raw)	0.862 / 4.88	0.877 / 5.87	0.873 / 6.88	0.827 / 9.64
CPIT (Raw, smooth)	0.919 / 5.48	0.922 / 5.83	0.898 / 7.36	0.857 / 10.18
CPIT (Raw, HDR)	0.814 / 3.96	0.828 / 4.35	0.834 / 5.44	0.794 / 8.23
CPIT (Raw, HDR, smooth)	0.910 / 5.09	0.912 / 5.25	0.892 / 6.82	0.865 / 9.76
CPIT (Global)	0.857 / 4.97	0.853 / 5.82	0.874 / 6.93	0.856 / 11.02
CPIT (Global, smooth)	0.918 / 5.68	0.915 / 5.83	0.898 / 7.39	0.874 / 11.07
CPIT (Global, HDR)	0.836 / 4.26	0.802 / 4.32	0.837 / 5.71	0.810 / 8.40
CPIT (Global, HDR, smooth)	0.908 / 5.22	0.907 / 5.25	0.889 / 6.88	0.873 / 10.25
CPIT (GAM)	0.870 / 5.43	0.890 / 4.84	0.889 / 6.68	0.895 / 10.46
CPIT (GAM, smooth)	0.923 / 6.13	0.941 / 5.48	0.929 / 7.48	0.917 / 10.64
CPIT (GAM, HDR)	0.844 / 4.51	0.857 / 4.06	0.863 / 5.89	0.844 / 8.26
CPIT (GAM, HDR, smooth)	0.913 / 5.76	0.930 / 5.17	0.921 / 7.11	0.903 / 10.01
QR (Raw)	0.891 / 4.65	0.915 / 4.23	0.892 / 6.15	0.882 / 9.08
QR (GAM)	0.882 / 5.32	0.922 / 4.68	0.911 / 6.57	0.895 / 9.29
SR (GAM)	0.887 / 5.25	0.929 / 4.84	0.910 / 6.67	0.887 / 9.07
eCRPS (GAM)	0.915 / 5.82	0.926 / 6.17	0.915 / 6.72	0.885 / 9.39

mm. These results show that smoothing is particularly useful when the finite 50-member ensemble limits the resolution of high-coverage intervals.

The QR, SR, and eCRPS baselines perform well under the fixed 90% interval criterion because each is calibrated specifically at that nominal level. Their outputs, however, are level-specific intervals rather than reusable predictive distributions. CPIT instead produces a single calibrated CDF from which quantiles, intervals, and exceedance probabilities at arbitrary thresholds can be derived coherently. For example, letting $\widehat{\pi}_i(c) = 1 - \widehat{F}_i(c)$ denote the predicted probability that precipitation exceeds c in forecast case i, the resulting probabilities are automatically monotone in the threshold:

$$\widehat{\pi}_i(10\,\text{mm}) \geq \widehat{\pi}_i(20\,\text{mm}) \geq \widehat{\pi}_i(40\,\text{mm}) \geq \widehat{\pi}_i(80\,\text{mm}).$$

6.3. Diagnostic plots

Figures 3 and 4 show the fitted calibration maps for selected test cases. The Europe GAM-adjusted profiles are close to the diagonal, while the Taiwan profiles show that the remaining departures after global correction vary by cell and are substantially reduced by the GAM correction. Figures 5 and 6 show the updated randomized PIT histograms. These plots agree with Tables 3 and 6. Specifically, bias correction removes much of the location-scale error, and CPIT flattens the remaining rank distortions.

7. Discussion

CPIT is a distributional calibration layer for sample-only predictors, not a fixed-level conformal interval method. Its primary output is a calibrated predictive CDF, from which

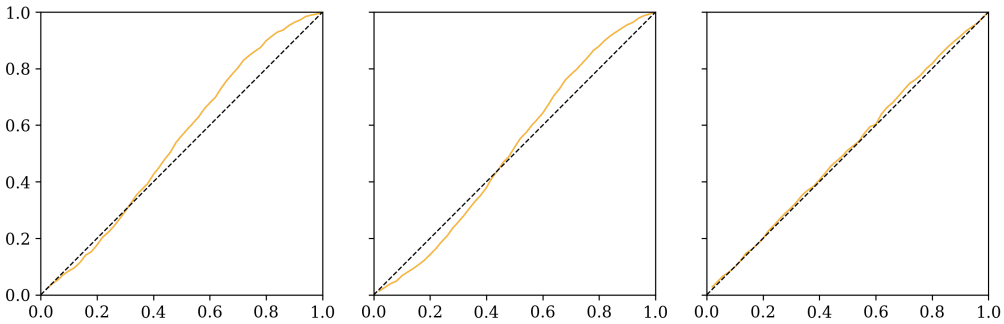

Figure 3: Europe empirical CPIT calibration profiles for the seed-123 log-millimeter mean-target analysis. Left to right: CPIT(Raw), CPIT(Global), and CPIT(GAM). The dashed line is the uniform reference $\widehat{C}(t) = t$.

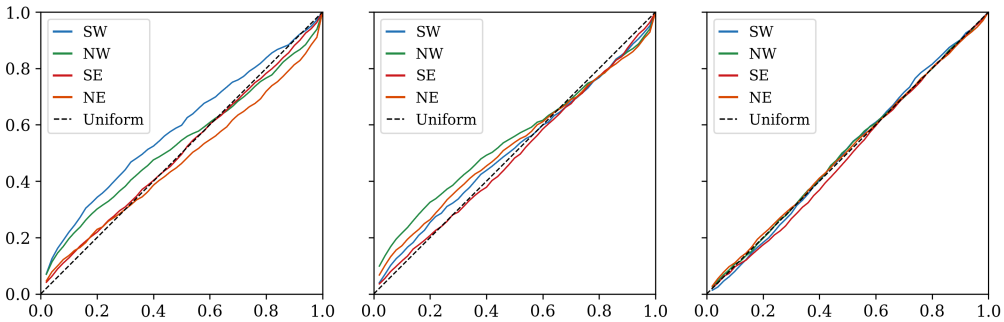

Figure 4: Taiwan empirical CPIT calibration profiles for the seed-123 log-millimeter mean-target analysis. Left to right: CPIT(Raw), CPIT(Global), and CPIT(GAM). Colours distinguish the SW, NW, SE, and NE cells.

one can compute threshold probabilities, quantiles at arbitrary levels, HDR intervals, tail probabilities, tail expectations, calibrated resamples, and decision-relevant risk summaries. This distinction is important in applications such as precipitation forecasting, where the same predictive law may be queried at many thresholds and risk levels, not just at one nominal coverage level. The finite-sample guarantee for CPIT is therefore stated in terms of calibrated PIT ranks, reflecting its goal of distributional reliability. When exact marginal coverage is also required at a specified level, the PIT-centrality conformal wrapper in Theorem 2 can be applied with a separate interval-calibration split to obtain a standard split-conformal prediction interval.

CPIT is also computationally simple. The calibrated predictive law is a weighted empirical distribution supported on the original generator draws, with optional smoothing for interpolation and more stable tail summaries. This makes the method easy to apply to black-box ensembles or simulation-based predictors, since it requires only predictive samples and a held-out calibration set. It also makes the graphical diagnostics transparent through PIT histograms, calibrated rank CDFs, interval-length comparisons, and threshold-event score curves, all of which evaluate different projections of the same calibrated distribution. This

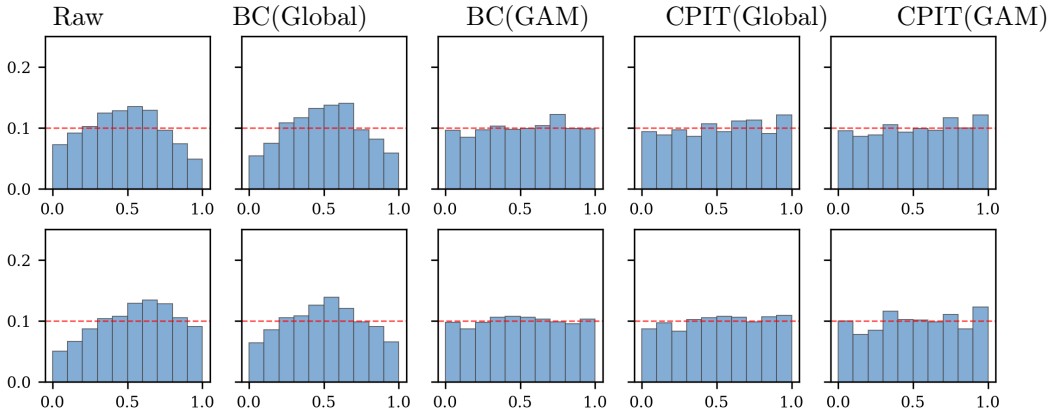

Figure 5: Europe randomized PIT histograms for an illustrative split. Top: area-weighted mean target. Bottom: regional p95 target. Each row compares the raw ensemble, bias-corrected distributions, and CPIT-calibrated distributions.

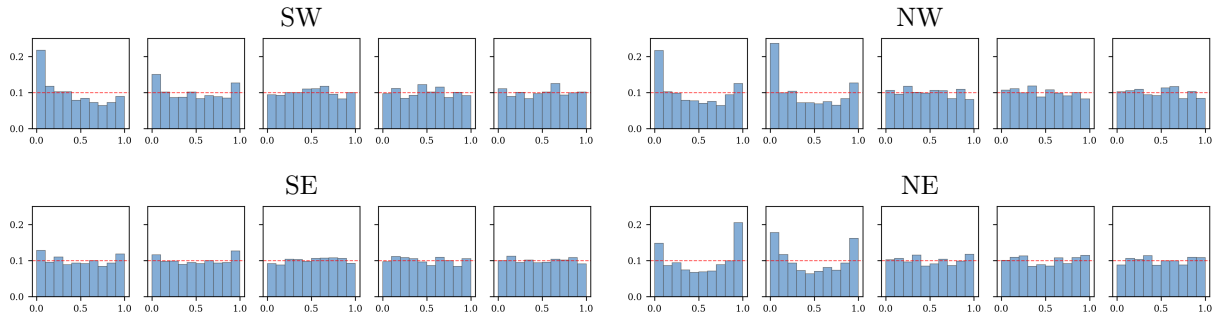

Figure 6: Taiwan randomized PIT histograms for the seed-123 log-millimeter analysis. The panels correspond to the SW, NW, SE, and NE cells. Within each panel, the columns are Raw, BC(Global), BC(GAM), CPIT(Global), and CPIT(GAM).

combination of reusable distributional output and graphical diagnostics is a key practical advantage of the method.

The number of generator draws m determines the resolution of the empirical CDF and the amount of available tail support. Larger m generally improves empirical quantiles, HDR intervals, calibrated resampling, and tail-risk summaries by providing a richer sample cloud. When m is small, or when the main targets are high-coverage intervals or rare-event probabilities, the smoothed weighted CDF can be more stable than the raw weighted empirical CDF. Smoothing should therefore be viewed as a numerical regularization step, not as a replacement for adequate ensemble diversity.

At the same time, CPIT cannot create information that is absent from the underlying sample cloud. If the generator misses important modes, has too few tail samples, or produces samples on a coarse finite support, the calibrated CDF can reweight and smooth those samples but cannot fully reconstruct the missing conditional distribution. This limitation is most visible for empirical equal-tail and HDR intervals when the number of generator draws m is small, or when very high coverage levels require extrapolation beyond the available

sample support. For this reason, interval coverage, interval length, PIT calibration, CRPS, and threshold-event scores should be reported together. Good performance on one summary need not imply good distributional calibration.

The present theory gives marginal calibration under exchangeability. This is the natural distribution-free target for a general post-processing layer, but it does not imply exact conditional calibration at each covariate value. More localized versions of CPIT could use covariate-dependent calibration maps, Mondrian partitions (Bostrom et al., 2021), or weighted calibration samples. Naive kernel- or nearest-neighbor-weighted PIT recalibration need not retain exact finite-sample marginal validity. Carefully constructed localized conformal procedures, however, can preserve finite-sample marginal coverage while improving local adaptivity (Guan, 2023). Exact distribution-free conditional coverage at every covariate value remains impossible without additional restrictions (Barber et al., 2021).

Another important direction is calibration under distribution shift. In many simulation-to-real and forecasting problems, the conditional distribution of the response given the predictive sample may be relatively stable, while the marginal distribution of covariates changes between calibration and deployment. Weighted conformal methods under covariate shift provide a natural route for adapting CPIT in this setting (Tibshirani et al., 2019). The same idea could be applied at the PIT level by weighting calibration cases based on their relevance to the deployment covariate distribution.

Overall, CPIT provides a computationally lightweight way to convert sample-only predictive output into a calibrated predictive distribution. Its main advantage is coherence. All reported quantities are derived from one CDF, so quantiles, intervals, exceedance probabilities, and resamples are mutually consistent. This makes CPIT especially useful as a post-processing step for modern ensemble, Monte Carlo, and generative prediction systems, where the predictor naturally returns samples, but downstream statistical analysis requires calibrated distributional summaries.

Data Availability Statement

The WeatherBench 2 data used in this study, including the ECMWF IFS-ENS forecasts and the corresponding ERA5 verification fields, are publicly available through the WeatherBench 2 data archive: https://weatherbench2.readthedocs.io/en/latest/data-guide.html.

Disclosure Statement

No potential conflict of interest was reported by the authors.

References

Altman, N. and Léger, C. (1995). Bandwidth selection for kernel distribution function estimation. *Journal of Statistical Planning and Inference*, 46, 195–214.

Barber, R. F., Candés, E. J., Ramdas, A., and Tibshirani, R. J. (2021). The limits of distribution-free conditional predictive inference. *Information and Inference*, 10, 455–482.

Boström, H., Johansson, U., and Löfström, T. (2021). Mondrian conformal predictive distributions. In *Proceedings of the Tenth Symposium on Conformal and Probabilistic Prediction and Applications*, PMLR 152, 24–38.

Bowman, A. W., Hall, P., and Prvan, T. (1998). Bandwidth selection for the smoothing of distribution functions. *Biometrika*, 85, 799–808.

Chernozhukov, V., Wüthrich, K., and Zhu, Y. (2021). Distributional conformal prediction. *Proceedings of the National Academy of Sciences*, 118, e2107794118.

Dvoretzky, A., Kiefer, J., and Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Annals of Mathematical Statistics*, 27, 642–669.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102, 359–378.

Guan, L. (2023). Localized conformal prediction: A generalized inference framework for conformal prediction. *Biometrika*, 110, 33–50.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113, 1094–1111.

López-de Ullibarri, I. (2015). Bandwidth selection in kernel distribution function estimation. *The Stata Journal*, 15, 784–795.

Massart, P. (1990). The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The Annals of Probability*, 18, 1269–1283.

Rasp, S., Hoyer, S., Merose, A., Langmore, I., Battaglia, P., Russell, T., Sanchez-Gonzalez, A., Yang, V., Carver, R., Agrawal, S., Chantry, M., Ben Bouallegue, Z., Dueben, P., Bromberg, C., Sisk, J., Barrington, L., Bell, A., and Sha, F. (2024). WeatherBench 2:

A benchmark for the next generation of data-driven global weather models. *Journal of Advances in Modeling Earth Systems*, 16, e2023MS004019.

Romano, Y., Patterson, E., and Candés, E. J. (2019). Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, 32.

Shafer, G. and Vovk, V. (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9, 371–421.

Tibshirani, R. J., Barber, R. F., Candés, E. J., and Ramdas, A. (2019). Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, 32.

Villani, C. (2009). *Optimal Transport: Old and New*. Springer.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic Learning in a Random World*. Springer.

Vovk, V., Shen, J., Manokhin, V., and Xie, M.-g. (2017). Nonparametric predictive distributions based on conformal prediction. In *Proceedings of the Sixth Workshop on Conformal and Probabilistic Prediction and Applications*, PMLR 60, 82–102.

Vovk, V., Petej, I., Toccaceli, P., Gammerman, A., Ahlberg, E., and Carlsson, L. (2020). Conformal calibrators. In *Proceedings of the Ninth Symposium on Conformal and Probabilistic Prediction and Applications*, PMLR 128, 84–99.

Wang, Z., Gao, R., Yin, M., Zhou, M., and Blei, D. (2023). Probabilistic conformal prediction using conditional random samples. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, PMLR 206, 8814–8836.

Appendix A. Proofs for Section 4

Proof of Theorem 1. Let

$$\mathcal{J} = \mathcal{I}_{cal} \cup \{n+1\}, \qquad |\mathcal{J}| = N+1.$$

Conditional on $\mathcal{D}_{\text{tr}}$ and $\mathcal{D}_{bias}$, the fitted generator and bias-correction rule are fixed. By assumption, $\{(\mathcal{O}_i, V_i) : i \in \mathcal{J}\}$ is exchangeable. The randomized PIT u_i is the same measurable function of $(\mathcal{O}_i, V_i)$ for every $i \in \mathcal{J}$, so $\{u_i : i \in \mathcal{J}\}$ is exchangeable. Moreover, each u_i lies in $(0, 1)$ almost surely.

For $r \in \mathcal{J}$, define its upper rank among the $N+1$ PIT values by

$$R_r^+ = \sum_{k \in \mathcal{J}} \mathbf{1}\{u_k \le u_r\}.$$

Because $u_{n+1} > 0$ almost surely, (9) gives

$$\widehat{C}(u_{n+1}) = \frac{1 + \sum_{i \in \mathcal{I}_{cal}} \mathbf{1}\{u_i \le u_{n+1}\}}{N+1} = \frac{R^+_{n+1}}{N+1}.$$

For any $k \in \{1, \dots, N+1\}$, at most k indices can have upper rank no larger than k. Exchangeability therefore implies

$$\begin{aligned}\mathbb{P}\{R^+_{n+1} \le k \mid \mathcal{D}_{\text{tr}}, \mathcal{D}_{bias}\} &= \frac{1}{N+1}\sum_{r \in \mathcal{J}} \mathbb{P}\{R^+_r \le k \mid \mathcal{D}_{\text{tr}}, \mathcal{D}_{\text{bias}}\} \\ &= \frac{1}{N+1}\mathbb{E}\left[\sum_{r \in \mathcal{J}} \mathbf{1}\{R^+_r \le k\} \,\middle|\, \mathcal{D}_{\text{tr}}, \mathcal{D}_{\text{bias}}\right] \le \frac{k}{N+1}.\end{aligned}$$

For $t \in [0,1]$, take $k = \lfloor (N+1)t \rfloor$. Then

$$\mathbb{P}\left\{\widehat{C}(u_{n+1}) \le t \,\middle|\, \mathcal{D}_{\text{tr}}, \mathcal{D}_{\text{bias}}\right\} \le \frac{k}{N+1} \le t,$$

which proves (26).

Conditional on all forecast-response objects, each u_i is a strictly increasing affine function of the independent continuous variable V_i. Hence the $N+1$ PIT values are almost surely distinct. Their ranks are therefore a uniformly random permutation of $\{1, \dots, N+1\}$, so R^+_{n+1} is uniform on this set. It follows that $\widehat{C}(u_{n+1})$ is uniform on $\{1/(N+1), \dots, 1\}$. □

Proof of Theorem 2. Conditional on $\mathcal{D}_{CPIT}$, the score rule $(x, y, \widehat{y}^{(1)}, \dots, \widehat{y}^{(m)}) \mapsto \left|2\widetilde{F}_x(y) - 1\right|$ is fixed and is applied identically to every case. Thus

$$\{S^{\text{pit}}_i : i \in \mathcal{I}_{int} \cup \{n+1\}\}$$

is exchangeable. By (21),

$$Y_{n+1} \in \mathcal{C}^{\text{pit}}_\alpha(X_{n+1}) \quad \Longleftrightarrow \quad S^{\text{pit}}_{n+1} \le q^{\text{pit}}_\alpha.$$

The standard split-conformal rank argument gives

$$\mathbb{P}\{S^{\text{pit}}_{n+1} \le q^{\text{pit}}_\alpha \mid \mathcal{D}_{CPIT}\} \ge \frac{k_\alpha}{N_{int}+1} \ge 1 - \alpha,$$

with the convention that the event is certain when $k_\alpha > N_{int}$ and $q^{\text{pit}}_\alpha = +\infty$. This proves (27).

If the $N_{\text{int}} + 1$ scores are almost surely distinct, the rank of the test score is uniform on $\{1, \dots, N_{\text{int}} + 1\}$. The coverage probability is then exactly $k_\alpha/(N_{\text{int}} + 1)$, which is at most $1 - \alpha + 1/(N_{\text{int}} + 1)$.

Finally, $\alpha_1 < \alpha_2$ implies $k_{\alpha_1} \ge k_{\alpha_2}$ and therefore $q^{\text{pit}}_{\alpha_1} \ge q^{\text{pit}}_{\alpha_2}$. The corresponding score sublevel sets are nested: $\mathcal{C}^{\text{pit}}_{\alpha_1}(x) \supseteq \mathcal{C}^{\text{pit}}_{\alpha_2}(x)$ for every x. □

Proof of Proposition 1. Fix x and write

$$z_j = \widehat{Y}_{\mathrm{adj},(j)}(x), \qquad j = 1, \dots, m.$$

The weights $w_1, \dots, w_m$ are nonnegative and sum to one because they are increments of the nondecreasing finite-support calibration map $\widehat{C}_m$. Consequently,

$$\widetilde{F}_x^{\tau}(y) = \sum_{j=1}^{m} w_j \Phi\left(\frac{y - z_j}{\tau_x}\right)$$

is a convex combination of Gaussian CDFs and hence is a proper CDF. Differentiating term by term gives (28). At least one weight is positive, and every Gaussian density is strictly positive on $\mathbb{R}$, so $\widetilde{f}_x^{\tau}(y) > 0$ for every y. Thus the distribution has full support and its CDF is strictly increasing.

Let J be a discrete random index with $\mathbb{P}(J = j) = w_j$, and let $Z \sim N(0,1)$ be independent of J. Define

$$\widetilde{Y}_x = z_J, \qquad \widetilde{Y}_x^{\tau} = z_J + \tau_x Z.$$

Then $\widetilde{Y}_x \sim \widetilde{P}_x$ and $\widetilde{Y}_x^{\tau} \sim \widetilde{P}_x^{\tau}$, so this construction is a coupling of the two distributions. Therefore

$$W_1(\widetilde{P}_x^{\tau}, \widetilde{P}_x) \le \mathbb{E}|\widetilde{Y}_x^{\tau} - \widetilde{Y}_x| = \tau_x \mathbb{E}|Z| = \tau_x \sqrt{2/\pi}.$$

If h is L-Lipschitz, the same coupling gives

$$\left|\mathbb{E}_{\widetilde{P}_x^{\tau}}\{h(Y)\} - \mathbb{E}_{\widetilde{P}_x}\{h(Y)\}\right| \le \mathbb{E}\left|h(\widetilde{Y}_x^{\tau}) - h(\widetilde{Y}_x)\right| \le L\tau_x\sqrt{2/\pi}.$$

□

Proof of Theorem 3. Let

$$\mathbb{C}_N(u) = \frac{1}{N} \sum_{i \in \mathcal{I}_{cal}} \mathbf{1}\{u_i \le u\}$$

be the ordinary empirical CDF of the calibration PIT values. Conditional on $\mathcal{D}_{\mathrm{tr}}$ and $\mathcal{D}_{bias}$, these values are independent with common CDF $C^\star$. The Dvoretzky-Kiefer-Wolfowitz (DKW) inequality (Dvoretzky et al., 1956; Massart, 1990) gives

$$\mathbb{P}\left\{ \sup_{u \in [0,1]} |\mathbb{C}_N(u) - C^\star(u)| > \varepsilon \,\middle|\, \mathcal{D}_{\mathrm{tr}}, \mathcal{D}_{bias} \right\} \le 2\exp(-2N\varepsilon^2).$$

For $u > 0$,

$$\widehat{C}(u) = \frac{1 + N\mathbb{C}_N(u)}{N+1},$$

and therefore

$$|\widehat{C}(u) - C^\star(u)| \le |\mathbb{C}_N(u) - C^\star(u)| + \frac{1}{N+1}.$$

At $u = 0$, both CDFs are zero. Hence, on the event

$$\mathcal{E}_\varepsilon = \left\{ \sup_{u\in[0,1]} |\mathbb{C}_N(u) - C^\star(u)| \le \varepsilon \right\},$$

we have

$$\Delta_N := \sup_{u\in[0,1]} |\widehat{C}(u) - C^\star(u)| \le \delta_{N,\varepsilon}. \tag{A.1}$$

On $\mathcal{E}_\varepsilon$, the assumption $\delta_{N,\varepsilon} < C^\star\{m/(m+1)\}$ implies

$$\widehat{C}\{m/(m+1)\} \ge C^\star\{m/(m+1)\} - \Delta_N > 0.$$

For $j = 0, \dots, m$, define the cumulative weight error

$$B_j = \sum_{\ell=1}^{j} (w_\ell - w_\ell^\star) = \frac{\widehat{C}\{j/(m+1)\}}{\widehat{C}\{m/(m+1)\}} - \frac{C^\star\{j/(m+1)\}}{C^\star\{m/(m+1)\}},$$

where the empty sum is zero. Thus $B_0 = B_m = 0$. For every $j = 0, \dots, m$,

$$\begin{aligned}
|B_j| &\le \frac{|\widehat{C}\{j/(m+1)\} - C^\star\{j/(m+1)\}|}{\widehat{C}\{m/(m+1)\}} \\
&\quad + C^\star\{j/(m+1)\} \left| \frac{1}{\widehat{C}\{m/(m+1)\}} - \frac{1}{C^\star\{m/(m+1)\}} \right| \\
&\le \frac{2\Delta_N}{\widehat{C}\{m/(m+1)\}} \\
&\le \frac{2\Delta_N}{C^\star\{m/(m+1)\} - \Delta_N}.
\end{aligned}$$

The second inequality uses the monotonicity of $C^\star$, which gives $C^\star\{j/(m+1)\} \le C^\star\{m/(m+1)\}$.

Fix x and y, and write

$$K_j(x, y) = \Phi\left(\frac{y - \widehat{Y}_{\mathrm{adj},(j)}(x)}{\tau_x} \right).$$

Because the adjusted samples are ordered, $K_1(x, y) \ge \cdots \ge K_m(x, y)$. Also, $w_j - w_j^\star = B_j - B_{j-1}$. Summation by parts therefore yields

$$\widetilde{F}_x^\tau(y) - F_{x,C^\star}^\tau(y) = \sum_{j=1}^{m} (B_j - B_{j-1}) K_j(x, y) = \sum_{j=1}^{m-1} B_j \{K_j(x, y) - K_{j+1}(x, y)\}.$$

Since the differences $K_j - K_{j+1}$ are nonnegative and telescope to at most one,

$$\left| \widetilde{F}_x^\tau(y) - F_{x,C^\star}^\tau(y) \right| \le \max_{0\le j\le m} |B_j| \le \frac{2\Delta_N}{C^\star\{m/(m+1)\} - \Delta_N}.$$

On $\mathcal{E}_\varepsilon$, (A.1) and the monotonicity of $z \mapsto 2z/[C^\star\{m/(m+1)\} - z]$ give the bound in (31), uniformly over $x \in \mathcal{X}_0$ and $y \in \mathbb{R}$. The DKW inequality bounds the probability of $\mathcal{E}_\varepsilon^c$ by $2\exp(-2N\varepsilon^2)$.

For the unsmoothed CDFs, replace $K_j(x, y)$ by $\mathbf{1}\{\widehat{Y}_{\mathrm{adj},(j)}(x) \le y\}$. This sequence is again nonincreasing in j, so the same summation-by-parts argument applies. □