Design-marginal calibration of Gaussian process predictive distributions: Bayesian and conformal approaches

Aurélien Pion, Emmanuel Vazquez

Design-marginal calibration of Gaussian process predictive distributions: Bayesian and conformal approaches *

Aurélien Pion† and Emmanuel Vazquez‡

Abstract. We study the calibration of Gaussian process (GP) predictive distributions in the interpolation setting from a design-marginal perspective. Conditioning on the data and averaging over a design measure μ, we formalize μ-coverage for central intervals and μ-probabilistic calibration through randomized probability integral transforms. We introduce two methods. CPS–GP adapts conformal predictive systems to GP interpolation using standardized leave-one-out residuals, yielding stepwise predictive distributions with finite-sample marginal calibration. BCR–GP retains the GP posterior mean and replaces the Gaussian residual by a generalized normal model fitted to cross-validated standardized residuals. A Bayesian selection rule—based either on a posterior upper quantile of the variance for conservative prediction or on a cross-posterior Kolmogorov–Smirnov criterion for probabilistic calibration—controls dispersion and tail behavior while producing smooth predictive distributions suitable for sequential design. Numerical experiments on benchmark functions compare CPS–GP, BCR–GP, Jackknife+ for GPs, and the full conformal Gaussian process, using calibration metrics (coverage, Kolmogorov–Smirnov, integral absolute error) and accuracy or sharpness through the scaled continuous ranked probability score.

Keywords: Gaussian processes; calibration; μ-calibration; PIT; conformal prediction; conformal predictive systems; generalized normal residuals; Bayesian calibration; interpolation; uncertainty quantification; predictive intervals; proper scoring rules.

1. Introduction.

Gaussian processes (GPs) are classical Bayesian models used to approximate an unknown real-valued deterministic function f from limited evaluations over a design space $\mathbb{X} \subset \mathbb{R}^d$ (see, e.g., Stein, 1999; Santner et al., 2003; Rasmussen and Williams, 2005). They provide not only point predictions but also a measure of uncertainty, which makes them central to tasks where decisions depend on both accuracy and risk. Applications include Bayesian optimization (see, e.g., Jones et al., 1998; Villemonteix et al., 2009; Feliot et al., 2017) and the estimation of excursion sets (Bect et al., 2017; Azzimonti et al., 2021), where uncertainty quantification directly drives the exploration of the design space.

In this work we focus on the interpolation setting, where observations are assumed exact. A GP prior with mean function m and covariance kernel k induces a Gaussian posterior predictive distribution at any location $x \in \mathbb{X}$. The posterior mean serves as an interpolator of the observed data, while its variance quantifies predictive uncertainty. Both quantities are available in closed form.

The combination of exact interpolation and quantified predictive uncertainty explains the success of GPs in applications where uncertainty guides the exploration of the input space. Yet, the quality of decisions relies on the calibration of predictive distributions: the empirical frequency with which $f(x)$ falls within nominal prediction intervals, when x varies, should match the intended coverage. We assess these frequencies with respect to a design measure μ on $\mathbb{X}$, which governs how observation points are drawn. Although this design-marginal notion

*This work was funded by Transvalor S.A.

†Transvalor S.A., Biot, France; Univ. Paris-Saclay, CNRS, CentraleSupélec, L2S, Gif-sur-Yvette, France.

‡Univ. Paris-Saclay, CNRS, CentraleSupélec, L2S, Gif-sur-Yvette, France. firstname.lastname@centralesupelec.fr

of calibration has appeared in the literature under various forms, we examine it more closely in this article and refer to it as *μ-calibration*. In practice, GP-based predictive intervals are often miscalibrated, leading to overconfident or overly conservative predictions. Such discrepancies have been documented in the interpolation setting, for instance with a constant mean function and Matérn covariance (Pion and Vazquez, 2025).

Figure 1 illustrates this issue. It displays the trade-off between predictive accuracy, measured by RMSE, and calibration quality, measured by the Kolmogorov–Smirnov (KS) metric of probability integral transform (PIT) values (calibration metrics are detailed in Section 3). The maximum-likelihood (ML) hyperparameters achieve low RMSE, consistent with the empirical findings of Petit et al. (2023), but poor calibration. In contrast, post-hoc calibration using the methods presented in this article improves calibration without degrading accuracy.

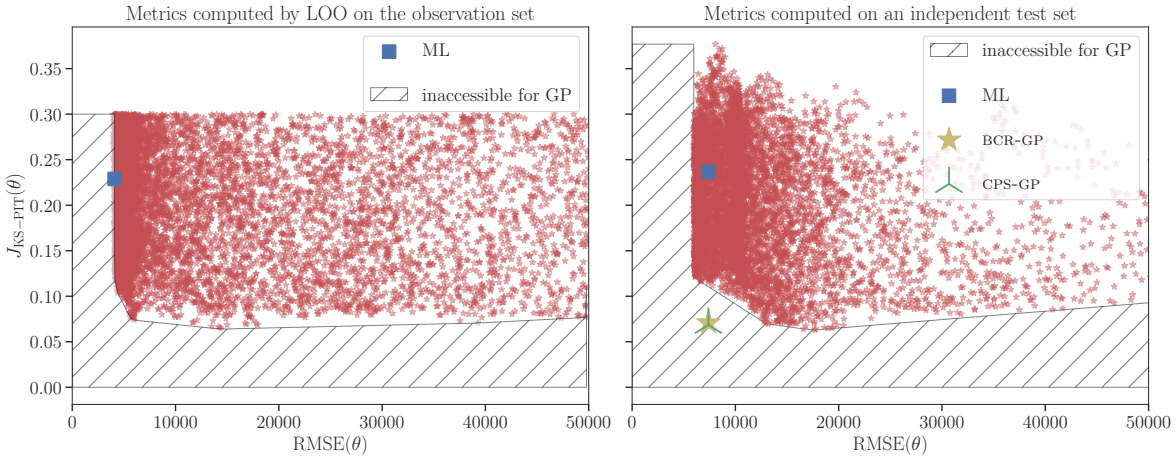


Figure 1: Trade-off between predictive accuracy (RMSE) and calibration quality (KS–PIT, smaller is better, see Section 3) for a uniform random sample of GP kernel hyperparameters (red points). We interpolate the Goldstein–Price function using a GP with constant mean and Matérn covariance. The left panel shows metrics computed by leave-one-out (LOO) on the observation set (150 points), and the right panel shows metrics computed on an independent test set drawn from μ (1500 points). The ML-selected hyperparameters (blue square) yield accurate but poorly calibrated predictions. Post-processed predictors using CPS–GP (green symbol) and BCR–GP (gold star) improve calibration on the test set without degrading accuracy. The hatched region corresponds to RMSE–KS–PIT pairs that cannot be attained by any GP posterior under the considered GP family.

Among the approaches developed to address miscalibration, conformal prediction (CP) is particularly attractive. It is model-agnostic and provides distribution-free guarantees on marginal coverage (Vovk et al., 2005). Several recent works adapt CP to Gaussian processes in order to construct calibrated predictive intervals at user-specified levels (Papadopoulos, 2024; Jaber et al., 2024). Beyond interval prediction, Vovk et al. (2019) introduced the *conformal predictive systems* (CPS), which extend CP to produce a calibrated predictive distribution at each test point.

Building on these ideas, we develop two calibration approaches for GP interpolation: an

adaptation of CPS to the interpolation setting (CPS–GP) and a new Bayesian parametric method (BCR–GP).

The CPS–GP construction yields predictive distributions whose associated prediction sets satisfy the standard conformal finite-sample marginal coverage guarantee under exchangeability, but their stepwise, non-differentiable form limits their direct use in standard Bayesian optimization or excursion-set estimation algorithms.

The second approach, BCR–GP, short for *Bayesian-calibrated residuals for Gaussian processes*, retains the GP posterior mean but replaces the predictive distribution with a parametric family fitted to normalized GP residuals. We adopt a generalized normal distribution, with shape and scale parameters selected through a Bayesian strategy inspired by tolerance-interval constructions (Meeker et al., 2017). The resulting predictive distributions support inference at arbitrary confidence levels and can be tuned to favour more or less conservative uncertainty quantification. Unlike CPS–GP, this parametric formulation produces smooth predictive CDFs, directly usable in sequential-design algorithms that rely on closed-form expressions (e.g. for expected improvement or related criteria).

We compare BCR–GP with CPS–GP, the Jackknife+ Gaussian process method of Jaber et al. (2024), and the full conformal approach of Papadopoulos (2024), using calibration diagnostics and proper scoring rules that assess both calibration and sharpness.

The remainder of the article is organized as follows. Section 2 reviews calibration of GP predictive distributions, including hyperparameter selection, conformal prediction, and post-hoc recalibration. Section 3 fixes notation and setup, introduces μ-calibration (coverage and PIT), recalls randomized PIT, and summarizes calibration metrics. Section 4 develops CPS–GP for interpolation and discusses its properties and practical limitations. Section 5 presents BCR–GP (generalized normal residual model), parameter estimation, and implementation details. Section 6 reports experiments on benchmarks, comparing all methods. Section 7 concludes with main findings and future directions.

2. Related work.

We summarize results on GP hyperparameter selection and post-hoc calibration, including conformal and distributional adjustments.

2.1. Effect of hyperparameter selection in GPs.

Calibration of GP predictive intervals is strongly influenced by how kernel hyperparameters are selected. Karvonen et al. (Karvonen et al., 2020) study noiseless GP interpolation with Sobolev/Matérn kernels when only a global scale parameter is estimated by maximum likelihood. In the RKHS setting, they show that the ratio between the interpolation error and the posterior standard deviation is bounded by a factor of order $n^{1/2}$ as the number of points n increases, with a complementary characterization of underconfidence in terms of a specific subspace of the RKHS. For Sobolev classes that are smoother or rougher than the RKHS, related polynomial bounds are obtained under additional assumptions on the design and smoothness. These results quantify a form of slow overconfidence in a worst-case sense (supremum over $\mathbb{X}$ and over the relevant function class) when only a global scale is fitted by maximum likelihood.

Beyond scale-only fitting, adaptive empirical Bayes also faces structural limitations. In the Gaussian white-noise model, Szabó et al. (2015) show that credible balls based on marginal-likelihood tuning of prior smoothness cannot achieve nominal frequentist coverage uniformly over a Sobolev class, even though the corresponding posteriors are (near) minimax rate-

adaptive. They also show that asymptotically correct coverage can be recovered after restricting attention to a suitably regular subclass of functions. For GPs with squared-exponential kernels, Hadji and Szabó (2021) show that empirical-Bayes L^2-credible balls, with length-scale chosen by marginal maximum likelihood, can be severely overconfident for a large subclass of truths, with frequentist coverage converging to zero. They further prove that coverage can be restored under additional regularity assumptions by inflating the credible sets or by modifying the empirical-Bayes estimator. Taken together, these results indicate that credible sets based on empirical-Bayes hyperparameter tuning do not generally provide reliable uncertainty quantification without additional corrections or restrictions.

Empirical studies also document the sensitivity of coverage to the selection criterion. Petit et al. (2023) compare likelihood-based and leave-one-out criteria cast as proper scoring rules and find that the choice of model family often matters more than the specific selection criterion, with several criteria yielding comparable performance for Matérn models. Marrel and Iooss (2024) survey estimation and validation diagnostics and introduce a multi-objective hyperparameter estimation algorithm targeting improved predictive distributions. Acharki et al. (2023) propose a two-step, coverage-oriented adjustment: they first tune hyperparameters to match a target leave-one-out coverage and then calibrate prediction intervals at a prescribed level. Implementing a $(1-\alpha)$ interval in this approach requires fitting two quantile-oriented GP models (for the $\alpha/2$ and $1-\alpha/2$ bounds).

2.2. Post-hoc calibration methods. Conformal prediction (CP) methods are a widely used class of post-hoc calibration techniques. They wrap around any base predictive model to produce valid marginal coverage guarantees without assuming model correctness. Jackknife+ and full conformal procedures have been adapted to GPs (Jaber et al., 2024; Papadopoulos, 2024), and are recommended for robust calibration (Pion and Vazquez, 2025). However, these methods construct prediction intervals only at fixed confidence levels and do not provide a full calibrated predictive distribution.

Vovk et al. (2017b, 2019) introduce the conformal predictive system (CPS), which provides a full predictive distribution and supports interval estimation at arbitrary levels. Vovk et al. (2017a) adapt CPS to kernel ridge regression, and in this work we extend this approach to GP interpolation. The CPS framework and its adaptation to our setting are detailed in Section 4.

2.3. Auxiliary-model and distributional adjustment methods. Another line of post-hoc approaches recalibrates predictions by fitting an auxiliary model. In the GP setting, Capone et al. (2023) keep the base GP for the mean and compute predictive quantiles with a second GP whose hyperparameters are tuned on a hold-out calibration set to meet coverage; this uses data splitting (or cross-fitting).

More generally, distributional recalibration learns a (typically monotone) mapping from predicted to empirical conditional distributions using a hold-out set. Examples include monotone quantile/CDF mappings (Kuleshov et al., 2018) and density-estimation–based distribution calibration (Kuleshov and Deshpande, 2022). These methods are model-agnostic and operate on pointwise marginals; they do not exploit GP prior structure and act only on marginal distributions. In the same spirit, Dey et al. (2024) propose LADaR, which applies a local

probability–probability map (Cal-PIT) for instance-wise calibration of conditional CDFs.

3. Background and setup.

3.1. Setting and notation.

We consider noise-free observations of a deterministic but unknown function $f : \mathbb{X} \subset \mathbb{R}^d \to \mathbb{R}$ of the form

$$Z_i = f(X_i), \quad i = 1, \dots, n,$$

where the design points X_i are i.i.d. from a probability measure μ on $\mathbb{X}$, that will be referred to as the *design measure*. The dataset is $\mathcal{D}_n = \{(X_i, Z_i)\}_{i=1}^n$.

Throughout, the Bayesian GP framework is used only as a *construction device* for building predictive distributions; we do not adopt a Bayesian interpretation (f is fixed, non-random). We write

$$\mathsf{P}_n(\,\cdot\,) := \mathsf{P}\big(\,\cdot \mid \mathcal{D}_n\big)$$

for probabilities conditional on the observed dataset.

Given $\mathcal{D}_n$, the remaining randomness under P_n comes solely from auxiliary draws (e.g., fresh test points $X_i^\star \sim \mu$, $i = 1, 2 \ldots$), all independent of $\mathcal{D}_n$.

3.2. Predictive distributions and prediction intervals.

Let $\hat{F}_n(\cdot \mid x)$ denote a predictive CDF for $f(x)$, at $x \in \mathbb{X}$, constructed from $\mathcal{D}_n$. In the GP framework, $\hat{F}_n(\cdot \mid x)$ is taken as a Gaussian CDF with mean and variance given by the kriging equations (see, e.g., Chilès and Delfiner, 1999; Stein, 1999).

For any predictive CDF (not necessarily continuous), we use the generalized (left-continuous) quantile

$$\hat{F}_n^{-1}(p \mid x) \;:=\; \inf\{z \in \mathbb{R} :\; \hat{F}_n(z \mid x) \geq p\}, \qquad p \in (0,1), \tag{3.1}$$

and define the central $(1-\alpha)$ interval

$$\mathcal{C}_{n,1-\alpha}(x) \;=\; \big[\hat{F}_n^{-1}(\alpha/2 \mid x),\, \hat{F}_n^{-1}(1-\alpha/2 \mid x)\big]. \tag{3.2}$$

If $V \sim \hat{F}_n(\cdot \mid x)$ independently of $\mathcal{D}_n$, then, under P_n,

$$\mathsf{P}_n\Big\{\hat{F}_n^{-1}(\alpha/2 \mid x) \;\leq\; V \;\leq\; \hat{F}_n^{-1}(1-\alpha/2 \mid x)\Big\} \;\geq\; 1-\alpha,$$

with equality when $\hat{F}_n(\cdot \mid x)$ is continuous at both endpoints. Thus, when $\hat{F}_n$ has atoms, $\mathcal{C}_{n,1-\alpha}(x)$ is conservative: its predictive mass under $\hat{F}_n(\cdot \mid x)$ exceeds $1-\alpha$ whenever an endpoint coincides with a jump of the CDF.

Exact mass via boundary randomization. To obtain exact $(1-\alpha)$ mass in the presence of discontinuities, introduce randomization at the jumps. Fix $x \in \mathbb{X}$ and let $\tau \sim \mathcal{U}(0,1)$ be independent of $\mathcal{D}_n$. Define the randomized predictive CDF

$$\hat{F}_{n,\tau}(z \mid x) = \hat{F}_n(z^- \mid x) + \tau\big(\hat{F}_n(z \mid x) - \hat{F}_n(z^- \mid x)\big),$$

and its (random) generalized inverse $\hat{F}_{n,\tau}^{-1}(\cdot \mid x)$, defined analogously to (3.1).

Proposition 3.1 (Boundary randomization and exact interval mass). *Let $V \sim \hat{F}_n(\cdot \mid x)$, $x \in \mathbb{X}$, be independent of $(\tau, \mathcal{D}_n)$, and set $U := \hat{F}_{n,\tau}(V \mid x)$. Then, $U \mid \mathcal{D}_n \sim \mathcal{U}(0,1)$.*

Moreover, the half-open randomized interval

$$\mathcal{C}_{n,\tau,1-\alpha}(x) := \big[\hat{F}^{-1}_{n,\tau}(\alpha/2 \mid x),\ \hat{F}^{-1}_{n,\tau}(1-\alpha/2 \mid x)\big)$$

satisfies

$$\mathsf{P}_n\{\, V \in \mathcal{C}_{n,\tau,1-\alpha}(x)\,\} = 1-\alpha\,.$$

Proof. See Appendix A.2. ∎

Note that the random variable $U = \hat{F}_{n,\tau}(V \mid x)$ in Proposition 3.1 is a *probability integral transform (PIT)*, a standard tool for calibration assessment. Its general role will be developed in Section 3.5.

3.3. μ-calibration: general principle. In this work, calibration is assessed relative to the design measure μ on $\mathbb{X}$. After conditioning on the data $\mathcal{D}_n$, we consider a fresh input $X \sim \mu$ (independent of $\mathcal{D}_n$) and the corresponding value $f(X)$. A predictive family $\{\hat{F}_n(\cdot \mid x) : x \in \mathbb{X}\}$ is said to be *μ-calibrated* if its predictive statements match the distribution of $f(X)$ under $X \sim \mu$. Two forms will be considered: *μ-coverage*, which concerns the frequency with which $f(X)$ falls inside prediction intervals $\mathcal{C}_{n,1-\alpha}(X)$, and *$\mu$-probabilistic calibration*, which requires that the PIT values $\hat{F}_n(f(X) \mid X)$ are uniform on $[0,1]$ under $X \sim \mu$.

The focus on μ-calibration is motivated by empirical testability. In the classical Bayesian GP framework, the GP posterior predictive distribution at any fixed x admits a density that is positive on every open interval (away from observation points where it degenerates in the noise-free case), so a single realization $f(x)$ cannot empirically falsify it. By contrast, μ-coverage and μ-probabilistic calibration are *spatial* properties: they concern the distribution of $f(X)$ under $X \sim \mu$ and can be estimated and tested on an independent test design, then rejected if they fail. This perspective is consonant with Matheron's theory of *regionalized variables*, where spatial, design-dependent properties are regarded as estimable and models are assessed through observable consequences under the sampling protocol (Matheron, 1989).

3.4. μ-coverage and integrated error. As introduced above, μ-coverage assesses interval calibration after conditioning on $\mathcal{D}_n$ and marginalizing over $X \sim \mu$: for a family of centered prediction intervals $\mathcal{C}_{n,1-\alpha}(x)$, constructed from predictive distributions $\hat{F}_n(\cdot \mid x)$, define its μ-coverage as

$$\begin{aligned}\delta_\alpha(\hat{F}_n;\mu) &= \mathsf{P}_n\big\{\, f(X) \in \mathcal{C}_{n,1-\alpha}(X)\,\big\}\,, \qquad X \sim \mu\,,\\ &= \mu\big(\{x :\ f(x) \in \mathcal{C}_{n,\,1-\alpha}(x)\}\big)\,.\end{aligned} \tag{3.3}$$

Calibration at level $1-\alpha$ means $\delta_\alpha(\hat{F}_n;\mu) = 1-\alpha$.

Given an independent test design $\{X^\star_j\}_{j=1}^m$ with $X^\star_j \sim \mu$, the natural Monte Carlo estimator of $\delta_\alpha(\hat{F}_n;\mu)$ is

$$\hat{\delta}_{\alpha,\,m}(\hat{F}_n) = \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}\{f(X^\star_j) \in \mathcal{C}_{n,\,1-\alpha}(X^\star_j)\}.$$

Remark 3.2. When no separate test set is available, estimating $\delta_\alpha(\hat{F}_n;\mu)$ on the design $\mathcal{D}_n$ is not informative. In GP interpolation, $\hat{F}_n(\cdot \mid X_i)$ is a Dirac mass at $Z_i = f(X_i)$, so for

all $\alpha \in (0,1)$

$$\mathcal{C}_{n,1-\alpha}(X_i) = \big[\hat{F}_n^{-1}(\alpha/2 \mid X_i),\, \hat{F}_n^{-1}(1-\alpha/2 \mid X_i)\big] = \{Z_i\},$$

and hence

$$\mathbb{1}\{f(X_i) \in \mathcal{C}_{n,1-\alpha}(X_i)\} = 1.$$

Empirical coverage computed on $\mathcal{D}_n$ is therefore identically equal to 1 for every α and carries no information about $\delta_\alpha(\hat{F}_n;\mu)$. Cross-validation (e.g., leave-one-out) can avoid this degeneracy by removing each (X_i, Z_i) when assessing its inclusion, but introduces additional biases due to data reuse and dependence.

To summarize deviations across all confidence levels, Marrel and Iooss (2024) introduced the *Integrated Absolute Error* (IAE),

$$J_{\mathrm{IAE},\mu}(\hat{F}_n) = \int_0^1 \big|\delta_\alpha(\hat{F}_n;\mu) - (1-\alpha)\big|\, d\alpha.$$

The empirical IAE $\hat{J}_{\mathrm{IAE},\,m}(\hat{F}_n)$ is obtained by substituting $\hat{\delta}_{\alpha,\,m}(\hat{F}_n)$ for $\delta_\alpha(\hat{F}_n;\mu)$ in the definition above.

3.5. Probability integral transform and μ-probabilistic calibration.

Probability integral transform (PIT). Beyond interval coverage, calibration can also be assessed through PIT, a classical notion in probabilistic forecasting (see Appendix B). In its simplest form, given a continuous predictive CDF $\hat{F}$ for a random variable Z, the PIT is

$$U_{\hat{F}}^{Z} = \hat{F}(Z).$$

If $\hat{F}$ coincides with the true distribution of Z, then $U_{\hat{F}}^{Z} \sim \mathcal{U}(0,1)$. This property underlies the standard use of PIT values in forecast evaluation: when multiple forecast–observation pairs $(\hat{F}_i, Z_i)$ are available, the empirical distribution of $\{\hat{F}_i(Z_i)\}$ is compared to the uniform distribution, typically through histograms or empirical CDF plots (Dawid, 1984; Gneiting and Resin, 2023). Departures from uniformity reveal systematic miscalibration such as over- or underdispersion (see Figure 2).

For discontinuous predictive CDFs, one may use the *randomized PIT*

$$U_{\hat{F}}^{Z,\tau} := \hat{F}(Z^-) + \tau\big(\hat{F}(Z) - \hat{F}(Z^-)\big),$$

with $\tau \sim \mathcal{U}(0,1)$ independent of $(\hat{F}, Z)$, which preserves the uniformity property (see Appendix B).

μ-probabilistic calibration. In our setting, after conditioning on the observed data $\mathcal{D}_n$, we do not have multiple forecast–observation pairs but instead a single fitted predictive family $\hat{F}_n(\cdot \mid x)$ indexed by x. To apply the PIT idea, we define a (possibly randomized) μ-PIT as

$$\begin{aligned} U_{\hat{F}_n}^{f,\mu,\tau} &:= \hat{F}_{n,\,\tau}(f(X) \mid X) \\ &= \hat{F}_n(f(X)^- \mid X) + \tau\big(\hat{F}_n(f(X) \mid X) - \hat{F}_n(f(X)^- \mid X)\big), \end{aligned} \tag{3.4}$$

with $X \sim \mu$, and where $\tau \sim \mathcal{U}(0,1)$ is independent of $(\hat{F}_n,\, X,\, \mathcal{D}_n)$. Randomization only takes effect when $\hat{F}_n(\cdot \mid x)$ has jumps; in the continuous and strictly increasing case (for μ-a.e. x),

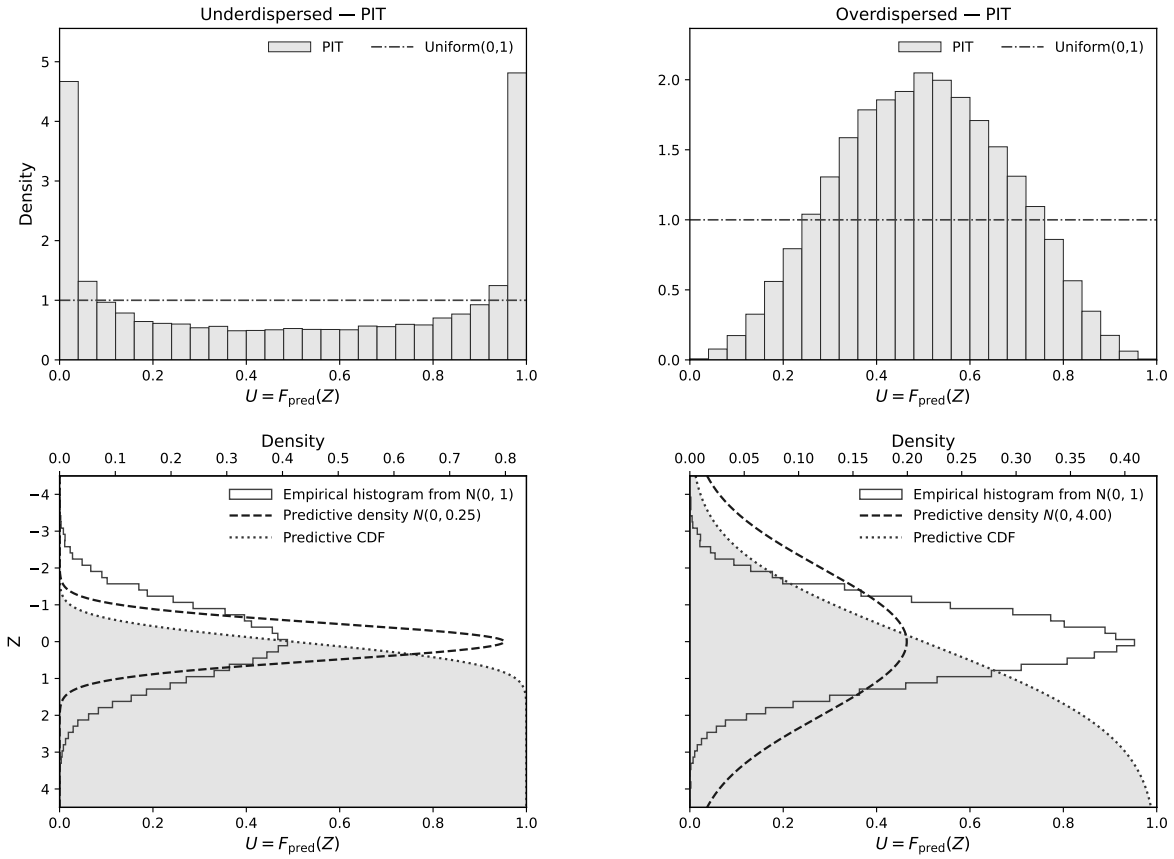


Figure 2: *First row*: PIT histograms with the uniform density (dashed line) as reference. A ∪-*shaped* PIT (mass near 0 and 1) indicates *underdispersion*; predictive intervals are too narrow and observations fall outside too often (optimistic coverage). A ∩-*shaped* PIT (mass near 1/2) indicates *overdispersion*; intervals are too wide and observations fall inside too often (pessimistic coverage). *Second row (rotated view):* vertical axis is z. Bottom axis shows the predictive CDF $u = F_{\text{pred}}(z)$ (shaded area), while top axis shows the density scale: horizontal empirical histogram (outline) and predictive pdf (dashed). A horizontal slice at a given z maps to a CDF value u on the bottom axis, which is precisely the PIT value contributing to the histogram in the first row. The empirical sample is drawn from the standard normal distribution $\mathcal{N}(0, 1)$. Predictive distributions are normal with the same mean but different scales: $\mathcal{N}(0, 0.5^2)$ (underdispersed) and $\mathcal{N}(0, 2^2)$ (overdispersed).

μ-PIT reduces to

$$U_{\hat{F}_n}^{f,\mu,\tau} = U_{\hat{F}_n}^{f,\mu} = \hat{F}_n(f(X) \mid X), \qquad X \sim \mu.$$

We will say that $\hat{F}_n$ is μ-*probabilistically calibrated* if

$$U_{\hat{F}_n}^{f,\mu,\tau} \sim \mathcal{U}(0, 1).$$

Remark 3.3. Define the distribution function

$$G_\mu(u) := \mathsf{P}_n\big(U_{\hat{F}_n}^{f,\mu,\tau} \le u\big), \qquad u \in [0,1]. \tag{3.5}$$

Thus, μ-probabilistic calibration is equivalent to

$$G_\mu(u) = u, \qquad \forall u \in [0,1],$$

In the continuous and strictly monotone case (for μ-a.e. x), randomization does not take effect, so that

$$\begin{aligned} G_\mu(u) &= \mathsf{P}_n\left(U_{\hat{F}_n}^{f,\mu} \le u\right) = \mathsf{E}_n\left[\mathbb{1}\{\hat{F}_n(f(X) \mid X) \le u\}\right] \\ &= \int_{\mathbb{X}} \mathbb{1}\{\hat{F}_n(f(x) \mid x) \le u\}\, \mathrm{d}\mu(x) \\ &= \mu\left(\{x :\ f(x) \le \hat{F}_n^{-1}(u \mid x)\}\right), \end{aligned}$$

where E_n denotes expectation under P_n. Thus, we obtain the following spatial characterization of μ-probabilistic calibration: for all $u \in (0,\, 1)$,

$$G_\mu(u) = \mu\big(\{x :\ f(x) \le \hat{F}_n^{-1}(u \mid x)\}\big) = u, \qquad u \in (0,1).$$

(Without strict monotonicity of $\hat{F}_n(\cdot \mid x)$, the following bounds hold for all $u \in (0,1)$:

$$\mu\big(\{x :\ f(x) < \hat{F}_n^{-1}(u \mid x)\}\big) \ \le\ G_\mu(u) \ \le\ \mu\big(\{x :\ f(x) \le \hat{F}_n^{-1}(u \mid x)\}\big).$$

See Appendix A.3 for the detailed justification.)

Example (constant f and non-degenerate μ-probabilistically calibrated $\hat{F}_n$). Assume $f(x) \equiv c$. Let $\mathbb{X} = [0,1]$, $\mu = \mathcal{U}[0,1]$, and define

$$\hat{F}_n(z \mid x) = \Phi\big(z - \big(c + \Phi^{-1}(x)\big)\big),$$

where Φ is the standard normal CDF. Then, with $X \sim \mu$,

$$U_{\hat{F}_n}^{f,\mu} \ =\ \hat{F}_n(c \mid X) \ =\ \Phi\big(c - (c + \Phi^{-1}(X))\big) \ =\ \Phi\big(-\Phi^{-1}(X)\big) \ =\ 1 - X \quad \text{a.s.}$$

Hence $U_{\hat{F}_n}^{f,\mu} \sim \mathcal{U}(0,1)$ and $\hat{F}_n$ is μ-probabilistically calibrated, while remaining non-degenerate (Gaussian predictive law with unit variance at every x). Note, however, that this predictive family is clearly misspecified: although f is constant, the predictive mean $c + \Phi^{-1}(x)$ varies with x and ranges over $\mathbb{R}$, so $\hat{F}_n(\cdot \mid x)$ does not represent a reasonable belief about f.

This example highlights a limitation: for a given f, many "incorrect" predictive families can still be μ-probabilistically calibrated. However, μ-probabilistic calibration alone does not penalize overly diffuse or uninformative predictive distributions. If meaningful uncertainty quantification is sought, calibration is necessary but not sufficient: one must also assess *sharpness*—the concentration of predictive distributions—using proper scoring rules (see Section 3.7), or enforce structural constraints on the predictive family, as in the BCR–GP method (see Section 5).

Remark 3.4. As already mentioned in Section 3.2 and Proposition 3.1, μ-coverage and the

μ-PIT are linked as follows (see Appendix A.4 for the detailed derivations).

(i) If, for μ-a.e. x, the map $z \mapsto \hat{F}_n(z \mid x)$ is continuous and strictly increasing, then

$$\delta_\alpha(\hat{F}_n;\mu) = \mathsf{P}_n\big\{\, a \le U^{f,\mu}_{\hat{F}_n} \le b \,\big\}, \qquad \text{with } a = \alpha/2,\ b = 1-\alpha/2.$$

(ii) Without continuity, one always has the bounds

$$\mathsf{P}_n\big\{\, a \le U^{f,\mu}_{\hat{F}_n} < b \,\big\} \le \delta_\alpha(\hat{F}_n;\mu)$$

and

$$\delta_\alpha(\hat{F}_n;\mu) \le \mathsf{P}_n\big\{\, a \le U^{f,\mu}_{\hat{F}_n} \le b \,\big\} + \mu\big(\{x:\ f(x) = q_b,\ \hat{F}_n(q_b \mid x) > b\}\big),$$

where $q_b = \hat{F}_n^{-1}(b \mid x)$. (The extra term captures boundary mass when $\hat{F}_n(\cdot \mid x)$ jumps at the upper endpoint.)

(iii) When $\hat{F}_n(\cdot \mid x)$ has jumps, boundary randomization restores exact equivalence: for every fixed τ,

$$\mu\big(\{x:\ f(x) \in [\,\hat{F}^{-1}_{n,\tau}(a \mid x),\ \hat{F}^{-1}_{n,\tau}(b \mid x)\,)\}\big) = \mathsf{P}_n\big\{\, a \le U^{f,\mu,\tau}_{\hat{F}_n} < b \mid \tau \big\}.$$

Thus, μ-coverage at level $1-\alpha$ can be expressed as a PIT inclusion probability, with boundary randomization ensuring exact equality in the discontinuous case.

3.6. Metrics for μ-probabilistic calibration. To evaluate μ-probabilistic calibration we use PIT values on an independent test design $\{X_j^\star\}_{j=1}^m$ with $X_j^\star \sim \mu$. Given predictive CDFs $\hat{F}_n(\cdot \mid x)$, define, for each j,

$$a_j := \hat{F}_n(f(X_j^\star) \mid X_j^\star), \qquad a_j^- := \hat{F}_n(f(X_j^\star)^- \mid X_j^\star).$$

The PIT values are then

$$U_j = \begin{cases} a_j, & \text{if } \hat{F}_n(\cdot \mid X_j^\star) \text{ is continuous,} \\ a_j^- + \tau_j(a_j - a_j^-), & \text{otherwise,} \end{cases}$$

with $\tau_j \sim \mathcal{U}(0,1)$ i.i.d., independent of $(\mathcal{D}_n, \{X_j^\star\})$.

Several metrics can be used to quantify deviations of $\{U_j\}$ from uniformity.

Variance-based metric. If $U \sim \mathcal{U}(0,1)$ then $\mathrm{var}(U) = 1/12$. A natural measure of departure from uniformity is

$$J_{\text{Var-PIT},\,m}(\hat{F}_n) = \frac{1}{m}\sum_{j=1}^m (U_j - \tfrac{1}{2})^2 - \tfrac{1}{12}.$$

When the PIT is uniform, $J_{\text{Var-PIT},\,m}(\hat{F}_n) \to 0$ as $m \to \infty$. In particular, for approximately symmetric PIT distributions, a positive deviation ($J_{\text{Var-PIT},\,m} > 0$) corresponds to a $\cup$-shaped PIT histogram: mass accumulates near 0 and 1, which means predictive intervals are too narrow and observed values fall outside too often (optimistic coverage). A negative deviation ($J_{\text{Var-PIT},\,m} < 0$) corresponds to a $\cap$-shaped histogram: mass concentrates near 1/2, which means predictive intervals are too wide and observed values fall inside too often (pessimistic

coverage).

KS–PIT metric. Another possibility is to compare the full empirical distribution of PIT values to the uniform law using the Kolmogorov–Smirnov (KS) distance. Applied to PIT values, this yields the *KS–PIT metric* (Diebold et al., 1998):

$$J_{\text{KS-PIT},\,m}(\hat{F}_n) = \sup_{u\in[0,1]} \left| \frac{1}{m} \sum_{j=1}^{m} \mathbb{1}\{U_j \le u\} - u \right|.$$

Smaller values indicate a PIT distribution closer to uniformity, hence better μ-probabilistic calibration.

It is natural to introduce the corresponding population quantity. As defined in Remark 3.3 (Equation (3.5)), $G_\mu(u)$ denotes the distribution function of the μ-PIT values. The population KS–PIT distance is then

$$J_{\text{KS-PIT},\,\mu}(\hat{F}_n) \;=\; \sup_{u\in[0,1]} \big| G_\mu(u) - u \big|.$$

Proposition 3.5. *Condition on $\mathcal{D}_n$. If $X_j^\star \overset{\text{i.i.d.}}{\sim} \mu$, $\tau_j \overset{\text{i.i.d.}}{\sim} \mathcal{U}(0,1)$, independent of $(\mathcal{D}_n, \{X_j^\star\})$, then*

$$J_{\text{KS-PIT},\,m}(\hat{F}_n) \xrightarrow[m\to\infty]{\text{a.s.}} J_{\text{KS-PIT},\,\mu}(\hat{F}_n).$$

Proof. See Appendix A.5. ∎

Remark 3.6 (On reuse of observation points). As with coverage, evaluating KS–PIT on observations using cross-validation can be misleading. The target is post-data, design-marginal uniformity under $X \sim \mu$, so an independent test design (or a proper sample split / cross-fitting) should be used.

Relation to the Integrated Absolute Error (IAE). The Integrated Absolute Error (IAE), recalled in Section 3.4, quantifies deviations between nominal and empirical coverage of central prediction intervals. The KS–PIT metric, by contrast, evaluates the full distribution of PIT values against the uniform law. The two notions are linked:

Proposition 3.7 (IAE bounded by KS–PIT). *For both empirical and population measures, the following inequalities hold:*

$$J_{\text{IAE},m}(\hat{F}_n) \;\le\; 2\, J_{\text{KS-PIT},m}(\hat{F}_n), \qquad J_{\text{IAE},\mu}(\hat{F}_n) \;\le\; 2\, J_{\text{KS-PIT},\mu}(\hat{F}_n).$$

Proof. See Appendix A.6. ∎

Thus, KS–PIT provides a uniformity-based calibration criterion that dominates IAE: small KS–PIT implies small IAE. In applications, it is useful to report both metrics: IAE to summarize coverage accuracy across nominal levels, and KS–PIT as a global indicator of probabilistic calibration.

3.7. Scoring rules. As noted above, calibration alone does not guarantee useful uncertainty quantification: overly diffuse forecasts can be perfectly calibrated yet uninformative. To assess both calibration and concentration (*sharpness*), we use *proper scoring rules* (Gneiting and Raftery, 2007), which assign a numerical score $S(\hat{F}, z)$ to a predictive CDF $\hat{F}$ and

outcome $z \in \mathbb{R}$. A scoring rule is *strictly proper* if, for the true distribution F,

$$\mathsf{E}_{Z\sim F}[S(F,Z)] \;\le\; \mathsf{E}_{Z\sim F}[S(\hat{F},Z)],$$

with equality if and only if $\hat{F} = F$.

Two standard examples are the (negative) log predictive density (log score) and the continuous ranked probability score (CRPS):

$$S_{\mathrm{NLPD}}(\hat{F},z) = -\log \hat{f}(z), \qquad S_{\mathrm{CRPS}}(\hat{F},z) = \int_{-\infty}^{\infty} \big(\hat{F}(u) - \mathbb{1}\{u \ge z\}\big)^2 \,\mathrm{d}u,$$

where $\hat{f}$ is the density corresponding to $\hat{F}$ (when it exists). The CRPS evaluates both calibration and sharpness by integrating the squared difference between the predictive and empirical CDFs across all thresholds. The CRPS also admits the expectation form

$$S_{\mathrm{CRPS}}(\hat{F},z) = \mathsf{E}\big[\,|Z - z|\,\big] - \tfrac{1}{2}\,\mathsf{E}\big[\,|Z - Z'|\,\big],$$

where Z and Z' are independent and distributed as $\hat{F}$.

To reduce the scale sensitivity of CRPS, Bolin and Wallin (2023) introduced the *scaled continuous ranked probability score* (SCRPS). For a random variable $Z \sim \hat{F}$ and an independent copy Z', it is defined as

$$S_{\mathrm{SCRPS}}(\hat{F},z) = -\frac{\mathsf{E}|Z - z|}{\mathsf{E}|Z - Z'|} - \tfrac{1}{2}\log\big(\mathsf{E}|Z - Z'|\big).$$

The SCRPS normalizes the CRPS by the expected dispersion $\mathsf{E}|Z-Z'|$, making it less sensitive to overall scaling and more suitable for comparing predictive distributions of different spread. Lower values correspond to sharper and better-calibrated predictions. Analytical expressions for computing SCRPS are given in Appendix D.

Proper scoring rules reward forecasts that are both calibrated (agreement between predicted probabilities and observed frequencies) and sharp (concentrated predictive distributions). Sharpness is meaningful only under calibration; otherwise, a forecast can be sharp but systematically biased. CRPS-type scores (including SCRPS) provide a single global measure of predictive performance and admit decompositions that separate calibration and sharpness contributions (Arnold et al., 2024). By contrast, PIT-based diagnostics (KS–PIT, IAE) focus solely on calibration. The two perspectives are complementary: scoring rules quantify overall predictive skill, while PIT-based tools reveal calibration errors more directly. When tail behavior is of primary interest, tail-sensitive scoring rules such as those proposed by Allen et al. (2023, 2025) offer additional, focused diagnostics.

Given a true distribution F, a forecast $\hat{F}$, and a scoring rule S, predictive performance is quantified by the expected score

$$J_S(\hat{F}) = \mathsf{E}_{Z\sim F}\big[S(\hat{F},Z)\big].$$

In our interpolation setting, forecasts are given by the family of predictive CDFs $\hat{F}_n(\cdot \mid x)$ indexed by $x \in \mathbb{X}$, and observation points are taken with respect to the design measure μ. The corresponding design-marginal score is

$$J_{S,\mu}(\hat{F}_n) = \mathsf{E}_{X\sim\mu}\Big[\, S\big(\hat{F}_n(\cdot \mid X),\, f(X)\big)\,\Big],$$

that is, the score averaged over the input space according to μ.

In practice, with an independent test design $X_1^\star, \dots, X_m^\star \overset{\text{i.i.d.}}{\sim} \mu$, the empirical estimator may be written as

$$J_{S,m}(\hat{F}_n) = \frac{1}{m} \sum_{j=1}^{m} S\big(\hat{F}_n(\cdot \mid X_j^\star),\, f(X_j^\star)\big)\,.$$

Such empirical scores are a standard tool for comparing predictive models and for guiding GP hyperparameter selection (Petit et al., 2023).

4. Conformal predictive systems for Gaussian processes.

4.1. Overview of conformal prediction. Conformal prediction (CP), introduced by Vovk et al. (2005), provides a distribution-free method for constructing prediction sets whose finite-sample coverage is guaranteed, on average over the randomness of the data, when the observations are exchangeable, which includes the common case of independent draws. Many variants exist (split/inductive CP, jackknife+, CV+, etc.); see Angelopoulos et al. (2025) for a recent and comprehensive survey. Here we recall the randomized full conformal prediction method (Angelopoulos et al., 2025, Chap. 9), implemented with leave-one-out (LOO) conformal scores, which serves as the theoretical foundation for the CPS–GP method.

Given i.i.d. data $\mathcal{D}_n = \{(X_i, Z_i)\}_{i=1}^n$ with $X_i \in \mathbb{X}$ and $Z_i \in \mathbb{R}$, let $s(x; \mathcal{D}_n)$ denote a regression function providing a point prediction of Z at $X = x$. In this section, we do not restrict to the interpolation setting: the responses may remain intrinsically random even conditionally on the covariates. Conformal prediction constructs a $(1-\alpha)$ prediction set for a new Z_{n+1} at covariate X_{n+1}, assuming that (X_{n+1}, Z_{n+1}) is an independent draw from the same distribution as the data $\mathcal{D}_n$. It relies on a *conformal score* $R(x, z; \mathcal{D}_n) \in \mathbb{R}$, which quantifies the agreement between a candidate pair (x, z) and the dataset $\mathcal{D}_n$. The score is required to be permutation-invariant in its dataset argument: for any permutation σ of $\{1, \dots, n\}$,

$$R(x, z; \sigma\mathcal{D}_n) = R(x, z; \mathcal{D}_n).$$

A common choice is the residual-based score

$$R(x, z; \mathcal{D}_n) = |z - s(x; \mathcal{D}_n)|\,. \tag{4.1}$$

Remark 4.1 (Terminology). In the CP literature, a *conformity score* assigns larger values to better agreement, while a *nonconformity score* (often called a *conformal score*) assigns smaller values to better agreement. We use the term conformal score in this text.

For a candidate value z at X_{n+1}, consider the augmented dataset

$$\mathcal{D}_{n+1}^z = \{(X_1, Z_1), \dots, (X_n, Z_n), (X_{n+1}, z)\},$$

and compute the LOO scores

$$R_i^z = R\big(X_i, Z_i; \mathcal{D}_{n+1}^z \setminus \{(X_i, Z_i)\}\big), \quad i = 1, \dots, n+1.$$

Then, for each z, we compare the score R_{n+1}^z to the n other scores $(R_i^z)_{i=1}^n$. More precisely,

introducing an independent random variable $\tau \sim \mathcal{U}(0,1)$, define the map

$$\pi(z) = \frac{1}{n+1}\sum_{i=1}^{n+1} \mathbb{1}\{R_i^z < R_{n+1}^z\} + \frac{\tau}{n+1}\sum_{i=1}^{n+1} \mathbb{1}\{R_i^z = R_{n+1}^z\} \tag{4.2}$$
$$= \frac{1}{n+1}\sum_{i=1}^{n} \mathbb{1}\{R_i^z < R_{n+1}^z\} + \frac{\tau}{n+1}\Big(1 + \sum_{i=1}^{n} \mathbb{1}\{R_i^z = R_{n+1}^z\}\Big).$$

The quantity $\pi(z)$ takes its values in $[0,1]$ and represents the normalized (randomized) rank of the test score R_{n+1}^z in the augmented dataset: it equals the proportion of scores strictly smaller than R_{n+1}^z, with ties accounted for using τ. While π is not a CDF in the strict sense, it can be interpreted as a predictive CDF proxy that maps each z to its randomized position within the augmented sample.

Proposition 4.2 (Uniform randomized rank and marginal coverage). *Assume $(X_i, Z_i)_{i=1}^{n+1}$ are i.i.d., and the conformal score $R(x, z; \mathcal{D})$ is permutation-invariant in its dataset argument. Let $\mathcal{D}_{n+1} = \mathcal{D}_{n+1}^{Z_{n+1}}$, and let π be defined as in (4.2) with a tie–breaker $\tau \sim \mathcal{U}([0,1))$ independent of the data. Then*

$$\pi(Z_{n+1}) \sim \mathcal{U}\big([0,1)\big).$$

In particular, for any $\alpha \in [0,1]$,

$$\mathsf{P}\big(\pi(Z_{n+1}) \le 1-\alpha\big) = 1-\alpha.$$

Proof. See Appendix A.7. ∎

The value $p(z) := 1-\pi(z)$ is known as the conformal p-value. As in hypothesis testing, $p(z)$ quantifies how extreme the candidate z is relative to the observed sample: values consistent with the data yield large $p(z)$, while discordant values yield smaller $p(z)$. Using this quantity, define the *conformal prediction set*

$$\Gamma_{n,\tau,1-\alpha}(X_{n+1}) = \{\, z \in \mathbb{R} : p(z) > \alpha \,\}, \tag{4.3}$$

which collects the candidate values sufficiently compatible with the data. Proposition 4.2 implies that

$$\mathsf{P}\big(Z_{n+1} \in \Gamma_{n,\tau,1-\alpha}(X_{n+1})\big) = 1-\alpha,$$

with probability taken over the joint randomness of $\mathcal{D}_{n+1}$ and τ. This construction thus provides prediction sets with exact finite-sample coverage, on average over the randomness of the data, for any regression function s, under the sole assumption that the observations are exchangeable, as is the case for independent and identically distributed draws.

Note that the set $\Gamma_{n,\tau,1-\alpha}(X_{n+1})$ need not be an interval in general, as the comparisons $R_i^z < R_{n+1}^z$ may switch more than once with z. It will, however, reduce to an interval in our CPS–GP construction (see Section 4.3).

For a comprehensive overview of conformal prediction (CP) and its variants, we refer the reader to Angelopoulos et al. (2025). The discussion above focused on the full conformal method, which forms the basis for CPS. Other variants have been developed to reduce computational cost while preserving finite-sample coverage. Split conformal methods avoid repeated model fitting by using a hold-out sample, at the cost of wider prediction sets. Jackknife-based

methods such as jackknife+ and CV+ provide an intermediate option, leveraging leave-one-out or cross-validation fits to improve efficiency without the full burden of refitting.

4.2. Conformal predictive systems. Standard conformal prediction produces prediction sets, but not full predictive distributions. To overcome this limitation, Vovk et al. (2017b, 2019) introduced the framework of *conformal predictive systems* (CPS), which extend conformal prediction to output predictive CDFs at each test covariate X_{n+1}. The approach builds on the framework of full conformal prediction and will be interpreted below through the lens of μ-probabilistic calibration introduced in Section 3.3.

Randomized predictive systems. CPS are formalized through the notion of a *randomized predictive system* (RPS). An RPS is intended to act as a predictive CDF that remains calibrated in finite samples. Given (i) a dataset $\mathcal{D}_n$ of n observations, (ii) a test covariate–candidate outcome pair $(x_{n+1}, z) \in \mathbb{X} \times \mathbb{R}$, and (iii) an auxiliary random variable $\tau \in [0,1]$ for tie-breaking, an RPS returns a value in $[0,1]$ that represents the predictive probability assigned to the event $\{Z_{n+1} \le z\}$ at covariate x_{n+1}.

Formally, an RPS is a function G satisfying the conditions in Definition 4.3.

Definition 4.3 (Shen et al. (2018); Vovk et al. (2019)). *A measurable function*

$$G : (\mathbb{X} \times \mathbb{R}) \times (\mathbb{X} \times \mathbb{R})^n \times [0,1] \to [0,1]$$

is called a randomized predictive system (RPS) if it satisfies:

(R.1) *(i) For each dataset $\mathcal{D}_n$ and test covariate $x_{n+1} \in \mathbb{X}$, the map*

$$(z, \tau) \mapsto G\big((x_{n+1}, z),\, \mathcal{D}_n,\, \tau\big)$$

is nondecreasing in both $z \in \mathbb{R}$ and $\tau \in [0,1]$.

(ii) For each $\mathcal{D}_n$ and $x_{n+1} \in \mathbb{X}$,

$$\lim_{z \to -\infty} G\big((x_{n+1}, z),\, \mathcal{D}_n,\, 0\big) = 0, \qquad \lim_{z \to +\infty} G\big((x_{n+1}, z),\, \mathcal{D}_n,\, 1\big) = 1.$$

(R.2) If $(X_i, Z_i)_{i=1}^n$ are i.i.d. from a distribution F, $(X_{n+1}, Z_{n+1}) \sim F$ independently, and $\tau \sim \mathcal{U}(0,1)$ is independent, then, with probability taken over the joint distribution of $(\mathcal{D}_n, (X_{n+1}, Z_{n+1}), \tau)$,

$$\forall \alpha \in [0,1], \quad \mathsf{P}\left(G\big((X_{n+1}, Z_{n+1}),\, \mathcal{D}_n,\, \tau\big) \le \alpha\right) = \alpha.$$

Requirement (R.1) ensures that, for fixed data and test covariate, the map $(z, \tau) \mapsto G((x_{n+1}, z), \mathcal{D}_n, \tau)$ behaves like a CDF: it is monotone in both arguments, and it ranges from 0 at $z \to -\infty$ (with $\tau = 0$) to 1 at $z \to +\infty$ (with $\tau = 1$). Randomization through τ provides principled tie handling in conformal ranking and yields exact uniformity of the randomized PIT and exact mass at CDF jumps; it does not remove the stepwise nature in z for a fixed τ.

Requirement (R.2) is a probabilistic calibration property: if $(X_i, Z_i)_{i=1}^n, (X_{n+1}, Z_{n+1})$ are i.i.d. and $\tau \sim \mathcal{U}(0,1)$ is independent, the randomized PIT

$$U := G\big((X_{n+1}, Z_{n+1}),\, \mathcal{D}_n,\, \tau\big)$$

is uniformly distributed on $[0,1]$.

Remark 4.4 (Averaging over the data). Condition (R.2) ensures that the randomized PIT associated with a CPS is uniform when averaged over the joint distribution of the observed

data, the test point, and the randomization variable τ. In the μ-probabilistic framework of Section 3.3, this would correspond to μ-probabilistic calibration *on average over the randomness of the observed data*:

$$\mathsf{E}\left[\mathsf{P}\left\{ G\big((X, f(X)), \mathcal{D}_n, \tau\big) \le u \;\middle|\; \mathcal{D}_n \right\}\right] = u, \qquad \forall u \in [0,1]$$

(where the expectation is taken over the randomness of $\mathcal{D}_n$). By contrast, μ-probabilistic calibration as defined in Section 3.3 is a *post-data* property: it requires the conditional equality

$$\mathsf{P}\left\{ G\big((X, f(X)), \mathcal{D}_n, \tau\big) \le u \;\middle|\; \mathcal{D}_n \right\} = u$$

to hold for each realization of $\mathcal{D}_n$, which CPS do not guarantee for finite n. This post-data notion aligns with the *training-conditional* calibration used in the conformal prediction literature (see, e.g., Angelopoulos et al., 2025, Ch. 4), while the property "on average over the data" corresponds to *marginal* calibration.

Conformal predictive distribution. Using the same setting as in Section 4.1, let $\mathcal{D}_{n+1}^z$ be the augmented dataset and $(R_i^z)_{i=1}^{n+1}$ the associated leave-one-out scores. Recall from (4.2) the function $\pi(z)$, which gives the (normalized, possibly randomized) rank of the test score among the augmented scores. For notational consistency, we introduce the equivalent notation

$$G\big((x,z), \mathcal{D}_n, \tau\big) := \frac{1}{n+1}\sum_{i=1}^{n+1} \mathbb{1}\{R_i^z < R_{n+1}^z\} + \frac{\tau}{n+1}\sum_{i=1}^{n+1} \mathbb{1}\{R_i^z = R_{n+1}^z\}, \tag{4.4}$$

with $\tau \sim \mathcal{U}(0,1)$ independent of $(\mathcal{D}_n, (X_{n+1}, Z_{n+1}))$. We now check whether G is an RPS. Requirement (R.2) holds provided the observation pairs $(X_i, Z_i)_{i=1}^{n+1}$ are exchangeable (e.g., i.i.d.) and the score $R(\cdot,\cdot;\cdot)$ is permutation-invariant in its dataset argument; then, by Proposition 4.2, $G\big((X_{n+1}, Z_{n+1}), \mathcal{D}_n, \tau\big) \sim U(0,1)$. The monotonicity condition (R.1) depends on the chosen conformal score. When using the absolute-residual score (4.1), G may fail to be monotone in z, and the construction does not define an RPS. Vovk et al. (2019) provide general sufficient conditions under which a function of the form (4.4) defines a valid RPS: broadly, the conformal scores must vary with the candidate label z in an order-preserving manner. We will later verify that these properties hold for the score structure used in the CPS–GP method.

When (R.1) and (R.2) hold, the map $z \mapsto G\big((x,z), \mathcal{D}_n, \tau\big)$ acts as a (randomized) predictive CDF at x; we then call it a *conformal predictive distribution (CPD)* and write

$$\hat{F}^{\mathrm{CPD}}_{n,\tau}(z \mid x) := G\big((x,z), \mathcal{D}_n, \tau\big).$$

Evaluating $\hat{F}^{\mathrm{CPD}}_{n,\tau}(\cdot \mid X_{n+1})$ at the realized outcome Z_{n+1} yields a randomized PIT value that is uniform, so the predictive distribution is probabilistically calibrated when averaging over the joint randomness of $\mathcal{D}_n$, (X_{n+1}, Z_{n+1}), and τ ("marginal" calibration).

Remark 4.5 (Terminology). In Vovk et al. (2019), the candidate CPD in (4.4) is called a *conformal transducer*. A *conformal predictive system (CPS)* is such a map that also satisfies the RPS axioms (Definition 4.3); for a CPS, the induced predictive CDF is termed a *conformal predictive distribution (CPD)*. For simplicity, we avoid "transducer" in the main text: we say "candidate CPD" for (4.4), and "CPD" once the RPS conditions are met.

4.3. CPS for GP interpolation. We derive the form of the CPS when the conformal score is constructed from Gaussian process (GP) interpolation, following the approach of Vovk et al. (2017a) for kernel ridge regression. In this setting, the CPS admits a closed-form expression and is a step function of the candidate outcome z, with jumps determined by thresholds derived from the GP posterior.

Setup and assumptions. Let $f : \mathbb{X} \subset \mathbb{R}^d \to \mathbb{R}$ be an unknown deterministic function. We observe a noiseless sample $\mathcal{D}_n = \{(X_i, Z_i)\}_{i=1}^n$ with $X_i \sim \mu$ i.i.d. and $Z_i = f(X_i)$, where μ is the design measure introduced in Section 3.1. A GP prior $\xi \sim \mathrm{GP}(m_\theta, k_\theta)$ is specified with mean m_θ and covariance k_θ depending on hyperparameters θ. In the derivation below, θ is treated as a fixed constant, independent of $\mathcal{D}_n$ (for instance, pre-specified from prior knowledge or selected once in a separate, external step). This assumption yields closed-form expressions and preserves the exchangeability property needed for CPS validity.

NB. Although $(X_i, Z_i)_{i=1}^n$ and X_{n+1} are random, in the algebra below we condition on a fixed dataset $\mathcal{D}_n = \{(x_i, z_i)\}_{i=1}^n$ and a fixed test location x_{n+1}. All GP posterior quantities such as m_n and σ_n are then deterministic functions of $(\mathcal{D}_n, x_{n+1}, \theta)$. The same holds for the auxiliary coefficients introduced later (e.g. the slopes and thresholds in Proposition 4.6). We return to the random-design setting when stating calibration results.

For a test covariate x_{n+1} and candidate outcome $z \in \mathbb{R}$, form the augmented dataset $\mathcal{D}_{n+1}^z = \mathcal{D}_n \cup \{(x_{n+1}, z)\}$ and use the standardized residual score (Papadopoulos, 2024). The score at the test point is

$$R_{n+1}^z = \frac{z - m_n(x_{n+1})}{\sigma_n(x_{n+1})}, \tag{4.5}$$

where m_n and σ_n denote the GP posterior mean and standard deviation given $\mathcal{D}_n$ (under the same fixed θ). For each observed pair (x_i, z_i), the leave–one–out score computed on $\mathcal{D}_{n+1}^z \setminus \{(x_i, z_i)\}$ is

$$R_i^z = \frac{z_i - m_{n+1,-i}(x_i)}{\sigma_{n+1,-i}(x_i)}. \tag{4.6}$$

Under these assumptions, the difference between the test score and each data–point score is affine in z.

Proposition 4.6 (Affine differences). *Fix $\mathcal{D}_n$, x_{n+1}, and θ. For each $i = 1, \ldots, n$, there exist a slope $\beta_i > 0$ and a threshold $c_i \in \mathbb{R}$, depending only on $(\mathcal{D}_n, x_{n+1}, \theta)$, such that*

$$R_{n+1}^z - R_i^z \;=\; \beta_i\,(z - c_i), \qquad z \in \mathbb{R}.$$

Proof. See Appendix A.8, which provides explicit expressions for β_i and c_i. ∎

This affine representation will be used in Proposition 4.9 to show that the resulting conformal predictive distribution is a step function with jumps at the thresholds $\{c_i\}_{i=1}^n$.

Remark 4.7 (Hyperparameters selected on $\mathcal{D}_n$). In practice, θ is often selected from the same data $\mathcal{D}_n$ (e.g., by maximum likelihood, Petit et al., 2023) and then kept fixed when evaluating conformal scores. With θ fixed after selection, Proposition 4.6 still applies.

This choice breaks the permutation invariance required by full conformal prediction: $\theta = \theta(\mathcal{D}_n)$ is a function of the n observed pairs only, not of the candidate (X_{n+1}, z). Under a

permutation of the $n+1$ pairs, θ should change to reflect the permuted first n pairs. Keeping $\theta(\mathcal{D}_n)$ fixed prevents this, so the joint distribution of conformity scores is no longer invariant under permutation, and exact finite–sample validity does not hold.

If formal validity is needed, one can either (i) use a split scheme (select θ on a disjoint subset, then conformalize on the remainder), or (ii) re-select θ for each augmented dataset $\mathcal{D}_{n+1}^z \setminus \{(X_i, Z_i)\}$ (or over a grid in z). The latter restores exchangeability but is computationally expensive and removes the simple affine dependence on z.

Remark 4.8 (Extensions to universal and intrinsic kriging). The framework extends beyond the case of a known mean function and strictly positive definite kernel. It also applies to *universal kriging* models with an unknown linear mean $m(x) = h(x)^\top \beta$ (for known regressors $h(x) \in \mathbb{R}^q$ and unknown coefficients β) and to *intrinsic kriging* based on conditionally positive definite kernels (Matheron, 1973; Chilès and Delfiner, 1999). The corresponding modifications to the leave-one-out formulas and to the affine representation of $R_{n+1}^z - R_i^z$ are outlined in Appendix A.8.

CPD construction. Specializing the map (4.4) to the standardized-residual scores (4.5)–(4.6), we define

$$\hat{F}_{n,\tau}^{\text{CPS-GP}}(z \mid x_{n+1}) \;:=\; G\big((x_{n+1}, z),\, \mathcal{D}_n,\, \tau\big), \tag{4.7}$$

with independent $\tau \sim \mathcal{U}(0,1)$.

The function $\hat{F}_{n,\tau}^{\text{CPS-GP}}(\,\cdot \mid x_{n+1})$ is a candidate CPD at the test location x_{n+1}. The next result shows that, in our GP interpolation setting, this function has a stepwise form with discontinuities at the thresholds $\{c_i\}_{i=1}^n$. We then verify that it satisfies the RPS axioms, showing that $\hat{F}_{n,\tau}^{\text{CPS-GP}}$ defines a valid CPD.

Proposition 4.9 (Stepwise form of $\hat{F}_{n,\tau}^{\text{CPS-GP}}$). *Fix $\mathcal{D}_n$ and x_{n+1}, and use the standardized residual score. Let $\{c_i\}_{i=1}^n$ be the thresholds from Proposition 4.6, and order them as*

$$-\infty =: c_{(0)} < c_{(1)} \le \cdots \le c_{(n)} < c_{(n+1)} := \infty.$$

Then, for any $\tau \in [0,1]$,

$$\hat{F}_{n,\tau}^{\text{CPS-GP}}(z \mid x_{n+1}) = \begin{cases} \dfrac{i+\tau}{n+1}, & \text{if } z \in (c_{(i)}, c_{(i+1)}), \quad i = 0, \dots, n, \\[2ex] \dfrac{i' - 1 + \tau\,(i'' - i' + 2)}{n+1}, & \text{if } z = c_{(i)} \text{ for some } i, \end{cases}$$

where, for a tie point $z = c_{(i)}$, the block endpoints

$$i' := \min\{r :\; c_{(r)} = c_{(i)}\}, \qquad i'' := \max\{r :\; c_{(r)} = c_{(i)}\}$$

denote the first and last indices of the tied block. In particular, $z \mapsto \hat{F}_{n,\tau}^{\text{CPS-GP}}(z \mid x_{n+1})$ is a step function with jumps at $\{c_i\}_{i=1}^n$.

Proof. Since $\beta_i > 0$, the relative ordering of R_{n+1}^z and R_i^z can change only at $z = c_i$. Sorting these thresholds partitions $\mathbb{R}$ into the intervals $(c_{(i)}, c_{(i+1)})$. If $z \in (c_{(i)}, c_{(i+1)})$, then z exceeds exactly i thresholds, so R_{n+1}^z exceeds i observed scores and is smaller than the others.

Thus

$$\hat{F}^{\text{CPS-GP}}_{n,\tau}(z \mid x_{n+1}) = \frac{i+\tau}{n+1}.$$

If $z = c_{(i)}$ lies in a block of ties $c_{(i')} = \cdots = c_{(i'')} = c_{(i)}$, then the test score exceeds $i'-1$ scores, equals $(i''-i'+1)$ scores, and is smaller than the rest. The randomized rank with tie–breaking is

$$\hat{F}^{\text{CPS-GP}}_{n,\tau}(c_{(i)} \mid x_{n+1}) = \frac{i'-1+\tau\,(i''-i'+2)}{n+1}.$$

■

Up to this point, we conditioned on a fixed dataset $\mathcal{D}_n = \{(x_i, z_i)\}_{i=1}^n$ and a fixed test location x_{n+1}. We now lift this conditioning: let $(X_i, Z_i)_{i=1}^n$ be i.i.d., let (X_{n+1}, Z_{n+1}) be an independent draw from the same distribution, and let $\tau \sim \mathcal{U}(0,1)$ be independent of everything else. Throughout the calibration statements below, the hyperparameters θ used in the scores are taken as fixed independently of the sample (or chosen on a disjoint split).

We refer to this construction as the *conformal predictive system for Gaussian-process interpolation* (CPS–GP). The next result shows that, under the assumptions of Section 4.3, the map $z \mapsto \hat{F}^{\text{CPS-GP}}_{n,\tau}(z \mid x_{n+1})$ satisfies the RPS axioms and therefore defines a valid CPD.

Proposition 4.10. *Assume the setup of Section 4.3: interpolation with known mean function m and strictly positive definite kernel k, the standardized residual score, and an independent tie–breaker $\tau \sim \mathcal{U}(0,1)$. If, in addition, $(X_i, Z_i)_{i=1}^n$ and (X_{n+1}, Z_{n+1}) are exchangeable (e.g., i.i.d.), then*

$$\hat{F}^{\text{CPS-GP}}_{n,\tau}(\,\cdot \mid x_{n+1})$$

satisfies the RPS axioms (Definition 4.3); hence CPS–GP *defines a valid CPD.*

Proof. (R.1): By Proposition 4.9, $z \mapsto \hat{F}^{\text{CPS-GP}}_{n,\tau}(z \mid x_{n+1})$ is nondecreasing, with the following properties: $\lim_{z\to-\infty} \hat{F}^{\text{CPS-GP}}_{n,0}(z \mid x_{n+1}) = 0$ and $\lim_{z\to+\infty} \hat{F}^{\text{CPS-GP}}_{n,1}(z \mid x_{n+1}) = 1$, and it is nondecreasing in τ. (R.2): By exchangeability and independence of $\tau \sim \mathcal{U}(0,1)$, $\hat{F}^{\text{CPS-GP}}_{n,\tau}(Z_{n+1} \mid X_{n+1}) \sim \mathcal{U}(0,1)$ (Proposition 4.2). ■

By Proposition 4.10, the randomized PIT based on $\hat{F}^{\text{CPS-GP}}_{n,\tau}(\cdot \mid X)$ is uniform when averaged over the joint randomness of $(\mathcal{D}_n, X, \tau)$. Thus, CPS–GP achieves design-marginal probabilistic calibration on average over the data. This guarantee is not data-conditional: for a fixed $\mathcal{D}_n$ the μ-PIT need not be exactly uniform. In numerical experiments on benchmark functions (Section 6), post-data calibration diagnostics (KS–PIT, IAE, and coverage) computed on an independent test design $X^\star \sim \mu$ remain close to their ideal targets.

In practical terms, Proposition 4.9 shows that $\hat{F}^{\text{CPS-GP}}_{n,\tau}(\cdot \mid x_{n+1})$ is a piecewise–constant function whose jumps occur at data–dependent thresholds $\{c_i\}_{i=1}^n$. Figure 3 contrasts this stepwise conformal predictive distribution with the smooth Gaussian posterior CDF.

Prediction intervals. For a test location x_{n+1} and a tie–breaking variable $\tau \sim \mathcal{U}(0,1)$, the central $(1-\alpha)$ interval is defined as the (randomized) half–open interval

$$\mathcal{C}^{\text{CPS-GP}}_{n,\tau,1-\alpha}(x_{n+1}) = \Big[\big(\hat{F}^{\text{CPS-GP}}_{n,\tau}\big)^{-1}(\alpha/2 \mid x_{n+1}),\ \big(\hat{F}^{\text{CPS-GP}}_{n,\tau}\big)^{-1}(1-\alpha/2 \mid x_{n+1})\Big).$$

Because $\hat{F}^{\text{CPS-GP}}_{n,\tau}(\cdot \mid x_{n+1})$ is stepwise, each endpoint is a threshold (see Figure 4).

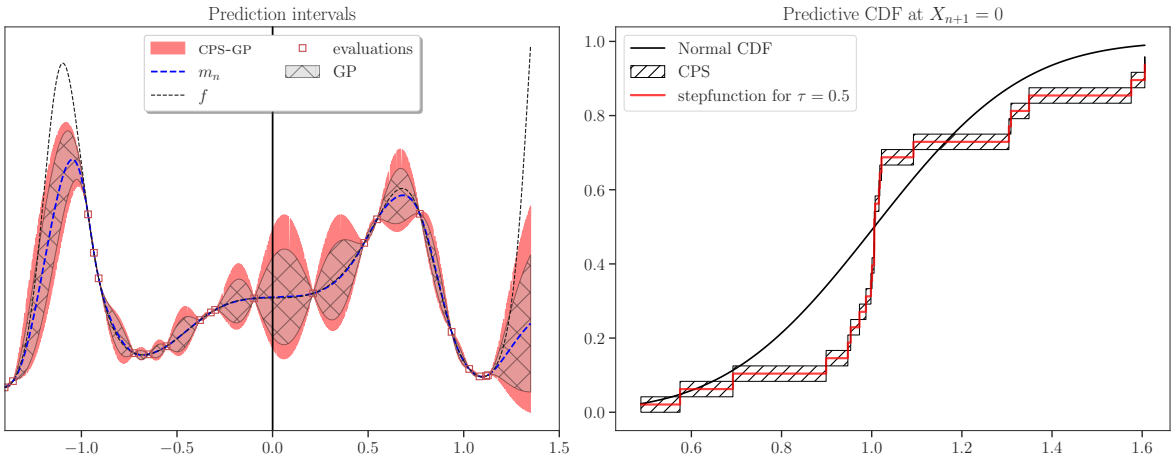


Figure 3: Stepwise CPD (with $\tau = 0.5$) compared to the Gaussian posterior CDF at $x_{n+1} = 0$. The hyperparameters of the GP are fitted on $\mathcal{D}_n$ and then kept fixed. The CPD has discrete jumps at thresholds determined by GP interpolation, while the Gaussian posterior yields a smooth curve.

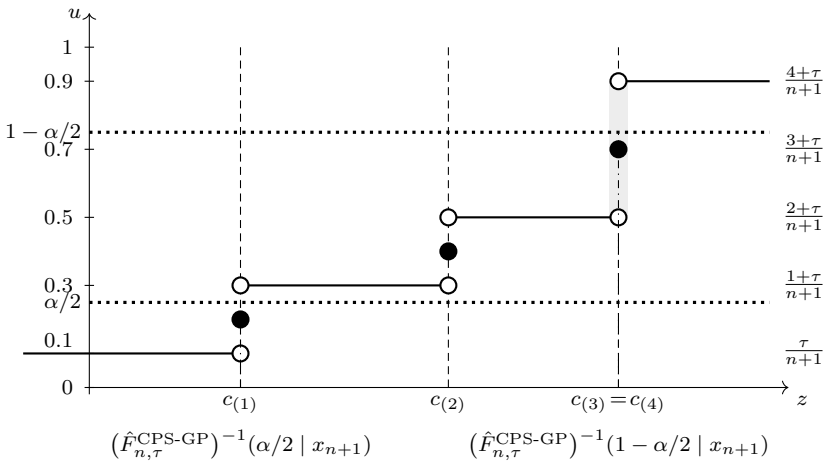


Figure 4: Conformal predictive distribution $\hat{F}^{\text{CPS-GP}}_{n,\tau}(\cdot \mid x_{n+1})$ for fixed τ. On each open interval $(c_{(i)}, c_{(i+1)})$ the value is $(i+\tau)/(n+1)$ (solid segments). At a singleton threshold $c_{(i)}$ the actual value is $(i-1+2\tau)/(n+1)$ (filled circle), with open circles marking the left and right limits. At a tie $c_{(i')} = \cdots = c_{(i'')}$, the value is $(i'-1+\tau(i''-i'+2))/(n+1)$ (filled circle), again between the open-circle limits.

Remark 4.11. From the definition of the generalized inverse $F^{-1}(p) = \inf\{z : \ F(z) \geq p\}$, we have

$$\big(\hat{F}^{\text{CPS-GP}}_{n,\tau}\big)^{-1}(p \mid x_{n+1}) = \min\Big\{ c_{(r)} : \ \hat{F}^{\text{CPS-GP}}_{n,\tau}\big(c^{+}_{(r)} \mid x_{n+1}\big) \ \geq \ p \Big\}, \qquad p \in (0,1),$$

where $F(z^+) := \lim_{t\downarrow 0} F(z+t)$. Equivalently, since on $(c_{(i)}, c_{(i+1)})$ the level is $(i+\tau)/(n+1)$,

$$\big(\hat{F}^{\text{CPS-GP}}_{n,\tau}\big)^{-1}(p \mid x_{n+1}) = c_{(i)}, \qquad i = \min\Big\{ r \in \{1,\dots,n\} : \ \frac{r+\tau}{n+1} \ \geq \ p \Big\}.$$

If there is a tie block $c_{(i')} = \cdots = c_{(i'')}$, the left and right limits at the block are $\frac{i'-1+\tau}{n+1}$ and $\frac{i''+\tau}{n+1}$, respectively, and

$$\big(\hat{F}^{\text{CPS-GP}}_{n,\tau}\big)^{-1}(p \mid x_{n+1}) \ = \ c_{(i')} \quad \text{for} \quad p \in \Big(\tfrac{i'-1+\tau}{n+1}, \ \tfrac{i''+\tau}{n+1}\Big].$$

(Any index in $\{i', \dots, i''\}$ yields the same threshold since $c_{(i')} = \cdots = c_{(i'')}$.)

By probabilistic calibration with randomized tie–breaking, the interval $\mathcal{C}^{\text{CPS-GP}}_{n,\tau,1-\alpha}(X_{n+1})$ achieves exact marginal coverage:

$$\mathsf{P}\left\{ Z_{n+1} \in \mathcal{C}^{\text{CPS-GP}}_{n,\tau,1-\alpha}(X_{n+1}) \right\} = 1-\alpha, \quad \text{averaged over } (\mathcal{D}_n, X_{n+1}, \tau).$$

Replacing the right-open interval $[a,b)$ by the closed interval $[a,b]$ yields conservative coverage, with equality restored by boundary randomization as in Proposition 3.1.

Remark 4.12. Fixing τ yields a deterministic, stepwise predictive distribution. Averaging over τ,

$$\bar{F}_n(z \mid x_{n+1}) \ = \ \mathsf{E}_\tau \left[\hat{F}^{\text{CPS-GP}}_{n,\tau}(z \mid x_{n+1}) \right],$$

keeps the same jump locations and replaces each plateau level by its mean over τ; it does not remove discontinuities in z. Exact PIT uniformity requires random τ.

Proposition 4.13 (Finite bounds). *Assume the thresholds $\{c_i\}$ are all distinct and fix $\tau \in (0,1)$. Then the endpoints of $\mathcal{C}^{\text{CPS-GP}}_{n,\tau,1-\alpha}(x_{n+1})$ are finite if and only if*

$$\frac{\alpha}{2} > \frac{\tau}{n+1} \quad \textit{and} \quad \frac{\alpha}{2} \geq \frac{1-\tau}{n+1},$$

or equivalently,

$$\alpha \ \geq \ \frac{2}{n+1} \max\{\tau, \, 1-\tau\},$$

with strict inequality required when the maximum is τ.

Proof. See Appendix A.9. ∎

This shows that for small n or small α, CPS intervals may have infinite endpoints, a common feature of conformal methods.

Computational complexity. We now examine the computational complexity of CPS–GP.

Proposition 4.14. *Let n denote the sample size and m the number of prediction points. The CPD based on GP interpolation can be computed with*

$$O(n^3) \textit{ precomputation} \quad \textit{and} \quad O(n^2 + n\log n) \textit{ per prediction point.}$$

The total computational cost is therefore

$$O(n^3) \ + \ m\, O(n^2 + n\log n).$$

If only interval endpoints are needed (without constructing the full stepwise CPD), the sorting

step can be omitted, reducing the per–point cost to $O(n^2 + n)$.

Once the Cholesky factorization of K_n is computed at cost $O(n^3)$, evaluating the GP posterior mean and variance at a new location x_{n+1} costs $O(n^2)$. CPS–GP involves the same $O(n^2)$ operations, plus $O(n)$ for computing the thresholds c_i and $O(n \log n)$ for sorting them. Thus, the dominant per–point complexity remains $O(n^2)$, with an additional $O(n \log n)$ term for sorting beyond standard GP prediction.

5. Bayesian calibration using a generalized normal distribution.

5.1. Generalized normal model for prediction errors. In Section 4.3, leave–one–out residuals were treated nonparametrically to construct a conformal predictive system. An alternative is to regard these residuals as samples from a parametric model and to apply a post–hoc calibration of the GP posterior. The posterior mean is kept as the point predictor, while the predictive variability is adjusted by fitting a generalized normal model to the residuals. In this way, one obtains explicit control over the tail behavior and predictive distributions that are, in principle, easier to use within GP–based sequential design algorithms. This parametric modeling of the residuals forms the basis of the BCR–GP method (Bayesian–calibrated residuals for Gaussian processes).

In more detail, let $X \sim \mu$ be a random input, where μ is a fixed design measure on $\mathbb{X}$. As above, we consider a Gaussian process prior on the unknown function $f : \mathbb{X} \to \mathbb{R}$, $\xi \sim \mathrm{GP}(m, k)$, and regard f as a fixed realization of ξ. Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mu$ be input locations. The observed dataset is then $\mathcal{D}_n = \{(X_i, f(X_i))\}_{i=1}^n$. For any $x \in \mathbb{X}$, the GP posterior has mean $m_n(x)$ and standard deviation $\sigma_n(x)$. Define the normalized prediction error

$$R_n(x, f(x)) = \frac{f(x) - m_n(x)}{\sigma_n(x)}, \qquad \text{with } R_n(x, f(x)) = 0 \text{ if } \sigma_n(x) = 0.$$

The *μ-marginal distribution* of the normalized error is the distribution of $R_n(X, f(X))$ when $X \sim \mu$ and $\mathcal{D}_n$ is fixed. This is the distributional object that will be calibrated.

For context, under the Bayesian GP model with fixed hyperparameters and for any fixed x with $\sigma_n(x) > 0$, the posterior predictive error satisfies

$$R_n(x, \xi(x)) \mid (x, \mathcal{D}_n) \ \sim \ \mathcal{N}(0, 1).$$

Thus $R_n(x, \xi(x))$ is a pivot: its conditional distribution does not depend on the parameters of the GP prior ξ (with the convention $R_n(x, \cdot) = 0$ when $\sigma_n(x) = 0$).

This pointwise pivot property does not imply that the normalized residual $R_n(X, f(X))$ is Gaussian when f is a fixed but unknown sample path and $X \sim \mu$. For such an f, $R_n(x, f(x))$ is deterministic at each x, and the only randomness comes from sampling X according to μ. The resulting distribution of $R_n(X, f(X))$ aggregates these deterministic values across x and generally deviates substantially from $\mathcal{N}(0, 1)$, especially in the tails, as illustrated in Figure 5.

Nevertheless, normalized residuals remain the natural object to model: the pivot property justifies using them as scale–free standardized errors, while their μ-marginal distribution, though not Gaussian for a fixed f, captures the relevant predictive variability to be calibrated. Our objective is to build a parametric approximation of this distribution.

In practice however, the μ-marginal distribution of $R_n(X, f(X))$, $X \sim \mu$, is not directly

observable. To approximate it, we use the leave–one–out standardized residuals

$$R_{n,-i} = \frac{Z_i - m_{n,-i}(X_i)}{\sigma_{n,-i}(X_i)}, \qquad i = 1, \ldots, n,$$

where $m_{n,-i}$ and $\sigma_{n,-i}$ are computed from $\mathcal{D}_n \setminus \{(X_i, Z_i)\}$. We adopt the working assumption that the conditional empirical distribution of $\{R_{n,-i}\}_{i=1}^n$, given $\mathcal{D}_n$, provides a proxy for the μ-marginal distribution of $R_n(X, f(X))$. This assumption is plausible when the GP fit is stable so that replacing (m_n, σ_n) by $(m_{n,-i}, \sigma_{n,-i})$ has negligible global effect. Intuitively, each $R_{n,-i}$ is an out-of-sample standardized error at a location distributed like a future input.

To obtain a tractable post-hoc calibration, we fit a parametric model to this proxy distribution. Specifically, we use the generalized normal (GN) family

$$\mathcal{GN}(\beta, 0, \lambda), \qquad \beta > 0,\ \lambda > 0,$$

with density given in Appendix C. The scale λ controls dispersion, while the shape parameter β tunes the tail behavior: $\beta = 2$ corresponds to the Gaussian case, and $\beta = 1$ to the Laplace case. The GP posterior mean m_n is retained as the point predictor, and uncertainty is recalibrated through the GN fit. This yields a closed-form predictive distribution that is easier to handle than a purely empirical residual distribution and integrates naturally into GP-based sequential search algorithms.

Figure 5 illustrates the motivation: empirical quantiles of leave-one-out residuals typically diverge from Gaussian quantiles, while a GN fit adapts to tail deviations more accurately.

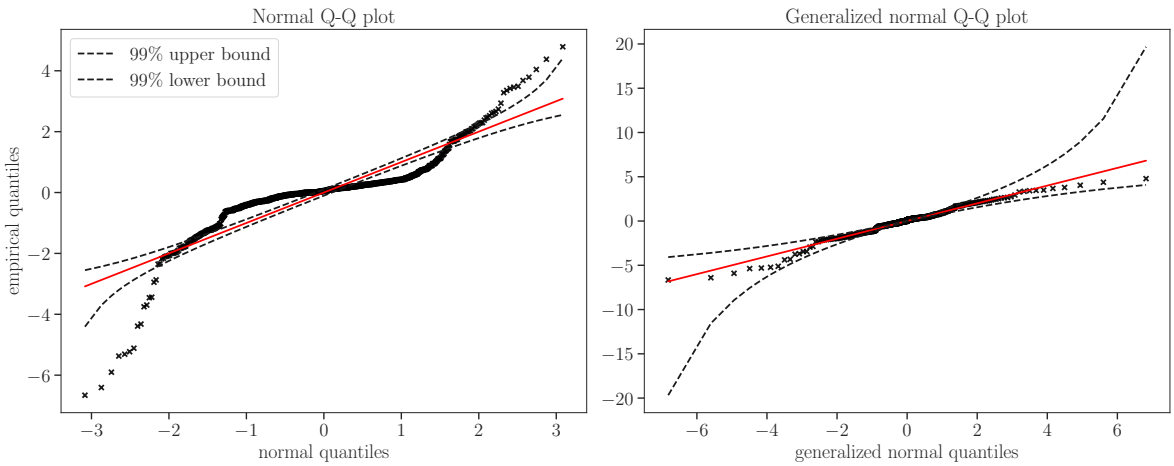


Figure 5: Empirical quantiles of standardized residuals $R = (Z - m_n(X))/\sigma_n(X)$ at 1000 test points, compared with those of the standard normal and a fitted generalized normal distribution. Shaded bands indicate 99% pointwise confidence intervals. Example: Goldstein–Price function in $d = 2$, $n = 40$.

Remarks. If the design points are not exactly μ-distributed, the same idea can be applied by reweighting the residuals according to importance weights that map the empirical design distribution to μ.

Alternative residual families, such as the Student-t, could also be considered. The generalized normal is adopted here as a convenient two-parameter choice, spanning Gaussian to

Laplace tails, with closed-form CDF and straightforward estimation.

5.2. Bayesian parameter selection. We now detail the Bayesian parameter selection strategy that completes the BCR–GP method. The generalized normal model involves two free parameters, the shape β and scale λ, which we group as $\theta = (\beta, \lambda)$. These must be estimated from the leave-one-out residuals to obtain a predictive distribution. Rather than relying on point estimation (e.g., maximum likelihood), we adopt a Bayesian framework: a posterior distribution over θ is constructed from the data, and a selection rule is applied to choose a representative parameter. This approach connects to the idea of *Bayesian tolerance intervals* (Meeker et al., 2017), where posterior draws are used to control conservatism in uncertainty quantification. Two complementary rules are considered below, depending on whether the goal is conservative coverage or μ-probabilistic calibration.

Formally, we place weakly informative priors

$$\beta \sim \mathcal{U}(0, a), \qquad \lambda \sim \mathcal{U}(0, b),$$

and define the posterior given the LOO residuals $\{R_{n,-i}\}_{i=1}^n$ as

$$p(\theta \mid \{R_{n,-i}\}_{i=1}^n) \;\propto\; \prod_{i=1}^n p(R_{n,-i} \mid \theta)\, \mathbb{1}_{\{0<\beta<a\}}\, \mathbb{1}_{\{0<\lambda<b\}}.$$

Posterior draws $\theta_j = (\beta_j, \lambda_j)$, $j = 1, \dots, K$, are obtained by MCMC sampling. Two rules for selecting θ^* are then defined.

Rule 1: Conservative prediction. For each posterior draw $\theta_j = (\beta_j, \lambda_j)$, compute the variance

$$v_j \;=\; \mathrm{var}\left(\mathcal{GN}(\theta_j)\right) \;=\; \lambda_j^2 \frac{\Gamma(3/\beta_j)}{\Gamma(1/\beta_j)}.$$

The values $\{v_j\}_{j=1}^K$ form a Monte Carlo sample from the posterior distribution of $\mathrm{var}(\mathcal{GN}(\theta))$ given the LOO residuals. Fix a tolerance level $\delta \in (0, 1)$ (small in practice) and select θ^* such that $v^* = \mathrm{var}(\mathcal{GN}(\theta^*))$ is the $(1-\delta)$-quantile of this distribution.

The rationale is that the true parameter θ is unknown. Selecting the $(1-\delta)$-quantile v^* of the posterior variance provides an upper credible bound: with posterior probability at least $1-\delta$, the variance of the true residual distribution is no larger than v^*. Equivalently, the resulting predictive distribution is at least as dispersed as the true one with probability $1-\delta$. Only a fraction δ of posterior draws correspond to even more dispersed (more conservative) models. Hence, smaller values of δ yield more pessimistic (wider) predictions, whereas larger values produce less conservative ones. This quantile rule parallels the construction of Bayesian tolerance bounds in Meeker et al. (2017).

Rule 2: Probabilistic calibration. Draw posterior samples $\{\theta_j\}_{j=1}^K$. For any parameter θ, let F_θ denote the CDF of $\mathcal{GN}(\theta)$ and F_θ^{-1} its quantile function.

Consider two models F_{θ_i} (generating) and F_{θ_j} (forecasting). For a residual $R \sim F_{\theta_i}$, the PIT under F_{θ_j} is $U = F_{\theta_j}(R)$, whose CDF is

$$G_{j|i}(u) = \mathsf{P}\big(F_{\theta_j}(R) \le u\big) = F_{\theta_i}\big(F_{\theta_j}^{-1}(u)\big), \qquad u \in [0, 1].$$

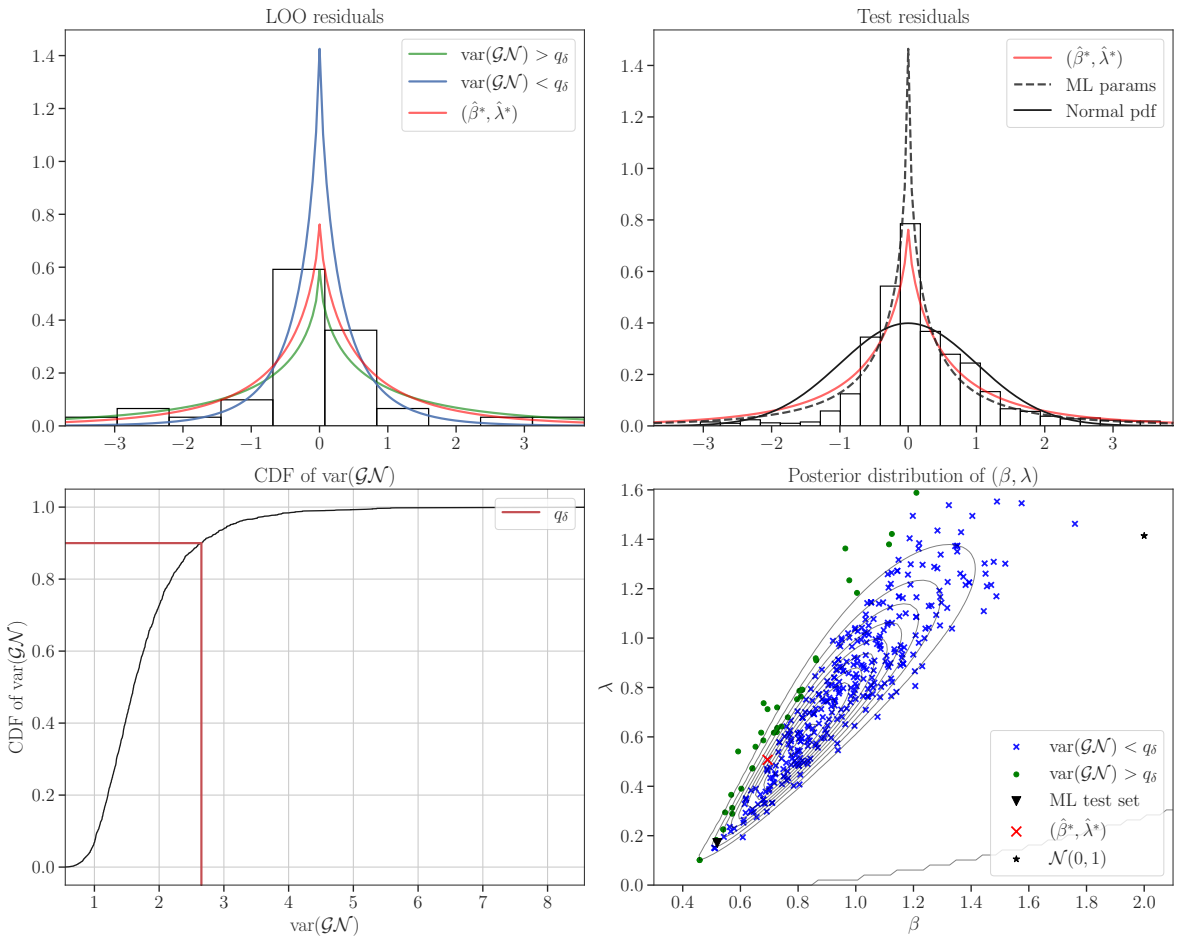


Figure 6: Illustration of the BCR–GP method on the Goldstein–Price function with $n = 40$ observations. **First row**: Left– Histogram of the LOO normalized residuals overlaid with predictive density functions corresponding to various parameter configurations. The red curve corresponds to the pdf with parameters $(\hat{\beta}^*, \hat{\lambda}^*)$; the green and blue curves represent more and less conservative forecasts, respectively. Right– Histogram of a sample of normalized residuals on a test set, where the dashed line represents the maximum likelihood fit, the red curve the pdf of the selected model and the black curve to the normal model. **Second row**: Left– CDF of variances from 3000 posterior samples. The selected $1 - \delta$ quantile is marked. Right: posterior samples in the (β, λ) space, colored by variance: more conservative than the selected pair $(\hat{\beta}^*, \hat{\lambda}^*)$ (green), less conservative (blue). The selected pair $(\hat{\beta}^*, \hat{\lambda}^*)$ is shown as a red cross; the maximum likelihood estimate (black triangle) is unavailable in practice; the parameter for normal model ($(\beta, \lambda) = (2, \sqrt{2})$) is shown as the black star.

The corresponding cross-posterior KS–PIT discrepancy is

$$D_{j|i} = \sup_{u \in [0,1]} |G_{j|i}(u) - u| = \sup_{u \in [0,1]} \left| F_{\theta_i}\left(F_{\theta_j}^{-1}(u)\right) - u \right|, \qquad i \neq j.$$

Since $u = F_{\theta_j}(z)$ is a one-to-one mapping on $\mathbb{R}$, this is equivalently the Kolmogorov distance

between the two CDFs:

$$D_{j|i} = \sup_{z\in\mathbb{R}} |F_{\theta_i}(z) - F_{\theta_j}(z)|.$$

Fix a small level $\delta \in (0,1)$ and compute

$$T_j(1-\delta) = \text{Quantile}_{1-\delta}\big(\{D_{j|i} \mid i \neq j\}\big).$$

Here, $T_j(1-\delta)$ is an upper posterior credible bound on the calibration error of model F_{θ_j}: with posterior probability at least $1-\delta$, the KS–PIT discrepancy between F_{θ_j} and a posterior-plausible generating model does not exceed $T_j(1-\delta)$. This quantity controls the worst-case lack of calibration over the central $(1-\delta)$ fraction of posterior uncertainty.

Select

$$\theta^* = \arg\min_{j=1,\dots,K} T_j(1-\delta).$$

The selected parameter θ^* minimizes this upper bound, yielding the forecast that remains closest to uniform calibration under nearly all posterior-plausible residual distributions. When the generalized normal family is well specified and the data are informative, the posterior concentrates and the selected θ^* yields PIT values close to uniform. Under model misspecification, the rule acts as a robust Bayesian procedure by controlling a high-probability upper bound on the calibration loss.

Remark 5.1 (KS–PIT, sharpness, and posterior constraint.). Section 3.5 shows that μ-probabilistic calibration alone does not preclude uninformative predictions: even for a constant f, many predictive families can yield $U^{f,\mu}_{\hat{F}_n} \sim \mathcal{U}(0,1)$ while differing substantially in dispersion. KS–PIT targets calibration only and ignores sharpness.

Rule 2 optimizes a calibration discrepancy *under the posterior constraint*: the search is restricted to residual models supported by the posterior built from the LOO residuals, so the selected distribution is not arbitrary. This does not, however, control sharpness or task-specific risk. In practice, one may augment or replace the KS–PIT objective by a proper scoring rule (e.g., NLPD, CRPS, SCRPS) or by a calibration functional tailored to the application (e.g., excursion-probability calibration, tail coverage at level $1-\alpha$). Formally, the same quantile-robust selection applies: replace $D_{j|i}$ by a loss $L_{j|i}$ of interest and minimize an upper posterior quantile $\text{Quantile}_{1-\delta}(\{L_{j|i} \mid i \neq j\})$ over j, or combine Rule 2 with Rule 1 to enforce a lower bound on dispersion (equivalently, an upper bound on sharpness) while keeping PIT-based calibration under posterior constraints.

Predictive distribution. Once a parameter $\theta^* = (\beta^*, \lambda^*)$ has been selected, define $G_{\beta,\lambda}$ as the CDF of the $\mathcal{GN}(\beta, 0, \lambda)$ distribution and $g_{\beta,\lambda}$ as its density. For any $x \in \mathbb{X}$, the BCR–GP predictive distribution is then

$$\hat{F}_n^{\text{BCR}}(z \mid x) \;=\; G_{\beta^*,\lambda^*}\left(\frac{z - m_n(x)}{\sigma_n(x)}\right), \qquad z \in \mathbb{R}, \tag{5.1}$$

with corresponding predictive density

$$\hat{f}_n^{\text{BCR}}(z \mid x) \;=\; \frac{1}{\sigma_n(x)}\, g_{\beta^*,\lambda^*}\left(\frac{z - m_n(x)}{\sigma_n(x)}\right). \tag{5.2}$$

The GP posterior mean $m_n(x)$ is preserved as the point predictor, while the uncertainty is recalibrated through the selected parameter θ^*, which governs both tail shape and dispersion.

6. Numerical study.

6.1. Outline. We conduct a numerical comparison between BCR–GP and CPS–GP, together with the full conformal approach of Papadopoulos (2024) and the Jackknife+ for GP (J^+–GP) method of Jaber et al. (2024), on a set of benchmark functions. Section 6.2 details the experimental protocol and evaluation metrics. Section 6.3 studies the influence of the tolerance parameter δ in the variance-based BCR–GP. Section 6.4 investigates the effect of the number of design points on empirical coverage and interval width, with emphasis on parameter selection in CPS–GP. Section 6.5 compares the coverage and width of prediction intervals across methods, and Section 6.6 analyzes their predictive distributions using KS–PIT and SCRPS.

6.2. Experimental setup. Predictions are based on a Gaussian process model ξ with a constant mean function and an anisotropic Matérn covariance kernel. Uncertainty quantification is then constructed using one of four procedures: BCR–GP, CPS–GP, J^+–GP, or FCP.

All methods are implemented using the Gaussian process micro package (`gpmp`) (Vazquez, 2025) for Gaussian-process computations in Python[1]. The implementation of FCP adapts the public code released by Papadopoulos (2024).

In the experiment below, the GP prior has a constant mean m and an anisotropic Matérn kernel covariance. The anisotropic Matérn kernel is defined as

$$k_{\sigma,\nu,\rho}(x,y) = \sigma^2 \kappa_\nu(h), \qquad h^2 = \sum_{i=1}^{d} \frac{(x_{[i]} - y_{[i]})^2}{\rho_i^2}, \quad x, y \in \mathbb{R}^d,$$

where σ^2 is the variance parameter, $\rho = (\rho_1, \ldots, \rho_d)$ are component-wise lengthscales, and κ_ν denotes the Matérn correlation function as defined in (Stein, 1999, Chapter 2.7). The smoothness parameter is set to $\nu = p + 1/2$ with $p \in \mathbb{N}^\star$.

The GP hyperparameters (m, σ^2, ρ) are selected by maximum likelihood on the dataset $\mathcal{D}_n$. In some experiments, they are instead selected on an independent dataset to assess the effect of data reuse on calibration.

For each test function of input dimension d, a design set $\mathcal{D}_n$ of size proportional to d is constructed, and the GP posterior is evaluated on $n_{\text{test}} = 4000$ test inputs sampled uniformly from the domain $\mathbb{X}$. Each experiment is repeated 100 times. The hyperparameters of the posterior distribution of (β, λ) in BCR–GP are fixed to $a = 10$ and $b = 10$. The analytical definitions and properties of the test functions are summarized in Table 9 in Appendix E.

As an initial illustration, Figure 7 presents a one-dimensional comparison of the prediction intervals produced by BCR–GP with variance-based selection, and CPS–GP.

6.3. Influence of the tolerance level for variance-based bcr–gp. We study the effect of the tolerance δ on the variance-based version of BCR–GP, where selection relies on the

[1]The code used to generate all results in this section will be made publicly available in a future version of this article.

variance of $\mathcal{GN}(\beta, 0, \lambda)$, Section 5.2, Rule 1, on $\mathcal{D}_n$. Results are shown in Figure 8.

For small tolerances, BCR–GP produces conservative prediction intervals at all confidence levels. As δ increases, intervals become narrower and undercoverage appears, particularly in the tails (e.g., at 90–95% central levels). The transition from conservative to optimistic behavior depends on the problem, the particular design and the GP specification. In these experiments, values of δ around 0.1 provided a good balance between coverage and interval width. When a single value must be fixed in practice, δ can be tuned by splitting $\mathcal{D}_n$ into calibration and validation subsets, or via a K-fold cross-validation scheme applied to $\mathcal{D}_n$.

6.4. Effect of design set size. This section examines how the number of design points n influences empirical coverage and the average width of prediction intervals. We first analyze this effect in relation to parameter selection strategies in the CPS–GP method, since its theoretical guarantees depend on the independence between parameter selection and prediction.

Parameter selection strategies for CPS–GP. The experiment was conducted on several test functions; results are reported here for the Hartmann6 function, which represents the most challenging case for CPS–GP among those considered. This benchmark is known to require a large number of design points for accurate parameter selection and approximation.

For each dataset $\mathcal{D}_n$, kernel parameters are either (i) selected using a subset of γn observations and predictions computed on the remaining $(1-\gamma)n$ points (split strategy), or (ii) selected and used on the full dataset $\mathcal{D}_n$. As a reference, an oracle configuration is also considered, where parameters are selected on an independent dataset of $10d$ points. Figure 9 reports empirical coverages obtained on an independent test grid.

When parameters are selected on an independent dataset, CPS–GP achieves empirical coverage whose average over repetitions matches the intended confidence levels, consistent with its theoretical guarantees. In contrast, selecting parameters on the same data used for prediction leads to poorer coverage, especially for small n. This effect diminishes as n increases: beyond roughly $20d$, results with or without data splitting become nearly identical. The improvement reflects the decreasing sensitivity of the selected hyperparameters to sample perturbations as the design size grows, which reduces the impact of the exchangeability violation.

It is worth noting that the Hartmann6 function represents a difficult case for CPS–GP, as parameter selection tends to be unstable for small n. For most other test functions considered in this study, CPS–GP is less sensitive to the design size and achieves calibration more easily.

Because data splitting reduces the number of points available for both parameter selection and prediction, some loss in performance is expected. Table 2 reports average RMSE and KS–PIT values for different split ratios γ. As expected, RMSE increases when fewer points are used for prediction, with the largest degradation at $\gamma = 0.5$. Calibration, measured by KS–PIT, deteriorates most for $\gamma = 0.8$. Using the same dataset $\mathcal{D}_n$ for both steps yields RMSE values close to the oracle configuration and calibration comparable to the best split strategy, though the average coverage remains biased. In practice, since any deterioration in RMSE relative to the standard GP model is undesirable, we adopt this configuration in the remainder of the paper to ensure a fair comparison across methods.

Remark 6.1. Liang and Barber (2025) introduced a notion of algorithmic stability in conformal prediction, showing that if a procedure is stable, its intervals can be slightly widened to

Method	$5d$		$10d$		$20d$	
	KS–PIT	RMSE	KS–PIT	RMSE	KS–PIT	RMSE
oracle CPS-GP	0.14	0.35	0.09	0.30	0.07	0.27
CPS-GP	0.17	0.39	0.11	0.30	0.08	0.22
CPS-GP $\gamma = 0.2$	0.18	0.47	0.13	0.40	0.10	0.31
CPS-GP $\gamma = 0.5$	0.16	1.11	0.11	1.03	0.08	0.48
CPS-GP $\gamma = 0.8$	0.22	0.42	0.19	0.38	0.14	0.33

Table 2: Average KS–PIT and RMSE over 100 repetitions for the Hartmann6 function at different design set sizes n. The GP regularity is fixed at $p = 2$. Predictive distributions are obtained with the CPS–GP method using either (i) a data split (γn points for parameter selection and $(1-\gamma)n$ for prediction), (ii) the same dataset $\mathcal{D}_n$ for both steps, or (iii) oracle parameters pre-selected on an independent dataset of $10d$ points. RMSE increases under data splitting, with the largest degradation observed for $\gamma = 0.5$, while calibration is poorest for $\gamma = 0.8$. The full-data configuration performs close to the oracle case but retains biased coverage.

recover exact coverage. Although their analysis focuses on predictive intervals, the same idea may apply to CPS–GP: if the selected hyperparameters vary little under sample perturbations, calibration is approximately preserved. A formal analysis of this stability for CPS–GP and its impact on finite-sample validity remains open.

Parameter selection on the prediction dataset. We next compare CPS–GP and BCR–GP when both use parameters selected on the same dataset $\mathcal{D}_n$ as that used for prediction. This configuration, already discussed above, breaks exchangeability and therefore invalidates the marginal coverage guarantees of Section 4.3. Figure 10 reports the resulting empirical coverages for the Hartmann6 and Goldstein–Price functions. For Hartmann6, BCR–GP with $\delta = 0.01$ attains coverage close to the intended confidence levels, whereas CPS–GP and the GP posterior remain noticeably miscalibrated. For Goldstein–Price, CPS–GP provides the most accurate coverage among the compared methods.

As shown in Proposition 4.13, the average interval width of CPS–GP becomes unbounded for $n = 5d$ at the 95% level, illustrating a finite-sample limitation of the procedure. In contrast, BCR–GP with $\delta = 0.01$ yields finite, slightly conservative intervals.

6.5. Comparison of coverage and width of prediction intervals. We evaluate empirical coverage and average interval width for the GP, FCP, J$^+$–GP, BCR–GP, and CPS–GP methods. Each result is averaged over 100 repetitions. Parameters are selected on the same dataset $\mathcal{D}_n$ used for prediction, so the marginal validity of conformal procedures no longer holds.

Figure 11 displays empirical coverages at 70%, 90%, and 95% for the Goldstein–Price function and for Matérn regularity orders $p = 0, 1, 2$. Tables 4 and 6 report, respectively, mean coverage and relative interval width across all test functions.

All conformal and Bayesian–conformal methods improve empirical coverage compared with the GP posterior. As expected, FCP and J$^+$–GP have similar results. Both CPS–GP and BCR–GP with $\delta = 0.1$ achieve coverage closest to the target levels, with BCR–GP slightly

conservative. When using the KS–PIT selection rule with $\delta = 0.1$, BCR–GP tends to become optimistic at high confidence (90–95%). None of the conformal variants reach exact coverage once parameters are reused.

Across test functions and for all values of p, BCR–GP remains near the intended coverage levels, showing stable behavior across problems.

Regarding interval width, BCR–GP with $\delta = 0.1$ produces the narrowest intervals among well-calibrated methods, while $\delta = 0.01$ yields wider, more conservative intervals. J$^+$–GP, FCP, and CPS–GP produce shorter intervals on average, consistent with their more optimistic coverage shown in Table 4.

Method	Goldstein –Price		Ackley4		Hartman3		Dixon –Price4		Rosenbrock6		Branin	
	90%	95%	90%	95%	90%	95%	90%	95%	90%	95%	90%	95%
bcr-gp (ks-pit)	0.86	0.90	0.89	0.93	0.89	0.94	0.87	0.92	0.87	0.92	0.91	0.95
bcr-gp (0.01)	0.92	0.95	0.94	0.96	0.94	0.97	0.93	0.96	0.91	0.96	0.96	0.99
bcr-gp (0.1)	0.89	0.93	0.92	0.95	0.92	0.96	0.90	0.94	0.89	0.94	0.94	0.97
fcp	0.87	0.93	0.87	0.93	0.89	0.94	0.87	0.93	0.86	0.93	0.90	0.96
gp	0.85	0.88	0.90	0.93	0.93	0.95	0.87	0.92	0.85	0.91	1.00	1.00
cps-gp	0.87	0.93	0.87	0.93	0.89	0.94	0.87	0.93	0.86	0.93	0.90	0.95
j+gp	0.87	0.93	0.88	0.93	0.89	0.94	0.88	0.93	0.87	0.93	0.90	0.95

Table 4: Average empirical coverage over 100 repetitions for several confidence levels for multiple test functions. The empirical coverage is computed on a test grid with 4000 points and $n = 20 \times d$ observations. For every method, the regularity of the covariance is $p = 2$. BCR–GP with the KS–PIT rule is used with $\delta = 0.1$.

Method	Goldstein –Price		Ackley4		Hartman3		Dixon –Price4		Rosenbrock6		Branin	
	90%	95%	90%	95%	90%	95%	90%	95%	90%	95%	90%	95%
bcr-gp (ks-pit)	3	3.41	1.03	1.10	1.21	1.32	1.60	1.60	1.27	1.28	1.55	1.64
bcr-gp (0.01)	4.43	5.36	1.31	1.44	1.64	1.90	1.91	1.92	1.50	1.52	2.17	2.40
bcr-gp (0.1)	3.55	4.06	1.16	1.24	1.40	1.56	1.73	1.72	1.38	1.40	1.81	1.91
fcp	1.22	1.59	0.89	0.99	0.84	0.99	1.03	1.06	1.05	1.08	0.50	0.60
cps-gp	3.37	4.41	0.99	1.12	1.27	1.59	1.61	1.67	1.29	1.34	1.66	2.09
j+gp	1.23	1.60	0.91	1.00	0.87	0.99	1.04	1.06	1.05	1.07	0.49	0.58

Table 6: Relative width of PIs compared with the width of PIs predicted by GP over 100 repetitions for several confidence levels for multiple test functions. The widths are computed on a test grid with 4000 points and $n = 20 \times d$ observations. For every method, the regularity of the covariance is $p = 2$. In bold are reported the width of the predictions intervals with the empirical coverage closest to the targeted confidence level. BCR–GP with the KS–PIT rule is used with $\delta = 0.1$.

6.6. Comparison of predictive distributions. We evaluate the predictive distributions of the posterior GP, BCR–GP, and CPS–GP using KS–PIT (uniformity of PIT) and SCRPS (proper scoring). Results averaged over 100 repetitions are reported in Table 8; the SCRPS definition and expressions for the generalized normal distribution are given in Appendix D. FCP and J$^+$–GP are excluded from this subsection because they do not define full predictive distributions (intervals only).

Predictive laws are obtained as follows: GP uses its Gaussian posterior; BCR–GP fits a generalized normal to normalized residuals; CPS–GP yields a stepwise predictive law (PIT computed with boundary randomization). All methods follow the same hyperparameter-selection protocol as in Section 6.2.

BCR–GP with the KS–PIT selection rule attains the lowest KS–PIT on most benchmarks, indicating better distributional calibration. The variance-based BCR–GP (e.g., $\delta \in \{0.1, 0.01\}$) achieves comparable or better SCRPS and slightly heavier tails. CPS–GP is competitive on both metrics but typically shows marginally larger KS–PIT when parameters are selected on the same dataset $\mathcal{D}_n$. Consistent with Section 6.5, the KS–PIT rule can be optimistic at high confidence, whereas the variance-based rule is more stable across problems.

Method	Goldstein –Price		Ackley4		Hartman3		Dixon –Price4		Rosenbrock6		Branin	
	KS–PIT	SCRPS	KS–PIT	SCRPS	KS–PIT	SCRPS	KS–PIT	SCRPS	KS–PIT	SCRPS	KS–PIT	SCRPS
bcr-gp (ks-pit)	0.11	5.78	0.09	0.86	0.09	-0.15	0.08	5.40	0.06	6.81	0.11	0.30
bcr-gp (0.01)	0.12	5.76	0.10	0.86	0.11	-0.12	0.09	5.40	0.07	6.81	0.13	0.34
bcr-gp (0.1)	0.12	5.75	0.09	0.85	0.10	-0.14	0.08	5.40	0.06	6.81	0.13	0.31
gp	0.16	5.78	0.10	0.87	0.16	-0.12	0.07	5.40	0.06	6.82	0.24	0.49
cps-gp	0.12	5.79	0.09	0.86	0.10	-0.14	0.09	5.41	0.08	6.82	0.14	0.31

Table 8: Average KS–PIT and SCRPS over 100 repetitions, computed on a test set of 4000 points and with $n = 20 \times d$ observations for GP, BCR–GP with $\delta = 0.1$ and $\delta = 0.01$, and CPS–GP. The regularity of the Matérn covariance function is $p = 2$. BCR–GP with the KS–PIT rule is used with $\delta = 0.1$.

7. Discussion. This work targets μ-calibration of Gaussian-process (GP) predictive distributions. We first presented CPS–GP, an adaptation of conformal predictive systems to GP interpolation that yields design-marginal probabilistic calibration (on average over the data). We then introduced BCR–GP, a Bayesian post-processing method that preserves the GP posterior mean and calibrates dispersion by fitting a parametric residual model, producing smooth predictive CDFs usable in standard sequential-design algorithms.

In empirical benchmarks, BCR–GP, CPS–GP, and J$^+$–GP improve empirical coverage relative to the Gaussian posterior, and BCR–GP and CPS–GP attain comparable probabilistic-calibration quality as measured by KS–PIT and SCRPS. Unlike FCP or J$^+$–GP, which only provide prediction intervals, BCR–GP and CPS–GP produce full predictive distributions. The additional computational cost relative to standard GP prediction remains modest in the

regimes considered.

Applicability and limitations. A practical limitation of CPS–GP is that it does not yield a simple parametric predictive distribution, which complicates its use in sampling criteria (aka acquisition functions) that rely on closed-form Gaussian formulas (e.g. expected-improvement Bayesian optimization). Another limitation, common to all conformal approaches, is that CPS–GP and J$^+$–GP guarantee design-marginal rather than training-conditional coverage. The formal marginal validity of CPS–GP further requires kernel hyperparameters to be chosen independently of the calibration step; when they are estimated on the same data, coverage remains approximately correct but is no longer exact, even marginally over the data.

By contrast, BCR–GP has no distribution-free guarantee: its calibration is purely model-based and relies on (i) the adequacy of the generalized-normal residual family and (ii) the working assumption that LOO standardized residuals provide a good proxy for the μ-marginal distribution of $R_n(X, f(X))$. In this work, these assumptions are only checked empirically; quantifying their deviation from the true residual law, and the resulting impact on coverage and PIT-based diagnostics, would be a natural and important direction for future study.

Noisy observations. Although both methods were developed for interpolation, many practical applications involve noisy observations $Z = f(X)+\varepsilon$, where the goal is to predict the latent function f. In such settings, distribution-free coverage guarantees for f are generally unattainable without modelling the noise. Once a noise model is specified, both CPS–GP and BCR–GP could be extended to yield model-conditional calibration, by applying the same constructions to the latent GP and treating the noise within the likelihood.

Sequential design. Applying μ-calibrated predictors within sequential algorithms raises the question of defining the design measure μ when design points depend on past data. In practice, for sampling criteria such as expected improvement, the adaptive design explores the domain, and calibration can be assessed empirically with respect to the empirical distribution of the visited points. Periodic recalibration, or calibration diagnostics computed with importance weights that map the empirical design distribution to a target μ, can be used to monitor and maintain μ-calibration as the design evolves.

Tail calibration. The BCR–GP framework can also be tailored toward upper-tail calibration by modifying the posterior selection rule or adopting a heavier-tailed residual model. Such adjustments make it possible to target exceedance probabilities or risk metrics while preserving smooth predictive CDFs and the GP mean structure.

Both proposed approaches enhance the design-marginal calibration of GP-based predictive distributions, a property often expected by practitioners but not guaranteed by the standard GP posterior. CPS–GP offers a distribution-free calibration mechanism suited to fixed designs, while BCR–GP provides smooth, parametric predictive distributions directly usable in sequential design and optimization. Together, they improve the practical reliability of GP modelling by producing calibrated uncertainty estimates that remain interpretable and computationally tractable.

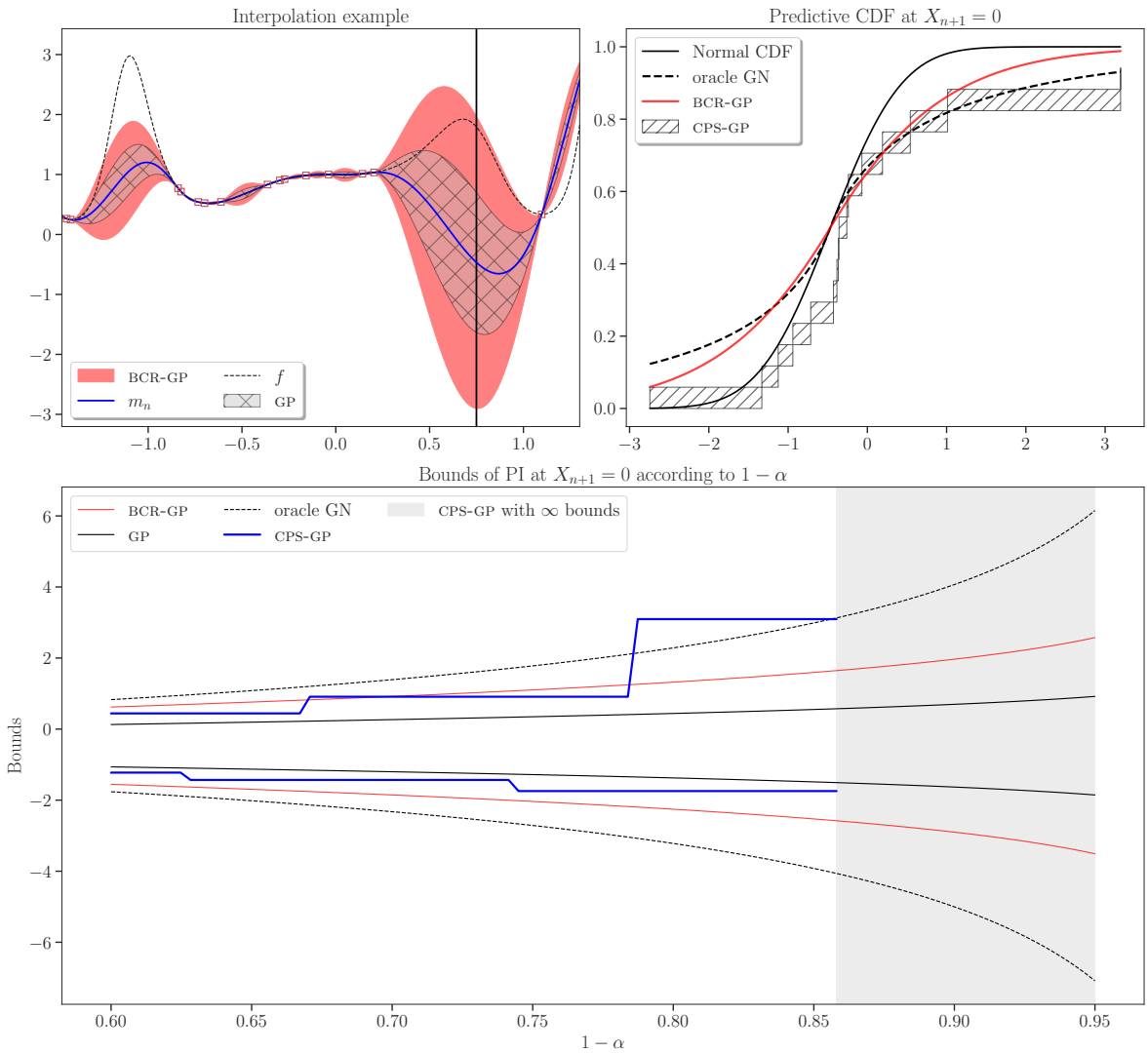
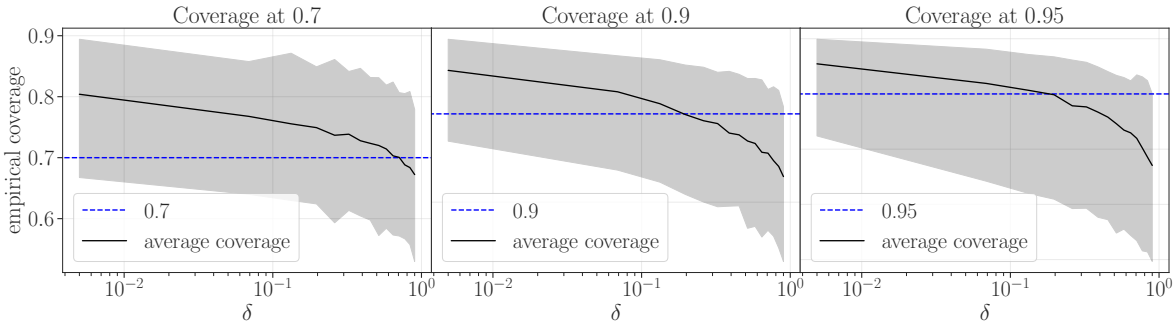
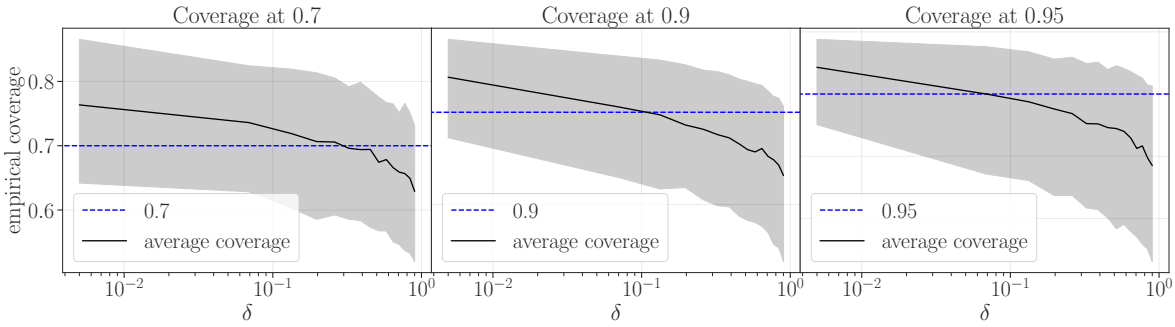


Figure 7: **Top left:** Prediction intervals constructed from the GP posterior distribution and from BCR–GP at confidence level $1 - \alpha = 0.9$. BCR–GP uses the variance of the generalized normal distribution for selection, with $\delta = 0.1$. **Top right:** Predicted CDFs at $x = 0.75$, comparing the GP posterior, the CDF from BCR–GP (red), the stepwise CDF from CPS–GP (black hatches), and an oracle CDF (black) obtained from a generalized normal model fitted on a test grid of $n_{\text{test}} = 2000$ points. This dataset exhibits strong miscalibration of the GP posterior predictive distributions. **Bottom:** Interval bounds as a function of $1 - \alpha$. The GP posterior underestimates uncertainty across confidence levels, while BCR–GP and CPS–GP produce larger intervals. CPS–GP yields unbounded interval widths for $1 - \alpha \gtrsim 0.85$.



(a) Empirical coverage for $\delta \in [0, 1]$ for the Goldstein–Price function.



(b) Empirical coverage for $\delta \in [0, 1]$ for the Dixon–Price function.

Figure 8: Empirical coverage of 70%, 90%, and 95% prediction intervals obtained with the variance-based BCR–GP method for $\delta \in [0, 1]$. The regularity parameter is fixed to $p = 2$. The top panel corresponds to the Goldstein–Price function ($n = 40$), and the bottom panel to the Dixon–Price function ($n = 60$). Evaluation is performed on 4000 test points. The black curve indicates the mean coverage across 100 repetitions, and the shaded area denotes the 0.05 and 0.95 quantiles.

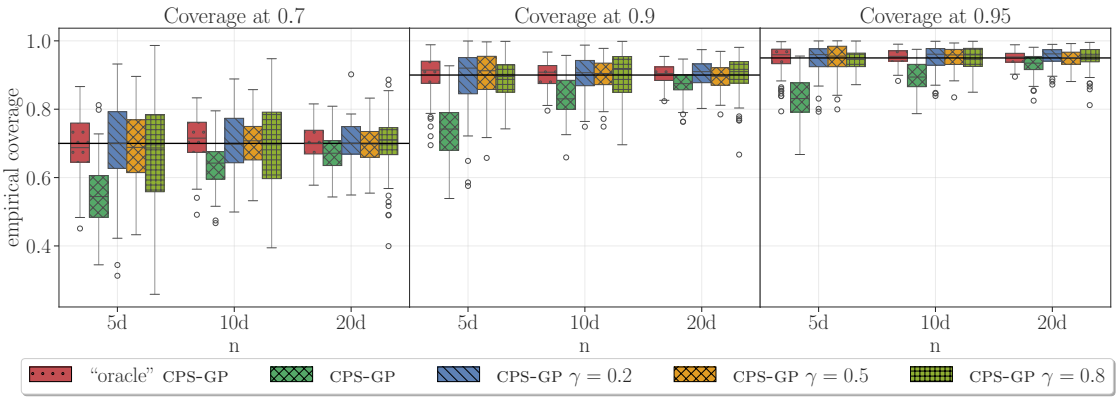


Figure 9: Empirical coverage of 70%, 90%, and 95% prediction intervals for the Hartmann6 function as a function of the design set size n. Horizontal lines indicate nominal coverage. Evaluation is based on 4000 test points and 100 repetitions. The GP regularity is fixed at $p = 2$. CPS–GP is applied using either a split of γn points for parameter selection and $(1-\gamma)n$ for prediction, or the same dataset $\mathcal{D}_n$ for both steps. The oracle CPS–GP variant uses parameters pre-selected on an independent dataset of $10d$ points.

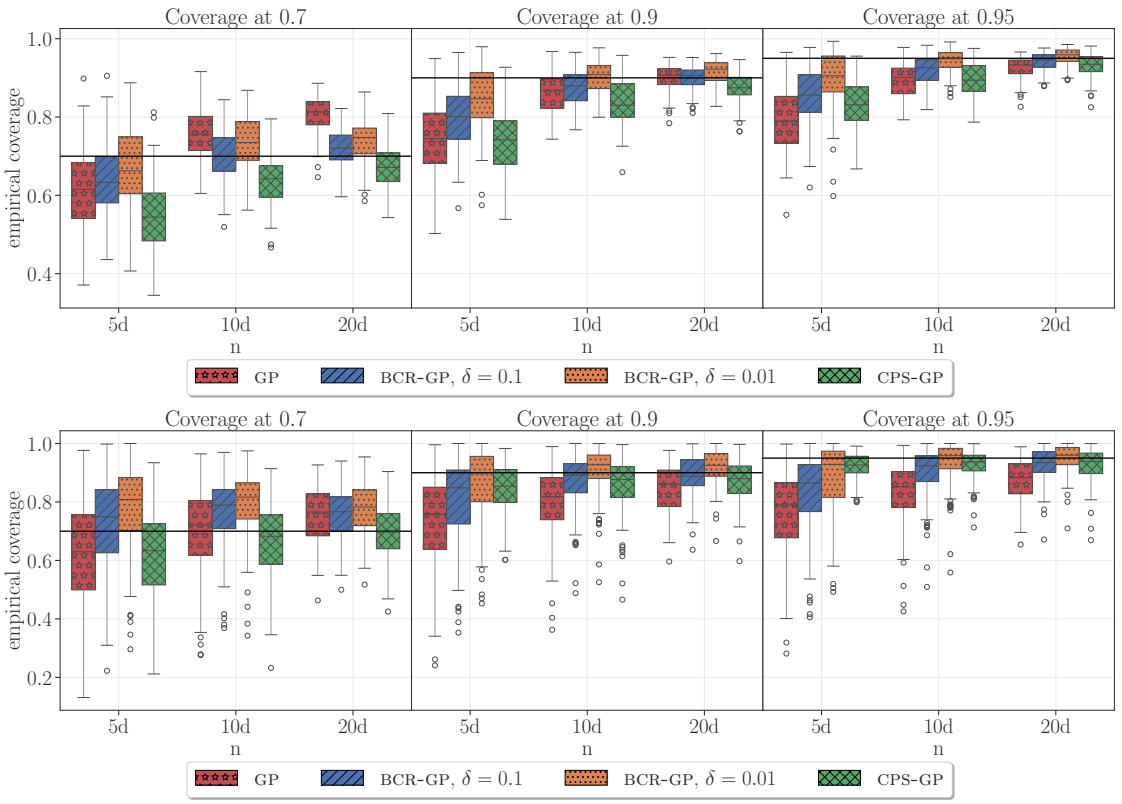


Figure 10: Empirical coverage of 70%, 90%, and 95% prediction intervals for the Hartmann6 (first row) and Goldstein–Price (second row) functions, for several design set sizes n. The GP regularity is fixed at $p = 2$. Prediction intervals are constructed with the variance-based BCR–GP method for $\delta = 0.1$ and $\delta = 0.01$, with parameters selected on the same dataset $\mathcal{D}_n$.

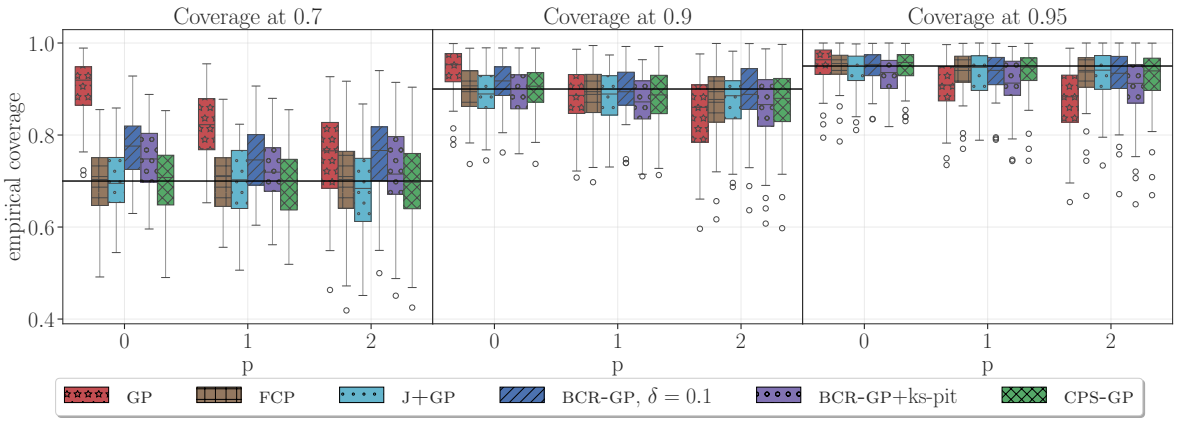


Figure 11: Empirical coverage of 70%, 90%, and 95% prediction intervals for the Goldstein–Price function ($n = 20d = 40$). Results are shown for Matérn regularity orders $p = 0, 1, 2$. Prediction intervals are computed with BCR–GP using either the variance-based rule ($\delta = 0.1$) or the KS–PIT rule ($\delta = 0.1$).

Appendix A. Proofs.

A.1. Generalized inverse identities.

Lemma A.1 (Generalized inverse identities). *Let $G : \mathbb{R} \to [0,1]$ be nondecreasing and right-continuous, and define its generalized inverse by*

$$G^{-1}(u) := \inf\{z \in \mathbb{R} : \ G(z) \geq u\}, \qquad u \in (0,1).$$

Then, for all $u \in (0,1)$ and $z \in \mathbb{R}$,

$$\{\, G(z) < u \,\} \iff \{\, z < G^{-1}(u) \,\}, \tag{A.1}$$

$$\{\, G(z^-) \leq u \,\} \iff \{\, z \leq G^{-1}(u) \,\}. \tag{A.2}$$

Moreover, for any $0 < p < q < 1$ and $v \in \mathbb{R}$,

$$\{\, G^{-1}(p) \leq v < G^{-1}(q) \,\} \iff \{\, p \leq G(v) < q \,\}. \tag{A.3}$$

Proof. The equivalences (A.1)–(A.2) follow directly from the definition of G^{-1} and the fact that $G(z^-) = \sup_{w<z} G(w)$. For (A.3), apply (A.1) with $u = q$ and its contrapositive with $u = p$ to relate $G^{-1}(p), G^{-1}(q)$ and $G(v)$. ∎

A.2. Proof of Proposition 3.1.

Fix $x \in \mathbb{X}$ and condition throughout on $\mathcal{D}_n$. Let

$$F(\cdot) := \hat{F}_n(\cdot \mid x), \qquad \Delta F(z) := F(z) - F(z^-),$$

and draw $V \sim F$ independently of $\tau \sim \mathcal{U}(0,1)$. Define

$$U := F(V^-) + \tau\, \Delta F(V).$$

Uniformity. Fix $t \in [0,1]$ and set the (left-continuous) quantile

$$z_t := F^{-1}(t) = \inf\{z \in \mathbb{R} : \ F(z) \geq t\}.$$

We have

$$\{U \leq t\} = \{V < z_t\} \ \cup \ \Big(\{V = z_t\} \cap \{\tau \leq \tfrac{t - F(z_t^-)}{\Delta F(z_t)}\}\Big),$$

with the convention that the second event is empty when $\Delta F(z_t) = 0$. Indeed:

(i) if $V < z_t$, then $F(V) \leq F(z_t^-) \leq t$, hence $U \leq F(V) \leq t$;
(ii) if $V > z_t$, then $F(V^-) \geq F(z_t) \geq t$, hence $U \geq F(V^-) > t$;
(iii) if $V = z_t$, then U is uniform on $[F(z_t^-), F(z_t)]$ given V.

Taking probabilities under $\mathsf{P}_n(\,\cdot\,)$ and using $V \sim F$ and independence of τ,

$$\begin{aligned}
\mathsf{P}_n(U \leq t) &= \mathsf{P}_n(V < z_t) \ + \ \mathsf{P}_n(V = z_t)\, \mathsf{P}_n\left(\tau \leq \frac{t - F(z_t^-)}{\Delta F(z_t)} \,\middle|\, V = z_t\right) \\
&= F(z_t^-) \ + \ \Delta F(z_t)\, \frac{t - F(z_t^-)}{\Delta F(z_t)} \\
&= t.
\end{aligned}$$

If $\Delta F(z_t) = 0$, then $\mathsf{P}_n(V = z_t) = 0$ and $F(z_t^-) = t$, yielding the same result. Thus $U \mid \mathcal{D}_n \sim$

$\mathcal{U}(0,1)$.

Exact mass. Set $a = \alpha/2$ and $b = 1 - \alpha/2$. Writing

$$\mathcal{C}_{n,\tau,1-\alpha}(x) := \big[\hat{F}^{-1}_{n,\tau}(a \mid x),\, \hat{F}^{-1}_{n,\tau}(b \mid x)\big),$$

we obtain by Lemma A.1 the event identity

$$\{\, V \in \mathcal{C}_{n,\tau,1-\alpha}(x) \,\} \iff \{\, a \le \hat{F}_{n,\tau}(V \mid x) < b \,\} \iff \{\, a \le U < b \,\}.$$

Taking probabilities under P_n and using $U \mid \mathcal{D}_n \sim \mathcal{U}(0,1)$ yields

$$\mathsf{P}_n\big\{\, V \in \mathcal{C}_{n,\tau,1-\alpha}(x) \,\big\} = \mathsf{P}_n\{\, a \le U < b \,\} = b - a = 1 - \alpha.$$

A.3. Proof of Remark 3.3. We want to justify the inequalities

$$\mu\big(\{x :\ f(x) < \hat{F}_n^{-1}(u \mid x)\}\big) \ \le\ G_\mu(u) \ \le\ \mu\big(\{x :\ f(x) \le \hat{F}_n^{-1}(u \mid x)\}\big), \qquad u \in (0,1),$$

where

$$G_\mu(u) := \mathsf{P}_n\big(U^{f,\mu,\tau}_{\hat{F}_n} \le u\big), \qquad U^{f,\mu,\tau}_{\hat{F}_n} := \hat{F}_n(f(X)^- \mid X) + \tau\big(\hat{F}_n(f(X) \mid X) - \hat{F}_n(f(X)^- \mid X)\big),$$

with $X \sim \mu$ and $\tau \sim \mathcal{U}(0,1)$ independent of $(\hat{F}_n,\, X,\, \mathcal{D}_n)$.

For each x, set $a_x := \hat{F}_n(f(x)^- \mid x)$ and $b_x := \hat{F}_n(f(x) \mid x)$. We have

$$\mathsf{P}_n\big(U^{f,\mu,\tau}_{\hat{F}_n} \le u \mid X = x\big) = \begin{cases} 0, & u \le a_x, \\ \dfrac{u - a_x}{b_x - a_x}, & a_x < u < b_x, \\ 1, & u \ge b_x. \end{cases}$$

Thus,

$$\mathbb{1}\{\, b_x \le u \,\} \ \le\ \mathsf{P}_n\big(U^{f,\mu,\tau}_{\hat{F}_n} \le u \mid X = x\big) \ \le\ \mathbb{1}\{\, a_x \le u \,\}.$$

Integrating over μ yields

$$\mu\big(\{x :\ b_x \le u\}\big) \ \le\ G_\mu(u) \ \le\ \mu\big(\{x :\ a_x \le u\}\big).$$

From Lemma A.1:

$$\{\, G(z) < u \,\} \Longleftrightarrow \{\, z < G^{-1}(u)\}, \qquad \{\, G(z^-) \le u \,\} \Longleftrightarrow \{\, z \le G^{-1}(u)\}\,.$$

Substituting $G(z) = \hat{F}_n(z \mid x)$ and $z = f(x)$, we obtain

$$\mu\big(\{x :\ f(x) < \hat{F}_n^{-1}(u \mid x)\}\big) = \mu\big(\{x :\ b_x < u\}\big) \ \le\ G_\mu(u)$$

and

$$G_\mu(u) \le\ \mu\big(\{x :\ a_x \le u\}\big) = \mu\big(\{x :\ f(x) \le \hat{F}_n^{-1}(u \mid x)\}\big),$$

as claimed.

A.4. Proof of Remark 3.4. Let $F_{n,\,x}(z) := \hat{F}_n(z \mid x)$ and $q_x(u) := \hat{F}_n^{-1}(u \mid x)$. Fix $a = \alpha/2$ and $b = 1 - \alpha/2$, and set $U := \hat{F}_n(f(X) \mid X)$ with $X \sim \mu$.

Lower bound for $\delta_\alpha(\hat{F}_n; \mu)$. By Lemma A.1, we have $\{U < b\} \iff \{f(X) < q_X(b)\}$ and $\{a < U\} \Rightarrow \{f(X) \ge q_X(a)\}$ (since $f(X) < q_X(a)$ would imply $U \le a$). Hence

$$\{\, a < U < b \,\} \ \subseteq\ \{\, q_X(a) \le f(X) < q_X(b) \,\} \ \subseteq\ \{\, q_X(a) \le f(X) \le q_X(b) \,\}.$$

Taking probabilities gives

$$\mathsf{P}_n\{\, a < U < b \,\} \;\le\; \delta_\alpha(\hat{F}_n; \mu).$$

Failure of the naive upper bound. In general, the inclusion

$$\{\, q_X(a) \le f(X) \le q_X(b) \,\} \subseteq \{\, a \le U \le b \,\}$$

may fail. At the upper endpoint, if $f(X) = q_X(b)$ and $F_{n,x}$ has a jump at $q_X(b)$, then $U > b$ is possible; similarly, at the lower endpoint $f(X) = q_X(a)$ may give $U > a$. Thus the upper inequality

$$\delta_\alpha(\hat{F}_n; \mu) \le \mathsf{P}_n\{a \le U \le b\}$$

is not always valid without regularity assumptions.

Safe two-sided bounds. We have

$$\mu\big(\{x :\; f(x) < q_x(b)\}\big) = \mathsf{P}_n\{U < b\}, \qquad \mu\big(\{x :\; f(x) < q_x(a)\}\big) = \mathsf{P}_n\{U < a\}.$$

(Since $\{\, F_{n,x}(z) < u \,\} \iff \{\, z < q_x(u) \,\}$, by Lemma A.1.) Hence

$$\begin{aligned}
\delta_\alpha(\hat{F}_n; \mu) &= \mu\big(\{x :\; f(x) \le q_x(b)\}\big) - \mu\big(\{x :\; f(x) < q_x(a)\}\big) \\
&= \mathsf{P}_n\{U < b\} - \mathsf{P}_n\{U < a\} + \mu\big(\{x :\; f(x) = q_x(b)\}\big) \\
&= \mathsf{P}_n\{a \le U < b\} + \mu\big(\{x :\; f(x) = q_x(b)\}\big) \\
&= \mathsf{P}_n\{a \le U \le b\} - \mathsf{P}_n\{U = b\} + \mu\big(\{x :\; f(x) = q_x(b)\}\big).
\end{aligned}$$

Notice that $\mu\big(\{x :\; f(x) = q_x(b)\}\big) = \mu(A_=) + \mu(A_>)$, with

$$A_= = \{x :\; f(x) = q_x(b),\; F_{n,x}(q_x(b)) = b\} \quad \text{and} \quad A_> = \{x :\; f(x) = q_x(b),\; F_{n,x}(q_x(b)) > b\}.$$

On $A_=$ ($F_{n,x}$ has no jumps), we have

$$U = \hat{F}_n(f(X) \mid X) = F_{n,X}(q_X(b)) = b,$$

so that

$$A_= \subseteq \{U = b\},$$

and

$$\mathsf{P}_n\{U = b\} \;\ge\; \mu(A_=).$$

Plugging in yields

$$\delta_\alpha(\hat{F}_n; \mu) \le \mathsf{P}_n\{a \le U \le b\} \;-\; \mu(A_=) \;+\; \mu(A_=) + \mu(A_>) = \mathsf{P}_n\{a \le U \le b\} \;+\; \mu(A_>).$$

Equivalently,

$$\delta_\alpha(\hat{F}_n; \mu) \;\le\; \mathsf{P}_n\{a \le U \le b\} + \mu\big(\{x :\; f(x) = q_x(b),\; F_{n,x}(q_x(b)) > b\}\big).$$

Therefore, we have the following two-sided bounds:

$$\mathsf{P}_n\{a \le U < b\} \;\le\; \delta_\alpha(\hat{F}_n; \mu) \;\le\; \mathsf{P}_n\{a \le U \le b\} \;+\; \mu\big(\{x :\; f(x) = q_x(b),\; F_{n,x}(q_x(b)) > b\}\big).$$

Restoring equality. If, for μ-a.e. x, $F_{n,x}$ is continuous and strictly increasing, then

$$\{\, q_X(a) \le f(X) \le q_X(b) \,\} \iff \{\, a \le U \le b \,\},$$

hence

$$\delta_\alpha(\hat{F}_n;\mu) = \mathsf{P}_n\{\, a \le U \le b \,\}.$$

In the discontinuous case, by Lemma A.1,

$$\{\, F^{-1}_{n,\,x,\,\tau}(a) \le f(x) < F^{-1}_{n,\,x,\tau}(b) \,\} \iff \{\, a \le F_{n,\,x,\,\tau}(f(x)) < b \,\}.$$

With $X \sim \mu$ and

$$U^{f,\,\mu,\,\tau}_{\hat{F}_n} := F_{n,\,X,\,\tau}(f(X)),$$

this gives

$$\mu\ \Big(\{x:\ f(x) \in [\,\hat{F}^{-1}_{n,\,\tau}(a \mid x),\ \hat{F}^{-1}_{n,\,\tau}(b \mid x)\,)\}\Big) = \mathsf{P}_n\left(\, a \le U^{f,\,\mu,\,\tau}_{\hat{F}_n} < b \;\middle|\; \tau \right).$$

Thus, boundary randomization restores exact equivalence in the general case (see also Proposition 3.1).

A.5. Proof of Proposition 3.5. Work throughout conditional on $\mathcal{D}_n$.

Since $(X^\star_j, \tau_j)$ are i.i.d., the U_j are i.i.d. on $[0,1]$ with distribution function

$$G_\mu(u) = \mathsf{P}\big(U \le u \mid \mathcal{D}_n\big), \quad U = \hat{F}_n\big(f(X)^- \mid X\big) + \tau\Big(\hat{F}_n\big(f(X) \mid X\big) - \hat{F}_n\big(f(X)^- \mid X\big)\Big),$$

where $X \sim \mu$ and $\tau \sim \mathcal{U}(0,1)$ are independent. Let

$$\hat{G}_m(u) = \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}\{U_j \le u\}.$$

By the Glivenko–Cantelli theorem,

$$\sup_{u\in[0,1]} \big|\hat{G}_m(u) - G_\mu(u)\big| \xrightarrow[m\to\infty]{\text{a.s.}} 0.$$

Now write

$$J_{\text{KS-PIT},\,m}(\hat{F}_n) = \sup_{u\in[0,1]} \big|\hat{G}_m(u) - u\big|, \qquad J_{\text{KS-PIT},\,\mu}(\hat{F}_n) = \sup_{u\in[0,1]} \big|G_\mu(u) - u\big|.$$

Then, for $A_m(u) = \hat{G}_m(u) - u$ and $A(u) = G_\mu(u) - u$,

$$\begin{aligned}
\big|J_{\text{KS}-\text{PIT},\,m}(\hat{F}_n) - J_{\text{KS}-\text{PIT},\,\mu}(\hat{F}_n)\big| &= \big|\sup_u |A_m(u)| - \sup_u |A(u)|\big| \\
&\le \sup_u |A_m(u) - A(u)| \\
&= \sup_u \big|\hat{G}_m(u) - G_\mu(u)\big| \xrightarrow{\text{a.s.}} 0.
\end{aligned}$$

Hence the claim.

A.6. Proof of Proposition 3.7. Let $(X^\star_i)_{i=1}^m \overset{\text{i.i.d.}}{\sim} \mu$ be an independent test design and set $Z_i := f(X^\star_i)$. For each i, write the predictive CDF at $X^\star_i$ as $\hat{F}_n(\cdot \mid X^\star_i)$ and define the PIT value

$$U_i = \begin{cases} \hat{F}_n(Z_i \mid X^\star_i), & \text{if } \hat{F}_n(\cdot \mid X^\star_i) \text{ is continuous,} \\ \hat{F}_n(Z_i^- \mid X^\star_i) + \tau_i\big(\hat{F}_n(Z_i \mid X^\star_i) - \hat{F}_n(Z_i^- \mid X^\star_i)\big), & \text{otherwise,} \end{cases}$$

with $\tau_i \overset{\text{i.i.d.}}{\sim} \mathcal{U}(0,1)$ independent of everything. Let $\hat{G}_m(u) := \frac{1}{m}\sum_{i=1}^m \mathbb{1}\{U_i \le u\}$ be the empirical CDF of the U_is.

Continuous case. If each $\hat{F}_n(\cdot \mid X_i^\star)$ is continuous and strictly increasing, the central $(1-\alpha)$ interval is

$$\mathcal{C}_{i,1-\alpha} = \big[\hat{F}_n^{-1}(\alpha/2 \mid X_i^\star),\ \hat{F}_n^{-1}(1-\alpha/2 \mid X_i^\star)\big],$$

and

$$\{Z_i \in \mathcal{C}_{i,1-\alpha}\} = \{\alpha/2 \le \hat{F}_n(Z_i \mid X_i^\star) \le 1-\alpha/2\} = \{\alpha/2 \le U_i \le 1-\alpha/2\}.$$

Hence

$$\hat{\delta}_{\alpha,m}(\hat{F}_n) = \frac{1}{m}\sum_{i=1}^m \mathbb{1}\{Z_i \in \mathcal{C}_{i,1-\alpha}\} = \hat{G}_m(1-\alpha/2) - \hat{G}_m(\alpha/2).$$

Since $(1-\alpha) = (1-\alpha/2) - (\alpha/2)$,

$$\big|\hat{\delta}_{\alpha,m}(\hat{F}_n) - (1-\alpha)\big| \le \big|\hat{G}_m(1-\alpha/2) - (1-\alpha/2)\big| + \big|\hat{G}_m(\alpha/2) - \alpha/2\big| \le 2 \sup_{u\in[0,1]} \big|\hat{G}_m(u) - u\big|.$$

Integrating over $\alpha \in [0,1]$ gives

$$J_{\text{IAE},m}(\hat{F}_n) = \int_0^1 \big|\hat{\delta}_{\alpha,m}(\hat{F}_n) - (1-\alpha)\big|\, d\alpha \le 2\, J_{\text{KS-PIT},m}(\hat{F}_n).$$

Population version. Let U denote the (randomized) μ–PIT and $G(u) := \mathsf{P}(U \le u)$ its CDF. Replacing $\hat{G}_m$ by G in the same argument yields

$$J_{\text{IAE},\mu}(\hat{F}_n) \ \le\ 2\, J_{\text{KS-PIT},\mu}(\hat{F}_n).$$

Discontinuous case. If some $\hat{F}_n(\cdot \mid X_i^\star)$ are discontinuous, using the randomized PIT above gives, almost surely,

$$\{Z_i \in \mathcal{C}_{i,1-\alpha}\} = \{\alpha/2 \le U_i < 1-\alpha/2\}.$$

Therefore the identity $\hat{\delta}_{\alpha,m}(\hat{F}_n) = \hat{G}_m(1-\alpha/2) - \hat{G}_m(\alpha/2)$ still holds (with half-open endpoints), and the inequalities in Proposition 3.7 follow unchanged.

A.7. Proof of Proposition 4.2. Assume $(X_i, Z_i)_{i=1}^{n+1}$ are i.i.d. and the conformal score $R((x,z);\mathcal{D})$ is permutation-invariant in its dataset argument. Let $\mathcal{D}_{n+1} = \mathcal{D}_{n+1}^{Z_{n+1}}$ and define π as in (4.2) with a tie–breaker $\tau \sim U(0,1)$ independent of the data.

For any candidate value z, set

$$A(z) := \#\{i \le n :\ R_i^z < R_{n+1}^z\}, \qquad B(z) := \#\{i \le n :\ R_i^z = R_{n+1}^z\}.$$

For the realized label Z_{n+1}, write $S_i := R_i^{Z_{n+1}}$. Let $S_{(1)} \le \cdots \le S_{(n+1)}$ be the order statistics, and let $v_1 < \ldots < v_m$ be the distinct values of $\{S_i\}_{i=1}^{n+1}$ with multiplicities $L_1, \ldots, L_m$. Define cumulative counts

$$A_1 := 0, \qquad A_r := \sum_{s<r} L_s \quad (r \ge 2).$$

Equivalently, in the ordered list $S_{(1)} \le \cdots \le S_{(n+1)}$, the rth tie block occupies indices $A_r + 1, \ldots, A_r + L_r$ and equals the constant value v_r.

Define

$$B_r \;:=\; \{\, S_{n+1} = v_r \,\} \quad \text{equivalently} \quad B_r \;:=\; \{\, \mathrm{rank}(S_{n+1}) \in \{A_r + 1,\, \ldots,\, A_r + L_r\}\}\,.$$

By i.i.d. sampling and permutation invariance of R, $(S_1, \ldots, S_{n+1})$ is exchangeable. Conditional on the ordered values, the index $n+1$ is equally likely to occupy any of the $n+1$ ranks. Hence

$$\mathsf{P}\left(B_r \;\middle|\; S_{(1)}, \ldots, S_{(n+1)}\right) = \frac{L_r}{n+1}.$$

On B_r we have $A(Z_{n+1}) = A_r$ and $B(Z_{n+1}) = L_r - 1$, so by (4.2)

$$(n+1)\,\pi(Z_{n+1}) \;=\; A_r + \tau\, L_r,$$

and conditionally, we have $(n+1)\pi(Z_{n+1}) \sim U\big([A_r,\, A_r + L_r]\big)$.

Therefore, for any $t \in [0, n+1]$,

$$\begin{aligned}\mathsf{P}\big((n+1)\pi(Z_{n+1}) \le t \;\big|\; B_r,\, S_{(1)}, \ldots, S_{(n+1)}\big) &= \mathsf{P}\Big(\tau \le \frac{t - A_r}{L_r} \;\Big|\; B_r,\, S_{(1)}, \ldots, S_{(n+1)}\Big) \\ &= \frac{(t - A_r)_+ \wedge L_r}{L_r}.\end{aligned}$$

By the law of total probability,

$$\mathsf{P}\big((n+1)\pi(Z_{n+1}) \le t \;\big|\; S_{(1)}, \ldots, S_{(n+1)}\big) = \sum_{r=1}^{m} \frac{L_r}{n+1} \cdot \frac{(t - A_r)_+ \wedge L_r}{L_r} = \frac{t}{n+1},$$

since the intervals $[A_r,\, A_r + L_r]$ partition $[0, n+1]$ and $(t - A_r)_+ \wedge L_r$ is the length of $[A_r,\, A_r + L_r] \cap [0, t]$. Hence $(n+1)\pi(Z_{n+1}) \sim U[0, n+1]$ unconditionally, and $\pi(Z_{n+1}) \sim U(0, 1)$.

In particular, for any $\alpha \in [0, 1]$,

$$\mathsf{P}\big(\pi(Z_{n+1}) \le 1 - \alpha\big) = 1 - \alpha.$$

A.8. Proof of Proposition 4.6.

Assumptions. We work in the noiseless interpolation setting with a fixed, centered Gaussian process prior $\xi \sim \mathrm{GP}(0, k)$, where the covariance kernel k is strictly positive definite on the set $\{x_1, \ldots, x_n, x_{n+1}\}$. The design points are pairwise distinct, so the Gram matrices below are symmetric positive definite (hence invertible). Throughout the derivation we condition on a fixed dataset $\mathcal{D}_n = \{(x_i, z_i)\}_{i=1}^n$ and a fixed test location x_{n+1}. If a known nonzero mean m is used, replace z_i by $z_i - m(x_i)$ and z by $z - m(x_{n+1})$; the algebra is unchanged. Extensions to universal and intrinsic kriging are discussed in Remark A.2.

Let $z_{1:n} = (z_1, \ldots, z_n)^\top$. For a test covariate x_{n+1}, set

$$K_n := \big(k(x_i, x_j)\big)_{i,j=1}^n, \qquad k_* := \big(k(x_1, x_{n+1}), \ldots, k(x_n, x_{n+1})\big)^\top, \qquad k_{**} := k(x_{n+1}, x_{n+1}).$$

Form the augmented covariance

$$K_{n+1} = \begin{bmatrix} K_n & k_* \\ k_*^\top & k_{**} \end{bmatrix}, \qquad u := K_n^{-1} k_*, \qquad v := k_{**} - k_*^\top K_n^{-1} k_*.$$

By the block inverse formula,

$$\bar{K}_{n+1} := K_{n+1}^{-1} = \begin{bmatrix} K_n^{-1} + u\, v^{-1} u^\top & -u\, v^{-1} \\ -v^{-1} u^\top & v^{-1} \end{bmatrix}.$$

Write $\bar{k}_{ij} = (\bar{K}_{n+1})_{ij}$ and $\bar{k}_i = (\bar{K}_{n+1})_{ii}$. Then

$$\bar{k}_{n+1} = \frac{1}{v}, \qquad \bar{k}_{i,n+1} = -\frac{u_i}{v}, \qquad \bar{k}_i = (K_n^{-1})_{ii} + \frac{u_i^2}{v}.$$

Let $z^{\mathrm{aug}} = (z_1, \ldots, z_n, z)^\top$ and set $w := \bar{K}_{n+1} z^{\mathrm{aug}}$.

Components of w. Block multiplication gives

$$w_{1:n} = K_n^{-1} z_{1:n} + u\, v^{-1}\big(u^\top z_{1:n} - z\big), \qquad w_{n+1} = v^{-1}\big(z - u^\top z_{1:n}\big).$$

Using $m_n(x_{n+1}) = k_*^\top K_n^{-1} z_{1:n} = u^\top z_{1:n}$,

$$w_j = \big(K_n^{-1} z_{1:n}\big)_j + \frac{u_j}{v}\big(m_n(x_{n+1}) - z\big), \quad j = 1, \ldots, n, \qquad w_{n+1} = \frac{z - m_n(x_{n+1})}{v}.$$

Standardized LOO scores. From GP LOO identities (e.g., Petit et al., 2023), the leave-one-out posterior at (x_i, z_i) in the augmented dataset satisfies

$$m_{n+1,-i}(x_i) = z_i - \frac{w_i}{\bar{k}_i}, \qquad \sigma_{n+1,-i}(x_i) = \frac{1}{\sqrt{\bar{k}_i}}.$$

Therefore the standardized residual score is

$$R_i^z = \frac{z_i - m_{n+1,-i}(x_i)}{\sigma_{n+1,-i}(x_i)} = \frac{w_i}{\sqrt{\bar{k}_i}}, \qquad i = 1, \ldots, n.$$

Substituting the expression of w_i gives

$$R_i^z = \frac{(K_n^{-1} z_{1:n})_i + \dfrac{u_i}{v}\,(m_n(x_{n+1}) - z)}{\sqrt{(K_n^{-1})_{ii} + \dfrac{u_i^2}{v}}}, \qquad i = 1, \ldots, n.$$

At the test point x_{n+1}, the GP posterior variance is $\sigma_n^2(x_{n+1}) = v$ and the mean is $m_n(x_{n+1})$, so

$$R_{n+1}^z = \frac{z - m_n(x_{n+1})}{\sqrt{v}} = \frac{w_{n+1}}{\sqrt{\bar{k}_{n+1}}}.$$

Affine form for $R_{n+1}^z - R_i^z$. Introduce the shorthands

$$a_i := (K_n^{-1} z_{1:n})_i, \qquad b_i := \frac{u_i}{v}, \qquad d_i := \sqrt{(K_n^{-1})_{ii} + \frac{u_i^2}{v}}, \qquad s := \sqrt{v}, \qquad m_n := m_n(x_{n+1}).$$

Then

$$R_i^z = \frac{a_i + b_i(m_n - z)}{d_i} = \frac{a_i + b_i m_n - b_i z}{d_i}, \qquad R_{n+1}^z = \frac{z - m_n}{s}.$$

Hence

$$R^z_{n+1} - R^z_i = \frac{z - m_n}{s} - \frac{a_i + b_i m_n - b_i z}{d_i}$$
$$= \left(\frac{1}{s} + \frac{b_i}{d_i}\right) z - \left(\frac{m_n}{s} + \frac{a_i + b_i m_n}{d_i}\right).$$

This is of the form $\beta_i z - \alpha_i$ with

$$\beta_i = \frac{1}{s} + \frac{b_i}{d_i} = \frac{1}{\sqrt{v}} + \frac{u_i}{v\sqrt{(K_n^{-1})_{ii} + u_i^2/v}}, \qquad \alpha_i = \frac{m_n}{s} + \frac{a_i + b_i m_n}{d_i}.$$

The difference of residual scores can be expressed in centered affine form,

$$R^z_{n+1} - R^z_i = \beta_i\,(z - c_i),$$

where

$$c_i = \alpha_i/\beta_i = \frac{m_n d_i + s a_i + s b_i m_n}{d_i + s b_i} = m_n + \frac{s a_i}{d_i + s b_i}\,.$$

Since $s b_i = u_i/\sqrt{v}$,

$$c_i = m_n + \frac{v\,a_i}{\sqrt{v\,(K_n^{-1})_{ii} + u_i^2} + u_i} = m_n(x_{n+1}) + \frac{v\,(K_n^{-1} z_{1:n})_i}{\sqrt{v\,(K_n^{-1})_{ii} + u_i^2} + u_i}\,.$$

Positivity of β_i. An equivalent expression for the slope is

$$\beta_i = \sqrt{\bar{k}_{n+1}} - \frac{\bar{k}_{i,n+1}}{\sqrt{\bar{k}_i}}, \qquad \text{since } \bar{k}_{n+1} = \frac{1}{v},\ \bar{k}_{i,n+1} = -\frac{u_i}{v},\ \bar{k}_i = (K_n^{-1})_{ii} + \frac{u_i^2}{v}.$$

Because $\bar{K}_{n+1} \succ 0$, Cauchy–Schwarz in the inner product $\langle x, y\rangle = x^\top \bar{K}_{n+1} y$ gives $|\bar{k}_{i,n+1}| < \sqrt{\bar{k}_i\,\bar{k}_{n+1}}$. Dividing by $\sqrt{\bar{k}_i}$ yields

$$-\sqrt{\bar{k}_{n+1}} < \frac{\bar{k}_{i,n+1}}{\sqrt{\bar{k}_i}} < \sqrt{\bar{k}_{n+1}},$$

hence $\beta_i = \sqrt{\bar{k}_{n+1}} - \frac{\bar{k}_{i,n+1}}{\sqrt{\bar{k}_i}} > 0$. (The inequality is strict since $\bar{K}_{n+1}$ is positive definite.)

Remark A.2 (Extensions to universal and intrinsic kriging). The same affine representation holds (with obvious substitutions) in two broader cases: (i) *universal kriging* with unknown linear mean $m(x) = h(x)^\top \beta$ and known regressors $h(x) \in \mathbb{R}^q$, by replacing K_n^{-1} with $Q^{-1} := K_n^{-1} - K_n^{-1}H(H^\top K_n^{-1} H)^{-1} H^\top K_n^{-1}$, where $H = [h(x_1), \ldots, h(x_n)]^\top$; and (ii) *intrinsic kriging* with conditionally positive definite kernels and drift space $\mathcal{P}$, by projecting onto $\mathrm{col}(H)^\perp$ (choose an orthonormal basis W and work with $W^\top K_n W$), assuming the design is unisolvent for $\mathcal{P}$.

A.9. Proof of Proposition 4.13. With distinct thresholds $-\infty = c_{(0)} < c_{(1)} < \cdots < c_{(n)} < c_{(n+1)} = \infty$, the stepwise CPD satisfies

$$\hat{F}^{\text{CPS-GP}}_{n,\tau}(z \mid x_{n+1}) = \frac{i + \tau}{n+1}, \qquad z \in (c_{(i)}, c_{(i+1)}),$$

so that the right limit at $c_{(i)}$ is $\hat{F}^{\text{CPS-GP}}_{n,\tau}(c^+_{(i)} \mid x_{n+1}) = (i+\tau)/(n+1)$.

Lower endpoint. The lower bound of $\mathcal{C}^{\text{CPS-GP}}_{n,\tau,1-\alpha}(x_{n+1})$ is finite if and only if $\alpha/2$ exceeds the leftmost plateau level $\hat{F}(c^+_{(0)} \mid x_{n+1}) = \tau/(n+1)$, i.e.

$$\frac{\alpha}{2} > \frac{\tau}{n+1}.$$

Upper endpoint. The upper bound is finite if and only if there exists a finite $c_{(j)}$ such that $\hat{F}(c^+_{(j)} \mid x_{n+1}) \geq 1-\alpha/2$. The last finite level is

$$\hat{F}(c^+_{(n)} \mid x_{n+1}) = \frac{n+\tau}{n+1} = 1 - \frac{1-\tau}{n+1},$$

so finiteness requires

$$1-\alpha/2 \leq \frac{n+\tau}{n+1} \quad \Longleftrightarrow \quad \frac{\alpha}{2} \geq \frac{1-\tau}{n+1}.$$

Combining both conditions yields

$$\alpha \geq \frac{2}{n+1}\max\{\tau,\, 1-\tau\},$$

with strict inequality required in the case $\tau > 1/2$, since the lower–endpoint condition involves a strict inequality.

A.10. Proof of Proposition 4.14. *Precomputation.* Compute the Cholesky factorization $K_n = LL^\top$ at a cost of $O(n^3)$. From L, obtain $\alpha = K_n^{-1} z_{1:n}$ via two triangular solves in $O(n^2)$. If needed, $\operatorname{diag}(K_n^{-1})$ can be computed once in $O(n^3)$ (for example by forming L^{-1}); this cost is included in the same precomputation order.

Per prediction point x_{n+1}.

(i) Compute $k_* \in \mathbb{R}^n$ and k_{**}: $O(n)$ kernel evaluations.

(ii) Solve $u = K_n^{-1} k_*$ using L: $O(n^2)$.

(iii) Compute $v = k_{**} - k_*^\top u$ and $m = k_*^\top \alpha$: $O(n)$.

(iv) Evaluate all thresholds c_i using their closed forms, based on u_i, v, m, $(K_n^{-1} z_{1:n})_i$, and $(K_n^{-1})_{ii}$: $O(n)$.

(v) If the full stepwise distribution is required, sort $\{c_i\}_{i=1}^n$: $O(n \log n)$.

The resulting per–point complexity is $O(n^2 + n\log n)$, dominated by the triangular solves, with sorting as the only extra operation beyond standard GP prediction ($O(n^2)$). If only the interval endpoints are needed, the relevant order statistics can be extracted in expected $O(n)$ time (e.g. by quickselect), yielding $O(n^2+n)$ per prediction point.

Appendix B. Forecasting primer: auto-, marginal-, and probabilistic calibration.

This section recalls standard calibration notions in a general prediction-space setting; they serve as background only. The main article evaluates calibration in a design-based sense with respect to μ (Sections 3.4–3.5).

Let $(\Omega, \mathcal{A}, \mathsf{P})$ be a probability space. A *predictive CDF* is a measurable random element $\hat{F} : \Omega \to \mathcal{D}$, where $\mathcal{D}$ is the set of all cumulative distribution functions on $\mathbb{R}$, equipped with the σ-algebra generated by finite-dimensional cylinders. We denote by $\sigma(\hat{F})$ the σ-algebra generated by $\hat{F}$.

Let $Z : \Omega \to \mathbb{R}$ be the outcome of interest with true CDF F. All statements below are made under the joint distribution of $(\hat{F}, Z)$ (and, when needed, an auxiliary $\tau \sim \mathcal{U}(0,1)$ independent of both).

Remark B.1. This setting differs slightly from the *prediction space* framework of Gneiting and Ranjan (2013), where elements of Ω are realizations of the triple $(\hat{F}, Z, \tau)$.

Auto-calibration. A predictive distribution is *auto-calibrated* (Tsyplakov, 2013) if

$$\mathsf{P}(Z \le z \mid \sigma(\hat{F})) = \hat{F}(z) \quad \text{a.s. for all } z \in \mathbb{R}.$$

This requires that, conditional on $\sigma(\hat{F})$, the true conditional distribution of Z coincides with $\hat{F}$. This is a strong requirement. Tests for auto-calibration exist (Strähl and Ziegel, 2015), but assessing it in practice is difficult, which motivates weaker notions.

Marginal calibration. A predictive distribution $\hat{F}$ is *marginally calibrated* if, for every $z \in \mathbb{R}$,

$$\mathsf{E}[\hat{F}(z)] = \mathsf{P}(Z \le z).$$

Auto-calibration implies marginal calibration: since $\mathsf{P}(Z \le z \mid \sigma(\hat{F})) = \hat{F}(z)$ a.s., taking expectations and using the law of total expectation gives the identity above. Marginal calibration is weaker than auto-calibration and is more tractable empirically, since both sides can be estimated by averages over forecast–observation pairs.

Probabilistic calibration. A second relaxation is *probabilistic calibration*, based on the *probability integral transform* (PIT). For a continuous predictive CDF $\hat{F}$ and an observation Z, define

$$U_{\hat{F}}^{Z} = \hat{F}(Z).$$

If $\hat{F} = F$, then $U_{\hat{F}}^{Z} \sim \mathcal{U}(0,1)$; this is the usual PIT property. A forecast $\hat{F}$ is called *probabilistically calibrated* if $U_{\hat{F}}^{Z}$ is uniformly distributed on $[0,1]$. In practice, the PIT property is assessed by comparing the empirical distribution of PIT values to the uniform law, for example via histograms or empirical CDF plots (Dawid, 1984). For general, possibly discontinuous, predictive distributions, one uses the randomized PIT

$$U_{\hat{F}}^{Z} = \hat{F}(Z^{-}) + \tau\big(\hat{F}(Z) - \hat{F}(Z^{-})\big), \qquad \tau \sim \mathcal{U}(0,1),$$

which restores uniformity when $\hat{F} = F$.

Proposition B.2. *Let Z be a real-valued random variable with CDF F. Let $\hat{F}$ be a random continuous CDF, independent of Z, and define $U_{\hat{F}}^{Z} = \hat{F}(Z)$. Then the following are equivalent:*

(i) For all $u \in [0,1]$,

$$\mathsf{P}\big(U_{\hat{F}}^{Z} \le u \mid \hat{F}\big) = u \quad \text{a.s.}$$

(ii) $\hat{F} = F$ almost surely.

Proof. (ii) $\Rightarrow$ (i). If $\hat{F} = F$ a.s., then conditional on $\hat{F}$ we have $U_{\hat{F}}^{Z} = F(Z)$, which is uniform on $[0,1]$. Hence

$$\mathsf{P}\big(U_{\hat{F}}^{Z} \le u \mid \hat{F}\big) = u \quad \text{a.s.}$$

(i) $\Rightarrow$ (ii). Fix $u \in [0,1]$ and define

$$q_u = \inf\{z : \ \hat{F}(z) \geq u\}.$$

Continuity of $\hat{F}$ gives $\hat{F}(q_u) = u$ and $q_u = \hat{F}^{-1}(u)$. Using independence of $\hat{F}$ and Z,

$$\mathsf{P}\big(U_{\hat{F}}^{Z} \leq u \mid \hat{F}\big) = \mathsf{P}\big(\hat{F}(Z) \leq u \mid \hat{F}\big) = \mathsf{P}(Z \leq q_u \mid \hat{F}) = F(q_u).$$

By assumption this equals u a.s., so $F(q_u) = u$ for all $u \in [0,1]$. Thus

$$F(\hat{F}^{-1}(u)) = u \quad \text{a.s. for all } u \in [0,1],$$

which implies $F = \hat{F}$ almost surely. ∎

Finally, auto-calibration implies probabilistic calibration. Indeed, if

$$\mathsf{P}(Z \leq z \mid \sigma(\hat{F})) = \hat{F}(z) \quad \text{a.s. for all } z \in \mathbb{R},$$

then for any $u \in [0,1]$, with $q_u = \inf\{z : \hat{F}(z) \geq u\}$,

$$\mathsf{P}(U_{\hat{F}}^{Z} \leq u \mid \sigma(\hat{F})) = \mathsf{P}(Z \leq q_u \mid \sigma(\hat{F})) = \hat{F}(q_u) = u,$$

so $U_{\hat{F}}^{Z}$ is uniform (for discontinuous $\hat{F}$, the same argument applies with the randomized PIT). By contrast, marginal calibration does not imply probabilistic calibration, and probabilistic calibration does not imply marginal calibration in general (see Gneiting and Ranjan (2013, §2) and Gneiting and Resin (2023)).

Appendix C. Generalized normal distribution.

The generalized normal distribution $\mathcal{GN}(\beta, \mu, \lambda)$, with shape parameter $\beta > 0$, location $\mu \in \mathbb{R}$ and scale parameter $\lambda > 0$, is a continuous distribution that extends the classical normal distribution. Its probability density function is

$$f(z) = \frac{\beta}{2\,\Gamma(1/\beta)\,\lambda} \exp\left[-\left(\frac{|z-\mu|}{\lambda}\right)^{\beta}\right], \qquad z \in \mathbb{R}.$$

All moments exist. By symmetry, $\mathsf{E}[Z] = \mu$, and the variance is

$$\mathrm{var}(Z) = \lambda^2 \frac{\Gamma(3/\beta)}{\Gamma(1/\beta)}.$$

The shape parameter β controls tail behavior. When $\beta = 2$, one recovers the Gaussian case:

$$\mathcal{GN}(2, \mu, \lambda) = \mathcal{N}\big(\mu, \lambda^2/2\big).$$

For $\beta < 2$, the distribution has heavier tails than the Gaussian, and for $\beta > 2$ it has lighter tails. Further analytical properties of the generalized normal family are given in Nadarajah (2005).

Appendix D. SCRPS formulas.

D.1. SCRPS for the generalized normal distribution.

Let F_β be the CDF of $\mathcal{GN}(\beta, 0, 1)$, and $\Gamma(\cdot,\cdot)$ is the upper incomplete gamma function. The SCRPS for $Z \sim \mathcal{GN}(\beta, 0, 1)$ can be approximated using the Propositions D.1 and D.2. Indeed, $\mathsf{E}(|Z' - Z|)$ for Z' a copy of Z, can be approximated for several values of β, then an interpolation technique

is used to learn $\beta \to \mathsf{E}(|Z' - Z|)$, and $\mathsf{E}(|Z - z'|)$ admits a close form formula.

Proposition D.1. *Let $z' \in \mathbb{R}$ and $Z \sim GN(\beta, \mu, \lambda)$, then, with $u = (z' - \mu)/(\lambda)$,*

$$\mathsf{E}(|Z - z'|) = \lambda \left[u\left(2F_\beta(u) - 1\right) + \frac{1}{\Gamma\left(1/\beta\right)} \Gamma\left(\frac{2}{\beta}, |u|^\beta\right)\right]. \tag{D.1}$$

Proof.

$$\begin{aligned}
\mathsf{E}(|Z - z'|) &= \int_{\mathbb{R}} |y - z'| \frac{\beta}{2\Gamma(1/\beta)\lambda} \exp\left[-\left(\frac{|y-\mu|}{\lambda}\right)^\beta\right] \mathrm{d}y \\
&= \lambda \frac{\beta}{2\Gamma(1/\beta)} \left(\int_{-\infty}^{u} (u - z) \exp\left(-|z|^\beta\right) \mathrm{d}z + \int_{u}^{+\infty} (z - u) \exp\left(-|z|^\beta\right) \mathrm{d}z \right) \\
&= \lambda \left[u(2F_\beta(u) - 1) + \frac{1}{\Gamma\left(1/\beta\right)} \Gamma\left(\frac{2}{\beta}, |u|^\beta\right)\right].
\end{aligned}$$

The last line is now proved, first notice that by symmetry of $\int_0^u z \exp\left(-|z|^\beta\right) \mathrm{d}z$:

$$\begin{cases}
\int_u^{+\infty} z \exp\left(-|z|^\beta\right) \mathrm{d}z &= \frac{1}{\beta}\Gamma\left(\frac{2}{\beta}\right) - \int_0^{|u|} z \exp\left(-|z|^\beta\right) \mathrm{d}z \\
\int_{-\infty}^{u} z \exp\left(-|z|^\beta\right) \mathrm{d}z &= \int_0^{|u|} z \exp\left(-|z|^\beta\right) \mathrm{d}z - \frac{1}{\beta}\Gamma\left(\frac{2}{\beta}\right),
\end{cases}$$

Finally,

$$\int_u^{+\infty} z \exp\left(-|z|^\beta\right) \mathrm{d}z - \int_{-\infty}^{u} z \exp\left(-|z|^\beta\right) \mathrm{d}z = 2\frac{1}{\beta}\Gamma(2/\beta, |u|^\beta)$$

which proves the results. ∎

Lemma D.2. *Let the random variables $Z', Z \sim GN(\beta, \mu, \lambda)$ and $U', U \sim GN(\beta, 0, 1)$, then* $\mathsf{E}(|Z' - Z|) = \lambda \mathsf{E}(|U' - U|)$.

Proof. Follows from the fact that $\lambda U + \mu \sim GN(\beta, \mu, \lambda)$. ∎

D.2. SCRPS for empirical distributions. Let an i.i.d sequence $Z_1, \ldots, Z_n$ and denote by $F_n(z) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}(Z_i \leq z)$ its empirical CDF. The SCRPS for a random variable with CDF F_n can be computed using Proposition D.3.

Proposition D.3. *Let $z \in \mathbb{R}$ and $Z', Z \sim F_n$, then:*

$$\mathsf{E}_{Z\sim F_n}(|Z - z|) = \frac{1}{n}\sum_{i=1}^n |Z_i - z|, \quad \textit{and} \quad \mathsf{E}_{Z',Z\sim F_n}(|Z' - Z|) = \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n |Z_i - Z_j|. \tag{D.2}$$

Proof. The proof is straightforward. ∎

Appendix E. Test functions used in the experiments.

Table 9: Test functions used in the numerical experiments. Each function is continuous, deterministic, and defined on the domain $\mathbb{X}$. Expressions and domains follow the standard definitions provided in Surjanovic and Bingham (2013–).

Name	**Domain**	**Expression of $f(x)$**
Branin	$[-5,10]\times[0,15]$	$\big(x_2-\frac{5.1}{4\pi^2}x_1^2+\frac{5}{\pi}x_1-6\big)^2+10\big(1-\frac{1}{8\pi}\big)\cos(x_1)+10$
Goldstein–Price	$[-2,2]^2$	$(1+(x_1+x_2+1)^2(19-14x_1+3x_1^2-14x_2+6x_1x_2+3x_2^2))(30+(2x_1-3x_2)^2(18-32x_1+12x_1^2+48x_2-36x_1x_2+27x_2^2))$
Rosenbrock	$[-5,10]^d$	$\sum_{i=1}^{d-1}\big[100(x_{i+1}-x_i^2)^2+(x_i-1)^2\big]$
Ackley	$[-32.168,32.168]^d$	$-20\exp\Big(-0.2\sqrt{\frac{1}{d}\sum_i x_i^2}\Big)-\exp\big(\frac{1}{d}\sum_i\cos(\pi x_i)\big)+20+e$
Beale	$[-4.5,4.5]^2$	$(1.5-x_1+x_1x_2)^2+(2.25-x_1+x_1x_2^2)^2+(2.625-x_1+x_1x_2^3)^2$
Dixon–Price	$[-10,10]^d$	$(x_1-1)^2+\sum_{i=2}^{d}i(2x_i^2-x_{i-1})^2$
Hartmann–3	$[0,1]^3$	$-\sum_{i=1}^{4}c_i\exp\big(-\sum_{j=1}^{3}a_{ij}(x_j-p_{ij})^2\big)$
Hartmann–6	$[0,1]^6$	$-\sum_{i=1}^{4}c_i\exp\big(-\sum_{j=1}^{6}a_{ij}(x_j-p_{ij})^2\big)$

Hartmann parameters.

$$a^{(3)}=\begin{bmatrix}3.0 & 10.0 & 30.0\\ 0.1 & 10.0 & 35.0\\ 3.0 & 10.0 & 30.0\\ 0.1 & 10.0 & 35.0\end{bmatrix},\quad c^{(3)}=\begin{bmatrix}1.0\\1.2\\3.0\\3.2\end{bmatrix},\quad p^{(3)}=10^{-1}\begin{bmatrix}1 & 1 & 1\\ 3 & 3 & 3\\ 5 & 5 & 5\\ 7 & 7 & 7\end{bmatrix}.$$

$$a^{(6)}=\begin{bmatrix}10 & 3 & 17 & 3.5 & 1.7 & 8\\ 0.05 & 10 & 17 & 0.1 & 8 & 14\\ 3 & 3.5 & 1.7 & 10 & 17 & 8\\ 17 & 8 & 0.05 & 10 & 0.1 & 14\end{bmatrix},\quad c^{(6)}=\begin{bmatrix}1.0\\1.2\\3.0\\3.2\end{bmatrix},\quad p^{(6)}=10^{-2}\begin{bmatrix}1312 & 1696 & 5569 & 124 & 8283 & 5886\\ 2329 & 4135 & 8307 & 3736 & 1004 & 9991\\ 2348 & 1451 & 3522 & 2883 & 3047 & 6650\\ 4047 & 8828 & 8732 & 5743 & 1091 & 381\end{bmatrix}.$$

References.

N. Acharki, A. Bertoncello, and J. Garnier. Robust prediction interval estimation for gaussian processes by cross-validation method. *Comput. Stat. Data Anal.*, 178:107597, 2023.

S. Allen, J. Bhend, O. Martius, and J. Ziegel. Weighted verification tools to evaluate univariate and multivariate probabilistic forecasts for high-impact weather events. *Weather and Forecasting*, 38(3):499–516, 2023.

S. Allen, J. Koh, J. Segers, and J. Ziegel. Tail calibration of probabilistic forecasts. *J. Am. Stat. Assoc.*, pages 1–20, 2025.

A. N. Angelopoulos, R. F. Barber, and S. Bates. Theoretical foundations of conformal prediction, 2025. URL https://arxiv.org/abs/2411.11824.

S. Arnold, E. Walz, J. Ziegel, and T. Gneiting. Decompositions of the mean continuous ranked probability score. *Electron. J. Stat.*, 18(2):4992–5044, 2024.

D. Azzimonti, D. Ginsbourger, C. Chevalier, J. Bect, and Y. Richet. Adaptive design of experiments for conservative estimation of excursion sets. *Technometrics*, 63(1):13–26, 2021.

J. Bect, L. Li, and E. Vazquez. Bayesian subset simulation. *SIAM/ASA J. Uncertain. Quantif.*, 5(1):762–786, 2017.

D. Bolin and J. Wallin. Local scale invariance and robustness of proper scoring rules. *Statistical Science*, 38(1):140–159, 2023.

A. Capone, S. Hirche, and G. Pleiss. Sharp calibrated gaussian processes. *Adv. Neural Inf. Process. Syst.*, 36:36579–36590, 2023.

J.-P. Chilès and P. Delfiner. *Geostatistics: modeling spatial uncertainty*. John Wiley & Sons, 1999.

A. P. Dawid. Present position and potential developments: Some personal views: Statistical theory: The prequential approach. *Journal of the Royal Statistical Society. Series A (General)*, 147(2):278–292, 1984.

B. Dey, D. Zhao, B. H. Andrews, J. A. Newman, R. Izbicki, and A. B. Lee. Towards instance-wise calibration: Local amortized diagnostics and reshaping of conditional densities (ladar), 2024. URL https://arxiv.org/abs/2205.14568. v6, 30 Dec 2024.

F. Diebold, T. Gunther, and A. Tay. Evaluating density forecasts with applications to financial risk management. *International Economic Review*, 39(4):863–883, 1998. ISSN 00206598, 14682354.

P. Feliot, J. Bect, and E. Vazquez. A bayesian approach to constrained single-and multi-objective optimization. *J. Global Optim.*, 67(1):97–133, 2017.

T. Gneiting and R. Ranjan. Combining predictive distributions. *Electron. J. Stat.*, 7, 2013.

T. Gneiting and J. Resin. Regression diagnostics meets forecast evaluation: conditional calibration, reliability diagrams, and coefficient of determination. *Electron. J. Stat.*, 17(2): 3226–3286, 2023.

T. G. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.*, 102:359–378, 2007.

A. Hadji and B. Szabó. Can we trust bayesian uncertainty quantification from gaussian process priors with squared exponential covariance kernel? *SIAM/ASA J. Uncertain. Quantif.*, 9 (1):185–230, 2021.

E. Jaber, V. Blot, N. Brunel, V. Chabridon, E. Remy, B. Iooss, D. Lucor, M. Mougeot, and A. Leite. Conformal approach to gaussian process surrogate evaluation with coverage guarantees, 2024. hal-04389163 (preprint submitted on 11 January 2024).

D. Jones, M. Schonlau, and W. Welch. Efficient global optimization of expensive black-box functions. *J. Global Optim.*, 13:455–492, 12 1998.

T. Karvonen, G. Wynne, F. Tronarp, C. Oates, and S. Sarkka. Maximum likelihood estimation and uncertainty quantification for gaussian process approximation of deterministic functions. *SIAM/ASA J. Uncertain. Quantif.*, 8(3):926–958, 2020.

V. Kuleshov and S. Deshpande. Calibrated and sharp uncertainties in deep learning via density estimation. In *Int. Conf. Mach. Learn.*, pages 11683–11693. PMLR, 2022.

V. Kuleshov, N. Fenner, and S. Ermon. Accurate uncertainties for deep learning using calibrated regression. *Proc. Int. Conf. Mach. Learn.*, 80:2796–2804, 2018.

R. Liang and R. F. Barber. Algorithmic stability implies training-conditional coverage for distribution-free prediction methods. *Ann. Stat.*, 53(4):1457–1482, aug 2025.

A. Marrel and B. Iooss. Probabilistic surrogate modeling by gaussian process: A review on recent insights in estimation and validation. *Reliab. Eng. Syst. Saf.*, page 110094, 2024.

G. Matheron. The intrinsic random functions and their applications. *Advances in applied probability*, 5(3):439–468, 1973.

G. Matheron. *Estimating and Choosing: An Essay on Probability in Practice*. Springer, hasofer, a.m. (translator) edition, 1989.

W. Q. Meeker, G. J. Hahn, and L. A. Escobar. *Statistical Intervals: A Guide for Practitioners and Researchers*. John Wiley & Sons, Hoboken, New Jersey, second edition, 2017. ISBN 978-0-471-68717-7.

S. Nadarajah. Analytical properties of generalized gaussian distributions. *J. Appl. Stat.*, 32 (7):685–694, 2005.

H. Papadopoulos. Guaranteed coverage prediction intervals with gaussian process regression. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(12):9072–9083, 2024.

S. J. Petit, J. Bect, P. Feliot, and E. Vazquez. Parameter selection in gaussian process interpolation: An empirical study of selection criteria. *SIAM/ASA J. Uncertain. Quantif.*, 11(4):1308–1328, 2023.

A. Pion and E. Vazquez. Gaussian process interpolation with conformal prediction: Methods and comparative analysis. In G. Nicosia, V. Ojha, S. Giesselbach, M. P. Pardalos, and R. Umeton, editors, *Mach. Learn. Optim. Data Sci.*, pages 218–228, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-82484-5.

C. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 11 2005. ISBN 9780262256834.

J. T. Santner, B. J. Williams, and W. I. Notz. *The design and analysis of computer experiments*. Springer, 2003.

J. Shen, R. Y. Liu, and M. Xie. Prediction with confidence—a general framework for predictive inference. *J. Stat. Plan. Inference*, 195:126–140, 2018.

M. L. Stein. *Interpolation of Spatial Data*. Springer Ser. Stat. Springer New York, 1999. ISBN 978-1-4612-7166-6 978-1-4612-1494-6.

C. Strähl and J. Ziegel. Cross-calibration of probabilistic forecasts, 2015.

S. Surjanovic and D. Bingham. Virtual library of simulation experiments: Test functions and datasets, 2013–. URL https://www.sfu.ca/~ssurjano/optimization.html. Simon Fraser University, accessed November 2025.

B. Szabó, A. van der Vaart, and H. van Zanten. Frequentist coverage of adaptive nonparametric bayesian credible sets. *Ann. Stat.*, 43(4):1391–1428, 2015.

A. Tsyplakov. Evaluation of probabilistic forecasts: Proper scoring rules and moments. *SSRN Electronic Journal*, 03 2013. doi: 10.2139/ssrn.2236605.

E. Vazquez. Gpmp: the gaussian process micro package, 2025. URL https://github.com/gpmp-dev/gpmp.

J. Villemonteix, E. Vazquez, and E. Walter. An informational approach to the global optimization of expensive-to-evaluate functions. *J. Global Optim.*, 44(4):509–534, 2009.

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

V. Vovk, I. Nouretdinov, V. Manokhin, and A. Gammerman. Conformal predictive distributions with kernels. In L. Rozonoer, B. Mirkin, and I. Muchnik, editors, *Braverman Readings in Machine Learning. Key Ideas from Inception to Current State, International Conference Commemorating the 40th Anniversary of Emmanuil Bravermanś Decease*, pages 103–121,

Boston, MA, USA, April 28-30 2017a. Springer International Publishing. ISBN 978-3-319-99492-5.

V. Vovk, J. Shen, V. Manokhin, and M. Xie. Nonparametric predictive distributions based on conformal prediction. In A. Gammerman, V. Vovk, Z. Luo, and Harris Papadopoulos, editors, *Proceedings of the Sixth Workshop on Conformal and Probabilistic Prediction and Applications*, volume 60 of *Proceedings of Machine Learning Research*, pages 82–102. PMLR, 13–16 Jun 2017b.

V. Vovk, J. Shen, V. Manokhin, and M. Xie. Nonparametric predictive distributions based on conformal prediction. *Machine Learning*, 108(3):445–474, 2019. ISSN 1573-0565.