CONFORMAL APPROACH TO GAUSSIAN PROCESS SURROGATE EVALUATION WITH MARGINAL COVERAGE GUARANTEES

Edgar Jaber,[1,3,4,*] Vincent Blot,[2,3] Nicolas J.B. Brunel,[2] Vincent Chabridon,[1] Emmanuel Remy,[1] Bertrand Iooss,[1] Didier Lucor,[3] Mathilde Mougeot,[4] & Alessandro Leite[3,5]

[1]*EDF R&D, 6 Q. Watier, 78401, Chatou, France*

[2]*Capgemini Invent, 147 Q. du Pdt. Roosevelt, 92130, Issy-les-Moulineaux, France*

[3]*Paris-Saclay University, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique, 91405, Orsay, France*

[4]*Paris-Saclay University, CNRS, ENS Paris-Saclay, Centre Borelli, 91190, Gif-sur-Yvette, France*

[5]*ENSIIE, 1 square de la Résistance 91000, Évry-Courcouronnes, France*

*Address all correspondence to: Edgar Jaber, EDF R&D, 6 Q. Watier, 78401, Chatou, France, E-mail: edgar.jaber@edf.fr

Gaussian processes (GPs) are a Bayesian machine learning (ML) approach widely used to construct surrogate models for the uncertainty quantification (UQ) of computer simulation codes in industrial applications. It provides both a mean predictor and an estimate of the posterior prediction variance, the latter being used to produce Bayesian credibility intervals. Interpreting these intervals relies on the Gaussianity of the simulation model and the well-specification of the priors, which may not be appropriate. We propose to address this issue with the help of conformal prediction (CP), which is a finite-sample and distribution-free technique for estimating prediction intervals with marginal coverage guarantees. In the present work, a method for building adaptive cross-conformal prediction intervals is proposed by weighting the nonconformity score with the posterior standard deviation of the GP. The resulting CP intervals exhibit a level of adaptivity akin to Bayesian credibility sets and display a significant correlation with surrogate model local approximation error while being free from the underlying model assumptions and having marginal frequentist coverage guarantees. These estimators can be used to evaluate the quality of a GP surrogate model and can assist a decisionmaker in choosing the best prior to the specific application of the GP. We illustrate the proposed method's performance through a panel of numerical examples based on various computer experiments, including the GP metamodeling of analytical functions and an expensive-to-evaluate simulator of the clogging phenomenon in steam generators of nuclear reactors.

KEY WORDS: *conformal prediction, uncertainty quantification, Gaussian process metamodel, surrogate modeling, nonconformity score, adaptivity*

Submitted: 8/2/24; Accepted: 6/11/25; Online: 6/12/25
https://doi.org/10.1615/JMachLearnModelComput.2025054687
2689–3967/25/$35.00

1. INTRODUCTION

1.1 Motivation and Overview

In the design and analysis of computer experiments (Fang et al., 2006), the "Verification, Validation, and Uncertainty Quantification" (VV&UQ) framework has become a gold standard in many engineering fields for assessing the impact of uncertainties in numerical simulation models (De Rocquigny et al., 2008; Ghanem et al., 2017; Sullivan, 2015). Uncertainty quantification (UQ) defines a computer model as a function g, mapping a d-dimensional input $X \in \mathcal{X} \subseteq \mathbb{R}^d$ to a scalar output $Y \in \mathcal{Y} \subseteq \mathbb{R}$ through the relationship $Y = g(X)$. These models are critical in engineering for decision-making tasks such as maintenance scheduling and risk assessment of industrial systems. Typically, g represents numerical solvers for partial differential equations or high-fidelity multi-physics models. In UQ, uncertainties are often treated probabilistically, allowing input samples to be drawn from the joint distribution of X and propagated through g [e.g., via Monte Carlo sampling, see Rubinstein and Kroese (2008)] to obtain the output distribution of Y. This process treats g as a "black box," requiring no modifications to the underlying code. However, when g is computationally expensive (e.g., requiring hours or days per evaluation), standard UQ techniques can become intractable. To address this, metamodels (or surrogate models), denoted by $\widehat{g}$, compute an estimation function thanks to observation data and are often employed to reduce computational costs. This paper focuses on Gaussian process (GP) regression metamodels, also known as "Kriging" metamodels (Gramacy, 2020; Rasmussen and Williams, 2006).

In GP regression, various validation metrics have been developed in the last decades to assess the predictive quality of the fitted GP metamodel (Demay et al., 2022; Marrel and Iooss, 2024). Some effort has been put into proposing validation metrics that enable one to go beyond the measure of the quality of the mean prediction (typically measured by the predictivity coefficient), for instance, by measuring the quality of the posterior predicted variance. As exemplified in De Carvalho et al. (2022) and Jaber et al. (2025), additional cross-validation [such as K-fold or leave-one-out (LOO)] techniques can be used for assessing the robustness of the estimation on these validation indicators. However, to the best of the authors' knowledge, validation is still an open question, and no strong consensus has been reached regarding the precise metrics that should be used for validating a GP metamodel or any other metamodel in general. An efficient surrogate model must be highly adaptive to local information, particularly the training data. By conditioning on the training data, the GP metamodel develops a deeper understanding of the underlying patterns and avoids overconfidence in regions with limited or no data. The reliability of GP predictions is evaluated through Bayesian credibility intervals, which reflect the confidence in these predictions. This reliability is significantly influenced by both the quality and quantity of the training data. In areas with more observed data that is less noisy, the GP predictions are more confident and reliable, resulting in narrower credibility intervals. Additionally, properly tuned hyperparameters and carefully chosen covariance kernels can improve the GP's trustworthiness. Conversely, poor choices in these parameters can lead to overly optimistic or overly conservative uncertainty estimates. Along these lines, an alternative approach, proposed by Acharki et al. (2023), aimed at enhancing the predictive capacity of a GP metamodel by optimizing the hyperparameters of the kernel in order to tackle model misspecification and obtain more robust Bayesian credibility intervals. However, this method still heavily relies on the assumption of Gaussianity of the original model.

In the present paper, the idea is to adapt the conformal prediction (CP) framework (Vovk et al., 2005) so as to ensure more reliable prediction intervals for GP metamodels. This approach

avoids relying heavily on strong Gaussian assumptions or having a well-defined prior for the covariance kernel of the process. At the same time, it leverages the flexibility and adaptability of the local approximation provided by GP models. These two key elements allow one to fully interpret the uncertainty given by the predictive intervals we propose. This complementary tool can thus be used to assist a decisionmaker in evaluating the general quality of a GP metamodel in the light of the application for which it is used. We should emphasize that the methodologies developed in this paper are more generic than the specific context of VV&UQ and can be applied to different scenarios of GP regression, but we choose to remain in the setting of computer experiments. The notations are used accordingly.

As for CP, it has gained in the last decade a huge popularity within the machine learning (ML) community since it allows for distribution-free UQ in both classification and regression applications (Angelopoulos and Bates, 2023; Vovk et al., 2005). The CP paradigm enables the estimation of frequentist prediction intervals for any ML models (and, consequently, any metamodel) that are agnostic to the specific family of models used during the learning step. The prediction sets come with frequentist coverage guarantees, meaning that, without any additional assumptions on the original model, the probability of the true computer code output value (at a new input point), lying within the metamodel prediction interval, will be above a chosen confidence threshold. The only key assumption necessary for constructing CP sets is the *exchangeability* of the dataset (Da Veiga, 2024), which means that the concatenation of the training data set with the new test point is interchangeable in law, which is typically the case when dealing with independent and identically distributed (i.i.d.) samples such as those obtained from a crude Monte Carlo design of experiments (DoE) in UQ of computer models, or as encountered in many standard ML datasets.

A primary challenge in CP lies in producing *adaptive* prediction intervals, which refers to the property of varying interval width for different test points. The concept of "adaptivity" (Romano et al., 2019) is intrinsically tied to the *expressivity* of the metamodel, as the interval width should be small when the metamodel prediction error is minimal (in particular around training points) and large otherwise. For the purpose of GP regression quality evaluation, the adaptivity of CP interval candidates is crucial.

Three main families of methods exist for building CP intervals: the historical "full-conformal" paradigm (Vovk et al., 2005), the "split-conformal," and the "cross-conformal" settings (Angelopoulos and Bates, 2023). For the standard CP estimators in these settings, adaptivity is often lacking, and the exploration of *nonconformity scores* (which measure how unusual a suggested outcome seems with respect to other output values in the training dataset) for ensuring this property has been predominantly studied and developed in the split-conformal case (Lei et al., 2021; Romano et al., 2019; Seedat et al., 2023). Nevertheless, this approach is not practical in cases with limited budgets and/or dataset size (which can be the case for costly-to-evaluate computer models in industrial applications). The split-conformal paradigm necessitates the allocation of a *calibration set*, dividing the available data into three parts for training the metamodel, calibrating the prediction sets, and testing, respectively. Conversely, the cross-conformal paradigm, and especially the "Jackknife+" interval estimators (Barber et al., 2021), allows for the utilization of the entire dataset but requires an additional computational budget since it implies training multiple LOO metamodels.

In this paper, we introduce a methodology to obtain distribution-free, finite-sample, and adaptive prediction intervals with a desired property of "marginal coverage" (which will be later detailed) for GP metamodels in computer experiments. We employ these estimators to evaluate the GP metamodel performance and demonstrate their effectiveness in differentiating between

various prior kernel choices. The proposed approach is illustrated through several numerical examples using reference analytical functions, as well as an industrial computer-code use case.

1.2 Related Works

Within the full-conformal paradigm, the concept of "conformalizing" GPs can be traced back to the *Burnaev-Wasserman program* (Vovk et al., 2005). A theorem establishes a theoretical comparison between Bayesian credibility sets and CP sets, assuming the Gaussianity and well-specification of the original model (Burnaev and Vovk, 2014). This limit theorem provides guarantees that the differences between the upper and lower endpoints of the two intervals follow a zero-mean Gaussian distribution asymptotically. The conclusion drawn is that conformalizing under the Gaussian hypothesis is not asymptotically "worse" than standard Bayesian credibility sets. Thorough numerical comparisons with Bayesian credibility sets in various scenarios are performed in Burnaev and Nazarov (2016).

The full-conformal paradigm extends to spatial Kriging as well, as demonstrated in Mao et al. (2024), where CP algorithms are developed for non-Gaussian data by establishing conditions for approximate exchangeability. However, it is important to note that full-conformal methods are computationally expensive, requiring a complete grid search on the output space, which can quickly become prohibitive (Barber et al., 2021; Papadopoulos, 2024; Vovk et al., 2005). To enhance the efficiency of full-conformal predictive intervals, a recent work from Papadopoulos (2024) explores the idea of conformalizing GPs. In this work, the structure of the GP prediction is used to shrink the output space and ease the computation of full-conformal intervals. Moreover, it makes use of a nonconformity score similar to the one proposed in the present paper. However, we stress that the motivations of our work are different: the use of GP metamodels is standard in UQ studies for computer experiments, and we propose a conformal method to lower the number of hypotheses required to interpret prediction intervals. This is different in Papadopoulos (2024), where GPs are used for efficiently estimating full-conformal prediction intervals.

The conformal paradigm finds application in enhancing Bayesian optimization, particularly when GPs serve as query functions (Stanton et al., 2023). This is especially relevant when Bayesian credible sets obtained are deemed unreliable due to model misspecification.

Finally, notice that an idea for combining the use of a calibration set and a specific CP estimator (the Jackknife+ one, detailed further) for obtaining adaptive intervals has been explored in the recent work of Deutschmann et al. (2023). However, to the best of the authors' knowledge, pure cross-conformal adaptive methods have not been found in recent literature.

1.3 Contributions and Organization

In this work, we introduce a nonconformity score tailored for the use of GP metamodels within the cross-conformal Jackknife paradigm (detailed further). Utilizing this score, we derive an adaptive prediction interval named "J+GP," along with its "min-max" variant, and establish the marginal coverage. By quantifying the adaptivity of these setestimators, we show that the length of these intervals is interpretable as a good proxy for surrogate precision. This interpretation is supported by the significant statistical correlation observed between the interval widths and the absolute values of the metamodel error, showcasing their capability to assess the quality of a GP.

We provide a reproducible and efficient implementation of the methodology through a GitHub repository.[†] This repository is based on two pre-existing Python libraries: open source

[†] Available at https://github.com/vincentblot28/conformalized_gp

initiative for the Treatment of Uncertainties, Risks'N Statistics (OpenTURNS), an open source UQ platform (Baudin et al., 2017), and Model Agnostic Prediction Interval Estimator (MAPIE), a library dedicated to CP (Cordier et al., 2023).

The paper is organized as follows. Section 2 recalls the definition and main principles of GP regression and conformal predictors. Section 3 provides the formal definition of the new "J+GP" conformal predictor, its estimator, and its variants. Moreover, a methodology for validating the link between the error spread and the adaptivity is also presented. Section 4 proposes numerical comparisons, through a panel of datasets for regression tasks, between usual GP-based credibility intervals and the proposed J+GP variants. Finally, Section 5 draws the main conclusions of this work and discusses a few perspectives.

2. NOTATIONS AND BACKGROUND

In the rest of this paper, suppose a fixed probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Random variables are denoted with capital letters. $\mathcal{N}(\mu, \sigma^2)$ denotes the Gaussian distribution with mean μ and standard deviation σ, while $\mathcal{T}(a, b, c)$ denotes the triangular distribution with mode $b \in [a, c]$. $g : \mathcal{X} \to \mathcal{Y}$ denotes a deterministic function where $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} \subseteq \mathbb{R}$ are regular domains. For a set D, 1_D denotes the indicator function of D. The cardinal of the output space will be denoted by $n_{\text{grid}} = \text{Card}(\mathcal{Y})$. The Cartesian product of the two spaces is denoted by $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, and $2^{\mathcal{L}}$ denotes the set of subspaces of a set $\mathcal{L}$. For a given $N \in \mathbb{N}$, we fix an i.i.d. dataset $\mathcal{D}_N$ of size N whose elements are written equivalently as $Z^{(i)} = \big(X^{(i)}, g\big(X^{(i)}\big)\big) = \big(X^{(i)}, Y^{(i)}\big)$ for all $i \in \{1, \ldots, N\}$. We denote the features (or inputs) by $\boldsymbol{X} = \big(X^{(1)}, \ldots, X^{(N)}\big)$ and similarly the outputs $g(\boldsymbol{X}) = \big(g\big(X^{(1)}\big), \ldots, g\big(X^{(N)}\big)\big)$, and the dataset is denoted by $\mathcal{D}_N = (\boldsymbol{X}, g(\boldsymbol{X}))$.

As customary in supervised ML, the dataset is split into training and testing subsets, typically $\mathcal{D}_N = \mathcal{D}_n \cup \mathcal{D}_m$, with $N = n + m$ and where n, m are the respective sizes of the two subsets. We denote by $\widehat{g}$ a metamodel of g trained on $\mathcal{D}_n$, and $\widehat{g}_{-i}$ is the corresponding LOO metamodel trained on $\mathcal{D}_n \backslash \big(X^{(i)}, g\big(X^{(i)}\big)\big)$ with $i \in \{1, \ldots, n\}$. The Spearman correlation coefficient between two random variables X and Y, corresponding to the Pearson coefficient in the rank space, is denoted by $r_{\text{Spearman}}(X, Y)$. The space of continuous k-differentiable functions on $\mathcal{L}$ is denoted by $\mathcal{C}^k(\mathcal{L})$. The space of square matrices of order n on $\mathbb{R}$ will be denoted by $\mathcal{M}_n(\mathbb{R})$. For an interval $I \subset \mathbb{R}$, we denote its length as $\ell(I)$. For any $m \in \mathbb{N}$, $\mathfrak{S}(m)$ denotes the set of permutations over $\{1, \ldots, m\}$. For any finite subset $\{v_i\}_{i=1,\ldots,n}$ of an ordered set, the $(1-\alpha)$-empirical quantile, with $\alpha \in (0, 1)$, is given by

$$\widehat{q}^{+}_{n,\alpha}\{v_i\} := \text{the } \lceil (1-\alpha)(n+1) \rceil \text{th smallest value of } v_1, \ldots, v_n, \tag{1}$$

where $\lceil \cdot \rceil$ denotes the ceil function. Similarly, the α-empirical quantile is given by

$$\widehat{q}^{-}_{n,\alpha}\{v_i\} := \text{the } \lfloor \alpha(n+1) \rfloor \text{th smallest value of } v_1, \ldots, v_n, \tag{2}$$

where $\lfloor \cdot \rfloor$ denotes the floor function such that

$$\widehat{q}^{-}_{n,\alpha}\{v_i\} = -\widehat{q}^{+}_{n,\alpha}\{-v_i\}. \tag{3}$$

2.1 Gaussian Process Metamodeling

2.1.1 General Definitions

Consider a computer model g and corresponding outputs on an i.i.d. design of n-experiments $\boldsymbol{X}$. These outputs can be perturbed by some additive noise ϵ, meaning that for all $i \in \{1, \ldots, n\}$,

we have

$$Y^{(i)} = g\big(X^{(i)}\big) + \epsilon_i, \tag{4}$$

with ϵ normally distributed for example. To build a GP metamodel (Rasmussen and Williams, 2006) of such a function g, suppose that g is the realization of a certain GP $\mathcal{G} \sim \mathcal{GP}(M, K)$, where $M : \mathcal{X} \to \mathcal{Y}$ is the *mean* of the process and $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is the *covariance kernel* of the process. Then, this process is *conditioned* on the available dataset $\mathcal{D}_n$. By doing so, this procedure amounts to performing Bayesian regression considering a Gaussian *prior* on the function g in order to then obtain a *posterior* distribution. By considering zero additive noise ϵ on the outputs, this special case is referred to as GP *interpolation*, and this is the usual path taken for building GP surrogates of deterministic codes. The general principle of this type of method is sketched in Fig. 1.

For simplicity, we choose that $M = 0$ (corresponding to the case usually called "ordinary Kriging"), and we use a Matérn-ν kernel defined, for all $\nu = (2k+1)/2,\ k \in \mathbb{N}$ and $x, x' \in \mathcal{X}$, by

$$\begin{aligned} K(x, x') &= K_{(\nu,\theta,\sigma)}(x, x') \\ &= \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} \frac{|x - x'|}{\theta} \right)^{\nu} K_{\nu} \left(\sqrt{2\nu} \frac{|x - x'|}{\theta} \right), \end{aligned} \tag{5}$$

where K_ν is a modified Bessel function of the second kind and Γ is the Euler gamma function. This kernel allows better control of the regularity of the process through its hyperparameter ν since the corresponding sample paths will lie in $\mathcal{C}^{\lfloor \nu - 1 \rfloor}(\mathcal{X})$ (Gu et al., 2018). The final conditional process $\widetilde{\mathcal{G}} := \mathcal{G}|\mathcal{D}_n$ is a GP with posterior mean and covariance functions defined for all $x, x' \in \mathcal{X}$ as

$$\begin{aligned} \widetilde{g}(x) &:= k(x)^\top \boldsymbol{K}^{-1} g(\boldsymbol{X}), \\ \widetilde{K}(x, x') &:= K(x, x') - k(x)^\top \boldsymbol{K}^{-1} k(x'), \end{aligned} \tag{6}$$

where for all $x \in \mathcal{X}$,

$$\begin{aligned} k(x) &:= \big(K\big(x, X^{(1)}\big), \ldots, K\big(x, X^{(N)}\big)\big)^\top \in \mathbb{R}^n, \\ \boldsymbol{K} &:= \big(K\big(X^{(i)}, X^{(j)}\big)\big)_{1 \le i,j \le n} \in \mathcal{M}_n(\mathbb{R}). \end{aligned} \tag{7}$$

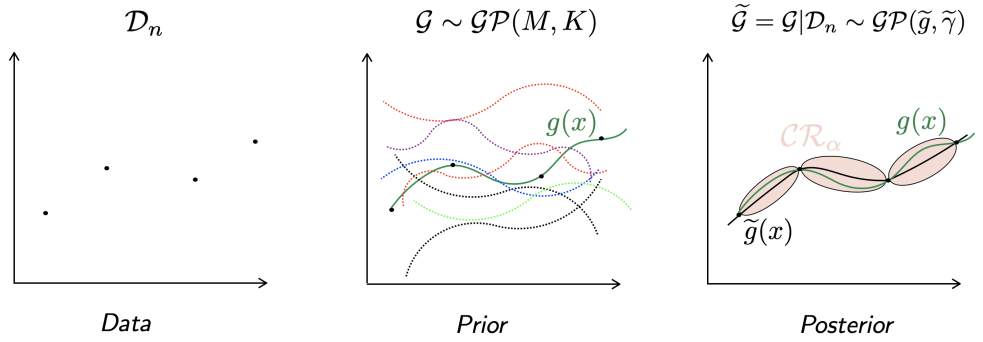

FIG. 1: GP interpolation metamodeling illustration. The data is an input–output DoE computed with the deterministic code g. Then, one assumes these data correspond to a function that is a trajectory of an underlying GP. In the absence of noise, the posterior process interpolates the data. In addition, this technique produces nonzero *credibility* intervals outside of the points in $\mathcal{D}_n$.

The mean of the posterior process $\widetilde{g}$ acts as a metamodel for the deterministic function g, thus in the case of GPs, $\widehat{g} = \widetilde{g}$. For a choice of the regularity parameter ν, the hyperparameters (σ^2, θ) of the Matérn kernel can be optimized using either maximum likelihood estimation (MLE) or a cross-validation (CV) strategy (Acharki et al., 2023). In the case of ordinary Kriging, the usual MLE optimization problem to be solved involves the negative log likelihood and can be written as

$$\big(\sigma^2_{\text{MLE}}, \theta_{\text{MLE}}\big) \in \underset{(\sigma^2,\theta)}{\arg\min} \big\{ g(\boldsymbol{X})^\top \boldsymbol{K}^{-1} g(\boldsymbol{X}) + \log(\det \boldsymbol{K}) \big\}. \tag{8}$$

According to Acharki et al. (2023), the MLE procedure yields better results if the kernel type is well specified, while the CV method is more robust in the case of misspecification. However, this is not necessarily the case for GP interpolation as outlined in Petit et al. (2023). For $x \in \mathcal{X}$, the posterior standard deviation is denoted by

$$\widetilde{\gamma}(x) := \widetilde{K}^{1/2}(x, x). \tag{9}$$

In this setting, it can be shown that for $x \in \big\{X^{(1)}, \ldots, X^{(n)}\big\}$, one has $\widetilde{\gamma}(x) = 0$ and $\widetilde{g}(x) = g(x)$, meaning that the GP metamodel is interpolating.

If the code is perturbed with an additive noise, then the usual GP regression setting is recovered, and the covariance matrix has to take into account a so-called *nugget effect*, meaning that one needs to add a regularization term modeling the noise dispersion in the covariance matrix, such that

$$\boldsymbol{K}_\epsilon := \boldsymbol{K} + \sigma^2_\epsilon \boldsymbol{I}_n, \tag{10}$$

with $\boldsymbol{I}_n$ being the identity matrix of size $(n \times n)$. The new hyperparameter σ_ϵ has to be tuned by constructing a full-likelihood, and the resulting metamodel is no longer interpolating (Rasmussen and Williams, 2006). The MLE optimization problem now becomes

$$\big(\sigma^2_{\text{MLE}}, \theta_{\text{MLE}}, \sigma^2_{\epsilon,\text{MLE}}\big) \in \underset{(\sigma^2,\theta,\sigma^2_\epsilon)}{\arg\min} \big\{ g(\boldsymbol{X})^\top \boldsymbol{K}_\epsilon^{-1} g(\boldsymbol{X}) + \log(\det \boldsymbol{K}_\epsilon) \big\}. \tag{11}$$

Moreover, a "full-Bayesian" approach exists to obtain the posterior distribution of the hyperparameters and update the predictive distribution. However, this method is out of the scope of this paper and would probably be cumbersome in the context of cross-conformal predictors since it would require tuning several complex Monte Carlo Markov chain algorithms on a large number of LOO-validated metamodels.

The use of ordinary Kriging with Matérn kernels is usually chosen to simplify the GP definitions outlined in the preceding paragraphs. It is important to note that the methodology introduced does not rely on the specific priors used. In fact, it is robust to prior misspecification, as will be demonstrated in the following sections.

2.1.2 Bayesian Credibility Intervals

In Bayesian inference, a credibility interval is related to the posterior distribution of a parameter. For a certain credibility level $(1 - \alpha) \in (0, 1)$, the value of the parameter should lie in this interval with probability $1 - \alpha$ given the available data. In the case of GPs, the parameter is the mean of the posterior GP, and for any new point $X^{(n+1)} \in \mathcal{X} \backslash \boldsymbol{X}$, the credibility prediction interval is given by

$$\mathcal{CR}_\alpha\big(X^{(n+1)}\big) = \Big[\widetilde{g}\big(X^{(n+1)}\big) \pm u_{1-\alpha/2} \widetilde{\gamma}\big(X^{(n+1)}\big) \Big], \tag{12}$$

where $u_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of the standard Gaussian distribution. Under the Gaussian assumption on the original function, if g is *truly* modeled by the posterior $\mathcal{G}|\mathcal{D}_n$, then one should have the exact training conditional coverage:

$$\mathbb{P}\left(g\big(X^{(n+1)}\big) \in \mathcal{CR}_\alpha\big(X^{(n+1)}\big) \mid \mathcal{D}_n\right) = 1-\alpha, \tag{13}$$

where $\mathcal{CR}_\alpha\big(X^{(n+1)}\big)$ is a *prediction* interval for the function g.

Even if the Gaussian assumption holds true, a common occurrence is the *misspecification* of the prior model. This implies that the set of prior mean and/or the family of kernels is proven to be incorrect. In essence, the practical reliability of Bayesian credibility intervals, particularly for GPs, can be significantly compromised.

2.2 Conformal Prediction Intervals

2.2.1 General Definitions

CP is a finite-sample and distribution-free framework for building prediction sets with a statistical guarantee on the coverage rate for any predictive algorithm (Vovk et al., 2005). Suppose a given training dataset $\mathcal{D}_n$ and a new test point $Z^{(n+1)} = \big(X^{(n+1)}, Y^{(n+1)}\big)$. It is assumed that the $n+1$ points are *exchangeable* (Vovk et al., 2005). Formally this means that for any permutation $\pi \in \mathfrak{S}(n+1)$, we have

$$\big(Z_1, \ldots, Z^{(n+1)}\big) \overset{\mathcal{L}}{=} \big(Z^{(\pi(1))}, \ldots, Z^{(\pi(n+1))}\big), \tag{14}$$

where $\overset{\mathcal{L}}{=}$ denotes an equality "in law" (also called "in distribution"). More concretely, this means that $Z^{(n+1)}$ could have been used as a training point and that any training point could have been a test point. An i.i.d. dataset is a special case of an exchangeable dataset. For any confidence level $\alpha \in (0,1)$, a conformal predictor of coverage $1-\alpha$ is any measurable function of the form (Vovk et al., 2005):

$$\begin{aligned} C_\alpha &\colon \mathcal{Z}^n \times \mathcal{X} \to 2^{\mathcal{Y}}, \\ (\mathcal{D}_n, X) &\mapsto C_\alpha(\mathcal{D}_n, X) =: C_{n,\alpha}(X), \end{aligned} \tag{15}$$

such that, for a new test point $Z^{(n+1)}$, one has the following *marginal* coverage:

$$\mathbb{P}\left(Y^{(n+1)} \in C_{n,\alpha}\big(X^{(n+1)}\big)\right) \geq 1-\alpha, \tag{16}$$

where the probability is taken over the data $\widetilde{\mathcal{D}}_n = \mathcal{D}_n \cup \big\{Z^{(n+1)}\big\}$. To build estimators of such set-functions, one relies on the use of a *nonconformity score*. This score is defined as a measurable function of the form (Vovk et al., 2005):

$$\begin{aligned} R &\colon \mathcal{Z}^n \times \mathcal{Z} \to \mathbb{R}, \\ (\mathcal{D}_n, Z) &\mapsto R(\mathcal{D}_n, Z), \end{aligned} \tag{17}$$

which quantifies how "not representative" the point Z is, compared to the dataset $\mathcal{D}_n$. For example, if a metamodel $\widehat{g}$ of a code g has been trained on $\mathcal{D}_n$, then a straightforward nonconformity score is given by the residuals:

$$R(\mathcal{D}_n, Z^{(n+1)}) = \big|g\big(X^{(n+1)}\big) - \widehat{g}\big(X^{(n+1)}\big)\big|. \tag{18}$$

It is noteworthy to insist that, in practice, the coverage property in Eq. (16) is called *marginal* since it holds on average over all possible realizations of the training set $\widetilde{\mathcal{D}}_n$. A more standard

coverage is the *training-conditional* coverage property (Angelopoulos and Bates, 2023), meaning that, for a conformal predictor estimator $\widehat{C}_{n,\alpha}$, one has

$$\mathbb{P}\left(Y^{(n+1)} \in \widehat{C}_{n,\alpha}\big(X^{(n+1)}\big) \;\middle|\; \mathcal{D}_n\right) \geq 1-\alpha. \tag{19}$$

There are mainly three ways of estimating such conformal predictors: "full-CP," also called "transductive CP" (Vovk et al., 2005); "split-CP," also called "inductive CP" (Papadopoulos et al., 2002a,b); and "cross-CP" (Vovk, 2015).

Transductive CP is historically the first CP method introduced by Vovk et al. (2005). For any choice of a nonconformity score, it implies computing the following set:

$$\begin{aligned}\widehat{C}_{n,\alpha}\big(X^{(n+1)}\big) = \Bigg\{ y \in \mathcal{Y} : \frac{1}{n}\mathrm{Card}\Big[\Big\{ i \in \{1,\dots,n\}, R\big(\widetilde{\mathcal{D}}_n, Z^{(i)}\big) \\ \geq R\Big(\widetilde{\mathcal{D}}_n, \big(X^{(n+1)}, y\big)\Big)\Big\}\Big] \;\geq \alpha\Bigg\}.\end{aligned} \tag{20}$$

As mentioned in the introduction, transductive CP is computationally intensive as it involves training one metamodel for each possible value in $\mathcal{Y}$. This conformal predictor can be made computationally effective in the cases of Ridge and Lasso regressions (Lei, 2019; Nouretdinov et al., 2001), k-nearest neighbors algorithm (Papadopoulos et al., 2008, 2011), and more recently, GP regression (Papadopoulos, 2024).

Inductive – or split-CP – on the other hand, has a very low computational cost as it requires a single training of the learning model. However, it needs to create (or save) a proper "calibration" (or "holdout") set that contains observations that have not been used during the training phase. This set is then used to estimate the quantiles of the evaluated nonconformity scores (usually, the residuals) on this set for constructing the intervals. However, for industrial applications where only a few hundred observations are available, such a calibration set may be very difficult to obtain.

Unlike the first two techniques, cross-CP has a relatively low computational cost and does not necessitate any holdout set. We now proceed to present the cross-conformal predictors.

2.2.2 Cross-Conformal Prediction Sets

The "standard" Jackknife prediction intervals require learning a metamodel $\widehat{g}$ on a training dataset $\mathcal{D}_n$ as well as n metamodels built using the LOO validation technique, denoted by $\widehat{g}_{-i}$, with $1 \leq i \leq n$. It then makes use of the $(1-\alpha)$-empirical quantile of the LOO residuals defined by

$$R_i^{\mathrm{LOO}} := \big|g\big(X^{(i)}\big) - \widehat{g}_{-i}\big(X^{(i)}\big)\big|. \tag{21}$$

For a new point $X^{(n+1)}$ and a coverage level $1-\alpha$, the standard Jackknife prediction interval is defined by

$$\widehat{C}^{\mathrm{J}}_{n,\alpha}\big(X^{(n+1)}\big) = \Big[\widehat{g}\big(X^{(n+1)}\big) \pm \widehat{q}^{\pm}_{n,\alpha}\big\{R_i^{\mathrm{LOO}}\big\}\Big]. \tag{22}$$

Unfortunately, this prediction interval does not fulfill the marginal coverage property in Eq. (16), especially in the case of a small dataset, as mentioned in Barber et al. (2021). To circumvent this limitation, a more robust cross-conformal estimator is the "Jackknife+" proposed by Barber et al. (2021). In this case, the interval is no longer centered on the prediction of the fully trained metamodel, but the LOO predictions are added in the empirical quantile. The estimator is thus given by

$$\widehat{C}^{\text{J+}}_{n,\alpha}\big(X^{(n+1)}\big) = \Big[\widehat{q}^{\pm}_{n,\alpha}\Big\{\widehat{g}_{-i}\big(X^{(n+1)}\big) \pm R_i^{\text{LOO}}\Big\}\Big]. \tag{23}$$

This estimator has the *marginal coverage* property of $1-2\alpha$. As mentioned in Theorem 5 of Barber et al. (2021), the factor "2" in $1-2\alpha$ can be removed if the metamodel satisfies the (ϵ,λ)-*out-sample stability* for $\epsilon > 0$ and $\lambda \in [0,1]$ if for all $i \in \{1,\ldots,n\}$,

$$\mathbb{P}(|\widehat{g}\big(X^{(i)}\big) - \widehat{g}_{-i}\big(X^{(i)}\big)| \leq \epsilon) \geq 1-\lambda. \tag{24}$$

In this case, Theorem 5 of Barber et al. (2021) states that the ϵ-inflation of the Jackknife+ interval [see Eq. (25)] achieves a marginal coverage of level $1-\alpha-2\sqrt{\lambda}$. Hence, if one can find a small enough value of λ, which satisfies Eq. (24), then a marginal coverage of level *approximately* $1-\alpha$ can be achieved, leading to the following prediction set:

$$\widehat{C}^{\text{J+},\epsilon}_{n,\alpha}\big(X^{(n+1)}\big) = \Big[\widehat{q}^{\pm}_{n,\alpha}\Big\{\widehat{g}_{-i}\big(X^{(n+1)}\big) \pm R_i^{\text{LOO}}\Big\} \pm \epsilon\Big]. \tag{25}$$

Nonetheless, the authors of Barber et al. (2021) indicate that, in numerical applications, the empirical coverage of the Jackknife+ barely drops below $1-\alpha$ unless the case study is somewhat "pathological." More recently, it has been established by Liang and Barber (2023) that the training-conditional property in Eq. (19) is achieved under the out-sample stability property. In most empirical cases, however, the marginal coverage property is respected.

Another way of achieving the $1-\alpha$ coverage is through the "Jackknife-minmax" method, again proposed by Barber et al. (2021), which is a more conservative implementation of the Jackknife+ method. This method differs from the latter as it does not use the prediction of each LOO-trained metamodel but the minimum (resp., the maximum) predicted value to compute the lower (resp., the upper) confidence bound. The Jackknife-minmax estimator is given by the following expression:

$$\begin{aligned}\widehat{C}^{\text{J-mm}}_{n,\alpha}\big(X^{(n+1)}\big) = \Big[&\min{}_{i\in\{1,\ldots,n\}}\Big\{\widehat{g}_{-i}\big(X^{(n+1)}\big)\Big\} - \widehat{q}^{-}_{n,\alpha}\big\{R_i^{\text{LOO}}\big\},\\ &\max{}_{i\in\{1,\ldots,n\}}\Big\{\widehat{g}_{-i}\big(X^{(n+1)}\big)\Big\} + \widehat{q}^{+}_{n,\alpha}\big\{R_i^{\text{LOO}}\big\}\Big].\end{aligned} \tag{26}$$

An illustration adapted from Barber et al. (2021) of the various Jackknife methods can be found in Fig. 2.

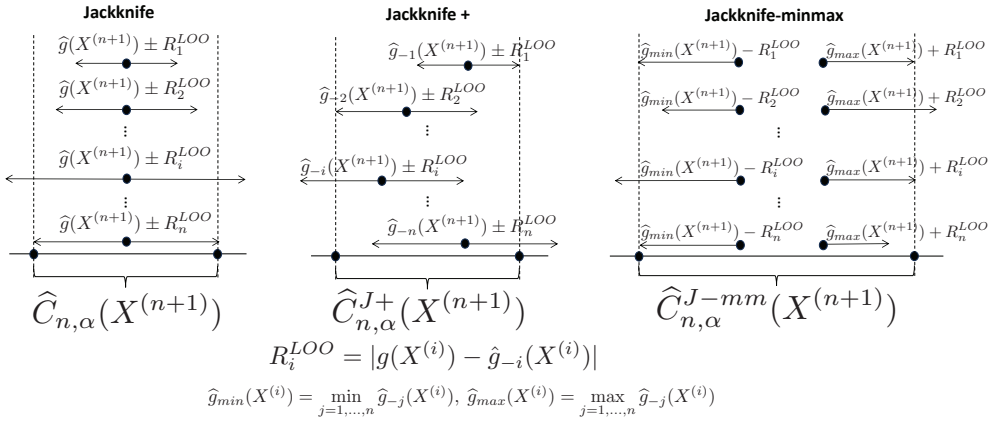

FIG. 2: Illustration of the various cross-CP methods (adapted from Barber et al., 2021)

Even if the Jackknife methods have a lower computational cost than the full-CP, if one disposes of a dataset with more than a few thousand observations, or in the case where the metamodel $\widehat{g}$ is long to train, the cost of cross-CP methods can still be quite expensive. A recap of the different cross-CP methods mentioned with their coverage and computational cost can be found in Table 1.

Finally, a major drawback of this CP method is that there is no theoretical guarantee of adaptivity. We recall that adaptivity (Romano et al., 2019) is the property of a CP interval to have nonconstant widths at different test points, and this property is related to the expressivity of the surrogate model studied. This topic will be addressed in the following using GPs.

3. CONFORMALIZED GAUSSIAN PROCESS REGRESSION: THE J+GP METHOD

3.1 Motivations and Proposed Estimator

The idea is to adapt the Jackknife+ method to GP metamodels to obtain adaptive prediction intervals. Suppose we have conditioned the GP on a dataset $\mathcal{D}_n$ by optimizing the hyperparameters $(\sigma^2_{\text{MLE}}, \theta_{\text{MLE}})$ of a Matérn-ν kernel with given ν. We thus have access to the posterior mean $\widetilde{g}$ as well as the posterior standard deviation denoted by $\widetilde{\gamma}$. For the respective LOO posteriors, we write $\widetilde{g}_{-i}$ and $\widetilde{\gamma}_{-i}$ for all $i \in \{1, \ldots, n\}$. We proceed in defining the LOO Gaussian nonconformity score, fixing a small $\delta > 0$, by

$$R_i^{\text{LOO}\gamma} := \frac{\left|g\big(X^{(i)}\big) - \widetilde{g}_{-i}\big(X^{(i)}\big)\right|}{\max\big(\delta, \widetilde{\gamma}_{-i}\big(X^{(i)}\big)\big)}, \quad \forall i \in \{1, \ldots, n\}. \tag{27}$$

The interest of δ is to avoid the zero division when the metamodel is interpolating (e.g., in the absence of a nugget effect). For a new prediction point $X^{(n+1)} \in \mathcal{X}$ and a coverage rate $(1-\alpha) \in (0,1)$, we define the "J+GP" conformal predictors, which are a variant of the Jackknife+ estimator adapted to the GP metamodeling setting:

$$\widehat{C}^{\text{J+GP}}_{n,\alpha}\big(X^{(n+1)}\big) = \left[\widehat{q}^{\pm}_{n,\alpha}\left\{\widetilde{g}_{-i}\big(X^{(n+1)}\big) \pm R_i^{\text{LOO}\gamma} \times \max\Big(\delta, \widetilde{\gamma}_{-i}\big(X^{(n+1)}\big)\Big)\right\}\right]. \tag{28}$$

This estimator enables adaptivity at different prediction points since the edges of the intervals are controlled by a function of $X^{(n+1)}$. Moreover, for this estimator, we achieve the same marginal coverage as the Jackknife+ as presented in the following theorem.

Theorem 1. *Assume $\mathcal{D}_n$ is exchangeable. For a new point $X^{(n+1)} \in \mathcal{X}$ and a coverage level $1-\alpha \in (0,1)$, one has*

$$\mathbb{P}\left(g\big(X^{(n+1)}\big) \in \widehat{C}^{\text{J+GP}}_{n,\alpha}\big(X^{(n+1)}\big)\right) \geq 1 - 2\alpha. \tag{29}$$

TABLE 1: Marginal coverage and reminder of both training and evaluation costs for CP methods adapted from Barber et al. (2021). n denotes the training sample size, n_{grid} is the cardinal of the output space, and m is the size of the test sample

Method	Marginal coverage	Training cost	Evaluation cost
Full	$\geq 1-\alpha$	$n \cdot n_{\text{grid}}$	$m \cdot n \cdot n_{\text{grid}}$
Split	$\geq 1-\alpha$	1	n
Jackknife+	$\geq 1-2\alpha$	n	$m \cdot n$
Jackknife-minmax	$\geq 1-\alpha$	n	$m \cdot n$

The proof is given in Appendix A and is based *mutatis mutandis* on the proof for the Jackknife+ in Barber et al. (2021). In the same spirit of Barber et al. (2021), we similarly propose the following "J-minmax-GP" estimator:

$$\widehat{C}_{n,\alpha}^{\text{J-mm-GP}}\big(X^{(n+1)}\big) = \Big[\min_i\Big\{\widetilde{g}_{-i}\big(X^{(n+1)}\big)\Big\} - \widehat{q}^{-}_{n,\alpha}\Big\{R_i^{\text{LOO}\gamma} \times \max\Big(\delta, \widetilde{\gamma}_{-i}\big(X^{(n+1)}\big)\Big)\Big\}, \max_i\Big\{\widetilde{g}_{-i}\big(X^{(n+1)}\big)\Big\} + \widehat{q}^{+}_{n,\alpha}\Big\{R_i^{\text{LOO}\gamma} \times \max\Big(\delta, \widetilde{\gamma}_{-i}\big(X^{(n+1)}\big)\Big)\Big\}\Big]. \tag{30}$$

Notice that this CP estimator inherits from the same coverage guarantee as the standard min-max estimator, as shown by the following theorem.

Theorem 2. *Assume $\mathcal{D}_n$ is exchangeable. For a new point $X^{(n+1)} \in \mathcal{X}$ and a marginal coverage level $\alpha \in (0,1)$, one has*

$$\mathbb{P}\left(g\big(X^{(n+1)}\big) \in \widehat{C}_{n,\alpha}^{\text{J-mm-GP}}\big(X^{(n+1)}\big)\right) \geq 1-\alpha. \tag{31}$$

The proof of the preceding theorem is given in Appendix B and is adapted from the proof for the Jackknife-minmax in Barber et al. (2021). The proposed adaptive cross-CP methods with their coverage and computational cost are summarized in Table 2.

3.2 Methodology Evaluation

A two-step approach is considered to assess the capabilities of the proposed J+GP and J-minmax-GP estimators in comparison with the usual cross-CP ones and the Bayesian credibility intervals. In the following, let $\widehat{C}^{*}_{n,\alpha}$ denote any type of prediction interval. The following computations are performed on the test subset $\mathcal{D}_m$. First, we check whether the empirical coverage property is achieved for different values of the coverage rate $1-\alpha \in [0,1]$:

$$\frac{1}{m}\sum_{i=1}^{m} 1\Big\{g\big(X^{(i)}\big) \in \widehat{C}^{*}_{n,\alpha}\big(X^{(i)}\big)\Big\} \geq 1-\alpha. \tag{32}$$

Second, the correlation of the interval width and the model error is computed. Indeed, for the intervals to be informative, they have to be small when the prediction error is small and large otherwise. Therefore, we could expect a significant correlation between the width of the interval and the residual. This metric will quantitatively reflect the adaptive nature of the proposed prediction intervals. It is thus valid to verify that, for a given coverage $1-\alpha \in (0,1)$, the Spearman correlation on the test data is nonzero is significantly different from 0 with a robustness analysis using Bootstrap estimation, i.e., that

TABLE 2: Marginal coverage, training, and evaluation costs for the intervals proposed. n denotes the training sample size, and m is the size of the test sample

Method	Marginal coverage	Training cost	Evaluation cost
J+GP	$\geq 1-2\alpha$	n	$m \cdot n$
J-minmax-GP	$\geq 1-\alpha$	n	$m \cdot n$

$$0 \ll r_{\text{Spearman}}\left(\left\{\ell\left(\widehat{C}^{*}_{n,\alpha}\left(X^{(i)}\right)\right), \left|g\left(X^{(i)}\right) - \widetilde{g}\left(X^{(i)}\right)\right|\right\}_{i\in\{n+1,\ldots,n+m\}}\right), \tag{33}$$

where $\ell\left(\widehat{C}^{*}_{n,\alpha}\left(X^{(i)}\right)\right)$ denotes the length of the prediction interval. Here, the Spearman correlation is chosen since it is more robust than the Pearson linear correlation coefficient to possible outliers and because it is able to measure monotonic dependency. It is computed in a similar fashion to the usual Pearson linear correlation, but it considers the rank transformation of data. To achieve statistical robustness, we compute bootstrap intervals on the estimation of the correlation metric.

Concerning the quality of the metamodel, it is assessed with the help of the usual *predictivity coefficient* computed from test data (Marrel et al., 2008):

$$Q^2 = 1 - \frac{\dfrac{1}{m}\displaystyle\sum_{i=n+1}^{n+m}\left|g\left(X^{(i)}\right) - \widetilde{g}\left(X^{(i)}\right)\right|^2}{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\left(g\left(X^{(i)}\right) - \dfrac{1}{n}\sum_{j=1}^{n} g(X^{(j)})\right)^2}. \tag{34}$$

This metric is widely used for assessing the predictive power of the surrogate model and for ensuring its validation (see, e.g., Fekhari et al., 2023). The closer to 1 the Q^2 is, the more predictive the mean metamodel is. Here, the analysis can be completed by computing the empirical coverage rates and the correlations between the lengths of the intervals and the residuals. Additionally, this strategy provides a way for decisionmakers to evaluate which model best suits their applications, since it can be used with different priors on the covariance kernel and the mean to further enhance the predictive power of the final metamodel. In the following numerical examples, we demonstrate the strength of our methodology by testing several GP metamodels with different values of Matérn regularity parameter ν and show that it allows one to discriminate between them in order to choose the best one. We focus on Matérn kernels because they are widely used in practice. However, this approach can be applied to other kernel families as well.

4. NUMERICAL RESULTS

In order to test our methodology, a series of numerical toy and use cases are carried out. We choose standard benchmark functions from the UQ and GP literature, as well as a real use case from nuclear engineering provided by EDF, the French national electric utility company. In the numerical results, all examples are treated as deterministic. Here, we suppose no noise in the data, which amounts to performing GP interpolation. Thus, the so-called nugget effect introduced in Eq. (10) is not considered here. Our goal is to assess the adaptivity of the estimators J+GP and J+GP-minmax using the Spearman correlation between the errors of the metamodel and the width of the prediction intervals.

As a preliminary step, standardization of the input data is recommended. Since, in these examples, we have access to the input distributions, the procedure only consists of subtracting the mean and dividing by the standard deviation. We start with an illustrative case of a misspecified one-dimensional GP and then proceed with a detailed study of three cases: namely, the wing weight function (Forrester et al., 2008), the Morokoff & Caflisch function (Morokoff and Caflisch, 1995), and an industrial use case provided by EDF (named "TPD" for "THYC-Puffer-DEPOTHYC" clogging simulation computer code; see Jaber et al. (2025) for more information about this use case). We start by going over the characteristics of the computer experiments used

and present the performance of each GP metamodel by computing its predictivity coefficient [recalled in Eq. (34)] and the *mean squared error* (MSE) given by

$$\text{MSE} = \frac{1}{m} \sum_{i=n+1}^{n+m} \left| \widetilde{g}\big(X^{(i)}\big) - g\big(X^{(i)}\big) \right|^2. \tag{35}$$

For each dataset, we present the performances of the different prediction intervals, namely the GP credibility intervals and the proposed J+GP and J-minmax-GP estimators. Three indices based on the prediction intervals are computed on the test dataset:

- the empirical coverage rate given in Eq. (32);
- the empirical average width:

$$\overline{\ell}(\widehat{C}^*_{n,\alpha}) = \frac{1}{m} \sum_{i=1}^{m} \ell\left(\widehat{C}^*_{n,\alpha}\big(X^{(i)}\big)\right); \tag{36}$$

- the Spearman correlation between the width and the error, as given in Eq. (33).

We compute the three metrics at all coverage levels $1-\alpha$ and display the plots for the Matérn kernel associated with the best performance. Then, we show tables for three different target-coverage levels (i.e., 90%, 95%, and 99%) and three different Matérn regularity parameters, such that $\nu \in \{1/2, 3/2, 5/2\}$.

In the rest of this section, we highlight in a series of tables, for each empirical coverage rate mentioned above, the target coverage rate, the kernel whose GP metamodel has the smallest average width, and the metamodel with the highest Spearman correlation (i.e., the correlation between the width of the interval and the residual error). In general, it is not the same kernel that performs best on both metrics. In this case, the decisionmaker must choose between more sensitivity to local information or more conservatism, depending on the intended application.

4.1 Code Description and Availability for Reproducible Results

The numerical results have been obtained with a Python code built upon two preexisting open source libraries: namely, OpenTURNS (Baudin et al., 2017), which is dedicated to UQ (especially, GP regression), and MAPIE (Cordier et al., 2023), which is dedicated to CP. A wrapper around OpenTURNS has been implemented to make the Scikit-learn (Pedregosa et al., 2011) GP constructors (i.e., with fit and predict methods) compatible with OpenTURNS's existing application programming interface since MAPIE handles such Scikit-learn objects. Only a few changes have been made to the MAPIE library to make it compatible with our methodology, and it preserves all of its standard conformal methods. For reproducibility purposes, the code can be found on the following GitHub repository: https://github.com/vincentblot28/conformalized_gp.

4.2 Numerical Results

4.2.1 Illustrative Function with Misspecified GP

To illustrate the performance of the J+GP prediction intervals in the misspecified case, we create an artificial function with an oscillatory regime and a strong discontinuity within its domain. To do this, we define the following one-dimensional function for $x \in [-10, 10]$:

$$f(x) = \begin{cases} \sin(x), & \text{if} \quad x > 1 \\ -x, & \text{if} \quad x \leq 1 \end{cases}. \tag{37}$$

We use 10 points for training and 90 for estimation on a 100-point discretization of the interval. We use zero-mean prior on the mean and a squared-exponential kernel and optimize the kernel hyperparameters. The resulting GP and predictive intervals are illustrated in Fig. 3. In this case, the GP is intentionally designed to be misspecified, resulting in Bayesian credibility intervals for $x \leq 0$ that fail to capture the true values of the function. In contrast, the J+GP prediction intervals have a more conservative size and, due to their adaptiveness, successfully capture a larger portion of the true values of the function f. This adaptiveness is absent in the classical J+ prediction intervals, which remain nearly constant across all input regions, as can be seen in the right-hand side plot. This example clearly demonstrates that the J+GP prediction intervals are more robust in quantifying prediction uncertainty when the GP model priors are misspecified. Moreover, the adaptive sizing of these intervals makes them more informative in local regions with denser training points, as seen in the region where $x \geq 0$.

4.2.2 Performance of the Trained GPs

In Table 3, we present the different dataset sizes and the percentages used for training and testing. For the three different Matérn regularity parameters, the corresponding predictivity coefficients and mean squared errors are displayed.

4.2.3 Wing Weight Function

The wing weight function was proposed in Forrester et al. (2008). It is an analytic function with 10 independent input variables representing the design parameters of the wing and a scalar output representing the weight of the wing. This engineering model is representative of a Cessna C172 Skyhawk wing aircraft and is used for UQ benchmarks in the aerospace field. If we denote the inputs as $X = (X_1, \ldots, X_{10})$, the function is given by

$$g(X) = 0.036(X_1)^{0.758}(X_2)^{0.0035}\left(\frac{X_3}{\cos^2(X_4)}\right)^{0.6}(X_5)^{0.006} \times (X_6)^{0.04}\left(\frac{100 \times X_7}{\cos(X_4)}\right)^{-0.3}(X_8X_9)^{0.49} + X_1X_{10}. \tag{38}$$

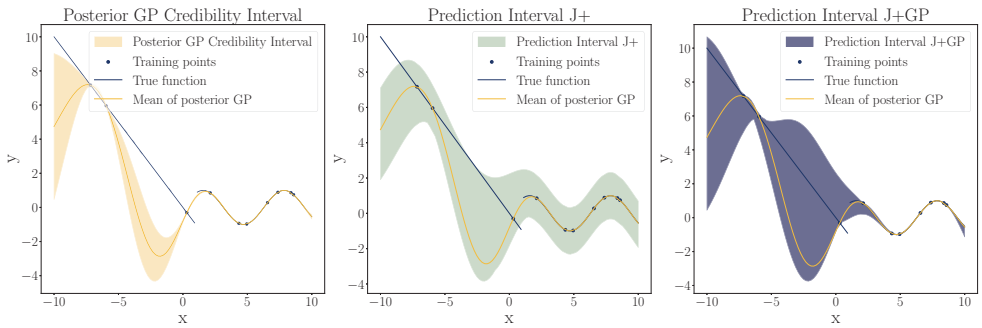

FIG. 3: Misspecified GP metamodel—with a squared-exponential kernel—of illustrative function f. The function f, as well as the training points and the resulting mean predictor, are plotted. On the left the posterior credibility intervals for $1 - \alpha = 90\%$ are plotted. In the middle the J+ prediction intervals for the same coverage level are plotted. On the right, for the same coverage level, the J+GP predictive intervals are obtained from our methodology.

TABLE 3: Summary of the performance metrics of the GP metamodels

ν		Wing Weight	Morokoff & Caflisch	TPD
—	d	7	10	7
	N	600	600	1000
	% train	75	75	80
	% test	25	25	20
1/2	Q^2	0.993	0.928	0.990
	MSE	16.32	2.19×10^{-3}	1.46
3/2	Q^2	0.998	0.940	0.996
	MSE	2.65	1.80×10^{-3}	0.54
5/2	Q^2	0.999	0.937	0.997
	MSE	0.82	1.89×10^{-3}	0.46

The response variable Y is obtained by $Y^{(i)} = g\big(X^{(i)}\big)$, where $X^{(i)} = (X_1^{(i)}, \ldots, X_{10}^{(i)})$ for all $i \in \{1, \ldots, N\}$, drawn using Monte Carlo sampling according to uniform probability distributions whose marginal supports are given in Table 4. The generated dataset consists of $N = 600$ realizations. We optimize the GP hyperparameters by MLE with 75% of the points and use the remaining samples to test our methodology.

In the results presented in both Fig. 4 and Table 5, the credibility intervals of the GP are larger than those of the conformal methods, thus providing empirical coverage higher than the desired one for all target coverage levels. The GP interval size as a function of the coverage level $1-\alpha$ is monotonically increasing since it is driven by the monotone normal quantile $u_{1-\alpha/2}$ in Eq. (12). Therefore, the ranking of the interval sizes does not change, and the Spearman correlation index remains constant for the GP intervals, as can be seen in the plot on the right side of Fig. 4. It is noteworthy that the J+GP conformal predictor in Fig. 4 achieves the smallest interval width and has, on average, the same Spearman correlations as the GP credibility intervals between its lengths and the metamodel approximation error. The J-minmax-GP achieves an even better average correlation of the interval width with the error, at the expense of being more conservative in size. In Table 5 we can see that the Matérn kernel with the regularity parameter $\nu = 5/2$ achieves the smallest width and the best Spearman correlation. More importantly, the sizes of the adaptive CP intervals are smaller than the GP credibility intervals and do not require any other hypothesis

TABLE 4: Supports of the uniform marginal input distributions for the wing weight function

Component	Domain	Component	Domain
X_1	$[150, 200]$	X_6	$[0.5, 1]$
X_2	$[220, 300]$	X_7	$[0.08, 0.18]$
X_3	$[6, 10]$	X_8	$[2.5, 6]$
X_4	$[-10, 10]$	X_9	$[1700, 2500]$
X_5	$[16, 45]$	X_{10}	$[0.025, 0.08]$

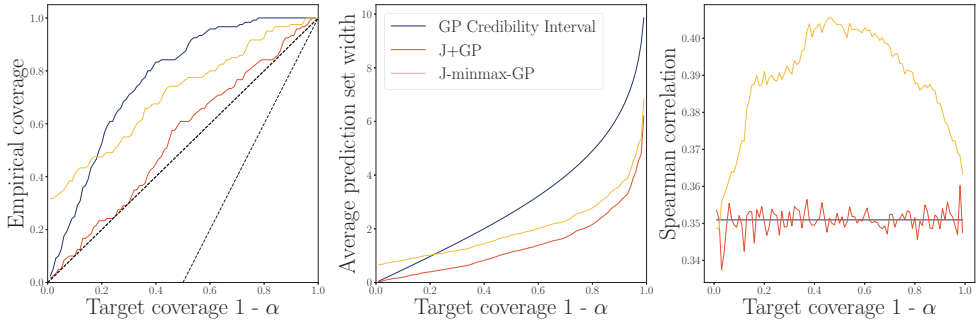

FIG. 4: GP metamodel of wing weight function with Matérn-5/2 kernel empirical coverage (with the $1-\alpha$ and $1-2\alpha$ marginal coverage in dotted lines), average width size, and Spearman correlation of the approximation error with the interval lengths for J+GP, J-minmax-GP, and GP credibility interval, as a function of the target coverage

TABLE 5: Wing weight analytical function. Empirical coverage rate, average width, and Spearman correlation for different predictive intervals (standard Bayesian credibility, cross-conformal, and the proposed estimator) for different Matérn kernels and for three confidence levels. In purple and underlined: the empirical coverage closest to the target coverage in absolute value. In red and bolded: lowest widths and highest Spearman correlations obtained under the target coverage condition

Method	Matérn	Coverage			Average Width			Spearman Corr.		
		90%	95%	99%	90%	95%	99%	90%	95%	99%
GP Credibility intervals	1/2	0.983	1.000	1.000	27.313	32.545	42.772	0.198	0.198	0.198
	3/2	1.000	1.000	1.000	11.319	13.487	17.725	0.326	0.326	0.326
	5/2	1.000	1.000	1.000	6.296	7.502	9.859	0.351	0.351	0.351
J+	1/2	0.917	0.958	0.992	12.922	18.227	32.359	–0.079	–0.092	–0.155
	3/2	0.925	0.975	0.992	5.660	8.523	14.084	0.127	–0.046	–0.243
	5/2	0.933	0.975	0.992	3.337	5.006	7.933	0.188	–0.172	–0.161
J-minmax	1/2	0.917	0.958	0.992	14.231	19.545	33.792	0.309	0.309	0.309
	3/2	0.942	0.983	1.000	6.563	9.463	15.101	0.349	0.349	0.349
	5/2	0.967	0.983	1.000	3.945	5.653	8.576	0.341	0.341	0.341
J+GP	1/2	0.917	0.950	0.992	11.962	16.731	28.458	0.198	0.188	0.205
	3/2	0.982	0.967	1.000	4.997	6.786	10.885	0.326	0.328	0.321
	5/2	0.933	0.967	1.000	**2.993**	**3.865**	**6.205**	0.349	0.352	0.348
J-minmax-GP	1/2	0.917	0.958	0.992	13.208	18.048	29.825	0.333	0.307	0.274
	3/2	0.942	0.983	1.000	5.881	7.685	11.746	0.361	0.346	0.339
	5/2	0.975	0.992	1.000	3.591	4.495	6.835	**0.382**	**0.372**	**0.363**

for interpretation. The good fit of the Matérn-5/2 could already be seen when inspecting the Q^2, but the study of the prediction intervals allows for a complementary uncertainty evaluation that does not require any further hypothesis for interpretation (as in the case for the interpretation

of the Bayesian credibility intervals). The proposed method also outperforms the standard J+, J-minmax cross-conformal approaches, as can be seen from the results in Appendix C, where the J+ interval lacks any adaptivity, and the J-minmax is too conservative.

4.2.4 Morokoff & Caflisch Function

The second example is the Morokoff & Caflisch function (Morokoff and Caflisch, 1995), defined on the unit hypercube $[0,1]^d$ by

$$g(X) = \frac{1}{2}\left(1+\frac{1}{d}\right)^d \prod_{i=1}^{d} (X_i)^{1/d}. \tag{39}$$

We choose $d = 10$ and use $N = 600$ samples drawn according to the multivariate normal distribution $\mathcal{N}(0, \boldsymbol{C})$ with the variance-covariance matrix $\boldsymbol{C}$ described in Acharki et al. (2023). We observe in the middle plot of Fig. 5 that the GP credibility intervals have a relatively small average width size at all target coverage levels. This size is closely followed by the J+GP estimator and becomes larger than the GP intervals after a certain coverage level threshold. However, it should be noted that the GP empirical coverage does not reach the high target coverage levels as seen in the left plot of Fig. 5. This indicates a possible misspecification of the metamodel since the empirical coverage here estimates the training-conditional Gaussian credibility interval in Eq. (12).

We observe a similar behavior for the Spearman correlation for the Morokoff & Caflisch function as for the wing weight function. Namely, the J+GP prediction interval has, on average, about the same size and error correlation properties as the GP credibility intervals. We still observe that the J-minmax-GP conformal predictor has more conservative interval sizes and stronger correlations. As can be seen in Appendix C, the adaptive J+GP and J-minmax-GP cross-conformal predictors have improved performance compared to their nonadaptive counterparts.

As can be seen in Table 3, the predictivity coefficient is high for all three regularity parameters, with only a small variation between $\nu = 3/2$ and $\nu = 5/2$. The proposed methodology is of particular interest here for completing the GP prior to kernel discrimination. In Table 6, under the target coverage levels 90%, 95%, and 99%, the Matérn-3/2 outperforms the Matérn-5/2. The J+ has the smallest width for the 99% target level, but its error correlation is very low.

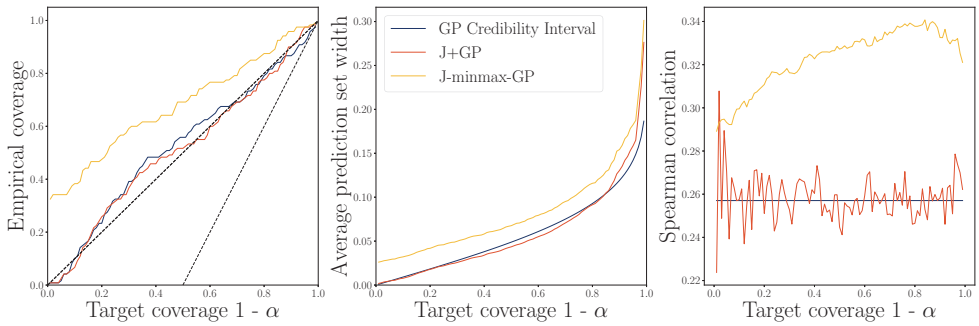

FIG. 5: GP metamodel of Morokoff & Caflisch function with Matérn-3/2 kernel empirical coverage (with the $1-\alpha$ and $1-2\alpha$ marginal coverage in dotted lines), average width size and Spearman correlation of the approximation error with the interval lengths for J+GP, J-minmax-GP and GP credibility interval, as a function of the target coverage.

TABLE 6: Morokoff & Caflisch analytical function. Empirical coverage rate, average width, and Spearman correlation for different predictive intervals (standard Bayesian credibility, cross-conformal, and the proposed estimator), for different Matérn kernels, and for three confidence levels. In purple and underlined: the empirical coverage closest to the target coverage in absolute value. In red and bolded: lowest widths and highest Spearman correlations obtained under the target coverage condition

Method	Matérn	Coverage			Average Width			Spearman Corr.		
		90%	95%	99%	90%	95%	99%	90%	95%	99%
GP Credibility intervals	1/2	0.925	0.958	0.983	0.161	0.192	0.252	0.258	0.258	0.258
	3/2	0.867	0.917	0.975	0.119	0.142	0.187	0.257	0.257	0.257
	5/2	0.842	0.900	0.950	0.108	0.129	0.170	0.241	0.241	0.241
J+	1/2	0.875	0.942	0.992	0.128	0.182	0.305	–0.215	0.012	0.096
	3/2	0.883	0.975	0.992	0.134	0.171	**0.288**	0.132	–0.155	–0.070
	5/2	0.892	0.942	0.983	0.136	0.175	0.282	0.084	0.199	0.001
J-minmax	1/2	0.900	0.950	0.992	0.144	0.197	0.321	0.132	0.132	0.132
	3/2	0.967	0.975	0.992	0.158	0.195	0.309	0.283	0.283	0.283
	5/2	0.942	0.975	0.992	0.164	0.201	0.310	0.238	0.238	0.238
J+GP	1/2	0.875	0.942	0.983	0.122	0.175	0.292	0.252	0.254	0.257
	3/2	0.875	0.958	0.983	0.125	**0.158**	0.271	0.254	0.264	0.259
	5/2	0.900	0.958	0.983	**0.132**	0.172	0.238	0.243	0.248	0.230
J-minmax-GP	1/2	0.883	0.958	0.992	0.137	0.190	0.309	0.253	0.261	0.272
	3/2	0.933	0.975	0.992	0.150	0.182	0.297	**0.317**	**0.318**	**0.305**
	5/2	0.942	0.983	0.983	0.160	0.202	0.265	0.285	0.276	0.273

4.2.5 Industrial Use Case: The THYC-Puffer-DEPOTHYC Code

The following industrial use case is related to the issue of clogging in steam generators of pressurized water nuclear reactors (Prusek, 2012). Over time, the steam generators of some reactors may face the challenge of clogging, a deposition phenomenon that increases the risk of mechanical and vibratory stresses on tube bundles and internal structures. It also affects their response during hypothetical accidental transients. To make maintenance planning more robust, EDF R&D has developed a multi-physics computational chain named "THYC-Puffer-DEPOTHYC" (TPD). This numerical tool uses specific physical models to reproduce the kinetics of clogging and to generate time-dependent clogging rate profiles for specific steam generators (Feng et al., 2023; Prusek, 2012). Some input parameters of this code are subject to uncertainties. To better understand the sensitivity of the output uncertainty with respect to the input variability, a metamodeling methodology based on polynomial chaos expansions and advanced global sensitivity techniques has recently been proposed in Jaber et al. (2025). Here, it is assumed that we dispose of a dataset of 10^3 Monte Carlo simulations, with $d = 7$ independent input variables to predict the clogging rate at a given time. The probability distributions of the inputs are listed in Table 7. More information about the physical nature of the variables can be found in Jaber et al. (2025).

TABLE 7: Distributions of the input variables of TPD

Component	Distribution	Component	Distribution
X_1	$\mathcal{N}(101.6, 4.0)$	X_5	$\mathcal{T}(0.5, 5.0, 10.0) \times 10^{-6}$
X_2	$\mathcal{N}(0.0233, 0.0005)$	X_6	$\mathcal{T}(1.0, 4.5, 8.0) \times 10^{-9}$
X_3	$\mathcal{T}(0.2, 0.3, 0.5)$	X_7	$\mathcal{T}(0.1, 7.8, 12) \times 10^{-4}$
X_4	$\mathcal{T}(0.01, 0.05, 0.3)$		

The results of the analysis are detailed in Table 8. The predictive capability of the posterior GP metamodel proves to be extremely high ($Q^2 \geq 0.99$) for all regularity parameters, making it again challenging to find the optimal candidate. To determine what leads to a robust GP metamodel of TPD to speed up industrial studies on clogging, the different conformal predictors reveal an advantage for a GP employing Matérn-3/2 and Matérn-5/2 prior kernels. As seen in Fig. 6, the GP credibility intervals show poor coverage rates above the target coverage threshold of ~ 0.8. The J+GP also shows poor empirical coverage above $1 - \alpha = 0.5$. This result is explained by Theorem 1 since coverage is only guaranteed above $1 - 2\alpha$, and it is

TABLE 8: THYC-Puffer-DEPOTHYC analytical function. Empirical coverage rate, average width, and Spearman correlation for different predictive intervals (standard Bayesian credibility, cross-conformal, and the proposed estimator), using various Matérn kernels and three confidence levels. In purple and underlined: the empirical coverage closest to the target coverage in absolute value. In red and bolded: lowest widths and highest Spearman correlations obtained under the target coverage condition

Method	Matérn	Coverage			Average Width			Spearman Corr.		
		90%	95%	99%	90%	95%	99%	90%	95%	99%
GP Credibility intervals	1/2	0.960	0.975	0.975	4.717	5.621	7.387	0.463	0.463	0.463
	3/2	0.915	0.940	0.950	2.000	2.384	3.133	0.353	0.353	0.353
	5/2	0.850	0.885	0.945	1.632	1.944	2.555	0.281	0.281	0.281
J+	1/2	0.855	0.900	0.975	2.438	3.610	7.391	0.266	–0.223	0.132
	3/2	0.840	0.905	0.975	1.529	2.031	3.943	–0.355	0.043	0.202
	5/2	0.840	0.920	0.965	1.353	1.836	3.109	–0.052	0.301	0.273
J-minmax	1/2	0.860	0.920	0.975	2.763	3.943	7.711	0.666	0.666	0.666
	3/2	0.890	0.920	0.980	1.857	2.350	4.260	0.653	0.653	0.653
	5/2	0.905	0.950	0.980	**1.763**	**2.233**	3.505	**0.606**	**0.606**	0.606
J+GP	1/2	0.845	0.895	0.975	2.314	3.198	6.367	0.469	0.466	0.458
	3/2	0.840	0.925	0.955	1.523	2.058	3.215	0.351	0.345	0.352
	5/2	0.845	0.905	0.970	1.509	2.072	3.689	0.279	0.280	0.288
J-minmax-GP	1/2	0.870	0.920	0.975	2.638	3.523	6.700	0.617	0.592	0.546
	3/2	0.900	0.950	0.985	1.852	2.387	3.543	0.519	0.489	0.449
	5/2	0.895	0.960	0.995	1.918	2.477	**4.080**	0.424	0.390	**0.349**

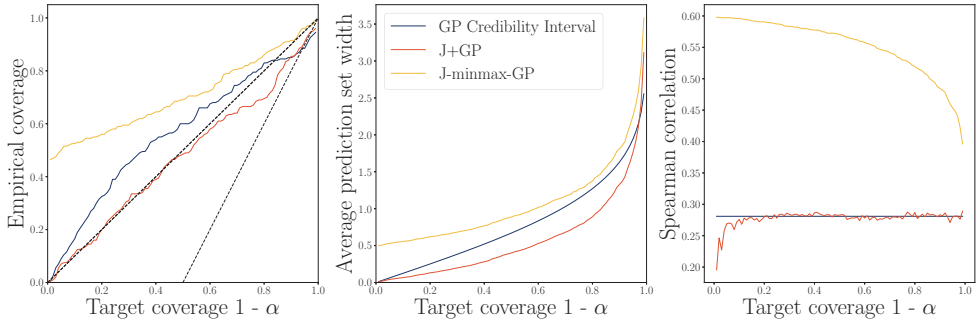

FIG. 6: GP metamodel of TPD dataset with Matérn-3/2 kernel empirical coverage, average width size and Spearman correlation of the approximation error with the interval lengths for J+GP, J-minmax-GP, and GP credibility interval, as a function of the target coverage

visible here on the TPD use case. These empirical coverage results are conditioned on the training dataset.

Therefore, to properly check the lower bound, it would be necessary to average all permutations of the train-test dataset to fully account for this result. Such scenarios where the coverage is strictly between $1-2\alpha$ and $1-\alpha$ are rather rare for standard J+ (as explained in Barber et al., 2021). However, the low empirical coverage rate observed for the credibility intervals may signal misspecification, and thus their interpretation may not be reliable for quantifying the metamodel uncertainty. Knowing this fact, the J+GP has smaller average widths and the same average correlation as the GP credibility intervals; these characteristics have already been observed in the previous examples. The J-minmax-GP also has stronger correlations at all coverage rates, but the average width is larger than both the credibility intervals and the J+GP. The correlation variation J-minmax has a different profile here, being regular and monotonically decreasing, whereas, in the previous examples, it was shaped as a parabola. This could be explained by the presence of more data ($N = 10^3$) and perhaps better regularity of the sample. In particular, the standard J-minmax estimator shows a remarkable degree of adaptivity, as seen in Table 8, especially for the coverage rates of 90% and 95%, surpassing the correlations obtained with the unreliable Bayesian credibility interval widths. Therefore, in terms of applications for speeding up industrial uncertainty studies of clogging, the GP metamodel with zero mean and Matérn-5/2 kernel optimized by MLE can be considered as the best candidate for metamodeling TPD.

4.3 Synthesis of the Results

We have shown that for a given target coverage, studying the average width of the prediction intervals and their Spearman correlation with the error improves the evaluation of the metamodel quality. This has been numerically exemplified on two different deterministic analytical UQ functions and on a complex industrial use case based on a real industrial computational chain, for which the selection of different metamodels, e.g., by different Matérn kernels, on the sole basis of the Q^2 is not fully conclusive. The proposed new cross-conformal J+GP estimator yields smaller widths on average while keeping the correlation of widths with the metamodel error close to that of the Bayesian credible intervals, and the J-minmax-GP achieves better correlations than the standard Bayesian credible intervals at the cost of larger intervals. We further show that inspection of our hypothesis-free CP intervals can help in choosing a more robust prior kernel for the GP metamodel of a computer code.

5. CONCLUSION AND PERSPECTIVES

In this work, we explore the idea of "conformalizing" metamodels based on GP regression in the cross-conformal prediction paradigm in order to make GP metamodel evaluation more robust for industrial applications. The idea is to make GP metamodels more reliable to improve prediction in the context of possible misspecification. To this end, we adapt the classical LOO nonconformity score by weighting it with the local GP posterior standard deviation. This method allows the CP intervals to have a better adaptivity, thus having a different interval span for different new test points. Moreover, the proposed J+GP prediction interval enjoys the same theoretical marginal coverage property as the Jackknife+ one and its min-max variant proposed by Barber et al. (2021). In order to quantify this adaptivity of the prediction interval, we evaluate the Spearman correlation between the width of the intervals and the absolute value of the metamodel approximation error. We show that our method achieves a better adaptivity than both standard cross-conformal prediction methods and GP credibility intervals.

We demonstrate the potential application of our methodology for GP model selection among different prior regularity parameters for the Matérn kernels. Furthermore, we show how the proposed methodology can help to evaluate the validity of a GP metamodel for industrial applications through a real use case related to nuclear engineering. A future line of research would be to generalize this methodology to families of deterministic metamodels, such as polynomial chaos expansions, which do not naturally come equipped with an inherent stochastic structure like GPs, or to more general statistical models that come with a quantifiable notion of dispersion. Moreover, recent work by Pion and Vazquez (2024) shows that our J+GP estimator generally outperforms the full-conformal approach and other variants in the setting of Gaussian interpolation (i.e., GP regression without the nugget effect).

ACKNOWLEDGMENTS

The PhD programs of the contributing authors are funded by the French National Association for Technological Research (ANRT) under Grants No. 2022/1412 and 2022/0667, respectively. Part of this work was supported by EDF and Quantmetry and initiated during the Summer Mathematical Research Center on Scientific Computing (whose French acronym is "CEMRACS"), which took place at CIRM in Marseille (France), from July 17 to August 25, 2023 (http://smai.emath.fr/cemracs/cemracs23/index.html). The authors from EDF R&D would like to thank Morgane Garo Sail and Charlotte Géry (project managers at EDF R&D) for their financial and organizational support through their projects.

REFERENCES

Acharki, N., Bertoncello, A., and Garnier, J., Robust Prediction Interval Estimation for Gaussian Processes by Cross-Validation Method, *Comput. Stat. Data Anal.*, vol. **178**, p. 107597, 2023.

Angelopoulos, A.N. and Bates, S., Conformal Prediction: A Gentle Introduction, *Found. Trends Mach. Learn.*, vol. **16**, no. 4, pp. 494–591, 2023.

Barber, R.F., Candès, E.J., Ramdas, A., and Tibshirani., R.J., Predictive Inference with the Jackknife+, *Annals Stat.*, vol. **49**, pp. 486–507, 2021.

Baudin, M., Dutfoy, A., Iooss, B., and Popelin, A., Open TURNS: An Industrial Software for Uncertainty Quantification in Simulation, *Handbook on Uncertainty Quantification*, R. Ghanem, D. Higdon, and H. Owhadi, Eds., Berlin: Springer, pp. 2001–2038, 2017.

Burnaev, E. and Nazarov, I., Conformalized Kernel Ridge Regression, *Proc. of the 15th IEEE Int. Conf. on Machine Learning and Applications*, Anaheim, CA, pp. 45–52, 2016.

Burnaev, E. and Vovk, V., Efficiency of Conformalized Ridge Regression, *Proc. Mach. Learn. Res.*, vol. **35**, pp. 605–622, 2014.

Cordier, T., Blot, V., Lacombe, L., Morzadec, T., Capitaine, A., and Brunel, N., Flexible and Systematic Uncertainty Estimation with Conformal Prediction via the MAPIE Library, *Proc. Mach. Learn. Res.*, vol. **204**, pp. 549–581, 2023.

Da Veiga, S., Tutorial on Conformal Prediction & Related Methods, *ETICS 2024 Research School*, accessed January 21, 2025, from https://sites.google.com/view/sebastien-da-veiga/etics-2024-tutorial-on-conformal-prediction-Ielated-methods, 2024.

De Carvalho, T.M., Van Rosmalen, J., Wolff, H.B., Koffijberg, H., and Coupé, V.M.H., Choosing a Metamodel of a Simulation Model for Uncertainty Quantification, *Med. Decision Making*, vol. **42**, no. 1, pp. 28–42, 2022.

De Rocquigny, E., Devictor, N., and Tarantola, S., Eds., *Uncertainty in Industrial Practice*, Hoboken, NJ: John Wiley & Sons, 2008.

Demay, C., Iooss, B., Le Gratiet, L., and Marrel, A., Model Selection Based on Validation Criteria for Gaussian Process Regression: An Application with Highlights on the Predictive Variance, *Qual. Reliab. Eng. Int.*, vol. **38**, no. 3, pp. 1482–1500, 2022.

Deutschmann, N., Rigotti, M., and Martinez, M.R., Adaptive Conformal Regression with Jackknife+ Rescaled Scores, arXiv Preprint arXiv:2305.19901, 2023.

Fang, K.T., Li, R., and Sudjianto, A., *Design and Modeling for Computer Experiments*, New York: CRC Press, 2006.

Fekhari, E., Iooss, B., Muré, J., Pronzato, L., and Rendas, J., Model Predictivity Assessment: Incremental Test-Set Selection and Accuracy Evaluation, *Studies in Theoretical and Applied Statistics*, N. Salvati, C. Perna, S. Marchetti, and R. Chambers, Eds., Berlin: Springer, pp. 315–347, 2023.

Feng, Q., Nebes, J., Bachet, M., Pujet, S., You, D., and Deri, E., Tube Support Plates Blockage of PWR Steam Generators: Thermal-Hydraulics and Chemical Modeling, *Proc. of the Int. Conf. on Nuclear Plant Chemistry*, Antibes, Juan-les-Pins, France, 2023.

Forrester, A.I.J., Sóbester, A., and Keane, A.J., *Engineering Design via Surrogate Modelling: A Practical Guide*, Chichester, UK: John Wiley & Sons, 2008.

Ghanem, R., Higdon, D., and Owhadi, H., Eds., *Handbook of Uncertainty Quantification*, Cham: Springer, 2017.

Gramacy, R.B., *Surrogates: Gaussian Process Modeling, Design, and Optimization for the Applied Sciences*, Boca Raton, FL: CRC Press, 2020.

Gu, M., Wang, X., and Berger, J.O., Robust Gaussian Stochastic Process Emulation, *Annals Stat.*, vol. **46**, no. 6A, pp. 3038–3066, 2018.

Jaber, E., Chabridon, V., Remy, E., Baudin, M., Lucor, D., Mougeot, M., and Iooss, B., Sensitivity Analyses of a Multi-Physics Long-Term Clogging Model for Steam Generators, *Int. J. Uncertainty Quant.*, vol. **15**, pp. 27–45, 2025.

Lei, J., Fast Exact Conformalization of the Lasso Using Piecewise Linear Homotopy, *Biometrika*, vol. **106**, no. 4, pp. 749–764, 2019.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R.J., and Wasserman, L., Distribution-Free Predictive Inference for Regression, *J. Am. Stat. Assoc.*, vol. **113**, pp. 1094–1111, 2021.

Liang, R. and Barber, R.F., Algorithmic Stability Implies Training-Conditional Coverage for Distribution-Free Prediction Methods, arXiv Preprint arXiv:2311.04295, 2023.

Mao, H., Martin, R., and Reich, B., Valid Model-Free Spatial Prediction, *J. Am. Stat. Assoc.*, vol. **119**, no. 546, pp. 904–914, 2024.

Marrel, A. and Iooss, B., Probabilistic Surrogate Modeling by Gaussian Process: A Review on Recent Insights in Estimation and Validation, *Reliab. Eng. Syst. Safety*, vol. **247**, p. 110094, 2024.

Marrel, A., Iooss, B., Van Dorpe, F., and Volkova, E., An Efficient Methodology for Modeling Complex Computer Codes with Gaussian Processes, *Comput. Stat. Data Anal.*, vol. **52**, pp. 4731–4744, 2008.

Morokoff, W.J. and Caflisch, R.E., Quasi-Monte Carlo Integration, *J. Comput. Phys.*, vol. **122**, no. 2, pp. 218–230, 1995.

Nouretdinov, I., Melluish, T., and Vovk, V., Ridge Regression Confidence Machine, *Proc. of the 18th Int. Conf. on Machine Learning*, Boca Raton, FL, pp. 385–392, 2001.

Papadopoulos, H., Guaranteed Coverage Prediction Intervals with Gaussian Process Regression, *IEEE Trans. Pattern Anal. Mach. Intel.*, vol. **46**, no. 12, pp. 9072–9083, 2024.

Papadopoulos, H., Gammerman, A., and Vovk, V., Normalized Nonconformity Measures for Regression Conformal Prediction, *Proc. of the 26th Int. Conf. on Artificial Intelligence and Applications*, Las Vegas, NV, pp. 64–69, 2008.

Papadopoulos, H., Proedrou, K., Vovk, V., and Gammerman, A., Inductive Confidence Machines for Regression, *Proc. of the 13th European Conference on Machine Learning*, Helsinki, Finland, pp. 345–356, 2002a.

Papadopoulos, H., Vovk, V., and Gammerman, A., Qualified Predictions for Large Data Sets in the Case of Pattern Recognition, *Proc. of the Int. Conf. on Machine Learning and Applications*, Las Vegas, NV, pp. 159–163, 2002b.

Papadopoulos, H., Vovk, V., and Gammerman, A., Regression Conformal Prediction with Nearest Neighbours, *J. Artif. Intell. Res.*, vol. **40**, pp. 815–840, 2011.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E., Scikit-Learn: Machine Learning in Python, *J. Mach. Learn. Res.*, vol. **12**, pp. 2825–2830, 2011.

Petit, S.J., Bect, J., Feliot, P., and Vazquez, E., Parameter Selection in Gaussian Process Interpolation: An Empirical Study of Selection Criteria, *SIAM/ASA J. Uncertainty Quant.*, vol. **11**, no. 4, pp. 1308–1328, 2023.

Pion, A. and Vazquez, E., Gaussian Process Interpolation with Conformal Prediction: Methods and Comparative Analysis, *Proc. of the 10th Int. Conf. on Machine Learning, Optimization, and Data Science*, Tuscany, Italy, pp. 1–13, 2024.

Prusek, T., Modélisation et Simulation Numérique du Colmatage à l'Échelle du Sous-Canal dans les Générateurs de Vapeur, PhD, Université Aix-Marseille, 2012.

Rasmussen, C.E. and Williams, C.K.I., *Gaussian Processes for Machine Learning*, Cambridge, MA: MIT Press, 2006.

Romano, Y., Patterson, E., and Candes, E., Conformalized Quantile Regression, *Proc. of the 33th Conf. on Neural Information Processing Systems (NeurIPS 2019)*, Vancouver, Canada, pp. 1–11, 2019.

Rubinstein, R.Y. and Kroese, D.P., *Simulation and the Monte Carlo Method*, 2nd ed., Hoboken, NJ: John Wiley & Sons, 2008.

Seedat, N., Jeffares, A., Imrie, F., and van der Schaar, M., Improving Adaptive Conformal Prediction Using Self-Supervised Learning, *Proc. of the 26th Int. Conf. on Artificial Intelligence and Statistics*, Valencia, Spain, pp. 10160–10177, 2023.

Stanton, S., Maddox, W., and Wilson, A.G., Bayesian Optimization with Conformal Prediction Sets, *Proc. of the 26th Int. Conf. on Artificial Intelligence and Statistics*, Valencia, Spain, pp. 959–986, 2023.

Sullivan, T.J., *Introduction to Uncertainty Quantification*, Berlin: Springer-Verlag, 2015.

Vovk, V., Cross-Conformal Predictors, *Annals Math. Artif. Intel.*, vol. **74**, nos. 1-2, pp. 9–28, 2015.

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

APPENDIX A. PROOF OF THEOREM 1

We prove a more general version of the theorem. We assume that we are in a regression setting, and we use a model $\widehat{g}$ that has an estimator of its standard deviation $\widehat{\sigma}(X)$. Moreover, we show that a slight modification of the scaled nonconformity score by taking powers $\beta > 0$ of the standard deviation does not change the main results [such powers are used in Papadopoulos (2024) in the full-conformal setting]. For GPs, the predictor is the posterior mean $\widehat{g} = \widetilde{g}$, and the estimated standard deviation is the posterior covariance $\widehat{\sigma} = \widetilde{\gamma}$.

Proof. Assume that

$$Y = g(X) + \epsilon, \tag{A.1}$$

with ϵ representing noise, and that a statistical learning model $\widehat{g}$ is trained on the dataset $\mathcal{D}_n = \{(X^{(i)}, Y^{(i)})\}_{i=1}^n$. Let $\left(X^{(n+1)}, Y^{(n+1)}\right) \in \mathcal{X} \times \mathcal{Y}$ be a new point. We denote by $\mathcal{D}_{n+1} := \mathcal{D} \cup \{(X^{(n+1)}, Y^{(n+1)})\}$. Let $\widehat{g}_{-(i,j)} \ \forall i \neq j \in \{1, ..., n+1\}$ be the statistical learning $\mathcal{D}_{n+1} \backslash \{(X^{(i)}, Y^{(i)}), (X^{(j)}, Y^{(j)})\}$. By exchangeability we have that $\widehat{g}_{-(i,j)} = \widehat{g}_{-(j,i)}$ and $\widehat{g}_{-i} = \widehat{g}_{-(i,n+1)}$. Let us denote by $\widehat{\sigma}\big(X^{(i)}\big)$ an estimator of the standard deviation of $\widehat{g}$ and assume without loss of generality that $\widehat{\sigma} > 0$ and similarly for the corresponding LOO (we could take the max function with a small $\delta > 0$ otherwise). Similarly, as for the Gaussian nonconformity score, we define

$$R_i^{\mathrm{LOO}\sigma} = \frac{\left|Y^{(i)} - \widehat{g}\big(X^{(i)}\big)\right|}{\widehat{\sigma}_{-i}^{\,\beta}\big(X^{(i)}\big)}. \tag{A.2}$$

We then proceed and define $R \in \mathcal{M}_{n+1}(\mathbb{R})$ as

$$R_{ij} = \begin{cases} +\infty & \text{if} \quad i = j, \\ \left|Y^{(i)} - \widehat{g}_{-(i,j)}\big(X^{(i)}\big)\right| \,/\, \widehat{\sigma}_{-(i,j)}^{\,\beta}\big(X^{(i)}\big) & \text{if} \quad i \neq j. \end{cases} \tag{A.3}$$

For simplifying the notations, we will now fix $\beta = 1$. We proceed in defining the matrix $A \in \mathcal{M}_{n+1}(\{0,1\})$:

$$A_{ij} = 1\{R_{ij} > R_{ji}\}. \tag{A.4}$$

It can be easily observed that $A_{ij} = 1 \Leftrightarrow A_{ji} = 0$ except when $j = i$. The strange set associated to A for $\alpha \in (0,1)$ is

$$\mathcal{S}(A) := \left\{ i \in \{1, \ldots, n+1\} \ : \ \sum_{j=1, j\neq i}^{n+1} A_{ij} \geq (1-\alpha)(n+1) \right\}. \tag{A.5}$$

In other words, a point i is *strange* if the residual R_{ij} compared with R_{ji} for all $j \neq i$ is larger for a given fraction of comparisons.

We start by bounding the cardinal of $\mathcal{S}(A)$. Let i be a strange point. $A_{ij} = 0$ for *at most* $\alpha(n+1)-1$ other strange points j since $A_{ij} = 1$ for at least $(1-\alpha)(n+1)$ and $i \neq j$. Let $s = |\mathcal{S}(A)|$; we now group pairs of strange points by $A_{ij} = 0$. For a chosen point i, there are at most s possibilities for the strange point j, and for each one, $A_{ij} = 0$ at most $\alpha(n+1)-1$ times. Thus there are at most $s \times (\alpha(n+1)-1)$ *pairs* of strange points.

We can now bound the number of ways we can choose two points in $\mathcal{S}(A)$, and we obtain

$$\frac{s(s-1)}{2} \leq s \times (\alpha(n+1)-1), \tag{A.6}$$

and rearranging:

$$s \leq 2\alpha(n+1). \tag{A.7}$$

By assumption, the dataset $\mathcal{D}_{n+1}$ is exchangeable. Thus, using permutation matrices Π, which maps a $j \in \{1,\ldots,n+1\}$ to $n+1$ (such that $\Pi_{j,n+1} = 1$), we prove that

$$\mathbb{P}(n+1 \in \mathcal{S}(A)) = \mathbb{P}\big(j \in \mathcal{S}\big(\Pi A \Pi^\top\big)\big) = \mathbb{P}(j \in \mathcal{S}(A)). \tag{A.8}$$

Therefore, any point is equally likely to be strange. We have, then:

$$\mathbb{P}(n+1 \in \mathcal{S}(A)) = \frac{1}{n+1}\sum_{j=1}^{n+1}\mathbb{P}(j \in \mathcal{S}(A)) = \frac{\mathbb{E}[|\mathcal{S}(A)|]}{n+1} \leq 2\alpha. \tag{A.9}$$

We can now reconnect with the definition of prediction intervals. We denote the generic version of our proposed J+GP prediction interval as

$$\widehat{C}^*_{n,\alpha}\big(X^{(n+1)}\big) = \left[\widehat{q}^{\pm}_{n,\alpha}\Big\{\widehat{g}_{-i}\big(X^{(n+1)}\big) \pm R_i^{\mathrm{LOO}\sigma} \times \widehat{\sigma}_{-i}\big(X^{(n+1)}\big)\Big\}\right]. \tag{A.10}$$

Let us suppose that $Y^{(n+1)} \notin \widehat{C}^*_{n,\alpha}$. Then, for at least $(1-\alpha)(n+1)$ values i in $\{1,\ldots,n+1\}$, we have

$$Y^{(n+1)} > \widehat{g}_{-i}\big(X^{(n+1)}\big) + R_i^{\mathrm{LOO}\sigma} \times \widehat{\sigma}_{-i}\big(X^{(n+1)}\big), \tag{A.11}$$

or

$$Y^{(n+1)} < \widehat{g}_{-i}\big(X^{(n+1)}\big) - R_i^{\mathrm{LOO}\sigma} \times \widehat{\sigma}_{-i}\big(X^{(n+1)}\big). \tag{A.12}$$

Finally, we can compute:

$$\begin{aligned}
(1-\alpha)(n+1) &\leq \sum_{i=1}^{n+1} 1\Big\{Y^{(n+1)} \notin \widehat{g}_{-i}\big(X^{(n+1)}\big) \pm R_i^{\mathrm{LOO}\sigma} \times \widehat{\sigma}_{-i}\big(X^{(n+1)}\big)\Big\} \\
&= \sum_{i=1}^{n+1} 1\Big\{R_i^{\mathrm{LOO}\sigma} \times \widehat{\sigma}_{-i}\big(X^{(n+1)}\big) < |Y^{(n+1)} - \widehat{g}_{-i}\big(X^{(n+1)}\big)|\Big\} \\
&= \sum_{i=1}^{n+1} 1\left\{R_i^{\mathrm{LOO}\sigma} < \frac{\big|Y^{(n+1)} - \widehat{g}_{-i}\big(X^{(n+1)}\big)\big|}{\widehat{\sigma}_{-i}\big(X^{(n+1)}\big)}\right\} \\
&= \sum_{i=1}^{n+1} 1\left\{\frac{\big|Y^{(i)} - \widehat{g}_{-i}\big(X^{(i)}\big)\big|}{\widehat{\sigma}_{-i}\big(X^{(i)}\big)} < \frac{\big|Y^{(n+1)} - \widehat{g}_{-i}\big(X^{(n+1)}\big)\big|}{\widehat{\sigma}_{-i}\big(X^{(n+1)}\big)}\right\} \\
&= \sum_{i=1}^{n+1} 1\{R_{i,n+1} < R_{n+1,i}\} = \sum_{i=1}^{n+1} A_{n+1,i},
\end{aligned}$$

where the last equality above is obtained with the identities $\widehat{\sigma}_{-i}\big(X^{(i)}\big) = \widehat{\sigma}_{-(i,n+1)}\big(X^{(i)}\big)$ and $\widehat{g}_{-i}\big(X^{(i)}\big) = \widehat{g}_{-(i,n+1)}\big(X^{(i)}\big)$. Therefore $n+1 \in \mathcal{S}(A)$ and:

$$\mathbb{P}\Big(g\big(X^{(n+1)}\big) \notin \widehat{C}^{*}_{n,\alpha}\big(X^{(n+1)}\big)\Big) \leq \mathbb{P}(n+1 \in \mathcal{S}(A)) \leq 2\alpha. \tag{A.13}$$

□

APPENDIX B. PROOF OF THEOREM 2

Proof. Assume the same hypothesis as in the previous theorem, and we make use of the same definitions and notations. We define the matrix $\widetilde{R} \in \mathcal{M}_{n+1}(\mathbb{R})$ as

$$\widetilde{R}_{ij} = \begin{cases} +\infty & \text{if} \quad i = j, \\ R_{ij} \times \widehat{\sigma}_{-(i,j)}\big(X^{(n+1)}\big) & \text{if} \quad i \neq j. \end{cases} \tag{B.1}$$

We redefine the matrix $A \in \mathcal{M}_{n+1}(\{0,1\})$:

$$A_{ij} = 1\{\min_{j'} \widetilde{R}_{ij'} \geq \widetilde{R}_{ji}\}, \tag{B.2}$$

where $\min_{j'} \widetilde{R}_{ij'}$ is the smallest residual for the point i when leaving out any point $j' \in \{1,\ldots,n\}$. We start by bounding the number of strange points, choose:

$$i_* \in \underset{i \in \mathcal{S}(A)}{\arg\min}\Big\{\min_{j'} \widetilde{R}_{ij'}\Big\}. \tag{B.3}$$

We can observe that for all strange points $j \in \mathcal{S}(A)$, the matrix element $A_{i_* j}$ is null. Indeed, this is because by definition,

$$\forall j \in \mathcal{S}(A),\ \widetilde{R}_{ji_*} \geq \min_{j'} \widetilde{R}_{jj'} \geq \min_{j'} \widetilde{R}_{i_* j'}. \tag{B.4}$$

We can then easily bound the number of strange points using $i_* \in \mathcal{S}(A)$:

$$n+1-|\mathcal{S}(A)| \geq \sum_{j=1}^{n+1} A_{i_* j} \geq (1-\alpha)(n+1), \tag{B.5}$$

and a rearrangement gives

$$|\mathcal{S}(A)| \leq \alpha(n+1). \tag{B.6}$$

Using the exchangeability property in the same fashion as the preceding proof, we have

$$\mathbb{P}\left(n+1 \in \mathcal{S}(A)\right) \leq \alpha. \tag{B.7}$$

Let us suppose now that $Y^{(n+1)} \notin \widehat{C}^{*-\text{minmax}}_{n,\alpha}$. Then, for at least $(1-\alpha)(n+1)$ values i in $\{1,\ldots,n+1\}$, we have

$$Y^{(n+1)} > \max_{i=1,\ldots,n} \widehat{g}_{-i}\big(X^{(n+1)}\big) + R_i^{\text{LOO}\sigma} \times \widehat{\sigma}_{-i}\big(X^{(n+1)}\big), \tag{B.8}$$

or

$$Y^{(n+1)} < \min_{i=1,\ldots,n} \widehat{g}_{-i}\big(X^{(n+1)}\big) - R_i^{\text{LOO}\sigma} \times \widehat{\sigma}_{-i}\big(X^{(n+1)}\big). \tag{B.9}$$

We denote $\widehat{g}_{\min}\big(X^{(i)}\big) := \min_{j=1\ldots,n}\widehat{g}_{-j}\big(X^{(i)}\big)$, and similarly for $\widehat{g}_{\max}$ and $\widetilde{R}_i\big(X^{(n+1)}\big) := R_i^{\mathrm{LOO}\sigma} \times \widehat{\sigma}_{-i}\big(X^{(n+1)}\big)$. Finally, we can compute:

$$
\begin{aligned}
&(1-\alpha)(n+1)\\
&\leq \sum_{i=1}^{n+1} 1\Big\{ Y^{(n+1)} \notin \Big[\widehat{g}_{\min}\big(X^{(n+1)}\big) - \widetilde{R}_i\big(X^{(n+1)}\big),\ \widehat{g}_{\max}\big(X^{(n+1)}\big) + \widetilde{R}_i\big(X^{(n+1)}\big)\Big]\Big\}\\
&= \sum_{i=1}^{n+1} 1\Big\{ \min_{j=1,\ldots,n}\big|Y^{(n+1)} - \widehat{g}_{-j}\big(X^{(n+1)}\big)\big| \geq R_i^{\mathrm{LOO}\sigma} \times \widehat{\sigma}_{-i}\big(X^{(n+1)}\big)\Big\}\\
&= \sum_{i=1}^{n+1} 1\left\{ \min_{j=1,\ldots,n}\frac{\big|Y^{(n+1)} - \widehat{g}_{-j}\big(X^{(n+1)}\big)\big|}{\widehat{\sigma}_{-j}\big(X^{(n+1)}\big)} \times \widehat{\sigma}_{-j}\big(X^{(n+1)}\big) \geq R_i^{\mathrm{LOO}\sigma} \times \widehat{\sigma}_{-i}\big(X^{(n+1)}\big)\right\}\\
&= \sum_{i=1}^{n+1} 1\left\{ \min_{j=1,\ldots,n}\frac{\big|Y^{(n+1)} - \widehat{g}_{-(n+1,j)}\big(X^{(n+1)}\big)\big|}{\widehat{\sigma}_{-(n+1,j)}\big(X^{(n+1)}\big)} \times \widehat{\sigma}_{-(n+1,j)}\big(X^{(n+1)}\big)\right.\\
&\qquad\qquad \left. \geq \frac{\big|Y^{(i)} - \widehat{g}_{-(i,n+1)}\big(X^{(i)}\big)\big|}{\widehat{\sigma}_{-(i,n+1)}\big(X^{(i)}\big)} \times \widehat{\sigma}_{-(i,n+1)}\big(X^{(n+1)}\big)\right\}\\
&= \sum_{i=1}^{n+1} 1\Big\{ \min_{j=1,\ldots,n} R_{n+1,j} \times \widehat{\sigma}_{-(n+1,j)}\big(X^{(n+1)}\big) \geq R_{i,n+1} \times \widehat{\sigma}_{-(i,n+1)}\big(X^{(n+1)}\big)\Big\}\\
&= \sum_{i=1}^{n+1} 1\Big\{ \min_{j=1,\ldots,n} \widetilde{R}_{n+1,j} \geq \widetilde{R}_{i,n+1}\Big\}\\
&= \sum_{i=1}^{n+1} A_{n+1,i}.
\end{aligned}
$$

Therefore $n+1 \in \mathcal{S}(A)$, and we conclude as in the preceding theorem. □

APPENDIX C. ADDITIONAL RESULTS

APPENDIX C.1 Branin Function

The Branin or Branin–Hoo function is a two-dimensional scalar function defined as

$$f(X_1, X_2) = a\big(X_2 - bX_1^2 + cX_1 - r\big)^2 + s(1-t)\cos(X_1) + s, \tag{C.1}$$

where the parameters are chosen as $a = 1$, $b = 5.1/(4\pi^2)$, $c = 5/\pi$, $r = 6$, $s = 10$, and $t = 1/(8\pi)$.

This function is effectively learned by Matérn GPs, as evidenced by the very high Q^2 values in Table C1. Despite the excellent predictivity coefficient, it is important to note that the Bayesian credibility intervals are quite large, resulting in a high level of empirical coverage. The J+GP method, on the other hand, provides significantly improved UQ through narrower interval

TABLE C1: Summary of the performance metrics of additional GP metamodels

ν		**Branin**	**Hartmann-3D**
—	d	2	3
	N	1000	1000
	% train	80	80
	% test	20	20
1/2	Q^2	0.999	0.918
	MSE	6×10^{-1}	3×10^{-3}
3/2	Q^2	0.999	0.930
	MSE	2×10^{-3}	2×10^{-3}
5/2	Q^2	0.999	0.932
	MSE	2×10^{-5}	2×10^{-3}

widths. As shown in Table C2, the Matérn-5/2 kernel achieves the smallest average interval width for critical coverage rates. In addition, the metamodel residuals show a strong correlation with both the GP credibility intervals and the interval width. Notably, the regular J-minmax method also shows a high level of correlation, further validating the quality of the Matérn-5/2 prior choice. In the middle plot of Fig. C1, it is evident that the J+GP average interval width is significantly smaller compared to the GP credibility intervals.

APPENDIX C.2 Hartmann-3D Function

The Hartmann-3D function is defined as

$$f(X_1, X_2, X_3) = -\sum_{i=1}^{4} \alpha_i \exp\left(-\sum_{j=1}^{3} A_{ij}(X_j - P_{ij})^2\right), \tag{C.2}$$

where

$$\alpha = \begin{bmatrix} 1.0 \\ 1.2 \\ 3.0 \\ 3.2 \end{bmatrix}, \quad A = \begin{bmatrix} 3.0 & 10.0 & 30.0 \\ 0.1 & 10.0 & 35.0 \\ 3.0 & 10.0 & 30.0 \\ 0.1 & 10.0 & 35.0 \end{bmatrix}, \quad P = 10^{-4} \begin{bmatrix} 3689 & 1170 & 2673 \\ 4699 & 4387 & 7470 \\ 1091 & 8732 & 5547 \\ 381 & 5743 & 8828 \end{bmatrix}. \tag{C.3}$$

The function is effectively learned by a GP, as evidenced by the high Q^2 values in Table C1. The Matérn-1/2 kernel shows the strongest correlations, as highlighted by the regular J-minmax interval in Table C3, underscoring the adaptiveness of this cross-conformal method. Similar conclusions hold for the Hartmann-3D GP metamodel: Bayesian credibility intervals are overly optimistic, with very large widths and excessively high coverage rates. As shown in Fig. C2, the conformalized GP prediction intervals (J+GP and J-minmax-GP) achieve much narrower widths compared to their Bayesian credibility counterparts.

APPENDIX C.3 Comparison with Standard CP Intervals

TABLE C2: Branin analytical function. Empirical coverage rate, average width, and Spearman correlation for different predictive intervals (standard Bayesian credibility, cross-conformal, and the proposed estimator), for different Matérn kernels, and for three confidence levels. In purple and underlined: the empirical coverage closest to the target coverage in absolute value. In red and bolded: lowest widths and highest Spearman correlations obtained under the target coverage condition

Method	**Matérn**	**Coverage**			**Average Width**			**Spearman Corr.**		
		90%	95%	99%	90%	95%	99%	90%	95%	99%
GP Credibility intervals	1/2	0.995	0.995	1.000	7.992	9.523	12.515	0.607	0.607	0.607
	3/2	1.000	1.000	1.000	0.418	0.498	0.655	0.568	0.568	0.568
	5/2	1.000	1.000	1.000	0.039	0.046	0.061	0.505	0.505	0.505
J+	1/2	0.915	0.965	0.990	0.875	1.873	7.466	0.118	–0.086	0.089
	3/2	0.935	0.975	0.990	0.039	0.103	0.354	0.230	0.309	0.033
	5/2	0.955	0.975	0.990	0.006	0.013	0.034	**0.890**	0.450	0.469
J-minmax	1/2	0.930	0.965	0.990	1.049	2.048	7.639	0.737	0.737	0.737
	3/2	0.970	0.980	0.990	0.053	0.116	0.368	0.855	0.855	0.855
	5/2	0.980	0.985	0.990	0.010	0.017	0.039	0.878	**0.878**	**0.878**
J+GP	1/2	0.920	0.955	0.990	0.797	1.666	6.554	0.606	0.605	0.603
	3/2	0.935	0.975	0.995	0.037	0.083	0.297	0.577	0.569	0.569
	5/2	0.955	0.970	1.000	**0.005**	**0.009**	**0.024**	0.579	0.570	0.537
J-minmax-GP	1/2	0.935	0.965	0.990	0.971	1.840	6.742	0.816	0.777	0.678
	3/2	0.975	0.980	0.995	0.051	0.097	0.311	0.759	0.703	0.627
	5/2	0.985	0.985	1.000	0.010	0.013	0.029	0.745	0.704	0.621

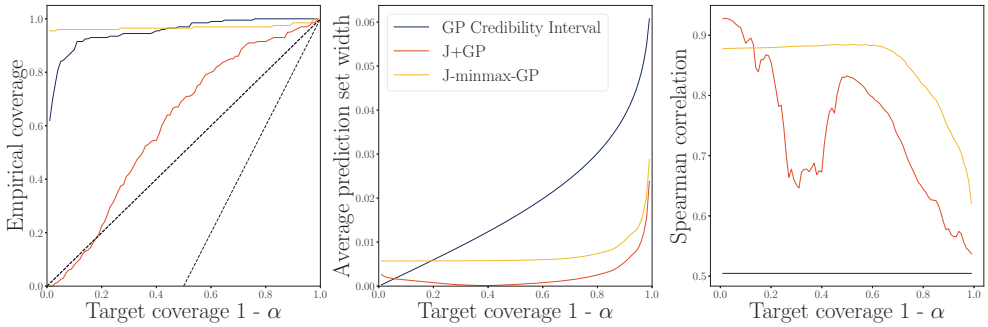

FIG. C1: GP metamodel of Branin function with Matérn-5/2 kernel empirical coverage (with the $1-\alpha$ and $1-2\alpha$ marginal coverage in dotted lines), average width size, and Spearman correlation of the approximation error with the interval lengths for J+GP, J-minmax-GP, and GP credibility interval, as a function of the target coverage.

TABLE C3: Hartmann-3 analytical function. Empirical coverage rate, average width, and Spearman correlation for different predictive intervals (standard Bayesian credibility, cross-conformal, and the proposed estimator), for different Matérn kernels, and for three confidence levels. In purple and underlined: the empirical coverage closest to the target coverage in absolute value. In red and bolded: lowest widths and highest Spearman correlations obtained under the target coverage condition

Method	Matérn	Coverage			Average Width			Spearman Corr.		
		90%	95%	99%	90%	95%	99%	90%	95%	99%
GP Credibility intervals	1/2	0.995	1.000	1.000	0.252	0.300	0.394	0.437	0.437	0.437
	3/2	0.995	1.000	1.000	0.048	0.057	0.075	0.459	0.459	0.459
	5/2	0.995	1.000	1.000	0.013	0.016	0.021	0.446	0.446	0.446
J+	1/2	0.930	0.975	0.990	0.073	0.120	0.221	–0.245	–0.138	0.133
	3/2	0.925	0.985	0.995	0.012	0.022	0.062	–0.194	0.243	–0.048
	5/2	0.915	0.970	0.995	0.004	0.007	0.016	0.072	0.141	–0.036
J-minmax	1/2	0.935	0.975	0.990	0.083	0.130	0.231	**0.714**	**0.714**	**0.714**
	3/2	0.950	0.990	0.995	0.015	0.025	0.066	0.709	0.709	0.709
	5/2	0.970	0.995	0.995	0.005	0.008	0.018	0.694	0.694	0.694
J+GP	1/2	0.920	0.970	0.990	0.068	0.105	0.193	0.438	0.431	0.438
	3/2	0.915	0.965	0.995	0.009	0.016	0.036	0.466	0.459	0.466
	5/2	0.915	0.965	0.990	**0.003**	**0.005**	**0.009**	0.450	0.448	0.453
J-minmax-GP	1/2	0.940	0.975	0.990	0.078	0.115	0.203	0.711	0.686	0.624
	3/2	0.945	0.985	0.995	0.012	0.019	0.039	0.709	0.663	0.588
	5/2	0.975	0.985	0.995	0.004	0.006	0.010	0.642	0.598	0.549

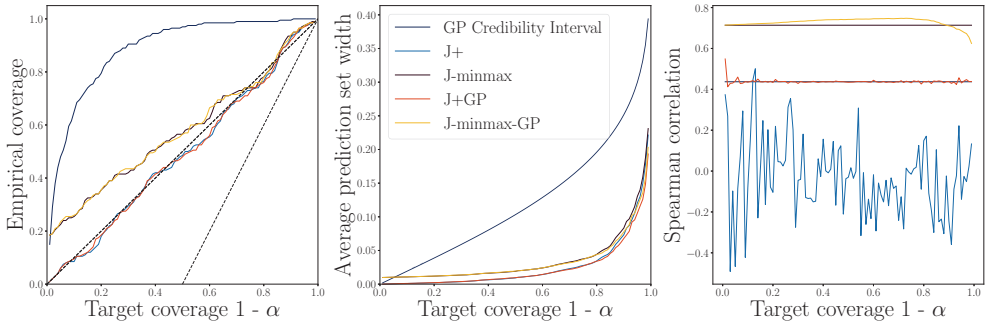

FIG. C2: GP metamodel of Hartmann-3D function with Matérn-1/2 kernel empirical coverage (with the $1-\alpha$ and $1-2\alpha$ marginal coverage in dotted lines), average width size, and Spearman correlation of the approximation error with the interval lengths for J+GP, J-minmax-GP, and GP credibility interval, as a function of the target coverage

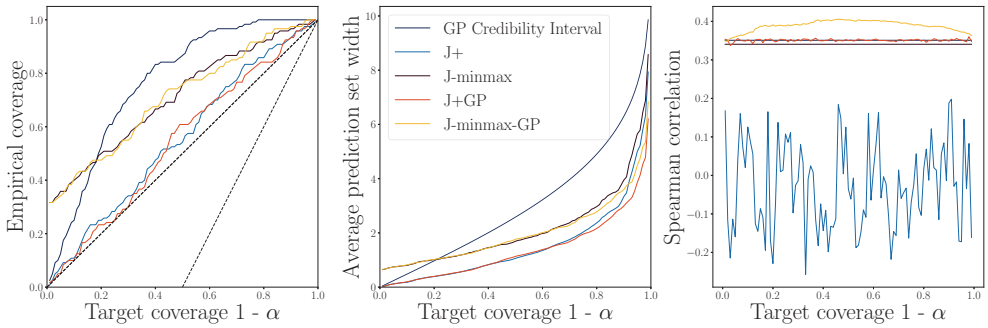

FIG. C3: GP metamodel of wing weight function with Matérn-5/2 kernel empirical coverage, average width size, and Spearman correlation of the approximation error with the interval lengths for J+GP, J-minmax-GP, GP credibility interval, and standard cross-conformal J+, J-minmax as a function of the target coverage

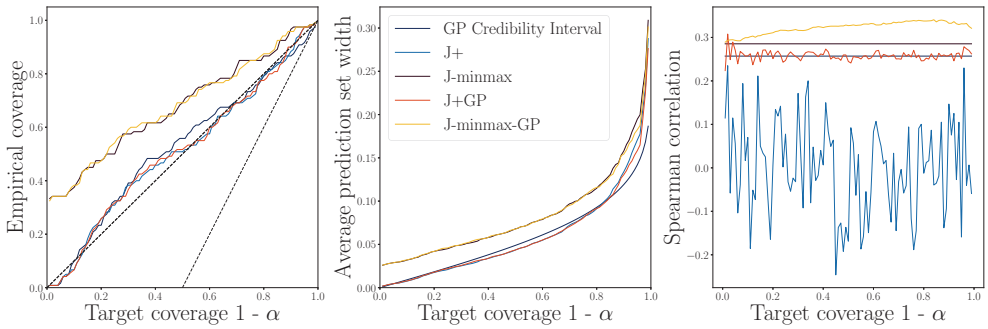

FIG. C4: GP metamodel of Morokoff & Caflisch function with Matérn-3/2 kernel empirical coverage, average width size, and Spearman correlation of the approximation error with the interval lengths for J+GP, J-minmax-GP, GP credibility interval, and standard cross-conformal J+, J-minmax as a function of the target coverage

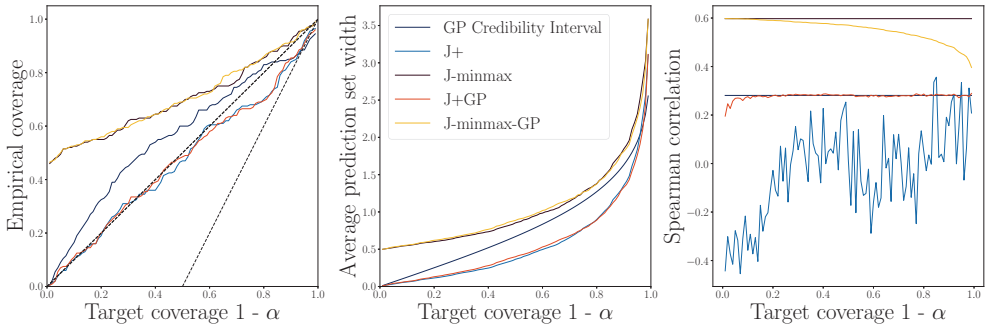

FIG. C5: GP metamodel of TPD computer code output with Matérn-5/2 kernel empirical coverage, average width size, and Spearman correlation of the approximation error with the interval lengths for J+GP, J-minmax-GP, GP credibility interval, and standard cross-conformal J+, J-minmax as a function of the target coverage