Probabilistic surrogate modeling by Gaussian process: A review on recent insights in estimation and validation

Amandine Marrel[a,c], Bertrand Iooss[b,c,d]

[a] *CEA, DES, IRESNE, DER, Cadarache, 13108 Saint-Paul-Lez-Durance, France*
[b] *EDF R&D, 6 quai Watier, 78400 Chatou, France*
[c] *Institut de Mathématiques de Toulouse, Toulouse, France*
[d] *Corresponding Author*

Abstract

In the framework of risk assessment, computer codes are increasingly used to understand, model and predict physical phenomena. As these codes can be very time-consuming to run, which severely limit the number of possible simulations, a widely accepted approach consists in approximating the CPU-time expensive computer model by a so-called "surrogate model". In this context, the Gaussian Process regression (also called kriging) is one of the most popular technique. It offers the advantage of providing a predictive distribution for all new evaluation points. An uncertainty associated with any quantity of interest (e.g., a probability of failure in reliability studies) to be estimated can thus be deduced and adaptive strategies for choosing new points to run with respect to this quantity can be developed. This paper focuses on the estimation of the Gaussian process covariance parameters by reviewing recent works on the analysis of the advantages and disadvantages of usual estimation methods, the most relevant validation criteria (for detecting poor estimation) and recent robust and corrective methods.

Keywords:
Computer experiments, Kriging, Machine learning, Metamodel, Uncertainty, Validation criteria

1. Introduction

In the framework of risk assessment, computer codes (or numerical simulators) are developed and increasingly used to understand, model and predict physical, engineering or biological phenomena [1]. They usually take a large number of input parameters driving the phenomenon of interest or related to its physical and numerical modeling. However, the available information about some of these parameters is often limited or uncertain. The uncertainties come mainly from the lack of knowledge about the underlying physics, the characterization of the input parameters of the model (e.g., due to the lack of experimental data) or to the choice of scenario parameters [2].Therefore, it is essential to take the uncertainties tainting the results of computer simulations into account in order to perform "Uncertainty Quantification" [3].

A probabilistic framework where the input uncertainties are modeled by fully or partially known probability distributions, based on available data, expert opinions or bibliographic databases is usually considered [2]. The uncertainty quantification process therefore relies on Monte Carlo techniques: a sample of code simulations is performed, where the inputs are drawn according to their probabilistic distributions. Estimators of the target statistical quantities, also called statistical quantities of interest (e.g., the variance, a probability of exceeding a threshold, or some quantiles) are then computed from the sample of code outputs. Depending on the nature of the quantities to be estimated and the expected confidence in the estimators, a very large number of simulations of the code can be necessary: from a few hundred to several tens of thousands for example. This number can also depends on the dimension of the inputs, when performing sensitivity analysis for instance [4]. There are numerous examples of sensitivity analysis and uncertainty quantification based on probabilistic approaches, particularly in the case of risk assessment using environmental models (see, e.g., [5, 6, 7]).

In this context, one key issue is that the numerical model under study can be very time-consuming to run, which can drastically limit the number of possible simulations. To solve this cost issue, a widely accepted approach consists in approximating the CPU-time expensive computer model by a CPU-time inexpensive mathematical function called "surrogate model" (or "metamodel", term that is used in the following). These metamodels can be based on polynomials, splines, random forests, neural networks, etc. [8, 9, 10], in fact on any machine learning techniques [11]. Built from a set of computer code simulations, they must be as representative as possible of the code outputs in the domain of variation of the uncertain parameters while having good prediction capabilities. Nowadays, metamodels are extensively used in several engineering fields to solve industrial issues as it provides a multi-purpose tool [12]: once fitted, the metamodel can be used, possibly in conjunction with the costly computer code, to perform sensitivity analysis, as well as uncertainty propagation, optimization, or calibration studies. Such techniques have been extensively developed for instance in nuclear engineering (see, e.g., [13, 14, 15, 16]). However, to be confident with this approximation-based approach in support of the different uncertainty quantification tasks, it is crucial to develop accurate and reliable metamodels to approximate the computer model.

Among the metamodels classically used for numerical experiments, the Gaussian Process (GP) regression, also called kriging model, is a popular tool for non-parametric function estimation. Historically introduced in the context of geostatistics for spatial interpolation (see, e.g., [17]), GP regression has been extended to interpolation of numerical simulation outputs [18, 19] and machine learning approximation (see, e.g., [20]). Its intuitive idea is to start from a prior over random functions (a GP is characterized by its mean and covariance functions), then the GP regression yields a posterior over functions given the observed data. On one hand, this makes it a very flexible non-parametric regression tool, suitable for modeling of numerical simulators and whose effectiveness has been illustrated in many applications [21]. On the other hand, it also offers a probabilistic framework: the GP metamodel yields a predictive distribution for the code output at each prediction point, with a simple analytic formulation. From this, a prediction but also an uncertainty via prediction intervals can be analytically derived.

Figure 1 illustrates this principle of using a (probabilistic) metamodel, by the way of the GP, to emulate computer code from a set of code simulations. In particular, its associated prediction intervals allow to develop so-called adaptive (also called "active learning" or "goal-oriented") strategies: the idea is to sequentially find, from the current set of simulations, a new set of points to run in the input space in order to most efficiently estimate a statistical quantity of interest (see the work of Jones et al. [22] which provides the initial idea for optimization purposes). This major advantage of GP is very appealing for risk and safety assessment applications because it has been shown to drastically increase convergence with respect to standard Monte Carlo or quasi-Monte Carlo algorithms (see Fuhg et al. [23], Moustapha et al. [24] for overviews and benchmarks of GP-based adaptive algorithms). Among the numerous algorithms that have been recently developed, one can distinguish those based on pointwise criteria, which use uniquely the conditional mean and variance of the GP at a given point (see, e.g., [25]) and those based on integral criteria, which integrate functions of the conditional mean and variance of the GP over the whole input domain (see, e.g., [26]). Among the different topics that use such algorithms, all recently published in the *Reliability Engineering and System Safety* journal, one can cite prediction-forecast [27, 28], calibration [29], functional risk curves [30], structural reliability [31, 32, 10, 33], reliability-based design optimization [34, 35] and robust optimization [36].

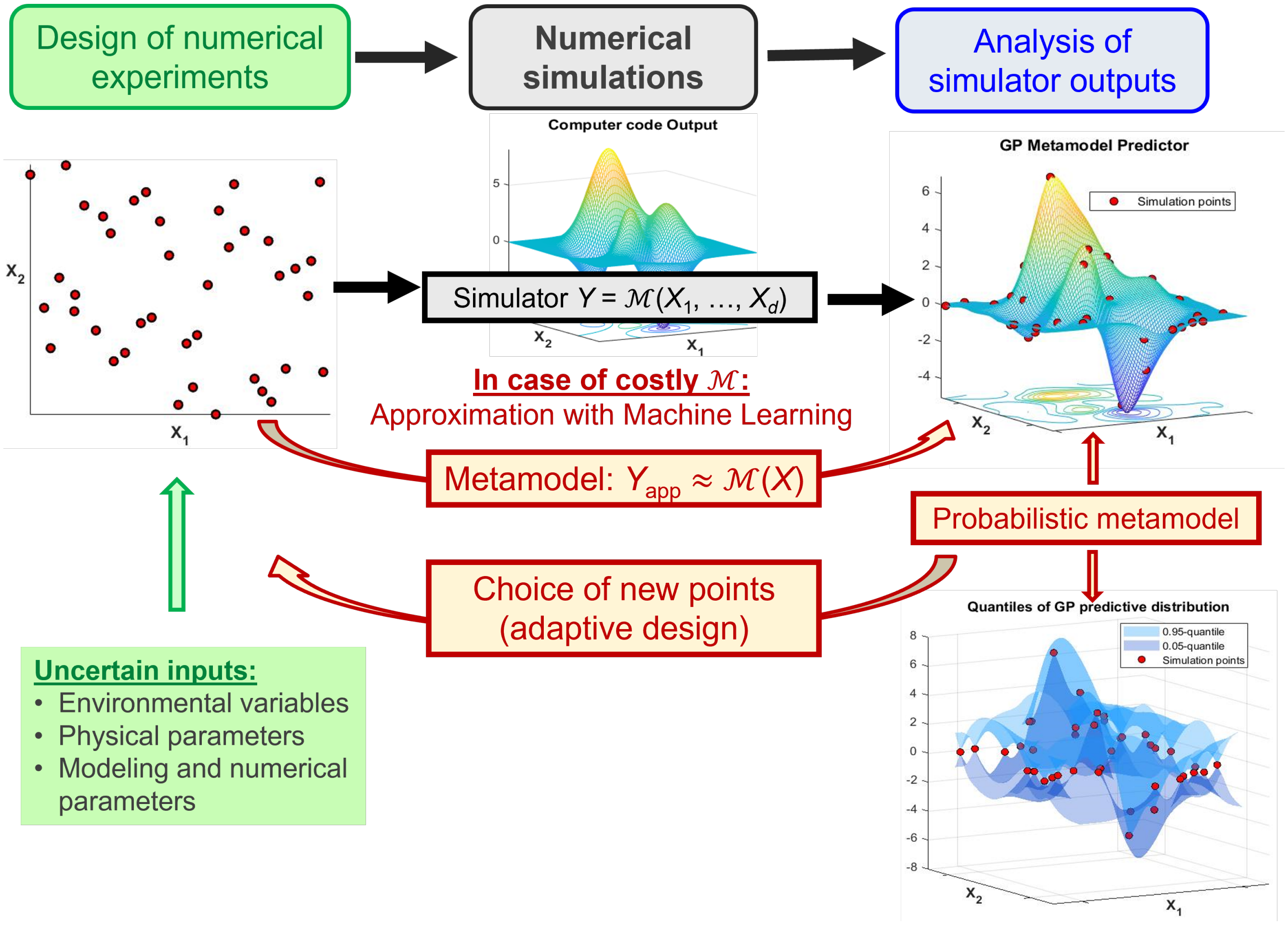
Design of numerical experiments
Numerical simulations
Analysis of simulator outputs
Computer code Output
GP Metamodel Predictor
Simulation points
x_2
x_1
Simulator $Y = \mathcal{M}(X_1, \ldots, X_d)$
In case of costly $\mathcal{M}$:
Approximation with Machine Learning
Metamodel: $Y_{app} \approx \mathcal{M}(X)$
Probabilistic metamodel
Choice of new points (adaptive design)
Quantiles of GP predictive distribution
0.95-quantile
0.05-quantile
Simulation points
Uncertain inputs:
Environmental variables
Physical parameters
Modeling and numerical parameters

Figure 1: Metamodel and probabilistic metamodel principles, illustrated with GP mean (GP metamodel predictor) and GP 90%-prediction intervals (quantiles of GP predictive distribution).

One aim of this paper is to focus strong attention on a key point, often omitted in the above mentioned studies, in the use of GP-based adaptive algorithms. Indeed, when fitting a GP metamodel on a given dataset, the specification of the covariance structure of the process is particularly sensitive. Related to this choice, the covariance parameters, often called the hyperparameters, have to be estimated with care. Mean and covariance functions are usually chosen within parametric families (for instance, class of Matérn covariance functions), and the estimation of GP only consists in the estimation of hyperparameters. For this, different methods exist based on either likelihood maximization, cross-validation technique or Bayesian approach. From a theoretical point of view and under the hypothesis of a well-specified covariance model, some results exist concerning the consistency of the different types of estimators [37, 38, 39]. From a practical standpoint, some studies [40, 41, 42] have proposed comparisons on analytical functions, but no consensus really emerges: the estimation of hyperparameters is often unstable regardless of the method. Moreover, most of the time, the authors only focus on the accuracy of GP predictor to assess the impact of hyperparameter estimates. The reliability of prediction intervals is rarely considered (except in Petit et al. [43]), while it is often affected by poor estimation of GP hyperparameters (as highlighted by Demay et al. [44]). To circumvent this problem, other authors such as Acharki et al. [45] propose to correct the hyperparameter estimates to obtain more reliable prediction intervals. Whatever the estimation method or correction considered, this calls for validation indicators to control the performance and robustness of GP regression.

In this framework, this paper reviews recent works dealing with the difficulties inherent in estimating GP hyperparameters, in order to analyze the advantages and disadvantages of estimation methods, to list and propose relevant validation criteria (to detect poor hyperparameter estimation), and to study some recently proposed robust and corrective methods. The rest of the document is organized as follows. Reminders on GP regression, parameterization and estimation of parameters are given in Section 2. Section 3 reviews the different hyperparameters estimation algorithms. Section 4 lists important criteria for GP validation. Section 5 then proposes a review of very recent papers dealing with the robust estimation of GP hyperparameters and allows to explain the orientation chosen for our research work, w.r.t. the application context. In particular, validation criteria to control the performance and robustness of GP regression are detailed. The last section gives some conclusions and prospects of this work. From this extensive review, a companion paper [46] proposes a new algorithm that solves some of the identified drawbacks of the previous ones.

2. Reminders on Gaussian process regression

Throughout the rest of this paper, the numerical model (computer code or simulator) is represented by the following input-output relationship:

$$\mathcal{M} : \left| \begin{array}{ccc} \mathcal{X} & \longrightarrow & \mathcal{Y} \\ \mathbf{X} & \longmapsto & Y = \mathcal{M}(\mathbf{X}) \end{array} \right. \tag{1}$$

where the uncertain output variable Y and the d input parameters $\mathbf{X} = (X_1, \ldots, X_d)^\top$ belong to some measurable spaces respectively denoted by $\mathcal{Y}$ and $\mathcal{X} \subset \mathbb{R}^d$. As part of the probabilistic approach, the inputs are considered as random variables with probability distributions denoted by $\mathbb{P}_{\mathbf{X}}$ on $\mathcal{X}$ [3]. It is therefore assumed that we have a n-size sample of inputs and associated outputs denoted by $(\mathbf{X_s}, \mathbf{Y}(\mathbf{X_s}))$ where $\mathbf{X_s} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\}$ with $\mathbf{x}^{(i)} = \left(x_1^{(i)}, \ldots, x_d^{(i)}\right)$ denotes the matrix of n-size sample locations (also called the "experimental design"), and $\mathbf{Y_s} = \{y^{(1)}, \ldots, y^{(n)}\}$ the corresponding outputs observations with $y^{(i)} = \mathcal{M}(\mathbf{x}^{(i)})$. $(\mathbf{X_s}, \mathbf{Y_s})$ constitutes the learning sample.

2.1. GP metamodel conditioned by the learning sample

In the GP regression framework [20, 47], the data are modeled as discrete observations of a GP sample path. The prior knowledge on observations is modeled by a GP completely specified by its mean function $m(\mathbf{x}) = \mathbb{E}[Y(\mathbf{x})]$ and its covariance function $k(\mathbf{x}, \tilde{\mathbf{x}}) = \mathbb{COV}(Y(\mathbf{x}), Y(\tilde{\mathbf{x}})) = \mathbb{E}[(Y(\mathbf{x}) - m(\mathbf{x}))(Y(\tilde{\mathbf{x}}) - m(\tilde{\mathbf{x}}))]$. $k(\cdot, \cdot)$ is also called the covariance kernel and is assumed to be a positive definite kernel. The predictive GP distribution is therefore naturally given by the GP conditioned by the known observations $\mathbf{Y_s}$, denoted $[Y(\mathbf{x}) | \mathbf{Y}(\mathbf{X_s}) = \mathbf{Y_s}]$. Its distribution can be obtained analytically from the following joint distribution:

$$\begin{pmatrix} Y(\mathbf{x}) \\ \mathbf{Y}(\mathbf{X_s}) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} m(\mathbf{x}) \\ \mathbf{m}(\mathbf{X_s}) \end{pmatrix}, \begin{pmatrix} k(\mathbf{x}, \mathbf{x}) & \mathbf{k}(\mathbf{x}, \mathbf{X_s})^T \\ \mathbf{k}(\mathbf{x}, \mathbf{X_s}) & \mathbf{K} \end{pmatrix} \right), \tag{2}$$

where:

- $\mathbf{Y}(\mathbf{X_s}) = (Y(\mathbf{x}^{(i)}))_{1 \le i \le n} \in \mathbb{R}^n$ is the vector of output value at sample locations,
- $\mathbf{m}(\mathbf{X_s}) = (m(\mathbf{x}^{(i)}))_{1 \le i \le n} \in \mathbb{R}^n$ is the vector of mean function evaluated at sample locations,
- $\mathbf{K} = (k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}))_{1 \le i,j \le n} \in \mathbb{R}^{n \times n}$ is the covariance matrix at sample locations,
- $\mathbf{k}(\mathbf{x}, \mathbf{X_s}) = (k(\mathbf{x}, \mathbf{x}^{(i)}))_{1 \le i \le n} \in \mathbb{R}^n$ is the covariance vector between $\mathbf{x}$ and sample locations.

By applying the conditioning theorem of Gaussian vectors to the joint distribution, the conditional field $Y(\mathbf{x})|\mathbf{Y}(\mathbf{X_s}) = \mathbf{Y_s}$ is still a GP whose mean is given by:

$$\hat{y}\left(\mathbf{x}\right) = \mathbb{E}[Y(\mathbf{x})|\mathbf{Y}(\mathbf{X_s}) = \mathbf{Y_s}] = m(\mathbf{x}) + \mathbf{k}(\mathbf{x}, \mathbf{X_S})^T \mathbf{K}^{-1}\Big(\mathbf{Y_s} - \mathbf{m}(\mathbf{X_s})\Big), \tag{3}$$

and its covariance function:

$$\hat{c}(\mathbf{x}, \tilde{\mathbf{x}}) = \mathbb{COV}[Y(\mathbf{x}), Y(\tilde{\mathbf{x}})|\mathbf{Y}(\mathbf{X_s}) = \mathbf{Y_s}] = k(\mathbf{x}, \tilde{\mathbf{x}}) - \mathbf{k}(\mathbf{x}, \mathbf{X_s})^T \mathbf{K}^{-1}\mathbf{k}(\tilde{\mathbf{x}}, \mathbf{X_s}). \tag{4}$$

In the following, the conditioning notation $[\bullet|\mathbf{Y}(\mathbf{X_s}) = \mathbf{Y_s}]$ will be reduced to $[\bullet|\mathbf{Y_s}]$ for the sake of brevity. Therefore, the predictive distribution for a new (unobserved) point $\mathbf{x}$ is the Gaussian distribution $\mathcal{N}(\hat{y}(\mathbf{x}), \hat{s}^2(\mathbf{x}))$ where $\hat{s}^2(\mathbf{x}) = \hat{c}(\mathbf{x}, \mathbf{x})$.

The conditional expectation $\hat{y}(\mathbf{x})$ is used as the predictor of the GP regression model and its mean-square error is given by the conditional variance $\hat{s}^2(\mathbf{x})$, while the Gaussian predictive distribution can be used to build predictive intervals of any level $\alpha \in]0, 1[$. More generally, conditional simulations (i.e. simulation of conditional GP trajectories) can be used to estimate, with a confidence interval, any statistical quantity of interest derived from the output (probability of exceeding a threshold, quantiles, etc.). The predictive distribution of some quantities of interest can also be defined analytically: this is obviously the case for a vector of prediction points, but also for derivatives or excursion sets. This possibility offered by GP regression is of particular interest in uncertainty quantification studies or for developing optimization strategies.

2.2. Covariance function, hyperparameters and nugget effect

The prior knowledge in GP regression consists in specifying the mean $m(\mathbf{x})$ and the covariance function $k(\mathbf{x}, \tilde{\mathbf{x}})$ which is certainly the most important ingredient of a GP regression as it describes the dependence structure and controls the smoothness of the approximation.

2.2.1. Usual covariance functions and consideration on a priori choice

In the GP regression of computer experiments, the most popular choice is undoubtedly the class of stationary ν-Matérn functions defined in one dimension ($x \in \mathbb{R}$) by:

$$k_{\sigma,\nu,\theta}(x, \tilde{x}) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \Big(\frac{\sqrt{2\nu}h}{\theta}\Big)^\nu K_\nu \Big(\frac{\sqrt{2\nu}h}{\theta}\Big), \tag{5}$$

where $h = |x - \tilde{x}|$, σ^2 and $\theta \in \mathbb{R}^+$ are respectively the variance parameter and the correlation hyperparameter (also called correlation length or length-scale). K_ν is a modified Bessel function of second kind with parameter $\nu \in \mathbb{R}^+$, and Γ is the Euler Gamma function. The parameter ν controls the smoothness of the GP: $Y(x)$ is at least k-time mean-square differentiable if and only if $\nu > k$. $\nu = 1/2$ corresponds to the exponential covariance with continuous but not differentiable GP trajectories, while the limiting case $\nu \to \infty$ yields the Gaussian covariance function with infinitely differentiable trajectories. Between these two extreme cases, two popular ν-Matérn covariances in the statistical learning community are the ones for $\nu = 3/2$ and $\nu = 5/2$, which respectively yield to GP trajectories once and twice differentiable (see Table 1).

Hence, choosing a correlation function most often consists in assuming a prior regularity for the model to be emulated since it directly defines the space of the possible trajectories in which the "real" function (or model) is supposed to belong. The success of the GP metamodel is conditioned to an adapted covariance model [48]. To ensure a relevant choice, one solution might be to consider the family of ν-Matérn functions and to integrate the regularity parameter ν in the set of hyperparameters to be estimated from the dataset. But, as pointed out among others by Gu et al. [49], this is not a very relevant idea in practice notably for the emulation of computer experiments

	$v = \frac{1}{2}$	$v = \frac{3}{2}$	$v = \frac{5}{2}$	$v = +\infty$
Usual name	exponential	3/2-Matérn	5/2-Matérn	Gaussian
$k_{\sigma,\nu,\theta}(x,\tilde{x})$	$\sigma^2 e^{-\frac{h}{\theta}}$	$\sigma^2(1+\sqrt{3}\frac{h}{\theta})e^{-\sqrt{3}\frac{h}{\theta}}$	$\sigma^2\left(1+\sqrt{5}\frac{h}{\theta}+\frac{5}{3}\left(\frac{h}{\theta}\right)^2\right)e^{-\sqrt{5}\frac{h}{\theta}}$	$\sigma^2 e^{-\frac{1}{2}\left(\frac{h}{\theta}\right)^2}$
Differentiability of GP trajectories	$\mathcal{C}^0$	$\mathcal{C}^1$	$\mathcal{C}^2$	$\mathcal{C}^\infty$

Table 1: Usual covariance functions and properties of associated GP trajectories.

in high dimension. First, most often a space-filling design [8] is used to generate $\mathbf{X_s}$ in order to have simulation points with good input space covering. Although this type of design optimizes in practice the predictivity of the GP metamodel, the absence of points' aggregates prevents from properly characterizing and therefore estimating the ν parameter. Indeed, the covariance functions mainly differ at the origin (i.e. for a distance between points tending towards 0) and in particular by the decay rate at this point. Furthermore, it turns out that jointly estimate ν with (σ^2, θ) may cause in practice computational and inferential difficulties in the estimation processes. It is therefore preferable (and commonly adopted) to estimate (σ^2, θ), conditionally to a specified value of ν.

Hence, as suggested by Petit [50] and in the direct line of Demay et al. [44], an interesting compromise (that we also recommend from our experience) is to consider a finite collection of covariance functions (those of Table 1), then estimate the hyperparameters (σ^2, θ) for each of them, and finally use a validation criterion (different from criterion used for the estimation) to select the best covariance.

Finally, it is reasonable to consider the 5/2-Matérn covariance as the default practical choice because, as highlighted by Gu et al. [49], it has very interesting behavior of the 5/2-Matérn covariance w.r.t. to the distance between two input points. On the one hand, when distance tends towards zero, 5/2-Matérn covariance behaves like Gaussian covariance, maintaining the smoothness for nearby inputs while ensuring better conditioning number of the covariance matrix. On the other hand, when distance tends to infinity, it behaves like the exponential covariance, preventing from decreasing quickly with distance (as does the Gaussian correlation). This can be useful for sparse data (as is often the case in numerical simulator emulation) or for non-influential inputs for which it is logical that the correlation is quasi-constant with distance.

2.2.2. Extension to multivariate case

In order to extend to multi-dimensional inputs $\mathbf{x} \in \mathbb{R}^d$, a widely used approach consists in considering a tensorized covariance defined as a product of univariate covariances:

$$k_{\sigma,\nu,\boldsymbol{\theta}}(\mathbf{x}, \tilde{\mathbf{x}}) = \sigma^2 \prod_{i=1}^{d} k_{1,\nu,\theta_i}(x_i - \tilde{x}_i). \tag{6}$$

The d 1-D covariance functions can be of different natures (with different smoothness parameters ν_i for instance). But in practice, given the large number of inputs and without any prior knowledge, the usual practice is to use the same function for all variables.

2.2.3. Additional variance modeled by nugget effect

An additional nugget effect can also be considered in the covariance: it assumes an additive white noise effect, whose variance denoted σ_ϵ^2 constitutes the nugget parameter. Most often, σ_ϵ^2 is assumed to be constant, independent from the inputs (homoscedastic hypothesis). The covariance matrix then becomes $\mathbf{K}^{'} = \mathbf{K} + \sigma_\epsilon^2 I_n$ where I_n is the identity matrix. From a purely parametric point of view, the variance of the nugget effect is often considered and parameterized relatively to the variance of the GP with $\lambda = \left(\frac{\sigma_\epsilon}{\sigma}\right)^2 \in \mathbb{R}^+$. Even for noiseless data as in the case of a deterministic simulator, the nugget effect is often used in GP metamodeling because its practical interest is twofold: both to relax the interpolation property of the GP regression and to improve the conditioning number of the covariance matrix (also referred to as GP regularization). Conceptually, it means that the model function $\mathcal{M}$ (numerical simulator in computer experiments) is supposed to be a slightly noised version of a smoother and deterministic simulator. This regularity aspect can also be considered in presence of sparse problems, characterized by a weak density of observations in the input parameter space.

2.2.4. Considerations around the GP trend

Finally, let us say now a brief word about the GP mean (or trend) $m(\mathbf{x})$. A constant $m(\mathbf{x}) = \beta_0$ or a one-degree polynomial trend $m(\mathbf{x}) = \beta_0 + \sum_i \beta_i x_i$ is usually considered in practice. But any linear regression model on a set of known basis functions could be used instead. For simplicity, it is assumed, in the rest of the section, that the prior mean is a constant and more exactly equals to zero (assuming that data are centered, for instance). This assumption is only made to simplify some equations (e.g. reminders on maximum likelihood estimation).

3. Estimation of Gaussian process hyperparameters

The type of covariance function is generally fixed among the usual choices of Table 1. It then remains to estimate the covariance parameters $(\sigma^2, \boldsymbol{\theta})$ (and eventually λ if a nugget effect is considered). The main estimation procedures are based either on minimization of the squared prediction error calculated by cross-validation (CV), or on maximization of likelihood (denoted MLE for maximum likelihood estimation). Note that Petit et al. [43] propose a review and comparison of a larger panel of criteria to be optimized to estimate hyperparameters. In particular, they suggest a generalization of the likelihood criterion called the Fasshauer's Hölderized likelihood and which is based on the orthogonal decomposition of the covariance matrix. In parallel to MLE and CV, a Bayesian estimation approach also exists: a prior distribution is assumed for the hyperparameters and combined with MLE to obtain a posterior distribution of the hyperparameters which is then propagated into the GP predictive distribution. The following subsections detail these different approaches and summarize recent works on the theoretical analysis and empirical comparison of these methods.

3.1. Cross-validation-based approach

A first estimation approach relies on the mean squared error (MSE) in prediction computed by cross-validation (CV) [11]. More precisely, in the case of *leave-one-out* (LOO) method, the GP hyperparameters are computed by minimizing the LOO-MSE:

$$\text{LOO-MSE}\left(\sigma^2, \boldsymbol{\theta}\right) := \frac{1}{n}\sum_{i=1}^{n} \left(\hat{y}_{-i} - y_i\right)^2,$$

where $\hat{y}_{-i}$ denotes the GP predictor (mean of predictive distribution) in $\mathbf{x}^{(i)}$ when $(\mathbf{x}^{(i)}, y^{(i)})$ is removed from the set of observations (this comes down to consider the GP conditioned by $\mathbf{Y}_{\mathbf{s},-\mathbf{i}}$).

Using the CV formulas of Dubrule [51], the predictive mean $\hat{y}_{-i}$ and variance $\hat{s}^2_{-i}$ are given by

$$\hat{y}_{-i} - y_i = \frac{(\mathbf{K}_{\sigma^2,\boldsymbol{\theta}}\, \boldsymbol{y})_i}{(\mathbf{K}_{\sigma^2,\boldsymbol{\theta}})_{i,i}}, \tag{7}$$

and

$$\hat{s}^2_{-i} = \frac{1}{(\mathbf{K}_{\sigma^2,\boldsymbol{\theta}})_{i,i}}. \tag{8}$$

The LOO-based estimators of GP hyperparameters are thus given by

$$\left(\hat{\sigma}^2_{MSE}, \hat{\boldsymbol{\theta}}_{MSE}\right) = \arg\min_{\sigma^2,\boldsymbol{\theta}} \; \boldsymbol{y}^\top \mathbf{K}_{\sigma^2,\boldsymbol{\theta}} \operatorname{Diag}(\mathbf{K}_{\sigma^2,\boldsymbol{\theta}})^{-2} \mathbf{K}_{\sigma^2,\boldsymbol{\theta}} \; \boldsymbol{y}. \tag{9}$$

In practice, there are no closed-form expressions for $\left(\hat{\sigma}^2_{MSE}, \hat{\boldsymbol{\theta}}_{MSE}\right)$, and the quantity to be optimized is not convex and may have several local optima. The optimization has to be done numerically. Note that under the hypothesis of a well-specified covariance model, some theoretical results exist concerning the consistency of LOO (and more generally CV) estimators [39].

3.2. Maximum likelihood-based approach

The most widely used approach is the MLE which consists in identifying the values of $\boldsymbol{\theta}$ which minimizes the negative log-likelihood of the dataset:

$$\ell(\mathbf{Y_s}) = \frac{1}{2}\left(n\log(2\pi) + \log\left|\mathbf{K}_{\sigma^2,\boldsymbol{\theta}}\right| + \mathbf{Y_s}^T\,\mathbf{K}^{-1}_{\sigma^2,\boldsymbol{\theta}}\mathbf{Y_s}\right)$$

with $|A|$ denoting the determinant of matrix A. Provided that $\boldsymbol{\theta}$ is known and writing $\mathbf{K}_{\sigma^2,\boldsymbol{\theta}} = \sigma^2\mathbf{R}_{\boldsymbol{\theta}}$, the MLE estimator of the variance parameter is given by:

$$\widehat{\sigma}^2_{MLE} = \frac{1}{n}\mathbf{Y_s}^T\,\mathbf{R}^{-1}_{\boldsymbol{\theta}}\mathbf{Y_s}. \tag{10}$$

Plugging back $\widehat{\sigma}^2_{MLE}$ into $\ell(\mathbf{Y_s})$ to get a concentrated (or profile) log likelihood involving just $\boldsymbol{\theta}$, MLE results in the following minimization problem for $\boldsymbol{\theta}$:

$$\widehat{\boldsymbol{\theta}}_{MLE} = \arg\min_{\boldsymbol{\theta}} \widehat{\sigma}^2_{MLE}\left|R_{\boldsymbol{\theta}}\right|^{\frac{1}{n}}. \tag{11}$$

Even if calculating the derivative of the above expression is analytic, solving it is not, and no closed form solution can be obtained. As for LOO-MSE approach, numerical methods are thus required to estimate $\widehat{\boldsymbol{\theta}}_{MLE}$. The interested reader can refer to Santner et al. [19] for the MLE equations in the case where a regression model is considered for the trend $m(\mathbf{x})$. In addition, Zhang [37] and Bachoc [38] provide theoretical results on the consistency of MLE (still under the hypothesis of a well-specified covariance model).

3.3. Maximum likelihood estimation in Gaussian process framework may be ill-posed

In their recent papers, Karvonen and Oates [48] and Gu et al. [49] discuss how the MLE is well- or conversely ill-posed and how this notion can be defined in the framework of GP hyperparameter estimation, and in the non-asymptotic setting. Note that the scalar notation will be used in what follows for θ without loss of generality for an extension to the case of a vector $\boldsymbol{\theta}$.

3.3.1. Infinite θ and lack of continuity with respect to the training dataset

Under the assumption of a well-specified prior (mean and covariance), the predictive performance of the GP is well understood in an asymptotic setting [52, 39, 53]. But, there is not much in the literature about the non-asymptotic setting when the hyperparameters $\boldsymbol{\theta}$ are estimated from the learning sample, and in a deterministic interpolation framework (interpolating GP to emulate a deterministic function).

First, Karvonen and Oates [48] defines the MLE ill-posedness as the occurrence of $\widehat{\theta}_{MLE} = \infty$, which yields $k(x, \tilde{x}) = 1 \; \forall \; (x, \tilde{x}) \in \mathbb{R} \times \mathbb{R}$ and estimated correlation matrix $\mathbf{R}_{\widehat{\theta}_{MLE}} = \mathbf{1}_n \mathbf{1}_n^T$. The GP predictive distribution is degenerated for each point of prediction (predictive variance is zero and all the probability is assigned to a single value). Having infinite precision from a finite dataset is undesirable and the GP metamodel loses its interest as a tool for uncertainty quantification. It is therefore important to have validation criteria that take into account the whole predictive distribution and thus enable this situation to be detected, such as those presented in section 4.3.

From this definition of ill-posedness, Karvonen and Oates [48, Theorem 2.3] demonstrate that if the data are m-constant, i.e. shifted from the mean function $m(x)$ by a constant $c \in \mathbb{R}$:

$$y_i = m\left(x_i\right) + c \quad \text{for} \quad i = 1, \ldots, n, \text{ with } n \geq 2,$$

and if the covariance function has a polynomial decaying Fourier transform (like the Matérn functions with smoothness $\nu > 0$), then $\widehat{\theta}_{MLE} = \infty$. On the contrary, if the data are not m-constant, then $\widehat{\theta}_{MLE} < \infty$ so that the predictive distributions are non-degenerate.

Then, the authors propose to consider the classical definition of well-posedness of Hadamard for an inference or estimation problem. More precisely, it is well-posed if (i) a solution exists, (ii) the solution is unique, and (iii) the solution depends continuously on the data. If these conditions are not met, the problem is ill-posed. The (iii) condition about the sensitivity of the GP estimate to the training dataset is particularly relevant for sensitive applications such as nuclear safety applications. From this definition, and under the same assumptions of regularity of the covariance function (polynomial decaying Fourier transform), Karvonen and Oates [48] demonstrate that if $\widehat{\theta} = \infty$, the GP regression problem is ill-posed by violation of (iii), in the sense that the resulting predictive distributions are not locally Lipschitz in the data w.r.t. the Hellinger distance, which means that predictive inference can be sensitive to small perturbations of the dataset.

In order to find solutions to this problem, the authors examine several alternative modelling and estimation methods and show that:

- the LOO-CV estimator of θ shares the same undesirable property of MLE when the data are m-constant;

- the addition of a parametric prior mean function $m(x)$, also estimated from the data (i.e., as in universal kriging), does not help;

- simultaneous MLE of σ and θ does not prevent ill-posedness (as usually done in GP regression, see Eqs. (10,11));

- the addition of a nugget effect does not prevent from having $\widehat{\theta}_{MLE} = \infty$ when the data are m-constant but prevents the GP predictive distribution from being degenerate. Moreover, when the data are not m-constant but close to it, $\widehat{\theta}_{MLE} \to \infty$ and the condition number of the correlation matrix $\mathbf{R}_{\widehat{\theta}_{MLE}}$ increases with a rate related to the smoothness of the covariance. This results in a numerical issue in the MLE process (likelihood for large value of θ cannot be computed). The introduction of a $\lambda > 0$ allows to mitigate this issue (by upper bounding the condition number). The price to pay is therefore to relax the interpolation constraint.

In conclusion, the practical recommendations to be retained from Karvonen and Oates [48] are:

▶ the addition of a nugget effect $\lambda > 0$, also estimated by MLE, is recommended as a regularization parameter, taking care to limit its value by a reasonable upper bound. Another solution can be to bound the parameter θ and find the MLE estimate in $(0, \theta_{\max}]$ (constrained MLE). However, it implies the arbitrary choice of a $\theta_{\max}$ which turn out to be the estimated value of $\widehat{\theta}_{MLE}$ if the data are m-constant.

- ▶ Finally, reasoning in a rather rudimentary way, this would argue for not assuming a too complex model for the GP trend $m(x)$ so that the observed data are not $m(x)$-constant. In practice, $m(x)$ is assumed to be a constant or a one-degree polynomial (universal kriging), with parameters also estimated by MLE. Another practical recommendation could be to check if the data are m-constant (or close to be) in order to detect problematic datasets which cause GP regression to be ill-posed. However, it is reasonable to expect that in our industrial applications, the probability of having strictly $m(x)$-constant data (without any deviation from $m(x)$ or noise) is relatively low due to the complexity of the considered models (or codes), the large input dimension and the sparsity of the data (low sample size).

3.3.2. Lack of robustness

Previously to Karvonen and Oates [48], Gu et al. [49] had already proposed to define the ill-posed MLE problem, which the authors prefer to refer to as a lack of robustness. Starting from the observation that the likelihood (Eq. (11)) is sometimes very flat in the tails, Gu et al. [49, Definition 3.1 and Lemma 3.2] propose to define the lack of robustness of GP hyperparameters by the occurrence of two cases:

- case (1), as in Karvonen and Oates [48]: $\widehat{\theta}=\infty$ and consequently $\mathbf{R}_{\widehat{\theta}}=\mathbf{1}_n\mathbf{1}_n^T$. In multidimensional case (i.e. $d>1$), this case corresponds to $\widehat{\theta}_i=\infty$ for all $1\leq i\leq d$;
- case (2): $\widehat{\theta}=0$ and consequently $\mathbf{R}_{\widehat{\theta}}=\mathbf{I}_n$. For $d>1$, this is encountered when $\exists i, 1\leq i\leq d$, for which $\widehat{\theta}_i=0$.

Case (2) is not more desirable than case (1) because it means that $\mathbf{R}_{\widehat{\theta}_{MLE}}$ is near $\mathbf{I}_n$ and the GP predictor is an impulse function interpolating the observations, while following the GP mean $m(\mathbf{x})$ elsewhere. The authors state and numerically show that even if "*such degeneracies are somewhat unusual in one-dimension, they are not particularly unusual with higher dimensional inputs*" (large dimension d).

The authors also mention that MLE instability can often be "*overcome by adding a nugget effect, but studies have found that the features of the emulator can significantly change when a nugget is added [54].*" So, to circumvent the problem of robustness, Gu et al. [49] prefer to focus on Bayesian approaches and demonstrate that certain prior and parameterizations for the GP parameters result in a more robust estimation than others (see Section 5.2 dedicated to the so-called *RobustGaSP* Bayesian approach).

3.4. Bayesian approach

A third approach is to consider a full-Bayesian approach where a prior is assumed on the GP hyperparameters. The marginal posterior distribution is then inferred by Bayes' rule from marginal likelihood of data, and with regard to the prior. The resulting posterior uncertainty is then integrated in the GP predictive distribution. More precisely, assuming a prior on the hyperparameters $(\sigma^2,\boldsymbol{\theta})\sim\pi(\sigma^2,\boldsymbol{\theta})$, their posterior distribution writes $\pi(\sigma^2,\boldsymbol{\theta}\mid\mathbf{Y_s})\propto\pi(\mathbf{Y_s}\mid\sigma^2,\boldsymbol{\theta})\,\pi(\sigma^2,\boldsymbol{\theta})$. The probability density function of the GP predictive distribution of $Y(\mathbf{x})$ is then given by:

$$p\left(y(\mathbf{x})|\mathbf{Y_s}\right)=\iint p\left(y(\mathbf{x})\mid\mathbf{Y_s},\sigma^2,\boldsymbol{\theta}\right)\pi\left(\sigma^2,\boldsymbol{\theta}\mid\mathbf{Y_s}\right)\mathrm{d}\sigma^2\;\mathrm{d}\boldsymbol{\theta}. \tag{12}$$

Full Bayesian approach thus allows to take into account the uncertainty on the estimation of the GP hyperparameters and to propagate it in the GP predictive law. It has been illustrated in dimension two by Wieskotten et al. [55] showing that it is relevant and can outperform the ordinary GP in terms of both predictivity and accuracy of predictive intervals. This is especially true when the sample size is small, the benefit decreasing as the size increases, as one might expect.

However, the tractability of the full Bayesian approach in higher dimension remains a major obstacle to its use. Indeed, in practice, the computation of $\pi\left(\sigma^2, \boldsymbol{\theta} \mid \mathbf{Y_s}\right)$ and $p\left(y(\mathbf{x})|\mathbf{Y_s}\right)$ requires the use of Markov chain Monte Carlo (MCMC) methods like Metropolis-Hastings algorithm [56] or Hamiltonian Methods [57]. The calculation cost of the predictive distribution with MCMC techniques becomes expensive in large dimension. To circumvent this limitation, some *plug-in* approaches can be considered as in Gu et al. [49]: $p\left(y(\mathbf{x})|\mathbf{Y_s}\right)$ is computed with the GP hyperparameters fixed at the maximum a posteriori probability (MAP) estimate (that equals the mode of the posterior distribution $\pi\left(\sigma^2, \boldsymbol{\theta} \mid \mathbf{Y_s}\right)$). In this case, the Bayesian framework is only used to compute this posterior distribution and is then discarded to calculate the predictive law $p\left(y(\mathbf{x})|\mathbf{Y_s}\right)$. Basically, this means replacing $\pi\left(\sigma^2, \boldsymbol{\theta} \mid \mathbf{Y_s}\right)$ with a Dirac distribution whose mass is concentrated on the MAP value. This somewhat brutal solution facilitates the intensive use of the predictive distribution. Only the problem of the estimation of $\pi\left(\sigma^2, \boldsymbol{\theta} \mid \mathbf{Y_s}\right)$ remains. Finally, the Bayesian approach (whether full or plug-in) also requires to define the prior distribution for the hyperparameters, choice which can be of prime importance as detailed in the work of Gu et al. [49] (See Section 5.2).

3.5. Discussion on the relative practical performance of the different approaches

Discussions on the choice of a method between MLE and CV (or LOO) methods are not new, but recent work and especially intensive benchmarks are shedding new light. Let us try to make a brief synthesis.

As shown by Bachoc [42], the MLE method is optimal when the covariance function is well-specified, i.e. when the "true" covariance function belongs to the assumed parametric set of covariance functions. If this is not the case (misspecification case), there is no more guarantee that the MLE method would perform correctly and optimally. Bachoc [42] illustrated that MLE may not be very robust in this case, especially if the number of data is small, while the CV-based approach performs better. On the other side, the availability of gradients in the MLE case (without significant additional computational cost) is an advantage in the implementation of numerical optimization algorithms required for hyperparameter estimation. However, this advantage must be qualified by the recent works on the computational complexity of cross-validation schemes and more precisely the fast computation of gradients of LOO criteria [58, 43, 50]. Still in Petit [50], an intensive benchmark on analytic functions of different dimensions shows that MLE is often preferable to its competitors (not only in well specified cases but also in case of overestimated regularity), and that the choice of regularity (ν in Matérn class) might be often more important than the estimation of GP hyperparameters. To conclude the comparison between MLE and LOO, let us mention the remark of Zhang and Wang [59] concerning the flatness of both MLE or LOO-based criteria around their optimal value. The authors argue that the flatness of LOO-based criteria is less damaging since it indicates that the predictive distribution is less sensitive to the hyperparameter value in the flatness region.

Faced with this lack of consensus between MLE and LOO, the full-Bayesian approach could appear as a relevant solution as it may yield more robust predictions (see, e.g., [60]). But, this approach strongly depends on the prior distribution of the hyperparameters, as highlighted by Muré [61], and has a much higher computational cost, especially if the number of hyperparameters is high. Besides, to the best of our knowledge, there are no applications using this approach to emulate numerical simulators in large dimension ($d \geq 10$ for instance). Only the RobustGaSP Bayesian method proposed by Gu et al. [49] and detailed in Section 5.2 with its specific priors and approximations, could overcome these limitations.

However, another solution would also be relevant: both the MLE and CV criteria would be integrated into the estimation of the hyperparameters. A multi-objective procedure could be developed where the MLE criterion would remain the main reference objective and another LOO-based criterion could be considered as a complementary criterion. This is the purpose of the work proposed in the companion paper [46] of the present article. More generally, we also think that a nugget effect has to be considered and estimated jointly with the other hyperparameters as it allows to enrich the family of covariance functions (this having not been considered in the benchmark of Petit et al. [43]). It also facilitates the MLE by regularizing the likelihood function, improving the conditioning of the correlation matrix and numerical convergence of algorithms. However, this nugget effect can be double-edged into a Bayesian approach as it may increase identifiability problems. Its use should be restricted to simple MLE.

4. Quantitative criteria for Gaussian process validation

Once the GP metamodel has been estimated, its predictive capabilities need to be checked to ensure confidence in its use (as a substitute for the simulator). Thus, validation criteria must be defined to assess the accuracy of the GP predictor, but also of its prediction variance, its covariance and more generally of the whole GP conditional distribution. For this, different quantitative criteria have been proposed (see, e.g., [59, 44, 43]) and are listed in the following. A new criterion, the IAEα criterion, is also proposed. Note that the validation criteria presented in the following are formulated in their cross-validation version (more precisely in their LOO version), but can of course be defined (and computed) in a similar way on a test basis, different and independent from the learning sample. Similar expressions can also be obtained with *K-fold* cross-validation [62].

4.1. Criteria to assess the accuracy of the Gaussian process predictor $\hat{y}(\mathbf{x})$

Classically, the root mean squared error (RMSE) writes

$$\text{RMSE} = \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(y(\mathbf{x}_i) - \hat{y}_{-i}(\mathbf{x}_i) \right)^2 \right\}^{1/2} \tag{13}$$

and its counterpart expressed in terms of the proportion of the variance explained, namely the *predictivity coefficient* Q^2 (see, e.g., [63, 64]):

$$Q^2 = 1 - \frac{RMSE^2}{\frac{1}{n}\sum_{i=1}^{n} \left(y_i - \frac{1}{n}\sum_{i=1}^{n} y_i \right)^2}. \tag{14}$$

The closer to one the Q^2, the better the accuracy of the metamodel predictor. On the contrary, a zero Q^2 indicates very poor predictive abilities, i.e. equivalent to what would be obtained with the empirical mean of the observations. Note that both RMSE and Q^2 correspond to averaged indicators and should be complemented with a plot of observed data versus predicted values for more detailed analysis.

4.2. Criterion to evaluate if the conditional Gaussian process variance is of the right order of magnitude

Other important indicators propose to deal with the model variance [65]. We focus here on the *predictive variance adequacy* (PVA) factor (see, e.g., [42, 44]):

$$\text{PVA} = \left| \log \left(\frac{1}{n} \sum_{i=1}^{n} \frac{(y_i - \hat{y}_{-i})^2}{\hat{s}_{-i}^2} \right) \right|. \tag{15}$$

In order to get reliable prediction intervals from the model, the prediction variances should be of the same order of the prediction errors so that the PVA should be close to zero. In summary, the smaller the PVA, the more reliable the prediction intervals. On the contrary, too low prediction variances w.r.t. the prediction errors (i.e. an "overconfident" predictive model) or too large prediction variances ("underconfident" or too uncertain predictive model) yield poor PVA. A more detailed analysis and interpretation of its values is available in Demay et al. [44, Table 1].

4.3. Criteria to assess the accuracy of the whole Gaussian process predictive distribution

The logarithmic score [66] is defined as the negative logarithm of the predictive density evaluated on the observations:

$$\mathrm{LogS} = \frac{n}{2}\log(2\pi) + \sum_{i=1}^{n}\left(\log \hat{s}_{-i} + \frac{1}{2}\frac{(y_i - \hat{y}_{-i})^2}{\hat{s}_{-i}^2}\right). \tag{16}$$

Some similarities exist between LogS and PVA, both depending on the standardized residuals. But Demay et al. [44] point out that PVA is preferable for GP validation since it will mitigate the effect of extreme values and it will similarly penalize models with too large or too small predictive variances. In contrast, the weighting of the two terms in LogS will less penalize the too large predictive variances.

Then, we have all the class of criterion based on the reliability of predictive intervals. As recalled by Zhang and Wang [59], from the Brier score defined for the predictive cumulative distribution F_{-i} by

$$\mathrm{BS}(y) = \frac{1}{n}\sum_{i=1}^{n}\left(F_{-i}(y) - 1\left\{Y\left(\mathbf{x}_i\right) \leq y\right\}\right)^2,$$

the continuous ranked probability score (CRPS, [67, 68]) is defined as the integration of BS:

$$\text{CRPS } = \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty}\left(F_{-i}(y) - 1\left\{Y\left(\mathbf{x}_i\right) \leq y\right\}\right)^2 \, \mathrm{d}y = \int_{-\infty}^{\infty}\mathrm{BS}(y)\mathrm{d}y. \tag{17}$$

For a GP metamodel, it can be demonstrated that:

$$\mathrm{CRPS} = \frac{1}{n}\sum_{i=1}^{n}\hat{s}_{-i}\left(\frac{(y_i - \hat{y}_{-i})}{\hat{s}_{-i}}\left(2\Phi\left(\frac{(y_i - \hat{y}_{-i})}{\hat{s}_{-i}}\right) - 1\right) + 2\phi\left(\frac{(y_i - \hat{y}_{-i})}{\hat{s}_{-i}}\right) - \frac{1}{\sqrt{\pi}}\right) \tag{18}$$

with ϕ and Φ respectively denoting the probability and cumulative functions of the standard Gaussian distribution. CRPS is more robust than LogS, but it will tend to favour models with small predictive variance, subject to similar calibration performance. This is not desirable for GP validation since the objective is to have the most reliable predictive distribution, more than the most concentrated, especially in a safety study framework. Moreover, as illustrated by Demay et al. [44] on their application, the CRPS-based criterion does not allow to identify an inaccurate covariance model when this mismodeling only affects the predictive variance (and not the predictor), unlike the PVA or other criteria based on the predictive interval and presented immediately afterwards.

By focusing on the validation of GP prediction intervals (PI), the level $\alpha \in]0,1[$ of any PI can be compared to the proportion of observations that actually lie within this interval. This proportion also called *empirical coverage function* [69] is defined as:

$$\hat{\Delta}(\alpha) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\left\{y_i \in PI_{\alpha,-i}\left(\mathbf{x}_i\right)\right\}, \tag{19}$$

where $\mathbf{1}\{A\}$ is the indicator function of A and $PI_{\alpha,-i}$ is the α-level prediction interval for the point $\mathbf{x}_i$ built from the Gaussian distribution $\mathcal{N}(\hat{y}_{-i}, \hat{s}_{-i}^2)$.

From this, a graphical tool referred to as α-PI plot can be built by plotting $\hat{\Delta}(\alpha)$ against α [70, 71, 44]. By definition, the more the points should be located around the $y = x$ line, the more reliable the GP predictive intervals are. In order to have a quantitative indicator summarizing the quality of the α-PI plot, we naturally propose to consider the following IAEα criterion:

$$\mathrm{IAE}\alpha = \int_0^1 \left|\hat{\Delta}(\alpha) - \alpha\right| \mathrm{d}\alpha. \tag{20}$$

This criterion denoted IAEα for integrated absolute error on α corresponds to the area between the alpha plot and the reference line. IAEα lies in $[0, 1]$ and the closer to zero the IAEα, better the PI in average. Note that this criterion is very close to the so-called "*mean squared error α*" of Wieskotten et al. [55], defined with L^2 norm instead of L^1 norm. As in Wieskotten et al. [55], IAEα will be computed in practice with a regular discretization of α over $]0, 1[$. The L^1 norm is preferred here to give a homogeneous weight whatever α, avoiding to give too much weight to the strongest deviations and having a direct interpretation w.r.t. the α-PI plot.

Remark 1. ***Practical recommendation.*** *The last group of criteria composed of $\hat{\Delta}(\alpha)$ and those derived from it, namely the α-PI plot and IAEα, are perfectly adapted to the control of the reliability of prediction intervals. However, they should not be used alone but in addition to a prior control of predictivity with Q^2 (or RMSE). To better understand this recommendation, let us take for example the extreme case with the following metamodel: a constant predictor corresponding to the empirical mean of the data, a constant prediction variance, equal to the empirical variance of the data, and a Gaussian predictive distribution. Finally, let us assume that the sample of observed data follows a distribution with a Gaussian shape. We would then obtain for the predictive law a Q^2 equal to 0 but a $\hat{\Delta}(\alpha)$ very close to the right level for each α. Consequently, $\hat{\Delta}(\alpha)$ is only of interest and should only be used if and only if the predictivity of the metamodel has already been checked and controlled (via RMSE or Q^2 for instance).*

4.4. Illustration of validation criteria

To illustrate the interest of considering several criteria in the validation process, let us consider the example of the "re-scaled" Branin function used by Picheny et al. [72]. It is defined for two independent uniform inputs X_1 and X_2 on $[0, 1]$ by:

$$\mathcal{M}_{\mathrm{Branin}}(X_1, X_2) = \frac{1}{51.95}\left(a(V_2 - bV_1^2 + cV_1 - r)^2 + s(1-t)\cos(V_1) + s - 44.81\right), \tag{21}$$

with $V_1 = 15X_1 - 5$, $V_2 = 15X_2$, $a = 1$, $b = \frac{5.1}{(4\pi^2)}$, $c = \frac{5}{\pi}$, $r = 6$, $s = 10$ and $t = \frac{1}{8\pi}$. This function is illustrated by Figure 2. Consider a GP metamodel with a 2-D tensorized Gaussian covariance and conditioned by a random learning sample of $n = 30$ points. A random test sample of 1000 points is used to compute the validation criteria. Considering three different sets of hyperparameters $[\theta_1, \theta_2]$, Figure 3 illustrates that they can lead to metamodels with a very similar $Q^2 \sim 0.9$, but significantly different IAEα, namely 0.2 for one and 0.05 for the other, and on the contrary with similar IAEα and very different Q^2. Moreover, considering the two first sets of hyperparameters, it also illustrates the resulting major difference in the level of the GP prediction intervals compared to the desired level. An increase of 0.15 in IAEα here results in much more conservative and less realistic intervals. For instance prediction intervals of level $\alpha = 0.7$ actually include on average nearly 95% of the data. Evaluating only Q^2 (or RMSE), as is often done, does not make it possible to dissociate the two metamodels, whereas one of the two offers a much better assessment of prediction uncertainty. In conclusion, the criteria should be used in a complementary way. Their complementarities will be further illustrated in in the companion paper [46].

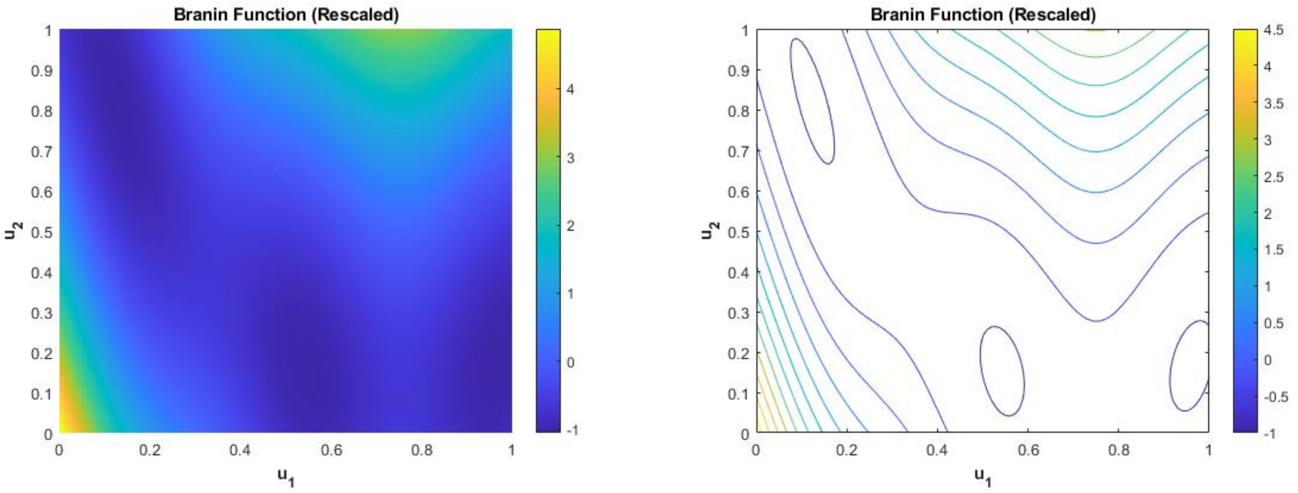

Figure 2: Analytical re-scaled Branin function on $[0, 1]^2$.

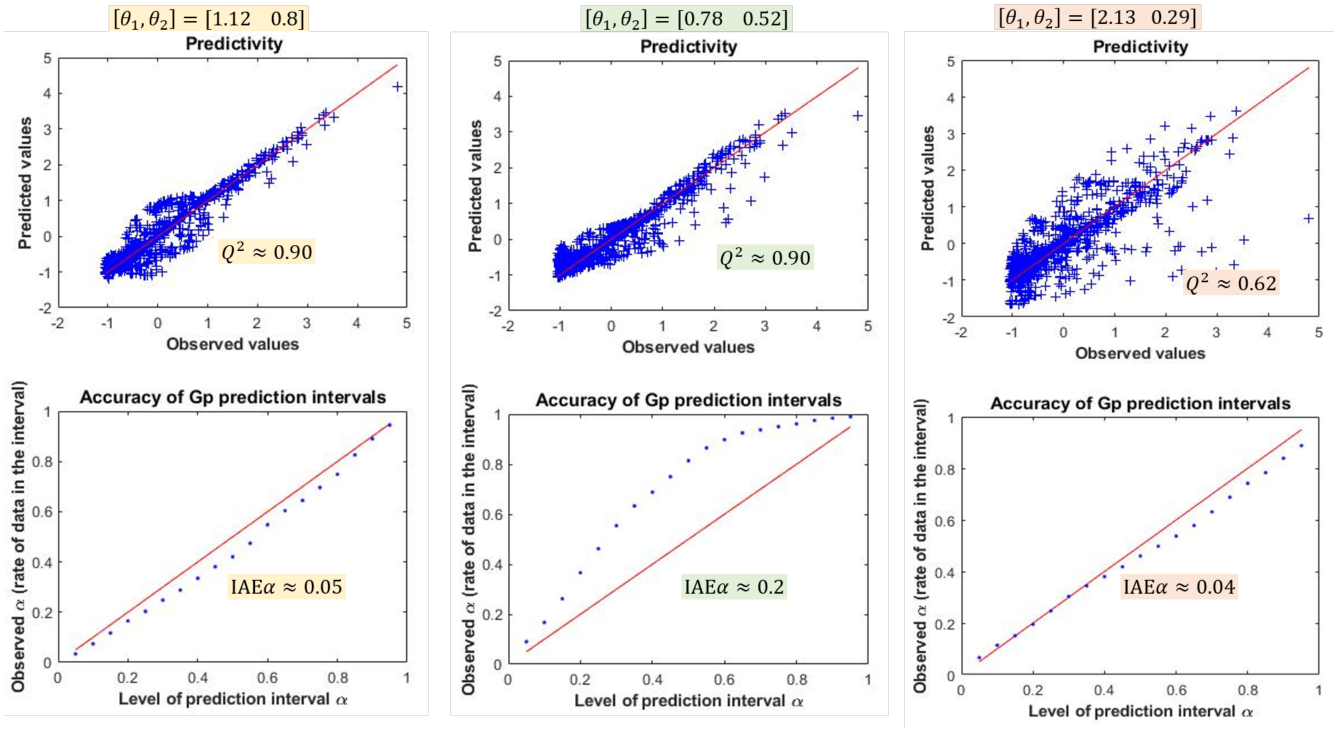

Figure 3: $\tilde{\mathcal{M}}_{\text{Branin}}$ Function – Illustration of the different impact of hyperparameters values on graphical and quantitative validation criteria for a GP with a tensorized Gaussian covariance and conditioned by a random sample of $n = 30$ simulations. Each column of two plots correspond to a specific set of hyperparameters (specified in the colored frame at the top). The first line plots the predicted values of the metamodel against the true values of $\tilde{\mathcal{M}}_{\text{Branin}}$ (on a test basis of 1000 points). The second line shows the α-plot of prediction intervals (again computed on the test basis).

5. Recent developments for more reliable Gaussian process predictions

In the case of emulation of deterministic functions from a numerical simulator, there is no guarantee that the function to be emulated is part of the set of possible trajectories generated by the assumed GP model (case of model misspecifications). As a result, there is no guarantee either that the MLE (or any other estimation approach) will work, or that the estimated prediction variance will control the metamodel error. More generally, the GP predictive distribution may not accurately cover unobserved data. Based on this statement, Acharki et al. [45] were interested in correcting the GP estimated hyperparameters to adjust the GP prediction intervals and ensure better coverage probabilities of the GP predictive distribution. In the following, we outline this approach, which is further detailed in Appendix A, before explaining the limits of the method in our application context.

5.1. A corrective approach to directly modify the bounds of Gaussian process prediction intervals

First of all, Acharki et al. [45] assume that he has a first set of estimated GP hyperparameters obtained by MLE or CV method and denoted $(\widehat{\sigma}_0^2, \widehat{\boldsymbol{\theta}}_0)$. Then, for a given level α of prediction intervals, they consider what they refer to as "Leave-One-Out Coverage Probability" and which corresponds to the $\hat{\Delta}(\alpha)$ given by Eq. (19). $\hat{\Delta}(\alpha)$ can be rewritten as a function of the quantiles of the GP predictive distribution, quantiles of level $(1-\alpha)/2$ and $(1+\alpha)/2$. So, to ensure a good value of $\hat{\Delta}(\alpha)$ (i.e. close to α), the authors propose to find two new sets of hyperparameters, respectively denoted $\left(\bar{\sigma}^2, \overline{\boldsymbol{\theta}}\right)$ and $(\underline{\sigma}^2, \underline{\boldsymbol{\theta}})$ which guarantee that the GP quantiles respectively of level $(1-\alpha)/2$ and $(1+\alpha)/2$ will yield $\hat{\Delta}(\alpha) = \alpha$. In addition to satisfying this condition, these two sets of hyperparameters are found to be as close as possible to the initial values $(\widehat{\sigma}_0^2, \widehat{\boldsymbol{\theta}}_0)$, in the sense of a similarity measure (see Appendix A for more details). Hence, the initial set $(\widehat{\sigma}_0^2, \widehat{\boldsymbol{\theta}}_0)$ is only used to build the GP predictor. $\left(\bar{\sigma}^2, \overline{\boldsymbol{\theta}}\right)$ and $(\underline{\sigma}^2, \underline{\boldsymbol{\theta}})$ (which can be viewed as corrections of the initial set) then yields two other GP metamodels which are used to compute the wto bounds (i.e. the lower and upper GP quantiles) corresponding to the α-prediction interval. The resulting method is called *Robust Prediction Intervals Estimation* (RPIE) by the authors.

The numerical tests proposed by the authors show that when the GP metamodel is well specified (good coverage probability of intervals obtained with initial MLE or CV-estimates $\widehat{\boldsymbol{\theta}}_0$), the RPIE method does not bring any added value. In the opposite case, the RPIE method is relevant and corrects efficiently the prediction intervals.

To the best of our knowledge, the RPIE method is the first to focus on a correction of the hyperparameters to control the quality of prediction intervals but, it seems perfectible on several points:

- ▶ First of all, the new sets of hyperparameter strongly depend on the set of initial values (estimated by MLE or CV), since the searched solutions are expressed as an isotropic shift of $\widehat{\boldsymbol{\theta}}_0$. In multidimensional case, the ratio between the different $\left(\widehat{\boldsymbol{\theta}}_{0,i}\right)_{i=1\ldots d}$ will thus be preserved, even if it was initially badly estimated. The method thus assumes that the initial estimation has been carried out correctly, under the assumption of a well-specified covariance.
- ▶ Secondly, the procedure proposed must be performed for the two bounds of any desired interval of level α, yielding two sets of corrected hyperparameters. RPIE does not look for a single hyperparameter correction that would provide a single GP that would both adjust the data well and provide reliable prediction intervals. Not having a single GP metamodel is not satisfactory in the perspective of using the GP metamodel as a multi-objective tool, i.e. to predict a quantile or a probability [16], estimate sensitivity indices [73, 74], perform an optimization [22], for inversion problem [75] or more generally to implement a SUR (Stepwise Uncertainty Reduction) approach [76]. These studies cannot be implemented directly from Acharki et al. [45]'s approach. We argue that it is simpler and preferable to have a single GP and to address the misspecification problem apart from the hyperparameter estimation, by testing and comparing different covariances.
- ▶ More fundamentally, it is not a correction of the GP predictive law but only of its quantiles (high and low) for a given level of prediction interval α It seems more interesting to us to correct the predictive law to simultaneously control $\hat{\Delta}(\alpha)$ whatever α.

- Finally, the numerical benchmark carried out by the authors is not exhaustive enough. Only 3 numerical examples are considered, all in dimension $d = 10$ and with a relatively large sample size $n = 600$. Moreover, only one Monte Carlo sample is drawn. More thorough tests appear necessary to evaluate the robustness of RPIE method to a poor estimation of the initial hyperparameters, to sampling variability, and to a smaller sample size (especially for the correction of high or very low quantiles).

5.2. A more efficient Bayesian approach

As mentioned in Section 3.4, the Bayesian approach used to estimate the GP hyperparameters relies on two key ingredients: first, the choice of a prior distribution $\pi\left(\sigma^2, \boldsymbol{\theta}\right)$ and second, the estimation of the posterior distribution and its propagation in the GP metamodel (see Eq.(12)).

5.2.1. Discussion on the choice of prior and focus on reference and Jeffreys priors

As highlighted by Muré [61], prior knowledge about the GP hyperparameters is often lacking. It then seems natural to use non-informative priors but these may fail to lead to a proper posterior (i.e., a distribution that integrates to a finite mass), this condition being necessary in our context of quantifying the uncertainty of GP parameters. Note that, for the 1-D or isotropic cases (i.e. $\theta \in \mathbb{R}$), proving that the reference posterior is proper amounts to finding appropriate upper bounds on the tail rates of $\pi\left(\theta \mid \mathbf{Y_s}\right)\pi(\theta)$ as $\theta \to 0$ and as $\theta \to +\infty$ (see Muré [61] for details).

With regard to this issue and still for the isotropic case, Berger et al. [54] first showed that among several prior distributions the reference prior of Bernardo [77] is the most satisfying default choice. Reference priors [78] aim to formalize the notion of "uninformative prior" and are defined so as to maximize a measure of distance or divergence between the posterior and prior, as data are observed (this choice allows the data to have maximum effect on the posterior estimates). Berger et al. [54] demonstrated that reference priors yield proper posterior for isotropic rough correlations that include the exponential correlation and the set of the Matérn family with smoothness parameter $\nu \geqslant 1$. Their demonstration notably relies on the fact that the correlation kernel cannot then be twice continuously differentiable at 0. Very recently, Muré [61] succeeds in extending this to a large class of smooth kernels, which includes the Gaussian and the Matérn family with smoothness parameter $\nu < 1$ [61, Theorem 4.4]. But this extension remains in the only case where the dimension of θ is equal to one (isotropic covariance). In a nutshell, the key behind Berger et al. [54] and Muré [61]'s demonstrations is that the reference prior should "compensate for the marginal likelihood" so that the integrated likelihood (i.e. posterior distribution) has the right decay rates on the distribution tails.

The extension of previous results to the anisotropic case ($\boldsymbol{\theta} \in \mathbb{R}^d$ with $d > 1$) is obtained by defining anisotropic correlations as products of one-dimensional rough correlations and a possible additional nugget effect [49]. The demonstration relies on the use of Jeffreys prior (obtained as the square root of the determinant of the Fisher information matrix) which is here a reference prior since the authors consider a separable product correlation function. Note therefore that this extension is not valid for anisotropic geometric covariances. Moreover, the one-dimensional correlation functions considered to build the tensorized covariance are again assumed to be rough. And, unfortunately, the proof used in Muré [61] to deal with smoother kernels cannot easily be adjusted to the much more complex prior considered by Gu et al. [49].

In summary, and to the best of our knowledge, there are only two configurations in which obtaining a proper posterior law $\pi\left(\boldsymbol{\theta} \mid \mathbf{Y_s}\right)$ from a reference prior $\pi(\boldsymbol{\theta})$ could be established:

- isotropic covariance ($\theta \in \mathbb{R}$) with the works of Berger et al. [54] for rough covariances (e.g. exponential, spherical or Matérn with $\nu < 1$) and Muré [61] for smooth covariances (e.g. Gaussian or Matérn with $\nu \geqslant 1$),
- anisotropic covariance defined from a product of one-dimensional correlation functions [49].

5.2.2. Focus on RobustGaSP method

Despite the fact that the demonstration of obtaining a proper posterior is not theoretically established for product of smooth covariances, the most relevant works in this context seem to be those of Gu et al. [49] which is referred to as the *RobustGaSP* method. So, to build their approach, the authors rely on the three following ingredients.

- A robust prior $\pi^R\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}\right)$ is assigned to the set of GP parameters (details in Appendix B). From this, a marginal likelihood can be deduced as a function of $\boldsymbol{\theta}$ alone. The marginal posterior distribution of $\boldsymbol{\theta}$ is then obtained.

- Sampling this marginal posterior distribution calls for the use of a Metropolis algorithm (MCMC sampling). Because of the cost of each likelihood evaluation (in $O\left(n^3\right)$) and the associated computational error which can be very large especially when the correlation matrix is close to the matrix $\mathbf{1}_n\mathbf{1}_n^T$, Gu et al. [49] do not advocate this method. Instead, they recommends to estimate (and consider) only the MAP (i.e. the mode) of the posterior distribution $\pi\left(\boldsymbol{\theta} \mid \mathbf{Y_s}\right)$. The Bayesian approach then stops at this step: the MAP is directly used to compute the GP predictive distribution in a so-called "*plug-in*" approach. This avoids a too much computationally prohibitive MCMC algorithm to estimate the posterior distribution of Eq. (12). In a nutshell, the reference prior distribution acts only as a penalizing factor on the likelihood and the Bayesian framework only ensures a robust estimation of the hyperparameters, hence the name *Robust* GaSP method (GaSP is the acronym for Gaussian Stochastic Process). Note that an alternative solution (not detailed here) is proposed by Muré, J. [79] to make the full-Bayesian procedure tractable: univariate conditional Jeffreys-rule posterior distributions and pseudo-Gibbs sampler are notably used.

- Finally, Gu et al. [49] consider some specific choices of reparameterization. Even if MLE is invariant under a reparameterization (injective transformation of GP hyperparameters), this is not the case of the MAP of posterior distribution because of the presence of the Jacobian for the prior. As a result, the authors consider other common ways of parameterizing the hyperparameters $\boldsymbol{\theta}$, in particular the inverse and log-inverse transformations.

From these elements, Gu et al. [49] establish two theorems related to the robustness of the estimated MAP, one theorem with nugget effect and another without. More precisely, under the assumption of the reference prior defined by the equation (B.2) with $a = 1$ and using the standard parametrization or the log-inverse reparametrisation, the authors demonstrate that the estimation of the MAP of the posterior distribution (Eq. (B.3)) is robust for the tensorized form of Matern, spherical and exponential correlation functions [49, Theorem 3.1]. Recall that the robustness defined by Gu et al. [49] refers to the two extreme cases described in Section 3.3 and leading to $\mathbf{R}_{\widehat{\theta}} = \mathbf{1}_n\mathbf{1}_n^T$ and $\mathbf{R}_{\widehat{\theta}} = \mathbf{I}_n$.

However, a serious drawback remains in RobustGaSP method: the computational cost required to compute the reference prior and, even more, to compute the mode of the posterior distribution $\pi\left(\boldsymbol{\theta} \mid \mathbf{Y_s}\right)$. Even if *in fine* the MAP estimate is used instead of the posterior sampling, the use of mode search algorithms, such as the quasi-Newton optimization method, typically relies on the information of the derivatives. Computing the derivative requires more evaluations of the likelihood and therefore more inversions of the correlation matrix. The total cost of the procedure then becomes prohibitive. To overcome this limitation, Gu [80] introduces an approximation of the reference prior $\pi^R\left(\boldsymbol{\beta_0}, \sigma^2, \boldsymbol{\theta}\right)$, which he calls the jointly robust prior. The author demonstrates that this prior retains the robustness property while being computationally simpler than the reference prior for the purpose of hyperparameter estimation.

Hence, the approach proposed by Gu et al. [49] combined with the approximation of the reference prior $\pi^R\left(\boldsymbol{\beta_0}\right)$ of Gu [80] seems to us very relevant and finally the most in line with the idea of having a more robust estimation of hyperparameters (rather than adopting an intractable full-Bayesian approach). Moreover, the proposed approximation of the reference prior allows to consider the application of the method even in the case of a large number of input variables. Finally, its availability in a dedicated software package [81] makes it the most interesting existing method. For all these reasons, RobustGaSP method will be compared with the new algorithm proposed in the companion paper [46].

6. Conclusion

The value of GP regression for emulating costly computational codes in the context of uncertainty management is well established, and explains why it is now widely used. Having a probabilistic metamodel, in the sense that it provides a predictive distribution for each new evaluation point, is of great added value, particularly for safety, reliability or risk assessment studies. It also enables the deployment of sophisticated GP-based approaches for active learning, robust optimization, reliability assessment, etc. In this context, it is essential to guarantee confidence in the GP predictive distribution, and not just in its mean value. This confidence requires, on the one hand, a reliable estimation of the GP metamodel and in particular of its hyperparameters, and, on the other hand, a rigorous validation of the entire GP predictive distribution.

The present paper has reviewed recent works dealing with the estimation of GP hyperparameters, from a theoretical and empirical point of view. It appears that the usual methods sometimes lead to poor-quality and not very robust estimates. MLE, the most widely used method, often leads to ill-posed problems. Although it leads in practice to good metamodel predictivity, the associated uncertainties and prediction intervals can be of poor quality. It is therefore essential to have validation indicators to detect this unreliability of the predictive distribution. Typically, it is insufficient to check only the GP's predictive capabilities: the accuracy of the entire GP predictive distribution needs to be assessed. To this end, we have reviewed the most relevant indicators and have proposed some derivatives. Emphasis has thus been put on GP validation that requires careful and informed consideration.

Concerning the estimation process, recent alternatives to standard estimation approaches have been explained. In particular, Bayesian approaches are theoretically very attractive, offering a kind of regularization of likelihood. However their cost in terms of complexity and required expertise, particularly in the definition of so-called robust priors and its tractability in large dimension (large number of inputs), refrain their use. Others approaches rely on ad-hoc corrections of the quantiles of the GP predictive distribution to ensure reliable prediction intervals for a given level, but these approaches do not necessarily seem relevant to our application context of multi-objective use of the metamodel.

In the companion paper [46], a new technique of using MLE for estimation is proposed, in particular by considering other criteria in the estimation procedure (criteria that until now have been reserved for the validation procedure). It also includes an intensive benchmark to test this new multi-objective optimization algorithm and compare its results with those obtained with other more standard estimation algorithms. The method is then applied on a real test case modeling an aquatic ecosystem, and used for environmental assessment.

Acknowledgments

This work was funded by the French ANR project SAMOURAI (ANR-20-CE46-0013).

Appendix A. Details on RPIE method

From $\hat{\Delta}(\alpha)$ given by Eq. (19), the authors introduce what they call the "quasi-Gaussian proportion" ψ_a to describe how close the a-quantile q_a of the standardized predictive distribution is to the level a (ideally, it should correspond to a). More precisely, ψ_a results from the rewriting of $\hat{\Delta}(\alpha)$ for GP as

$$\hat{\Delta}(\alpha) = \frac{1}{n}\sum_{i=1}^{n} h\left(q_{(1+\alpha)/2} - \frac{y_i - \hat{y}_{-i}}{\hat{s}_{-i}^2}\right) - \frac{1}{n}\sum_{i=1}^{n} h\left(q_{(1-\alpha)/2} - \frac{y_i - \hat{y}_{-i}}{\hat{s}_{-i}^2}\right),$$

where q_a denotes the $a-$quantile of the standard normal distribution and h is the Heaviside step function

$$h(x) = \mathbf{1}\{x \geq 0\} = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0. \end{cases}$$

Considering a nominal level of quantile a, ψ_a is then defined as a map from $[0, +\infty) \times (0, +\infty)^d$ to $[0, 1]$:

$$\psi_a\left(\sigma^2, \boldsymbol{\theta}\right) = \frac{1}{n}\sum_{i=1}^{n} h\left(q_a - \frac{y_i - \hat{y}_{-i}}{\hat{s}_{-i}^2}\right).$$

The objective is then to find the two sets of hyperparameters $(\sigma^2, \boldsymbol{\theta})$ such as to obtain the ideal values of ψ_a for the two bounds of the α predictive interval. These two pairs denoted $\left(\bar{\sigma}^2, \overline{\boldsymbol{\theta}}\right)$ and $(\underline{\sigma}^2, \underline{\boldsymbol{\theta}})$ are respectively defined by $\psi_{(1+\alpha)/2}\left(\bar{\sigma}^2, \overline{\boldsymbol{\theta}}\right) = (1+\alpha)/2$ and $\psi_{(1-\alpha)/2}\left(\underline{\sigma}^2, \underline{\boldsymbol{\theta}}\right) = (1-\alpha)/2$. These GP parameters allow to get the optimal value $\hat{\Delta}(\alpha) = \alpha$. The authors modify the function ψ_a into a new $\psi_a^{(\delta)}$ (with $\delta > 0$ to obtain one formulation for the optimization problem, whatever the value of α, and define the set of solution $\mathcal{A}_{a,\delta}$ w.r.t. $\hat{\Delta}(\alpha)$:

$$\mathcal{A}_{a,\delta} := \left\{\left(\sigma^2, \boldsymbol{\theta}\right) \in [0, +\infty) \times (0, +\infty)^d, \psi_a^{(\delta)}\left(\sigma^2, \boldsymbol{\theta}\right) = a\right\}.$$

Finally, as a correction of the initial estimated hyperparameters $(\widehat{\sigma}_0^2, \widehat{\boldsymbol{\theta}}_0)$, the authors propose to find the hyperparameters in $\mathcal{A}_{a,\delta}$ which are the closest to $(\widehat{\sigma}_0^2, \widehat{\boldsymbol{\theta}}_0)$ in the sense of a continuous similarity measure d_{sim} between the multivariate Gaussian distributions $\mathcal{N}(\mathbf{m}(\mathbf{X_s}), \mathbf{K})$ generated with the two sets of parameters. This results in the following optimization problem:

$$\operatorname{argmin}_{(\sigma^2,\theta)\in\mathcal{A}_{a,\delta}} d_{sim}^2\left(\left(\sigma^2, \boldsymbol{\theta}\right), \left(\widehat{\sigma}_0^2, \widehat{\boldsymbol{\theta}}_0\right)\right). \tag{A.1}$$

This problem should then be solved for $a = (1+\alpha)/2$ and $a = (1-\alpha)/2$ to obtain estimates of $\left(\bar{\sigma}^2, \overline{\boldsymbol{\theta}}\right)$ and $(\underline{\sigma}^2, \underline{\boldsymbol{\theta}})$, respectively.

However, two problems arise. First, the resolution of the problem (A.1) may be too costly and heavy to solve, especially as dimension d increases. Secondly, depending on the metric chosen, there is no guarantee that the barycenters of the two prediction intervals (generated from the two GP predictive distributions) are close. Similarly, there is no control over the sharpness of the prediction intervals obtained with the optimal solutions of Problem (A.1). This could result in a solution with a good coverage function but a poor Q^2, cf. Remark 1. To mitigate the second drawback, the authors first recommend the use of second Wasserstein distance W_2 for d_{sim}. Then, to address the two drawbacks, the authors propose a relaxed problem denoted $\mathcal{P}_\zeta$ where the optimal $\boldsymbol{\theta}$ is defined as shifted values of $\widehat{\boldsymbol{\theta}}_0$: $\zeta\widehat{\boldsymbol{\theta}}_0$ with $\zeta > 0$. Moreover, considering that σ^2 should be as small as possible to reduce the uncertainty of the GP predictions, the author define the optimal value of σ^2 w.r.t. λ as:

$$\forall\zeta \in (0, +\infty) : \sigma_{\text{opt}}^2(\zeta) := \min\left\{\sigma^2 \in [0, +\infty), \psi_a^{(\delta)}\left(\sigma^2, \zeta\boldsymbol{\theta}_0\right) = a\right\}. \tag{A.2}$$

The optimization problem is finally reformulated as a one-dimensional problem[1] $\mathcal{P}_\zeta$:

$$\mathcal{P}_\zeta : \quad \text{argmin}_{\zeta\in(0,+\infty)} \mathcal{L}(\zeta) := d^2\left(\left(\sigma^2_{\text{opt}}(\zeta), \zeta\widehat{\boldsymbol{\theta}}_0\right), \left(\widehat{\sigma}^2_0, \widehat{\boldsymbol{\theta}}_0\right)\right) \tag{A.3}$$

with $d^2\left(\left(\sigma^2,\boldsymbol{\theta}\right),\left(\widehat{\sigma}^2_0,\boldsymbol{\theta}_0\right)\right) = W_2^2\left(\mathcal{N}(\boldsymbol{m},\mathbf{K}),\mathcal{N}\left(\boldsymbol{m}_0,\mathbf{K}_0\right)\right)$. The purpose is to find the two solutions $\bar{\zeta}^*$ and $\underline{\zeta}^*$ of $\mathcal{P}_\zeta$, for $a = (1+\alpha)/2$ and $a = (1-\alpha)/2$ respectively. These two solutions yields two GP metamodels which are used to build the two bounds of the predictive interval whose coverage probability is demonstrated to be optimal. The resulting method is called *Robust Prediction Intervals Estimation* (RPIE) by the authors. The numerical tests propsed by the authors show that when the GP metamodel is well specified (good coverage probability of intervals obtained with initial MLE or CV-estimates $\widehat{\boldsymbol{\theta}}_0$), the RPIE method does not bring any added value. In the opposite case, the RPIE method is relevant and corrects efficiently the prediction intervals.

Appendix B. Details on prior and marginal posterior for the RobustGaSP method

To simplify, we suppose a constant GP mean $m(\mathbf{x}) = \beta_0$. The formulas generalized to the case of a q-dimensional vector of basis functions with parameter $\boldsymbol{\beta}\in\mathbb{R}^{q+1}$ for $m(\mathbf{x})$ are available in Gu et al. [49].

To define their "robust" prior, Gu et al. [49] first assign the objective prior for the regression and variance parameters:

$$\pi\left(\boldsymbol{\beta}_{\mathbf{0}},\sigma^2\right) \propto \frac{1}{\left(\sigma^2\right)^a}, \tag{B.1}$$

with $a > 0$. $a = 1$ corresponds to the standard reference prior. Then, the authors consider the Jeffrey's prior which is a reference prior for their parametric model (separable product of 1-D correlation functions as in Eq. (6)):

$$\pi^R\left(\boldsymbol{\beta}_{\mathbf{0}},\sigma^2,\boldsymbol{\theta}\right) \propto \frac{|\mathbf{I}(\boldsymbol{\theta})|^{1/2}}{\left(\sigma^2\right)^a}, \tag{B.2}$$

where $\mathbf{I}(\cdot)\in\mathbb{R}^{d\times d}$ is the expected Fisher information matrix as below,

$$\mathbf{I}(\boldsymbol{\theta}) = \begin{pmatrix} n-1 & \text{tr}\left(\mathbf{W}_1\right) & \text{tr}\left(\mathbf{W}_2\right) & \dots & \text{tr}\left(\mathbf{W}_d\right) \\ & \text{tr}\left(\mathbf{W}_1^2\right) & \text{tr}\left(\mathbf{W}_1\mathbf{W}_2\right) & \dots & \text{tr}\left(\mathbf{W}_1\mathbf{W}_d\right) \\ & & \text{tr}\left(\mathbf{W}_2^2\right) & \dots & \text{tr}\left(\mathbf{W}_2\mathbf{W}_d\right) \\ & & & \ddots & \vdots \\ & & & & \text{tr}\left(\mathbf{W}_d^2\right) \end{pmatrix},$$

where $\mathbf{W}_l = \dfrac{\delta\mathbf{R}_{\boldsymbol{\theta}}}{\delta\theta_l}\mathbf{R}_{\boldsymbol{\theta}}^{-1}\mathbf{P}_{\mathbf{R}_{\boldsymbol{\theta}}}$ with $\mathbf{P}_{\mathbf{R}_{\boldsymbol{\theta}}} = \mathbf{I}_n - \mathbf{1}_n\left\{\left(\mathbf{1}_n\right)^T\mathbf{R}_{\boldsymbol{\theta}}^{-1}\mathbf{1}_n\right\}^{-1}\left(\mathbf{1}_n\right)^T\mathbf{R}_{\boldsymbol{\theta}}^{-1}$, for $1\le l\le d$. The proofs of Eq. (B.2) and of the formula of $\mathbf{I}(\cdot)$ are given in Paulo [82] (precisely in Proposition 2.1 and Appendix A.0.2).

The prior on mean and variance parameter (Eq. (B.1)) allows to marginalize out these parameters in the likelihood function to obtain the marginal likelihood according to $\boldsymbol{\theta}$: $L\left(\mathbf{Y_s}\mid\boldsymbol{\theta}\right)$. The marginal posterior of $\boldsymbol{\theta}$ is then obtained by:

$$\pi\left(\boldsymbol{\theta}\mid\mathbf{Y_s}\right) \propto L\left(\mathbf{Y_s}\mid\boldsymbol{\theta}\right)|\mathbf{I}(\boldsymbol{\theta})|^{1/2} \tag{B.3}$$

where $L\left(\mathbf{Y_s}\mid\boldsymbol{\theta}\right) \propto |\mathbf{R}_{\boldsymbol{\theta}}|^{-\frac{1}{2}}\left|\left(\mathbf{1}_n\right)^T\mathbf{R}_{\boldsymbol{\theta}}^{-1}\mathbf{1}_n\right|^{-\frac{1}{2}}\left(S^2\right)^{-\left(\frac{n-1}{2}+a-1\right)}$ and $S^2 = \left(\mathbf{Y_s}\right)^T\mathbf{R}_{\boldsymbol{\theta}}^{-1}\mathbf{P}_{\mathbf{R}_{\boldsymbol{\theta}}}\mathbf{Y_s}$.

[1] Under some hypothesis on the GP trend, the authors demonstrate for all $\lambda > 0$, $H_\zeta(\lambda) := \left\{\sigma^2\in[0,+\infty), \psi_a^{(\delta)}\left(\sigma^2,\zeta\theta_0\right) = a\right\}$ is a non-empty and compact subset of $\mathbb{R}^+$. Assuming additional assumption of regularity of $H_\zeta(\lambda)$, they deduce that it provides the continuity of σ^2_{opt} on $(0,+\infty)$ and that $\mathcal{P}_\zeta$ admits at least one global minimizer ζ^* in $(0,+\infty)$.

References

[1] J. Sokolowski, C. Banks, Modeling and Simulation Fundamentals: Theoretical Underpinnings and Practical Domains, Wiley, 2010.

[2] E. de Rocquigny, N. Devictor, S. Tarantola (Eds.), Uncertainty in industrial practice, Wiley, 2008.

[3] R. Smith, Uncertainty quantification, SIAM, 2014.

[4] S. Da Veiga, F. Gamboa, B. Iooss, C. Prieur, Basics and Trends in Sensitivity Analysis. Theory and Practice in R, SIAM, 2021.

[5] C. Ciric, P. Ciffroy, S. Charles, Use of sensitivity analysis to discriminate non-influential and influential parameters within an aquatic ecosystem model, Ecological Modelling 246 (2012) 119–130.

[6] A. Marrel, N. Perot, C. Mottet, Development of a surrogate model and sensitivity analysis for spatio-temporal numerical simulators, Stochastic Environmental Research and Risk Assessment 29 (2015) 959–974.

[7] Y. Richet, V. Bacchi, Inversion algorithm for civil flood defense optimization: Application to two-dimensional numerical model of the Garonne river in France, Frontiers in Environmental Science 7 (2019) 160.

[8] K.-T. Fang, R. Li, A. Sudjianto, Design and Modeling for Computer Experiments, Chapman & Hall/CRC, 2006.

[9] N. Villa-Vialaneix, M. Follador, M. Ratto, A. Leip, A comparison of eight metamodeling techniques for the simulation of N2O fluxes and N leaching from corn crops, Environmental Modelling & Software 34 (2012) 51–66.

[10] S. Afshari, F. Enayatollahi, X. Xu, X. Liang, Machine learning-based methods in structural reliability: A review, Reliability Engineering & System Safety 219 (2022) 108223.

[11] T. Hastie, R. Tibshirani, J. Friedman, The elements of statistical learning, second ed., Springer, 2009.

[12] A. Forrester, A. Sobester, A. Keane (Eds.), Engineering design via surrogate modelling: a practical guide, Wiley, 2008.

[13] B. A. Khuwaileh, P. J. Turinsky, Surrogate based model calibration for pressurized water reactor physics calculations, Nuclear Engineering and Technology 49 (2017) 422–436.

[14] X. Wu, T. Kozlowski, H. Meidani, Kriging-based inverse uncertainty quantification of nuclear fuel performance code BISON fission gas release model using time series measurement data, Reliability Engineering and System Safety 169 (2018) 422–436.

[15] R. Christian, A. U. A. Shah, H. G. Kang, Dynamic PRA-Based Estimation of PWR Coping Time Using a Surrogate Model for Accident Tolerant Fuel, Nuclear Technology 207 (2021) 376–388.

[16] A. Marrel, B. Iooss, V. Chabridon, The ICSCREAM methodology: Identification of penalizing configurations in computer experiments using screening and metamodel – Applications in thermal-hydraulics, Nuclear Science and Engineering 196 (2022) 301–321.

[17] J.-P. Chilès, P. Delfiner, Geostatistics: Modeling spatial uncertainty, Wiley, New-York, 1999.

[18] J. Sacks, W. Welch, T. Mitchell, H. Wynn, Design and analysis of computer experiments, Statistical Science 4 (1989) 409–435.

[19] T. Santner, B. Williams, W. Notz, The Design and Analysis of Computer Experiments, Springer, 2003.

[20] C. Rasmussen, C. Williams, Gaussian Processes for Machine Learning, MIT Press, 2006.

[21] R. Ghanem, D. Higdon, H. Owhadi (Eds.), Springer Handbook on Uncertainty Quantification, Springer, 2017.

[22] D. Jones, M. Schonlau, W. Welch, Efficient global optimization of expensive black-box functions, Journal of Global Optimization 13 (1998) 455–492.

[23] J. N. Fuhg, A. Fau, U. Nackenhorst, State-of-the-art and comparative review of adaptive sampling methods for kriging, Archives of Computational Methods in Engineering 28 (2021) 2689–2747.

[24] M. Moustapha, S. Marelli, B. Sudret, Active learning for structural reliability: Survey, general framework and benchmark, Structural Safety 96 (2022) 102174.

[25] B. Echard, N. Gayton, M. Lemaire, Ak-mcs: An active learning reliability method combining kriging and monte carlo simulation, Structural Safety 33 (2011) 145–154.

[26] J. Bect, D. Ginsbourger, L. Li, V. Picheny, E. Vazquez, Sequential design of computer experiments for the estimation of a probability of failure, Statistics and Computing 22 (2012) 773–793.

[27] J. Betancourt, F. Bachoc, T. Klein, D. Idier, R. Pedreros, J. Rohmer, Gaussian process metamodeling of functional-input code for coastal flood hazard assessment, Reliability Engineering & System Safety 198 (2020) 106870.

[28] A. F. López-Lopera, D. Idier, J. Rohmer, F. Bachoc, Multioutput Gaussian processes with functional data: A study on coastal flood hazard assessment, Reliability Engineering & System Safety 218 (2022) 108139.

[29] G. Perrin, Adaptive calibration of a computer code with time-series output, Reliability Engineering & System Safety 196 (2020) 106728.

[30] B. Iooss, L. Le Gratiet, Uncertainty and sensitivity analysis of functional risk curves based on Gaussian processes, Reliability Engineering & System Safety 187 (2019) 58–66.

[31] N. Lu, Y.-F. Li, H.-Z. Huang, J. Mi, S. G. Niazi, AGP-MCS+D: An active learning reliability analysis method combining dependent Gaussian process and Monte Carlo simulation, Reliability Engineering & System Safety 240 (2023) 109541.

[32] S.-Y. Huang, S.-H. Zhang, L.-L. Liu, A new active learning kriging metamodel for structural system reliability analysis with multiple failure modes, Reliability Engineering & System Safety 228 (2022) 108761.

[33] J. Zhou, J. Li, IE-AK: A novel adaptive sampling strategy based on information entropy for kriging in metamodel-based reliability analysis, Reliability Engineering & System Safety 229 (2023) 108824.

[34] Z. Song, H. Zhang, Z. Liu, P. Zhu, A two-stage kriging estimation variance reduction method for efficient time-variant reliability-based design optimization, Reliability Engineering & System Safety 237 (2023) 109339.

[35] Y.-Z. Ma, X.-X. Jin, X.-L. Wu, C. Xu, H.-S. Li, Z.-Z. Zhao, Reliability-based design optimization using adaptive kriging-a single-loop strategy and a double-loop one, Reliability Engineering & System Safety 237 (2023) 109386.

[36] M. Ribaud, C. Blanchet-Scalliet, C. Helbert, F. Gillot, Robust optimization: A kriging-based multi-objective optimization approach, Reliability Engineering & System Safety 200 (2020) 106913.

[37] H. Zhang, Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics, Journal of the American Statistical Association 99 (2004) 250–261.

[38] F. Bachoc, Asymptotic analysis of maximum likelihood estimation of covariance parameters for gaussian processes: An introduction with proofs, in: A. Daouia, A. Ruiz-Gazen (Eds.), Advances in Contemporary Statistics and Econometrics: Festschrift in Honor of Christine Thomas-Agnan, Springer International Publishing, Cham, 2021, pp. 283–303.

[39] F. Bachoc, A. Lagnoux, T. M. N. Nguyen, Cross-validation estimation of covariance parameters under fixed-domain asymptotics, Journal of Multivariate Analysis 160 (2017) 42 – 67.

[40] J. Martin, T. Simpson, Use of kriging models to approximate deterministic computer models, AIAA Journal 43 (2005) 853–863.

[41] S. Sundararajan, S. S. Keerthi, Predictive Approaches for Choosing Hyperparameters in Gaussian Processes, Neural Computation 13 (2001) 1103–1118.

[42] F. Bachoc, Cross validation and maximum likelihood estimations of hyperparameters of Gaussian processes with model misspecification, Computational Statistics and Data Analysis 66 (2013) 55–69.

[43] S. Petit, J. Bect, P. Feliot, E. Vazquez, Model parameters in Gaussian process interpolation: an empirical study of selection criteria, 2023. URL: https://hal-centralesupelec.archives-ouvertes.fr/hal-03285513.

[44] C. Demay, B. Iooss, L. L. Gratiet, A. Marrel, Model selection for Gaussian process regression: an application with highlights on the model variance validation, Quality and Reliability Engineering International Journal 38 (2022) 1482–1500.

[45] N. Acharki, A. Bertoncello, J. Garnier, Robust prediction interval estimation for Gaussian processes by cross-validation method, Computational Statistics & Data Analysis 178 (2023) 107597.

[46] A. Marrel, B. Iooss, Probabilistic surrogate modeling by Gaussian process: A new estimation algorithm for more reliable prediction, Preprint (2023).

[47] R. B. Gramacy, Surrogates: Gaussian Process Modeling, Design and Optimization for the Applied Sciences, Chapman Hall/CRC, Boca Raton, Florida, 2020.

[48] T. Karvonen, C. J. Oates, Maximum likelihood estimation in Gaussian process regression is ill-posed, Journal of Machine Learning Research 24 (2023) 1–47.

[49] M. Gu, X. Wang, J. O. Berger, Robust Gaussian stochastic process emulation, The Annals of Statistics 46 (2018) 3038 – 3066.

[50] S. Petit, Improved Gaussian process modeling: Application to Bayesian optimization, Thèse de l'Université Paris-Saclay, 2022.

[51] O. Dubrule, Cross validation of kriging in a unique neighborhood, Journal of the International Association for Mathematical Geology 15 (1983) 687–699.

[52] M. Stein, Interpolation of spatial data, Springer, 1999.

[53] T. Karvonen, G. Wynne, F. Tronarp, C. Oates, S. Särkkä, Maximum likelihood estimation and uncertainty quantification for Gaussian process approximation of deterministic functions, SIAM/ASA Journal on Uncertainty Quantification 8 (2020) 926–958.

[54] J. Berger, V. De Oliveira, B. Sansó, Objective Bayesian analysis of spatially correlated data, Journal of the American Statistical Association 96 (2001) 1361–1374.

[55] M. Wieskotten, M. Crozet, B. Iooss, C. Lacaux, A. Marrel, A comparison between Bayesian and ordinary kriging based on validation criteria: application to radiological characterisation, Mathematical Geosciences (2023) DOI:10.1007/s11004–023–10072–y.

[56] C. Robert, G. Casella, Monte Carlo statistical methods, Springer, 2004.

[57] R. M. Neal, Priors for infinite networks, Technical Report, Departement of Computer Science - University of Toronto, Canada, 1994.

[58] S. Petit, J. Bect, S. da Veiga, P. Feliot, E. Vazquez, Towards new cross-validation-based estimators for Gaussian process regression: efficient adjoint computation of gradients, in: 52èmes Journées de Statistique de la SFdS (JdS 2020), Nice, France, 2021, pp. 633–638.

[59] H. Zhang, Y. Wang, Kriging and cross-validation for massive spatial data, Environmetrics 21 (2010) 290–304.

[60] R. Benassi, J. Bect, E. Vazquez, Robust Gaussian process-based global optimization using a fully Bayesian expected improvement criterion, in: 5th International Conference on Learning and Intelligent Optimization (LION 5), volume 6683 of *Lecture Notes in Computer Science*, Springer, Rome, Italy, 2011, pp. 176–190.

[61] J. Muré, Propriety of the reference posterior distribution in Gaussian process modeling, The Annals of Statistics 49 (2021) 2356 – 2377.

[62] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer, 2002.

[63] A. Marrel, B. Iooss, F. Van Dorpe, E. Volkova, An efficient methodology for modeling complex computer codes with Gaussian processes, Computational Statistics and Data Analysis 52 (2008) 4731–4744.

[64] E. Fekhari, B. Iooss, J. Muré, L. Pronzato, J. Rendas, Model predictivity assessment: incremental test-set selection and accuracy evaluation, in: N. Salvati, C. Perna, S. Marchetti, R. Chambers (Eds.), Studies in Theoretical and Applied Statistics, SIS 2021, Pisa, Italy, June 21-25, Springer, 2023, pp. 315–347.

[65] L. Bastos, A. O'Hagan, Diagnostics for Gaussian process emulators, Technometrics 51 (2009) 425–438.

[66] J. M. Bernardo, Expected Information as Expected Utility, The Annals of Statistics 7 (1979) 686–690.

[67] T. Gneiting, A. E. Raftery, A. Westveld, T. Goldman, Calibrated probabilistic forecasting using ensemble model output statistics and minimum crps estimation, Monthly Weather Review 133 (2005) 1098–1118.

[68] T. Gneiting, F. Balabdaoui, A. E. Raftery, Probabilistic forecasts, calibration and sharpness, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69 (2007) 243–268.

[69] L. W. Schruben, A coverage function for interval estimators of simulation response, Management Science 26 (1980) 18–27.

[70] A. Marrel, B. Iooss, S. Da Veiga, M. Ribatet, Global sensitivity analysis of stochastic computer models with joint metamodels, Statistics and Computing 22 (2012) 833–847.

[71] B. Iooss, A. Marrel, Advanced methodology for uncertainty propagation in computer experiments with large number of inputs, Nuclear Technology 205 (2019) 1588–1606.

[72] V. Picheny, T. Wagner, D. Ginsbourger, A benchmark of kriging-based infill criteria for noisy optimization, Structural and Multidisciplinary Optimization 48 (2013) 607–626.

[73] A. Marrel, B. Iooss, B. Laurent, O. Roustant, Calculations of the Sobol indices for the Gaussian process metamodel, Reliability Engineering & System Safety 94 (2009) 742–751.

[74] L. Le Gratiet, C. Cannamela, B. Iooss, A Bayesian approach for global sensitivity analysis of (multifidelity) computer codes, SIAM/ASA Journal on Uncertainty Quantification 2 (2014) 336–363.

[75] R. Ait Abdelmalek-Lomenech, J. Bect, V. Chabridon, E. Vazquez, Bayesian sequential design of computer experiments to estimate reliable sets, 2023. URL: https://hal-centralesupelec.archives-ouvertes.fr/hal-03835704, preprint.

[76] C. Chevalier, J. Bect, D. Ginsbourger, V. Picheny, Y. Richet, E. Vazquez, Fast kriging-based stepwise uncertainty reduction with application to the identification of an excursion set, Technometrics 56 (2014) 455–465.

[77] J. M. Bernardo, Reference analysis, in: D. Dey, C. Rao (Eds.), Bayesian Thinking, volume 25 of *Handbook of Statistics*, Elsevier, 2005, pp. 17–90.

[78] J. Berger, J. Bernardo, D. Sun, The formal definition of reference priors, The Annals of Statistics 37 (2009) 905 – 938.

[79] Muré, J., Optimal compromise between incompatible conditional probability distributions, with application to objective Bayesian kriging, ESAIM: PS 23 (2019) 271–309.

[80] M. Gu, Jointly Robust Prior for Gaussian Stochastic Process in Emulation, Calibration and Variable Selection, Bayesian Analysis 14 (2019) 857 – 885.

[81] M. Gu, J. Palomo, J. Berger, RobustGaSP: Robust Gaussian Stochastic Process Emulation, 2022. URL: https://CRAN.R-project.org/package=RobustGaSP, R package version 0.6.5.

[82] R. Paulo, Default priors for Gaussian processes, The Annals of Statistics 33 (2005) 556 – 582.