Reliability-based design optimization using stochastic emulators: current developments and perspectives

Jaad Bel Houari-Durand[†,1], Maliki Moustapha[§,1], Bruno Sudret[§,1]

[†] PhD student (presenting author). [§] PhD supervisor

PhD expected duration: Nov. 2025 – Oct. 2029

[1] Chair of Risk, Safety and Uncertainty Quantification, ETH Zürich, Stefano-Franscini-Platz 5, 8093 Zurich, Switzerland

jaad.bel.houari-durand@ibk.baug.ethz.ch
moustapha@ibk.baug.ethz.ch
sudret@ethz.ch

Abstract

Reliability-based design optimization (RBDO) aims at optimizing engineering systems under uncertainty while enforcing probabilistic safety constraints. In practical applications, the computational cost of RBDO remains a major bottleneck, as reliability assessments must be performed throughout the optimization process. This difficulty is further exacerbated in the presence of high-dimensional uncertainty and expensive-to-evaluate computational models.

Computational models used in RBDO are in general deterministic. In this work, we reformulate the RBDO problem using stochastic simulators. Unlike deterministic models, whose response is fully determined by the input parameters, stochastic simulators generate random outputs even when evaluated multiple times with the same input parameters. This behavior reflects the presence of intrinsic stochasticity that cannot be controlled or directly observed, and which is instead driven by latent random variables. As a result, the simulator output must be understood as a random quantity.

In this contribution, we reformulate the RBDO problem using *stochastic emulators*, which are surrogate models designed to approximate the conditional output distributions of stochastic simulators. More precisely, the input parameters are decomposed into deterministic design variables and random variables. The emulator is then built in the space of design variables, while the effect of random variables is lumped as a stochastic simulator. This formulation implicitly reduces dimensionality by capturing the influence of high-dimensional uncertainty through a latent representation, while simultaneously providing direct access to conditional response distributions. As a result, reliability measures such as failure probabilities or quantiles can be evaluated in a semi-analytical manner, avoiding the repeated use of classical Monte Carlo simulation or First/Second Order Reliability Method (FORM/SORM) procedures within the optimization loop.

Building on an existing proof of concept relying on stochastic polynomial chaos expansions (SPCE) [2] and generalized lambda models (GLaM) [1], we analyze the performance of stochastic emulators as a scalable solution for high-dimensional RBDO. Numerical benchmarks illustrate the potential of these models to significantly reduce computational cost by several orders of magnitude, while maintaining accurate reliability estimates, even for low failure probabilities.

The focus of the ongoing PhD work is to further improve the robustness and efficiency of this approach. Current research directions include (i) improving emulator accuracy through sparse regression techniques, with particular emphasis on accurately capturing the tails of conditional response distributions, and (ii) developing active learning strategies that adaptively enrich the surrogate's experimental design in regions most relevant to reliability assessment and optimization.

Short biography (PhD student)

Jaad Bel Houari-Durand is a PhD student at the Chair of Risk, Safety and Uncertainty Quantification (ETH Zurich) within the ORACLES project. His research focuses on uncertainty quantification and reliability-based design optimization, with emphasis on stochastic surrogate modeling and active learning. He holds a dual degree in applied mathematics from ENSIIE and Université Paris-Saclay and worked at Saint-Gobain Research Paris on data-efficient inverse design under uncertainty.

References

[1] X. Zhu and B. Sudret. Emulation of stochastic simulators using generalized lambda models. *SIAM/ASA Journal on Uncertainty Quantification*, 9(4):1345–1380, 2021.

[2] X. Zhu and B. Sudret. Stochastic polynomial chaos expansions to emulate stochastic simulators. *International Journal for Uncertainty Quantification*, 13(2):31–52, 2023.

Computational Bayesian Optimal Sensor Placement for Ocean Models: A Majorize-Then-Optimize Strategy

M. Doumbouya [†,1,2], A. Vidard[§,2], O. Zahm[§,2]

[†] PhD student (presenting author). [§] PhD supervisor

PhD expected duration: Oct. 2025 – Sep. 2028

[1] Univ. Grenoble Alpes, LJK
mohamed.doumbouya@univ-grenoble-alpes.fr

[2] Inria, AIRSEA Team, Centre Grenoble Alpes
{arthur.vidard,olivier.zahm}@inria.fr

Abstract

Optimal experimental design (OED) addresses a fundamental question: where should one observe a given system in order to maximize the information gained about it? This question is particularly critical in ocean modeling. Observations are costly to acquire, while our ability to understand and predict large-scale geophysical processes strongly depends on how and where data are collected.

Classical approaches to Bayesian optimal experimental design (BOED) are well established for linear models with Gaussian priors and observation operators [1, 8]. However, when moving toward realistic operational ocean models, several challenges arise simultaneously. These include strong nonlinearities, non-Gaussian features, and the high computational cost associated with forward simulations, all of which make standard BOED formulations difficult to apply in practice.

Recent gradient-based strategies have emerged as a promising alternative. Instead of minimizing the Expected Information Gain (EIG) directly, these approaches rely on optimizing a computable bound on the EIG [3]. Although approximate, this bound can be evaluated and differentiated at a much lower cost, thereby significantly reducing the number of expensive model runs required during optimization. In the linear–Gaussian setting, variance-based criteria such as A-optimality and D-optimality, combined with low-rank structure in the prior-to-posterior update [7], further enhance computational efficiency.

The objective of this project is to investigate key numerical aspects of this gradient-based BOED framework in the context of ocean modeling. In particular, we focus on three complementary research directions:

(i) Theoretical validation and acceleration. We systematically compare designs obtained from bound-based objectives with those computed using conventional EIG-based approaches [3, 1]. In addition, we explore the use of bound-based solutions as initializations or preconditioners for full BOED optimization. This majorize-then-minimize strategy aims to retain the accuracy of EIG-driven designs while benefiting from the computational savings of surrogate objectives.

(ii) Large-scale computational efficiency. To make the approach scalable, we employ tools from randomized numerical linear algebra, including randomized SVD and low-rank posterior updates [6, 7]. These methods take advantage of the fast spectral decay and low intrinsic

dimensionality commonly observed in geophysical inverse problems [1, 4], and are well suited to problems involving tens of thousands of discretized parameters.

(iii) Integration of realistic constraints. Finally, we incorporate physical, technical, and financial constraints directly into the sensor placement problem [5]. This step is essential to ensure that the resulting designs are not only statistically efficient, but also feasible within real operational oceanographic settings.

Overall, this work lies at the intersection of Bayesian statistics, inverse problems for PDEs, and high-performance scientific computing. While motivated by oceanographic applications, the proposed methodology is more broadly applicable to large-scale parameter estimation problems governed by expensive forward models, where gradient-based BOED and low-rank randomized techniques can provide substantial computational gains [2, 6].

Short biography (PhD student)

I am a PhD student in Applied Mathematics at Université Grenoble Alpes, and a member of the AIRSEA research team at Inria Grenoble. My research focuses on computational methods for Bayesian optimal experimental design, with a particular emphasis on sensor placement strategies for ocean models and large-scale inverse problems.

I am especially interested in the development of numerical and statistical tools that make BOED tractable for complex geophysical systems, where forward simulations are expensive and the parameter space is high dimensional. My work combines ideas from numerical linear algebra, optimization, and uncertainty quantification.

I hold a degree in Applied Mathematics and Modeling from Polytech Lyon, as well as a post-master's degree in High Performance Computing from Mines Paris – PSL. This training in scientific computing and HPC provides a strong foundation for addressing the computational challenges encountered in ocean and climate modeling, particularly in the context of large-scale inference and data assimilation.

References

[1] A. Alexanderian. Optimal experimental design for infinite-dimensional Bayesian inverse problems governed by PDEs: A review. *SIAM Review*, 63(3):307–403, 2021.

[2] J. M. Bardsley and K. Solna. Randomized sampling-based model order reduction for structural dynamics applications. *International Journal for Numerical Methods in Engineering*, 102(5):1243–1263, 2015.

[3] Q. Chen, E. Arnaud, R. Baptista, and O. Zahm. Coupled input-output dimension reduction: Application to goal-oriented bayesian experimental design and global sensitivity analysis. *SIAM Journal on Scientific Computing*, 47(5):A2403, 2025.

[4] J. Gao and P. Chen. Accurate, scalable, and efficient Bayesian optimal experimental design with derivative-informed neural operators. *arXiv preprint*, 2024.

[5] J. P. Kaipio, E. Somersalo, et al. Goal-oriented optimal experimental design for large-scale Bayesian linear inverse problems. Bibliographic details to be completed.

[6] P.-G. Martinsson and J. A. Tropp. Randomized numerical linear algebra: foundations and algorithms. *Acta Numerica*, 29:403–572, 2020.

[7] A. Spantini, A. Solonen, T. Cui, J. Martin, L. Tenorio, and Y. Marzouk. Optimal low-rank approximations of Bayesian linear inverse problems. *arXiv preprint*, 2015.

[8] A. M. Stuart. Inverse problems: a Bayesian perspective. *Acta Numerica*, 19:451–559, 2010.

Conformal Predictors for Polynomial Chaos Expansions

Arthur Hatstatt[†,1], Xujia Zhu[§,2], Bruno Sudret[§,1]

[†] PhD student (presenting author) [§] PhD supervisor

PhD expected duration: Jun. 2025 – Jun. 2029

[1] ETH Zürich, Chair of Risk, Safety and Uncertainty Quantification,
Stefano-Franscini-Platz 5, 8093 Zürich, Switzerland
arthur.hatstatt@ibk.baug.ethz.ch
sudret@ethz.ch

[2] Université Paris-Saclay, CNRS,
CentraleSupélec, Laboratoire de Sigaux et Systèmes,
3 rue Joliot Curie, 91190 Gif-sur-Yvette, France
xujia.zhu@l2s.centralesupelec.fr

Abstract

Polynomial chaos expansions (PCEs) are widely used surrogate models in uncertainty quantification due to their strong approximation properties and favorable computational efficiency. They are particularly attractive in applications where repeated model evaluations are required, such as sensitivity analysis, reliability assessment, or uncertainty propagation. Despite their widespread use, the quantification of local predictive uncertainty in PCE-based surrogates remains an open and insufficiently addressed problem. Most existing approaches focus on global error metrics or rely on resampling techniques, which do not provide rigorous guarantees at individual prediction points.

Bootstrap resampling is commonly employed to approximate local prediction uncertainty in PCEs, especially in active learning contexts [3]. While easy to implement, bootstrap-based intervals lack finite-sample coverage guarantees and can perform poorly when the available training data are limited. This limitation is critical in many engineering and scientific applications, where data acquisition is expensive and surrogate models are often trained on small experimental designs. In such settings, there is a clear need for uncertainty quantification methods that provide statistically valid prediction intervals with minimal assumptions on the underlying surrogate model.

Conformal prediction [4] has recently emerged as a powerful framework for uncertainty quantification in machine learning. It offers distribution-free, finite-sample guarantees on prediction interval coverage under minimal assumptions, typically exchangeability of the data. Its model-agnostic nature makes it particularly appealing for surrogate modeling, as it can be combined with a wide range of regression techniques without altering the underlying predictor. Despite these advantages, the integration of conformal prediction with PCE-based surrogates has received little attention so far.

This contribution investigates the application of conformal prediction to polynomial chaos expansions, with a focus on both full and sparse PCE formulations. Two conformal prediction strategies are considered: the full conformal method [4] and the Jackknife+ approach [1]. Both methods are adapted to the specific structure of PCE regression and evaluated in terms of statistical validity, calibration, and computational efficiency.

For full PCEs, which rely on standard least-squares regression, the structure of the regression problem allows for significant computational simplifications. These are exploited to reduce the computational cost for both full conformal and Jackknife+ prediction intervals, without compromising their theoretical guarantees.

Sparse PCEs introduce additional challenges. Their construction relies on non-symmetric regression algorithms, which violate assumptions commonly used in conformal prediction. Naive extension of the conformal methods for full PCEs in this context leads to invalid or poorly calibrated prediction intervals. To address this issue, appropriate modifications to the construction procedure are introduced, restoring the validity of the conformal framework while maintaining computational efficiency [2].

The proposed methods are assessed through numerical experiments on benchmark functions commonly used in surrogate modeling. The results show that conformal prediction yields well-calibrated prediction intervals for both full and sparse PCEs. Compared to bootstrap-based approaches, the conformal prediction intervals achieve coverage levels much closer to the targets. The Jackknife+ method tends to produce slightly conservative intervals, whereas the full conformal approach provides tighter intervals with coverage that more closely matches the prescribed confidence level.

From a practical perspective, the choice between full conformal and Jackknife+ depends on the application requirements. The full conformal method is better suited for risk-critical scenarios where reliable, locally adaptive prediction intervals are required at specific input locations, especially in regions of sparse data. This increased accuracy comes at a higher computational cost, which may limit its applicability in large-scale problems. The Jackknife+ method offers a more computationally efficient alternative, providing robust coverage guarantees and a useful global characterization of predictive uncertainty across the input space.

Overall, this work demonstrates that conformal prediction provides a principled and effective framework for enhancing uncertainty quantification in PCE-based surrogate models. It bridges an important gap between the efficiency of polynomial chaos methods and the need for statistically sound local uncertainty estimates.

Short biography (PhD student)

Arthur Hatstatt obtained his Bachelor's degree in Civil Engineering from EPF Lausanne and his Master's degree from ETH Zürich. He joined the Chair of Risk, Safety and Uncertainty Quantification to investigate the integration of conformal prediction with polynomial chaos expansions. He later became a member of the *ORACLES*[1] project, which focuses on developing optimal and data-efficient strategies for calibrating stochastic emulators for uncertainty quantification and optimization.

References

[1] R. Barber, E. Candès, A. Ramdas, and R. Tibshirani. Predictive inference with the Jackknife+. *The Annals of Statistics*, 49(1):486–507, 2021.

[2] J. Lei. Fast exact conformalization of the lasso using piecewise linear homotopy. *Biometrika*, 16(4):749–764, 2019.

[3] S. Marelli and B. Sudret. An active-learning algorithm that combines sparse polynomial chaos expansions and bootstrap for structural reliability analysis. *Structural Safety*, 75:67–74, 2018.

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

[1] `https://data.snf.ch/grants/grant/10004826`

Reducing dimensionality of MCMC methods for hierarchical Bayesian inference using stochastic emulators

A. Herr[†,1], S. Marelli[§,1], B. Sudret[§,1],

[†] PhD student (presenting author). [§] PhD supervisor

PhD expected duration: Oct. 2025 – Sep. 2029

[1] Chair of Risk, Safety & Uncertainty Quantification, ETH Zürich
{anherr,marellis,sudret}@ethz.ch

Abstract

Hierarchical Bayesian inference aims to infer the population-level hyperparameters $\boldsymbol{\theta}_{\boldsymbol{X}}$, which determine joint distribution of the input parameters of a computational model $\mathcal{M}$. A hierarchical (multilevel) model can be defined as a system composed of deterministic and stochastic submodels, which jointly represent the physical behavior of the system and the uncertainties inherent to it [1]. The forward model of the system is given as:

$$\mathcal{M}: \ (\boldsymbol{m}, \boldsymbol{x}, \boldsymbol{\zeta}, \boldsymbol{d}) \mapsto \tilde{\boldsymbol{y}} = \mathcal{M}(\boldsymbol{m}, \boldsymbol{x}, \boldsymbol{\zeta}, \boldsymbol{d}) \tag{1}$$

where $\tilde{\boldsymbol{y}}$ are predictions of some observable response $\boldsymbol{y}$ as a function of the input parameters $\{\boldsymbol{m}, \boldsymbol{x}, \boldsymbol{\zeta}, \boldsymbol{d}\}$. Here, $\boldsymbol{d}$ represents known experimental conditions, $\boldsymbol{m}$ denotes fixed but unknown constants (epistemic uncertainty), and $\boldsymbol{\zeta}$ have unknown realizations $\boldsymbol{\zeta}_i$ drawn from a *known distribution* $f_{\boldsymbol{Z}}(\boldsymbol{\zeta}; \boldsymbol{\theta}_{\boldsymbol{Z}})$ (aleatory uncertainty). The parameters $\boldsymbol{x}$ have unknown realizations $\boldsymbol{x}_i$ drawn from a population distribution governed by *unknown hyperparameters* $\boldsymbol{\theta}_{\boldsymbol{X}}$.

Prior information on all unknown quantities is encoded through the joint prior distribution $\pi(\boldsymbol{m}, \boldsymbol{x}, \boldsymbol{\zeta}, \boldsymbol{\theta}_{\boldsymbol{X}})$. This information is then conditioned on available experimental observations through the joint likelihood distribution $f_E(\boldsymbol{y}_i - \mathcal{M}(\boldsymbol{m}, \boldsymbol{x}_i, \boldsymbol{\zeta}_i, \boldsymbol{d}_i); \boldsymbol{\Sigma}_i)$, where $\boldsymbol{\Sigma}_i$ is a set of model discrepancy parameters that can represent both experimental noise and model error, often a covariance matrix in case of assumed Gaussian noise. This yields the joint posterior distribution:

$$\pi(\boldsymbol{m}, \boldsymbol{x}, \boldsymbol{\zeta}, \boldsymbol{\theta}_{\boldsymbol{X}} \mid \boldsymbol{y}) \propto \left(\prod_{i=1}^{n} f_E(\boldsymbol{y}_i - \mathcal{M}(\boldsymbol{m}, \boldsymbol{x}_i, \boldsymbol{\zeta}_i, \boldsymbol{d}_i); \boldsymbol{\Sigma}_i) \right) \pi(\boldsymbol{m}, \boldsymbol{x}, \boldsymbol{\zeta}, \boldsymbol{\theta}_{\boldsymbol{X}}). \tag{2}$$

To infer the hyperparameters $\boldsymbol{\theta}_{\boldsymbol{X}}$, the posterior distribution is typically sampled through Markov-Chain Monte Carlo (MCMC), using one of two main approaches. The first involves sampling the joint posterior distribution over $\boldsymbol{x}$ and $\boldsymbol{\theta}_{\boldsymbol{X}}$ by updating the full set of (hyper-)parameters $(\boldsymbol{m}, \boldsymbol{x}, \boldsymbol{\zeta}, \boldsymbol{\theta}_{\boldsymbol{X}})$ in a single MCMC scheme. In the second approach, inference is performed using a double-loop MCMC scheme: the outer loop samples the hyperparameters of interest $\boldsymbol{\theta}_{\boldsymbol{X}}$, while an inner loop samples $\boldsymbol{x} \sim f_{\boldsymbol{X}|\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x} \mid \boldsymbol{\theta}_{\boldsymbol{X}})$ to calculate the marginal likelihood $f_E(\boldsymbol{y} \mid \boldsymbol{\theta}_{\boldsymbol{X}})$. Both approaches suffer from high computational costs, particularly for for high-dimensional or expensive forward models. This can be in principle mitigated by using classical surrogate models $\widehat{\mathcal{M}}^d(\boldsymbol{m}, \boldsymbol{x}_i, \boldsymbol{\zeta}_i, \boldsymbol{d}_i)$, at the cost of constructing accurate surrogates over the entire input parameter space, a task that can become intractable in high-dimensional settings [2].

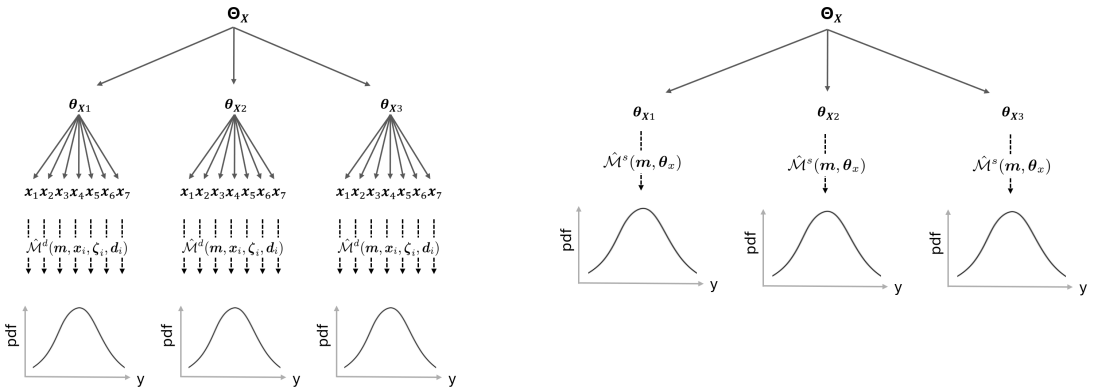


Figure 1: Schematic representation of the inference framework using deterministic (left) vs. stochastic (right) emulators.

In this contribution, we exploit recent developments in stochastic emulators to eliminate the inner MCMC loop and directly emulate the *conditional distribution of model outputs* as a function of the hyperparameters $\boldsymbol{\theta}_X$. Unlike deterministic surrogates, which approximate the input–output mapping at the level of individual parameters, stochastic emulators (such as generalized lambda models [3]) aim at approximating the stochastic model response conditioned on a set of explicit parameters, in this case $\widehat{\mathcal{M}}^s(\boldsymbol{m}, \boldsymbol{\theta_X}) \approx f_E(\boldsymbol{y}, \boldsymbol{m}, \boldsymbol{\theta_X})$. In other words, they capture the population-level response induced by variability in the latent parameters $\boldsymbol{X}$, effectively marginalizing over them. The conceptual difference between the deterministic and stochastic inference frameworks is illustrated in Figure 1.

This stochastic emulator-based approach enables direct calculation of the marginal likelihood $f_E(\boldsymbol{y} \mid \boldsymbol{\theta_X})$ and thereby substantially reduces the dimensionality and overall complexity of the problem. This framework makes previously intractable hierarchical Bayesian inference problems computationally accessible, particularly in settings where inference targets population-level parameters, or where the computational forward model is intrinsically stochastic.

Short biography (PhD student)

Anna received her Bachelor's degree in Natural Sciences and Master's degree in Physics from the University of Cambridge. She has gained research experience through internships at ETH and the University of Oxford. Her PhD is part of the ORACLES project (Optimization, Reliability And Calibration using Emulators of Stochastic computational models), which is funded by the Swiss National Science Foundation.

References

[1] J. B. Nagel and B. Sudret. A unified framework for multilevel uncertainty quantification in Bayesian inverse problems. *Probabilistic Engineering Mechanics*, 43:68–84, 2016.

[2] Y. M. Marzouk and H. N. Najm. Dimensionality reduction and polynomial chaos acceleration of Bayesian inference in inverse problems. *Journal of Computational Physics*, 228(6):1862–1902, 2009.

[3] X. Zhu and B Sudret. Emulation of stochastic simulators using generalized lambda models. *SIAM/ASA Journal on Uncertainty Quantification*, 9(4):1345–1380, 2021.

Multi-Fidelity Gaussian Processes for Time-Series Prediction with application to Tire Manufacturing

U. Labbé[†,1,2], J. Garnier[§,2], M. Binois[§,3], A. Chorfi[§,1], M. Hernandez[§,1]

[†] PhD student (presenting author). [§] PhD supervisor

PhD expected duration: Oct. 2025 – Sep. 2028

[1] Manufacture de Pneu Michelin, 63000 Clermont-Ferrand, France
{ugo.labbe,amina.chorfi,mayra.hernandez}@michelin.com

[2] Centre de Mathématiques Appliquées, Ecole polytechnique, Institut Polytechnique de Paris, 91120 Palaiseau, France
josselin.garnier@polytechnique.edu

[3] Université Côte d'Azur, Inria, CNRS, LJAD, 06902 Sophia Antipolis, France
mickael.binois@inria.fr

Abstract

Accurate forecasting and uncertainty quantification (UQ) of time series are of extreme importance for process control and safety in tire manufacturing. These processes rely on strict thermodynamic constraints, where deviations can significantly impact product quality and energy efficiency.

In this context, High-Fidelity (HF) measurements are often scarce due to acquisition costs or availability, while Low-Fidelity (LF) simulation data is abundant but imperfect. Multi-Fidelity Gaussian Process (MFGP) frameworks can leverage this cheap data to enhance predictions as well as giving accurate uncertainty quantifications [1, 2, 4, 5] .

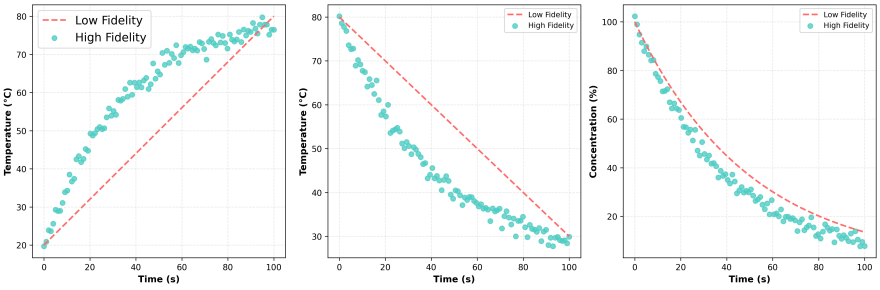


Figure 1: Time Series with High-Fidelity noisy measurements and Low-Fidelity simulations

Common MF formulation assumes an autoregressive relationship of the form

$$f_{high}(\mathbf{x}) = \rho(\mathbf{x}) \cdot f_{low}(\mathbf{x}) + \delta(\mathbf{x})$$

where the HF response $f_{high}(\mathbf{x})$ is approximated by scaling the low-fidelity Gaussian Process posterior $f_{low}(\mathbf{x})$ by $\rho(\mathbf{x})$ and adding a bias term $\delta(\mathbf{x})$ modeled by a GP.

However, this formulation inherently assumes a static, linear correlation between fidelities. Consequently, it often proves insufficient for dynamic industrial systems, where the discrepancy between simulation and physical reality is non-stationary and evolves over time.

To overcome these limitations, we propose a time-aware Multi-Fidelity Gaussian Process framework specifically designed for time-series modeling by decomposing the input space $\mathbf{x} = [t, \mathbf{z}]$, where t represents the temporal dimension and $\mathbf{z}$ represents the physical state variables.

Our approach generalizes the standard autoregressive formulation by introducing time-dependent correlation coefficients $\rho(t, \mathbf{z})$ and bias terms $\delta(t, \mathbf{z})$, enabling the model to capture non-stationary and evolving fidelity relationships.

To capture global trends and decouple the temporal dimension, we also investigate flexible time-scaled covariance structures [3, 6] using a product kernel formulation:

$$k(\mathbf{x}, \mathbf{x}') = k(t, t') \times k(\mathbf{z}, \mathbf{z}')$$

We highlight the efficiency of these methods, demonstrating improvements in both prediction accuracy ($RMSE$ and Q^2) and the quality of the confidence interval (IAE_α) compared to standard static-MF approaches, while remaining computationally efficient.

Short biography (PhD student)

Ugo Labbé is a CIFRE PhD Student at Michelin and the CMAP at Ecole Polytechnique and INRIA Côte d'Azur and is directed by Prof. Josselin Garnier. He obtained a Master's degree in Machine Learning and Data Mining from Université Jean Monnet, his end study internship was done in the R&D department at Airbus. He also obtained a Bachelor's degree in Statistics from Oregon State University.

References

[1] Nils Baillie, Baptiste Kerleguer, Cyril Feau, and Josselin Garnier. Efficient multi-fidelity Gaussian process regression for noisy outputs and non-nested experimental designs, November 2025. arXiv:2511.20183.

[2] Loïc Brevault, Mathieu Balesdent, and Ali Hebbal. Overview of Gaussian process based multi-fidelity techniques with variable relationship between fidelities, application to aerospace systems. *Aerospace Science and Technology*, 107:106339, December 2020.

[3] David Duvenaud, James Lloyd, Roger Grosse, Joshua Tenenbaum, and Ghahramani Zoubin. Structure discovery in nonparametric regression through compositional kernel search. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1166–1174, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

[4] M. C. Kennedy and A. O'Hagan. Predicting the Output from a Complex Computer Code When Fast Approximations Are Available. *Biometrika*, 87(1):1–13, 2000. Publisher: [Oxford University Press, Biometrika Trust].

[5] Loic Le Gratiet. *Multi-fidelity Gaussian process regression for computer experiments*. Theses, Université Paris-Diderot - Paris VII, October 2013.

[6] Andrew Wilson and Ryan Adams. Gaussian process kernels for pattern discovery and extrapolation. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1067–1075, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

Large-dimensional reliability-oriented Shapley effects estimation with Normalizing Flows

L. Monteiro[†,1,2,3], J. Morio[§,2], J. Demange-Chryst[§,2], F. Bachoc[§,4]

[†] PhD student (presenting author). [§] PhD supervisor

PhD expected duration: Jan. 2025 – Jan. 2028

[1] Institut de Mathématiques de Toulouse, UMR5219 CNRS, 31062 Toulouse, France
lucas.monteiro@math.univ-toulouse.fr

[2] ONERA/DTIS, Université de Toulouse, F-31055 Toulouse, France
{lucas.monteiro, jerome.morio, julien.demange-chryst}@onera.fr

[3] ANITI, Toulouse, France

[4] Laboratoire Paul Painlevé UMR8524 CNRS, Université de Lille
francois.bachoc@univ-lille.fr

Abstract

We will consider a numerical code modeled by a function $\phi : \mathbb{R}^d \to \mathbb{R}$, considered as black-box and deterministic, costly to evaluate and with no regularity assumptions. The inputs are modeled by the d-dimensional random vector $\mathbf{X} = (X_1, \dots, X_d)$ with values in $\mathcal{X} \subset \mathbb{R}^d$ and d being moderately large, typically between 10 and 30. Moreover $\mathbf{X}$ is supposed to be absolutely continuous with probability density function $f_{\mathbf{X}}$ and with no independence assumption. The output $Y = \phi(\mathbf{X})$ is a random variable with values in $\mathbb{R}$ and is supposed to be square integrable.

The reliability analysis of a black-box numerical model consists in estimating the failure probability. Without loss of generality, the failure event is defined by $\{\phi(\mathbf{X}) > t\}$ with $t \in \mathbb{R}$ and its associated probability is $p_t := \mathbb{P}(\phi(\mathbf{X}) > t) = \mathbb{E}[\mathbf{1}_{F_t}(\mathbf{X})]$, where $F_t = \{x \in \mathbb{R}^d \mid \phi(x) > t\}$ is the failure domain. When $p_t \ll 1$, more suited methods than crude Monte Carlo can be used developed such as importance sampling, subset sampling, moving particle, FORM/SORM or surrogate-based procedures. From these estimation schemes, it is possible to recover failing samples, meaning samples $\widetilde{\mathbf{X}}^{(n)}$ of $\mathbf{X}$ satisfying $\phi(\widetilde{\mathbf{X}}^{(n)}) > t$ and distributed according to the conditional density $f_{\mathbf{X}|F_t}$, defined by $f_{\mathbf{X}|F_t}(x) := f_{\mathbf{X}}(x)\mathbf{1}_{F_t}(x)/p_t$.

When studying the failure of a system, the reliability analysis brings no information on why the failure occurs and how the uncertainty in the input variables is related to the failure. We hence are interested in a sensitivity analysis performed on $\mathbf{1}_{F_t}(\mathbf{X})$. As the inputs may be correlated, the target Sobol indices loose their interpretability power. To overcome this limitation, one approach is to consider the target Shapley effects, which are defined using the target closed Sobol indices [2]. The existing estimation schemes are based either on Monte Carlo and require too much calls to the model ϕ when $p_t \ll 1$ [2], or on importance sampling to overcome this previous limitation [1]. However, this last method suffers from the curse of dimensionality as it relies on a nearest-neighbor approximation. As a consequence, we present a new estimation scheme based on normalizing flows to estimate the target Shapley effects when the dimension exceeds 10.

The proposed methodology is as follows. After having obtained an estimate of p_t and recovered failing samples $(\widetilde{\mathbf{X}}^{(n)})$, we first propose a rewriting of the closed target Sobol indices following

the same principle as in [4]. These rewritings may involve conditional densities of large dimension, which we propose to estimate with normalizing flows [3] using the failing samples. These conditional densities being estimated, we then estimate the target closed Sobol indices with crude Monte Carlo. Finally, suited aggregation procedures are used to avoid the complexity due to the dimension. As a result, this methodology allows to obtain estimates of target Shapley effects without additional call to ϕ than those used to estimate p_t. In addition to this estimation scheme based on normalizing flows, we propose a procedure to quantify the error made on the estimation, also without additional call to ϕ. This procedure allows to take into account the error made by the normalizing flows, by the estimation of the target closed Sobol indices and by the aggregation procedure. Promising results have been obtained for a Gaussian linear case in dimension 15, which are displayed in Figure 1 below.

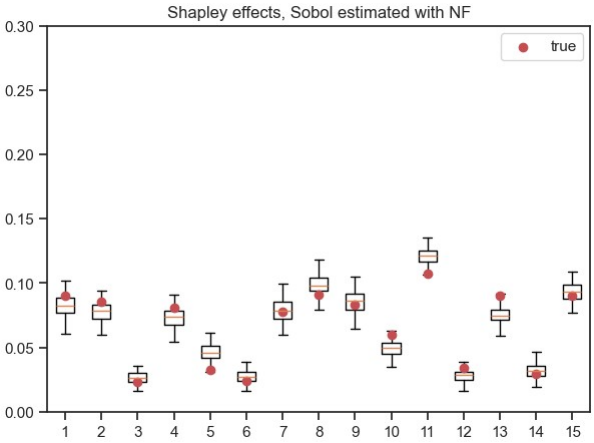


Figure 1: Estimation of target Shapley effects for the Gaussian Linear case, $d = 15$, with reference values, taking into account all sources of error.

Short biography (PhD student)

Lucas Monteiro is a doctoral student who earned a Licence in Economics followed by a Master in Statistics and Econometrics at TSE. He previously worked for Airbus in Toulouse. He is mainly located at ONERA, Toulouse with interactions at the Institut de Mathématiques de Toulouse. His PhD is funded half by ONERA and half by the AI institute ANITI.

References

[1] Julien Demange-Chryst, François Bachoc, and Jérôme Morio. Shapley effect estimation in reliability-oriented sensitivity analysis with correlated inputs by importance sampling. *International Journal for Uncertainty Quantification*, 13(3):1–37, 2023.

[2] Marouane Il Idrissi, Vincent Chabridon, and Bertrand Iooss. Developments and applications of shapley effects to reliability-oriented sensitivity analysis with correlated inputs. *Environmental Modelling & Software*, 143:105115, 2021.

[3] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

[4] Guillaume Perrin and Gilles Defaux. Efficient evaluation of reliability-oriented sensitivity indices. *Journal of Scientific Computing*, 79:1433–1455, 2019.

Optimization Under Uncertainties for Multi-Fidelity Black-Box Simulators

G. Pierron[†,1,2], M. R. El Amri[✠,2], C. Helbert[§,1], G. Perrin[§,3], D. Sinoquet[✠,2]

[†] PhD student (presenting author). [§] PhD supervisor. [✠] PhD advisor.

PhD expected duration: Oct. 2025 – Sep. 2028

[1] Institut Camille Jordan, Centrale Lyon
{celine.helbert, guillaume.pierron}@ec-lyon.fr

[2] IFP Énergies Nouvelles
{mohamed-reda.el-amri, guillaume.pierron, delphine.sinoquet}@ifpen.fr

[3] Université Gustave Eiffel
guillaume.perrin@univ-eiffel.fr

Abstract

We consider a complex physical system $\mathcal{S}$, depending on a controllable parameter $\mathbf{x} \in \mathcal{D}_X \subset \mathbb{R}^{d_X}$ and on an uncontrollable parameter $\mathbf{u} \in \mathcal{D}_U \subset \mathbb{R}^{d_U}$. We consider that $\mathbf{u}$ is an output of a random variable $\mathbf{U}$, defined on a universe $(\Omega, \mathcal{A}, \mathbb{P})$ and with support equal to D_U.

We want to optimize the system $\mathcal{S}$, by tuning the parameter $\mathbf{x}$. More precisely, we consider a feature $F^{(\infty)}(\mathbf{x}, \mathbf{u})$ of our system, and we aim to solve the optimization problem (1).

$$\arg\min_{\mathbf{x} \in \mathcal{D}_\mathbf{x}} \mathbb{E}\left[F^{(\infty)}(\mathbf{x}, \mathbf{U})\right] \tag{1}$$

For a given $(\mathbf{x}, \mathbf{U})$, we cannot directly compute $F^{(\infty)}(\mathbf{x}, \mathbf{U})$ as long as it depends on the system $\mathcal{S}$, whose physical equations are too complex to solve in closed form. Thus we must approximate our function of interest $F^{(\infty)}$ through a set of computer codes $\left(F^{(1)}, \cdots, F^{(L)}\right)$, so that, for $\ell \in \{1, \cdots, L\}$, the code $F^{(\ell)}$ simulates $F^{(\infty)}$ with the fidelity level ℓ, with the associated computational cost w_ℓ. We assume that $w_1 < \cdots < w_L$ (the lower the fidelity, the cheaper the simulator evaluation).

The practical optimization problem that we will try to solve is the following (2).

$$\arg\min_{\mathbf{x} \in \mathcal{D}_\mathbf{x}} \mathbb{E}\left[F^{(L)}(\mathbf{x}, \mathbf{U})\right] \tag{2}$$

An existing algorithm to address such a problem is a combination of two intricated loops :

- an outer optimization loop, which relies on the Efficient Global Optimization (EGO, see [2]) with objective function $f^{(L)} = \mathbb{E}\left[F^{(L)}(\cdot, \mathbf{U})\right]$, using a metamodel that can integrate uncertainties on each new observation (as we never can compute the exact value of the desired expectation).
- an inner estimation loop, which aims to estimate $f^{(L)}(\mathbf{x})$ for a new point $\mathbf{x}$, reducing the uncertainty of the estimation by evaluating many times the (costly) simulator $F^{(L)}$.

This methodology is very costly. The goal of this thesis is to learn how to use the low-fidelity codes, in order to have a better control on the uncertainties with the same computational expense for each iteration and to have better optimization results with the same total cost.

Several strategies could allow us to make use of these lower-fidelity simulators.
One possible way could be the modelization of all the fidelity levels, taking into account the difference of costs, in a mutli-fidelity meatmodel (similar to that introduced in the PhD thesis of Loic Le Gratiet [3]) on which it could be possible to build specific acquisition functions.

Another approach keeps stucking to the only modelization of $F^{(L)}$, but uses the low-fidelity codes as control variates [1]. The first part of my PhD work is to explore this approach. An existing elegant method, named Multi-Fidelity Monte-Carlo (MFMC), and developed in [4], allows to minimize the variance of the estimator of our objective function, with a fixed cost budget, by tuning a control-variate coefficent and an allocation of the budget to the different fidelity levels. We want to plug this method in our bayesian optimization loop.

This existing method assumes that we already know at least some coefficients of the covariance matrix of the random vector $\left(F^{(\ell)}(\mathbf{x}, \mathbf{U})\right)$ for any design point $\mathbf{x}$. We will try to move beyond this strong hypothesis, and we aim to see if it is still possible to obtain better optimization results than with the simple Monte-Carlo method on the high-fidelity simulator.

Short biography (PhD student)

My background is a master's degree on statistics and probability. My PhD thesis is partly funded by the CIROQUO consortium, which is a partnership between researchers and industrial companies to address issues raised by the high costs of data in optimization and uncertainty quantification. I am an employee of IFP Énergies Nouvelles for 3 years, and I am also enrolled as a PhD student in Centrale Lyon.

References

[1] Zdravko Botev and Ad Ridder. Variance Reduction. In *Wiley StatsRef: Statistics Reference Online*, pages 1–6. John Wiley & Sons, Ltd, Chichester, England, UK, November 2017.

[2] Donald R. Jones, Matthias Schonlau, and William J. Welch. Efficient Global Optimization of Expensive Black-Box Functions. *J. Global Optim.*, 13(4):455–492, December 1998.

[3] Loic Le Gratiet. *Multi-fidelity Gaussian process regression for computer experiments*. PhD thesis, Université Paris-Diderot - Paris VII, Paris, France, October 2013.

[4] B. Peherstorfer, K. Willcox, and M. Gunzburger. Optimal Model Management for Multifidelity Monte Carlo Estimation. *SIAM J. Sci. Comput.*, 2016.

Environmental impact minimization of computer experiments for aircraft design

G. Plat[†,1,2,3], P. Saves[§,4], N. Bartoli[§,2,3], T. Lefebvre[§,2,3], J. Morlier[§,5]

[†] PhD student (presenting author). [§] PhD supervisor

PhD expected duration: Oct. 2025 – Sep. 2028

[1] ISAE-SUPAERO, DMSM, Université de Toulouse, France
{gaston.plat, joseph.morlier}@isae-supaero.fr

[2] ONERA, DTIS, Université de Toulouse, France
{gaston.plat, nathalie.bartoli, thierry.lefebvre}@onera.fr

[3] Fédération ENAC ISAE-SUPAERO ONERA, Université de Toulouse, 31000, Toulouse, France

[4] IRIT, UMR 5505 CNRS, Université Toulouse Capitole
paul.saves@irit.fr

[5] ICA, Université de Toulouse, ISAE-SUPAERO, MINES ALBI, UPS, INSA, CNRS, France

Abstract

Expensive-to-evaluate blackbox simulations play a key role for many engineering and industrial applications. In this context, surrogate models have been widely used to address a large range of applications, *e.g.*, aircraft design [6], deep neural networks [5], coastal flooding prediction, agriculture forecasting, or seismic imaging. Most blackbox simulations are complex and computationally expensive. Typically, multidisciplinary design of aircraft leads to calling solvers, such as computational fluid dynamics or finite element, that can take days to be executed across clusters. As a result, there has been a growing interest in more efficient surrogate models, particularly in the context of Bayesian optimization based on Gaussian processes.

This work follows the PhD thesis of Paul Saves (2020-2023) in which high-dimensional multidisciplinary design optimization methods were developed for aircraft eco-design. Contributions on dimension reduction and mixed-variable techniques [8] for Gaussian processes were added to the open source software SMT (https://smt.readthedocs.io/en/latest/) and applied on both large sets of industrial or academic tests [7].

While mixed-categorical and hierarchical variables [9] enable more complex modeling, they also significantly increase computational overhead. To mitigate the resulting increase in execution time, parallel computing has emerged to be the most effective solution. Consequently, high-performance computing is now of main interest, even more in the machine learning field. Large-scale tasks can require energy-intensive exascale systems infrastructure, such as the european supercomputer Jupiter, whereas smaller-scale operations may be more efficiently executed on regional clusters or standard workstations. As such, we evaluate an energy consumption minimization problem by distributing workloads across a network of computers and their cores, using numerical simulations with various hyperparameter configurations. This problem can be formulated as a System Architecture Optimization problem [1] subject to an environmental budget threshold. Additionnaly, we plan to develop an acquisition function to be explicitly aware of the remaining budget, in a "non-myopic" look-ahead approach [2].

Existing works around hyperparameter optimization minimizing energy consumption of neural network's training have already shown that energy can be saved while keeping the same accuracy [10]. The eco-design approach will be further improved by integrating spatial and time [4] dependencies into the network of available computers. The targeted application of this PhD will be the eco-design of High Altitude Long Endurance [3] drone by including the architectural choice of the computational infrastructure in the overall process.

Short biography (PhD student)

The PhD is funded by the ONERA - ISAE - ENAC joint resarch group. The subject became part of my interests as I have graduated a MEng in Aeronautical Engineering at the Arts et Métiers Institute of technology following several internships in the aircraft design industry.

References

[1] Jasper H. Bussemaker, Paul Saves, Nathalie Bartoli, Thierry Lefebvre, and Rémi Lafage. System architecture optimization strategies: dealing with expensive hierarchical problems. *Journal of Global Optimization*, 91(4):851–895, April 2025.

[2] Francesco Di Fiore and Laura Mainini. NM2-BO: Non-Myopic Multifidelity Bayesian Optimization. *Knowledge-Based Systems*, 299:111959, September 2024.

[3] Edouard Duriez, Víctor Manuel Guadaño Martín, and Joseph Morlier. CO2 footprint minimization of solar-powered HALE using MDO and eco-material selection. *Scientific Reports*, 13(1):11994, July 2023. Publisher: Nature Publishing Group.

[4] Nicolas Gonel, Paul Saves, and Joseph Morlier. Frequency-aware Surrogate Modeling With SMT Kernels For Advanced Data Forecasting. In *Proceedings of ECCOMAS AeroBest 2025*, Lisbonne, Portugal, April 2025. Instituto Superior Técnico of the University of Lisbon.

[5] Edward Hallé-Hannan, Charles Audet, Youssef Diouane, Sébastien Le Digabel, and Paul Saves. A distance for mixed-variable and hierarchical domains with meta variables. *Neurocomputing*, 653:131208, 2025.

[6] Rémy Priem, Hugo Gagnon, Ian Chittick, Stephane Dufresne, Youssef Diouane, and Nathalie Bartoli. An efficient application of Bayesian optimization to an industrial MDO framework for aircraft design. In *AIAA AVIATION 2020 FORUM*, June 2020.

[7] Paul Saves, Nathalie Bartoli, Youssef Diouane, Thierry Lefebvre, Joseph Morlier, Christophe David, Eric Nguyen Van, and Sébastien Defoort. Constrained bayesian optimization over mixed categorical variables, with application to aircraft design. In *AeroBest 2021*, Lisbonne, Portugal, July 2021.

[8] Paul Saves, Youssef Diouane, Nathalie Bartoli, Thierry Lefebvre, and Joseph Morlier. A mixed-categorical correlation kernel for Gaussian process. *Neurocomputing*, 550, September 2023.

[9] Paul Saves, Edward Hallé-Hannan, Jasper Bussemaker, Youssef Diouane, and Nathalie Bartoli. Modeling hierarchical spaces: A review and unified framework for surrogate-based architecture design. *Structural and Multidisciplinary Optimization*, 2026.

[10] Dimitrios Stamoulis, Ermao Cai, Da-Cheng Juan, and Diana Marculescu. Hyperpower: Power-and memory-constrained hyper-parameter optimization for neural networks. In *2018 Design, Automation & Test in Europe Conference & Exhibition*, pages 19–24. IEEE, 2018.

Data-Driven Discovery of Dimensionless Numbers and Governing Laws: A Scaling Technique for Nuclear Test Facility Design

C. Razaire[†,1,2], A. Marrel[§,1,2], B. Iooss[§,3,4], S. Renaudière de Vaux[§,1], J. Cardolaccia[§,1]

[†] PhD student (presenting author). [§] PhD supervisor

PhD expected duration: Oct. 2025 – Sep. 2027

[1] CEA, DES, IRESNE, Cadarache
{camille.razaire, amandine.marrel, sebastien.renaudieredevaux, jerome.cardolaccia}@cea.fr

[2] Avignon Université, LMA, UPR 2151, Avignon

[3] EDF R&D, Chatou
{bertrand.iooss}@edf.fr

[4] SINCLAIR AI Lab, Palaiseau

Abstract

Designs of water-cooled reactors including new thermal-hydraulics systems undergo safety analysis during the licensing process. For economic reasons and safety concerns, systems are firstly tested on reduced scale test facilities. Appropriate scaling guarantees that safety-relevant phenomena are accurately represented in experiments. However, conventional equation-based scaling techniques rely heavily on expert knowledge of the governing equations, which constitutes a major limitation.

As an alternative, this work proposes a data-driven method to design nuclear test facilities, called D4NL for Data-Driven Discovery of Dimensionless Numbers and Governing Law. From a dataset, D4NL method aims to identify a governing law that describes the dominant safety-related phenomenon. The law is a function of dimensionless numbers, which are physically meaningful combinations of dimensional[1] variables identified by the algorithm. Dimensionless numbers are commonly used in scaling techniques, as they benefit from scale-invariant properties. They can be interpreted as dimensionless ratios of characteristic times, forces, or energies, and thus highlight the relative competition between different effects in a same phenomenon [6]. More generally, the D4NL method is part of the dimensionless numbers learning methods [2, 5, 1, 7] and adopts an approach conceptually similar to that of Xie et al. [7], while introducing significant methodological extensions.

In short, the D4NL method objective is to find, from a sample of physical dimensional variables, a set of constrained parameters that fully defines the nondimensional governing law. The law is assumed to be a polynomial function of a single dominant dimensionless number–an assumption that will be relaxed in future work. The constraints ensure both the physical interpretability of the law and the dimensionless nature of its terms. Two variants of the method are proposed.

[1] There are only seven fundamental dimensions in physics: L (length), T (time), M (mass), Θ (temperature), I (electric current), N (amount of matter), and J (luminous intensity). If one of the variable is a velocity v, its dimension, noted $[\cdot]$ in physics, is consequently $[v] = \mathrm{L}^1\mathrm{T}^{-1}$.

In the first variant, the sample includes dimensional inputs (like a pipe diameter or the density of the fluid in the pipe) and a known dimensionless output. The optimal set of parameters defining the law is obtained by minimizing a constrained loss function through a gradient-descent-based algorithm. To enhance robustness against small datasets (common in the nuclear context due to the high costs of experiments) and to mitigate the risk of convergence to local minima, the loss is estimated by K-Fold Cross-Validation and minimized via a multistart process based on space-filling-design methods. As a proof of concept, our method is tested on a noisy simulated dataset, which is representative of the single-phase natural circulation established in a passive heat removal system (a system studied for advanced nuclear reactors designs [4]). The method yields satisfactory results as it recovers the dominant dimensionless number and the governing law from the natural circulation physical model. In addition, a bootstrap approach is proposed to estimate confidence intervals on the parameters identified.

The second variant extends the D4NL method by no longer assuming that the dimensionless output is known. Instead, the algorithm estimates it jointly with the input dimensionless number and the associated governing law. This leads to a nested optimization relying on the D4NL architecture. Functional analysis based on variance decomposition and Sobol' indices [3] is incorporated into the algorithm to guide the construction of the output dimensionless number and ensure that the discovered law remains physically meaningful with respect to the phenomenon of interest. This D4NL extension is tested on similar natural circulation data: it identifies several governing laws that depend on variants of the physical model input and output.

Short biography (PhD student)

After preparatory classes, I obtained an energy and nuclear engineering degree from Phelma (Grenoble-INP school). I then pursued my cursus with a PhD at CEA Cadarache, in applied mathematics for nuclear engineering and thermal-hydraulics, under the supervision of Amandine Marrel and Bertrand Iooss. The thesis is funded by CEA and linked to EDF R&D.

References

[1] Joseph Bakarji, Jared Callaham, Steven L. Brunton, and J. Nathan Kutz. Dimensionally consistent learning with buckingham pi, 2022.

[2] Paul G. Constantine, Zachary del Rosario, and Gianluca Iaccarino. Data-driven dimensional analysis: algorithms for unique and relevant dimensionless groups, 2017.

[3] Sébastien Da Veiga, Fabrice Gamboa, Bertrand Iooss, and Clémentine Prieur. *Basics and Trends in Sensitivity Analysis*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2021.

[4] IAEA. Passive Safety Systems and Natural Circulation in Water Cooled Nuclear Power Plants. Technical Report 1624, IAEA, 2009.

[5] Lluís Jofre, Zachary R. Del Rosario, and Gianluca Iaccarino. Data-driven dimensional analysis of heat transfer in irradiated particle-laden turbulent flow. *International Journal of Multiphase Flow*, 125:103198, 2020.

[6] Paul S. Lykoudis. Non-dimensional numbers as ratios of characteristic times. *International Journal of Heat and Mass Transfer*, 33(7):1568–1570, 1990.

[7] Xiaoyu Xie, Arash Samaei, Jiachen Guo, Wing Kam Liu, and Zhengtao Gan. Data-driven discovery of dimensionless numbers and governing laws from scarce measurements. *Nature Communications*, 13(1):7562, 2022.

Toward Stochastic Digital Twins of Large-Scale Environments for Radar Sensor Design in EMPRISE software

Saou Aurélien Houret Thomas Bonnet Pierre
PhD expected duration: Jan.2026 – Dec.2028
ONERA AID
Université de Clermont-Ferrand

Abstract

Nowadays, simulators are ubiquitous in various research and industrial sectors, enabling exploration, prediction, model validation, cost and risk reduction, etc... The field of electromagnetics is no exception, allowing engineers to optimize parameters and configuration before physically testing them. These simulators rely on the use of digital twins of complex environments (terrain, vegetation, vehicles, etc.). At ONERA we have a software: EMPRISE [1] that act like a radar simulator and simulate electromagnetic scene. However, the reliability of the predictions provided by digital twins depends on their representativeness, which by design involves a certain degree of uncertainty. Therefore, given that reliability and accuracy are critical, it is essential to quantify uncertainties in EM environment and to move beyond deterministic digital twin models toward stochastic ones. The difficulty with EM environments is the plurality of uncertainties and their highly heterogeneous nature. Environmental uncertainties such as changes in vegetation, the presence or absence of vehicles or humidity and temperature, which will impact how materials reflect radar waves. There are also uncertainties due to measurements, such as noises and imperfect georeferencing. It should be noted that most of these uncertainties are highly correlated spatially and sometimes very difficult to observe directly. Radar scenes are therefore complex environments, which multiplies the possible uncertainties. Given the issues raised by determinist digital twins, it is reasonable to ask how to build and exploit digital twins of electromagnetic environments whose uncertainties are explicitly represented in order to produce more reliable and interpretable RADAR simulations.

The initial data input step is itself subject to uncertainties [3], as the data are only valid at a specific point in time; therefore, a deterministic approach is not the most appropriate. The idea is therefore to find a probabilistic but accurate representation, notably by incorporating stochastic processes into the selection of parameters or spatial variability. The question is also to determine and quantify the extent to which uncertainties in our digital twin impact simulated radar measurements. Imagine we have taken measure of Toulon environment thanks to SETHI [2], we can use EMPRISE to simulate the environment and then do a simulation measurement comparison, which give us an accurate measure of representativeness of our product. But what if we want to track a drone in an environment for which we have no data bases of real measure? How can we estimate how reliable it is; for practical use we must be able to measure the trustworthiness without comparison to simulation.

To conclude, my thesis lies at the interface between statistics, uncertainty measurements, and electromagnetic physics. Above all, we are seeking an unified approach to uncertainty measurements [4] that we will apply to realistic EM environments. Although EM is our field of application, the idea is to make methodological contributions that can be replicated in other environments Accordingly, this project aims to improve the reliability

of RADAR simulations used for decision support, as well as to increase the robustness of digital twins as their adoption continues to grow. It's a real challenge in the field of uncertainty due to the complexity of the simulated environments and the imperfect nature of the data used.

Short biography (PhD student)

My name is Aurelien Saou, after a very classical mathematical background with a bachelor's degree in mathematics followed by master's degree in applied mathematics, I chose to do a thesis in order to fully understand the world of research. However, as I value the practical side of mathematics, I naturally turned to this thesis founded by both AID and ONERA to contribute to our EMPRISE software in radar simulation. My goal is to be able to quantify the reliability of the simulation. We just started in January so we're still in an exploratory stage

[1] https://www.emprise-em.fr/

[2] "Sethi : Review Of 10 Years Of Development And Experimentation Of The Remote Sensing Platform" Rémi Baqué, Philippe Dreuillet, Hélène Oriot https://hal.science/hal-02502425/document

[3] Kessels, B.M., Fey, R.H.B. & van de Wouw, N. Uncertainty quantification in real-time parameter updating for digital twins using Bayesian inverse mapping models. *Nonlinear Dyn* **113**, 7613–7637 (2025)

[4] Daniel P. Thunnissen, "Uncertainty Classification for the Design and Development of Complex Systems", Proceedings of the 3rd Annual Predictive Methods Conference, Veros Software, Santa Ana, CA, 2003

Hybrid Models for seismic hazard quantification

I. Seydi[†,1,2], S. Donnet[§,1], M. Keller[§,2], J. Muré[§,2], J. Stoehr[§,1,3]

[†] PhD student (presenting author). [§] PhD supervisor

PhD expected duration: May 2025 – May 2028

[1] MIA Paris-Saclay, Université Paris-Saclay, AgroParisTech, INRAE
first.last@agroparistech.fr ; first.last@inrae.fr

[2] Électricité de France, EDF R&D Lab Chatou
first.last@edf.fr

[3] CEREMADE, Université Paris Dauphine-PSL, CNRS
last@ceremade.dauphine.fr

Abstract

Probabilistic Seismic Hazard Analysis (PSHA) traditionally relies on seismotectonic zoning models based on homogeneous Poisson processes [2]. While robust and conceptually transparent, these models struggle to capture the heterogeneous, uncertain, and spatially variable nature of seismicity [3]. Particularly in low-seismicity contexts such as metropolitan France, where limited data, catalog incompleteness, and diffuse earthquake patterns challenge classical assumptions. Zoneless approaches such as Kernel Density Estimation (KDE) attempted to address these shortcomings, yet they remain limited by their frequentist rigidity, sensitivity to bandwidth choices, and reduced effectiveness in sparse-data environments. These limitations highlight the need for a probabilistic framework capable of flexibly adapting to data while integrating decades of geological and geophysical expertise.

A promising direction arises from **spatio-temporal Hawkes processes** [1], which model seismicity as a combination of background activity and clustered aftershock sequences through a conditional intensity

$$\lambda(t,x,y) \;=\; \mu(x,y) \;+\; \sum_{i\,:\,t_i<t} g(t-t_i, x-x_i, y-y_i).$$

This formulation captures well-established empirical laws such as Omori decay and Gutenberg–Richter scaling and provides a coherent structure for distinguishing mainshocks from aftershocks. However, standard Hawkes models rely on **fixed parametric forms** for both the background rate $\mu(x)$ and the triggering kernel $g(\cdot)$. Such rigid specifications constrain their ability to reflect complex spatial structures or accommodate the large epistemic uncertainties inherent to low-activity catalogs. As a result, parameter estimates may become unstable or overly dependent on arbitrary modeling choices.

This motivates a **Bayesian nonparametric approach**, which offers several decisive advantages. First, by modeling the spatial background intensity $\mu(x)$ with flexible priors — such as Dirichlet Processes [5], Gaussian Processes [4] — the model's complexity adapts directly to the data, avoiding the need to predefine the number or geometry of seismic sources. Second,

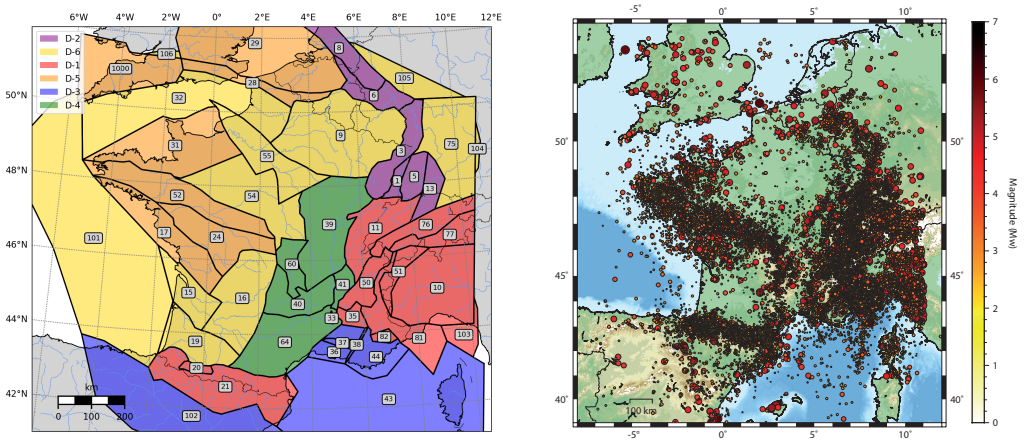


Figure 1: (Left) A **seismotectonic model** of France proposed by EDF. (Right) A **French earthquake catalog** for the PSHA.

Bayesian inference provides a principled mechanism for incorporating **informative priors derived from traditional zoning**, thereby embedding expert knowledge within a fully probabilistic learning framework. This allows geological structures, fault systems, and domain knowledge accumulated by EDF and the scientific community to guide inference without enforcing rigid parametric constraints. Third, the Bayesian paradigm naturally quantifies **both epistemic and aleatory uncertainties**, enabling richer risk assessments and transparent propagation of uncertainties into PSHA outputs. Ultimately, this research seeks to deliver a new generation of seismic source models that enhance the robustness, interpretability, and practical usability of PSHA for low-to-moderate seismicity contexts.

Short biography (PhD student)

I completed both my bachelor's and master's degrees in mathematics at Sorbonne University. I then undertook a PhD in applied mathematics funded through a CIFRE contract from the ANRT, in collaboration between EDF R&D MIA Paris-Saclay laboratories. My research is motivated by EDF's fundamental need to assess and enhance the safety of its critical industrial facilities with respect to seismic risk.

References

[1] Alba Bernabeu Atanasio, Jiancang Zhuang, and Jorge Mateu. Spatio-temporal hawkes point processes: A review. *Journal of Agricultural, Biological and Environmental Statistics*, 2024.

[2] Stéphane Drouet, Gabriele Ameri, Kristell Dortz, Ramon Secanell, and Gloria Senfaute. A probabilistic seismic hazard map for the metropolitan france. *Bulletin of Earthquake Engineering*, 18, 03 2020.

[3] Merlin Keller, Sanaa Zannane, Clara Duverger, and Jessie Mayor. Bayesian estimation and selection of seismic source models.

[4] Christian Molkenthin, Christian Donner, Sebastian Reich, Gert Zöller, Sebastian Hainzl, Matthias Holschneider, and Manfred Opper. Gp-etas: semiparametric bayesian inference for the spatio-temporal epidemic type aftershock sequence model. *Statistics and Computing*, 2022.

[5] Gordon J. Ross and Aleksandar A. Kolev. Semiparametric bayesian forecasting of spatio-temporal earthquake occurrences. *Annals of Applied Statistics*, 16(4):2083–2100, 2022.

Asymmetric conformal prediction with penalized kernel sum-of-squares

L. Allain[†,1,2], S. Da Veiga[§,2], B. Staber[§,1],

[1] Safran Tech, Digital Sciences & Technologies, 78114 Magny-Les-Hameaux, France
{louis.allain,brian.staber}@safrangroup.com

[2] Univ Rennes, Ensai, CNRS, CREST - UMR 9194, F-35000 Rennes, France
sebastien.da-veiga@ensai.fr

[†] PhD student (presenting author). [§] PhD supervisors

PhD expected duration: Dec. 2024 – Nov. 2027

Abstract

Conformal prediction (CP) [3, 6] is a distribution-free method to construct reliable prediction intervals that has gained significant attention in recent years. Despite its success and various proposed extensions, a significant practical feature which has been overlooked in previous research is the potential skewed nature of the noise, or of the residuals when the predictive model exhibits bias [4, 2, 7]. In this work, we leverage recent developments in CP [1] to propose a new asymmetric procedure that bridges the gap between skewed and non-skewed noise distributions, while still maintaining adaptivity of the prediction intervals.

More precisely, let us consider two RKHSs $\mathcal{H}_{\text{low}}$ and $\mathcal{H}_{\text{up}}$ with respective feature maps ϕ_{low} and ϕ_{up}. For $\mathcal{A}_{\text{low}} \in \mathcal{S}_+(\mathcal{H}_{\text{low}})$ and $\mathcal{A}_{\text{up}} \in \mathcal{S}_+(\mathcal{H}_{\text{up}})$ two positive semi-definite (PSD) operators from $\mathcal{H}_{\text{low}}$ (resp. $\mathcal{H}_{\text{up}}$) to $\mathcal{H}_{\text{low}}$ (resp. $\mathcal{H}_{\text{up}}$), we define two non-negative functions $f_{\text{low}}(X) = \langle \phi_{\text{low}}(X), \mathcal{A}_{\text{low}}\phi_{\text{low}}(X)\rangle_{\mathcal{H}_{\text{low}}}$ and $f_{\text{up}}(X) = \langle \phi_{\text{up}}(X), \mathcal{A}_{\text{up}}\phi_{\text{up}}(X)\rangle_{\mathcal{H}_{\text{up}}}$, called *kernel sum-of-squares*. These two functions, thanks to their non-negativity property, are key components of our proposed new asymmetric score function for CP:

$$S(X,Y) = \max\big(\widehat{m}_n(X) - f_{\text{low}}(X) - Y, Y - \widehat{m}_n(X) - f_{\text{up}}(X)\big).$$

From there, we propose to estimate the functions $f_{\text{low}}(X)$ and $f_{\text{upp}}(X)$ defining the prediction bands by solving the following learning problem:

$$\begin{aligned} \inf_{\substack{\mathcal{A}_{\text{low}}\in\mathcal{S}_+(\mathcal{H}_{\text{low}})\\ \mathcal{A}_{\text{up}}\in\mathcal{S}_+(\mathcal{H}_{\text{up}})}} \quad & \frac{b}{n}\sum_{i=1}^{n}\left(f_{\text{low}}(X_i) + f_{\text{up}}(X_i)\right) + \Omega_{\text{low}}(\mathcal{A}_{\text{low}}) + \Omega_{\text{up}}(\mathcal{A}_{\text{up}}) + \lambda\mathcal{P} \\ \text{s.t.} \quad & m(X_i) - Y_i - f_{\text{low}}(X_i) \le 0,\ i \in [n] \\ & Y_i - m(X_i - f_{\text{up}}(X_i) \le 0,\ i \in [n]\,. \end{aligned} \tag{1}$$

The new statistical learning problem in Equation (1) constructs adaptive and asymmetric prediction bands, with a unique feature based on a penalty which promotes symmetry: when the penalty intensity varies, the intervals smoothly change from symmetric to asymmetric ones. We study two possible penalties to achieve this, one on the operators and one on the training points:

$$\mathcal{P} = \|\mathcal{A}_{\text{low}} - \mathcal{A}_{\text{up}}\|_\star + \|\mathcal{A}_{\text{low}} - \mathcal{A}_{\text{up}}\|_F^2, \qquad \mathcal{P} = \sum_{i=1}^{n}\left(f_{\text{low}}(X_i) - f_{\text{up}}(X_i)\right)^2.$$

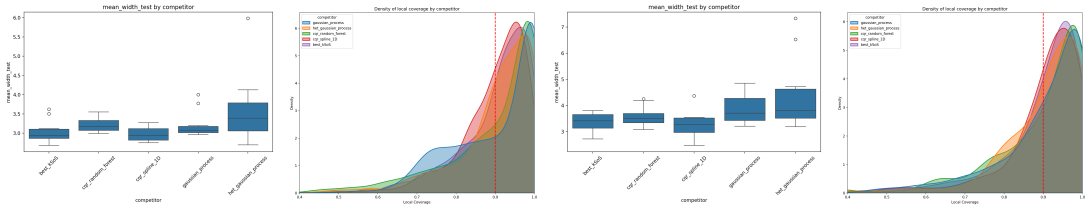


Figure 1: Comparison of prediction interval width (lower is better) and local coverage (vertical line indicates target level $1-\alpha$) for a symmetric datasets (left) and an asymmetric one (right) with $n = 100, m = 1000$ and 20 repetitions.

This statistical learning problem is based on reproducible kernel Hilbert spaces and the recently introduced kernel sum-of-squares framework [5]. First, we establish representer theorems to make our problems tractable in practice, and derive dual formulations which are essential for scalability to larger datasets. We also provide theoretical guarantees that these penalties indeed provide a continuum between asymmetric and symmetric prediction bands. Second, the intensity of the penalty is chosen using a novel data-driven method which automatically identifies the symmetric nature of the noise. We show that consenting to some asymmetry can let the learned prediction bands better adapt to small sample regimes, outliers or biased predictive models. Finally, our experiments illustrate the efficiency of such penalized kernel sum-of-squares to construct adaptive prediction bands.

Short biography (PhD student)

Following an engineering degree in statistics from ENSAI, my PhD focuses on quantifying uncertainty in machine learning using conformal prediction and kernel methods. The thesis is a joint initiative between the CREST laboratory and Safran Tech, supported by Safran funding. This industrial partnership ensures the research addresses critical uncertainty quantification gaps in metamodel-assisted robust design.

References

[1] Louis Allain, Sébastien Da Veiga, and Brian Staber. Scalable and adaptive prediction bands with kernel sum-of-squares. In *NeurIPS*, 2025.

[2] Matt Y. Cheung, Tucker J. Netherton, Laurence E. Court, Ashok Veeraraghavan, and Guha Balakrishnan. Regression conformal prediction under bias, 2024.

[3] A Gammerman, V Vovk, and V Vapnik. Learning by transduction. In *Conference on Uncertainty in Artificial Intelligence*, 1998.

[4] Henrik Linusson, Ulf Johansson, and Tuve Löfström. Signed-Error Conformal Regression. In *Advances in Knowledge Discovery and Data Mining*. Springer International Publishing, 2014.

[5] Ulysse Marteau-Ferey, Francis Bach, and Alessandro Rudi. Non-parametric models for non-negative functions. In *NeurIPS*, 2020.

[6] Harris Papadopoulos, Kostas Proedrou, Vladimir Vovk, and Alexander Gammerman. Inductive confidence machines for regression. In *ECML*, 2002.

[7] Thomas Pouplin, Alan Jeffares, Nabeel Seedat, and Mihaela Van Der Schaar. Relaxed quantile regression: Prediction intervals for asymmetric noise. In *ICML*, 2024.

Multi-fidelity Gaussian processes for noisy outputs and non-nested experiment designs

N. Baillie[†,1,2], J. Garnier[§,2], B. Kerleguer[§,3], C. Feau[§,1]

[†] PhD student (presenting author). [§] PhD supervisor

PhD expected duration: Nov. 2024 – Nov. 2027

[1] Université Paris-Saclay, CEA, Service d'Études Mécaniques et Thermiques, 91191 Gif-sur-Yvette, France
{nils.baillie,cyril.feau}@cea.fr

[2] CMAP, CNRS, École polytechnique, Institut Polytechnique de Paris, 91120 Palaiseau, France
josselin.garnier@polytechnique.edu

[3] CEA, DAM, DIF, F-91297 Arpajon, France
baptiste.kerleguer@cea.fr

Abstract

Surrogate models provide fast approximations of costly computer codes or experiments and are trained with a limited set of observations from said code. In the multi-fidelity framework, we assume that two computer codes of different costs and accuracy levels are available. The high-fidelity code z_H is the most precise, but also the most expensive, whereas the low-fidelity code z_L is much cheaper but less accurate.

In that context, several types of models exist, but we focus in this work on Gaussian processes (GP) [4] and notably, on the auto-regressive model which supposes a linear relation between the two codes:

$$\forall \mathbf{x} \in \mathbb{R}^D, \quad z_H(\mathbf{x}) = \rho(\mathbf{x}) \cdot z_L(\mathbf{x}) + \delta_H(\mathbf{x}),$$

where ρ is the scaling factor and the functions z_L and δ_H are approximated with GPs. This model was initially developed by [2] and improved afterwards by [3] with the more computationally efficient recursive formulation. However, two important assumptions are made in these works: the observed outputs are deterministic, and the experiment designs are nested, i.e. at the input level, every high-fidelity point coincides with a low-fidelity point. Several aspects are simplified in this framework: the equations for the predictive mean and covariance, and in particular, the likelihood function for the parameter optimization.

It is possible to relax one of these assumptions by relying on the approach proposed by [5] which utilizes the EM (expectation-maximization) algorithm to infer the parameters when the experimental designs are not nested, but for noise-free outputs. We generalize this approach to the case of noisy outputs and to the case where ρ is a linear predictor defined with a vector parameter β_ρ: $\rho(\mathbf{x}) = g_L(\mathbf{x})^\top \beta_\rho$ and not a fixed function. We apply the developed model to several cases of varying difficulty [1].

Short biography (PhD student)

After graduating from Université Paris-Saclay in applied mathematics (master MVA), I started my PhD thesis in November 2024, which is jointly organized by the CEA (French Alternative Energies and Atomic Energy Commission) and École polytechnique. I am supervised by Josselin Garnier (CMAP), Cyril Feau and Baptiste Kerleguer (CEA). The PhD thesis is funded by the CEA and the Paris-Saclay SEISM Institute (`https://www.institut-seism.fr/en/`).

References

[1] Nils Baillie, Baptiste Kerleguer, Cyril Feau, and Josselin Garnier. Efficient multi-fidelity Gaussian process regression for noisy outputs and non-nested experimental designs. *arXiv preprint, arXiv:2511.20183*, 2025.

[2] Marc Kennedy and Anthony O'Hagan. Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, 87, 10 2000.

[3] Loic Le Gratiet and Josselin Garnier. Recursive co-kriging model for design of computer experiments with multiple levels of fidelity. *International Journal for Uncertainty Quantification*, 4(5):365–386, 2014.

[4] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 11 2005.

[5] Federico Zertuche. *Assessment of uncertainty in computer experiments when working with multifidelity simulators*. PhD thesis, Université Grenoble Alpes, October 2015.

Control variates for variance-reduced ratio of means estimators

L. Bocquet-Nouaille[†,1,2], J. Morio[§,1,2], B. Bobbia[§,2],

[†] PhD student (presenting author). [§] PhD supervisor

PhD expected duration: Oct. 2024 – Sep. 2027

[1] ONERA/DTIS, Université de Toulouse, F-31055 Toulouse
{louison.bocquet_nouaille,jerome.morio}@onera.fr

[2] Fédération ENAC ISAE-SUPAERO ONERA, Université de Toulouse, 31000 Toulouse
benjamin.bobbia@isae-supaero.fr

Abstract

Many quantities of interest in engineering, physics, and risk analysis can be expressed as ratios of expectations $R = \mathbb{E}[A]/\mathbb{E}[C]$, where A and C are arbitrary real random variables. The standard Monte Carlo (MC) estimator, defined as the ratio of empirical means $\widehat{R}_{\frac{MC}{MC}} = \overline{A_n}/\overline{C_n}$, suffers from high variance when the number of samples n is small, which is often the case in practice.

To reduce variance without increasing n, the Control Variates (CV) method [1] is applied to both numerator and denominator. The idea is to exploit auxiliary random variables B and D, correlated with A and C, whose expectations are known. The CV ratio estimator is defined as

$$\widehat{R}_{\frac{CV}{CV}} = \frac{\overline{A_n} + \alpha(\mathbb{E}[B] - \overline{B_n})}{\overline{C_n} + \beta(\mathbb{E}[D] - \overline{D_n})} \tag{1}$$

where $(A_i, B_i, C_i, D_i)_{i=1\dots n}$ are i.i.d. real samples from the joint distribution of the random variables $A, B, C, D \in L^2$. In practice, the coefficients α and β are estimated.

In applied settings, the expectations $\mathbb{E}[B]$ and $\mathbb{E}[D]$ are often unknown, but they may be estimated from a larger set of $n + m$ samples of B and D, resulting in the Approximate Control Variates (ACV) ratio estimator [5]. We consider this framework, referred to as semi-supervised, where n joint samples of A, B, C, D are available with m additional samples of B and D.

For single mean estimation, the CV approach guarantees variance reduction without added bias. However, extending the method to ratio of means estimators is delicate. We show that ill-chosen coefficients can lead to a variance increase, in particular the "classical" coefficients [1] for single mean estimators, and the coefficients proposed in [4] optimized sequentially for ratio estimators.

We propose new jointly optimized coefficients that guarantee variance reduction.

$$(\alpha_o, \beta_o) := \underset{(\alpha,\beta)\in\mathbb{R}^2}{\operatorname{argmin}} \operatorname{Var}\left(\frac{\overline{A_n} + \alpha(\mathbb{E}[B] - \overline{B_n})}{\overline{C_n} + \beta(\mathbb{E}[D] - \overline{D_n})}\right) \tag{2}$$

Closed-form expressions of these optimal coefficients are derived [2], by minimizing an asymptotic approximation of the ratio variance which is convex if $|\mathrm{Corr}(B, D)| < 1$. Alternative expressions of the optimal coefficients are known if that condition is not fulfilled. Significant variance reduction can be achieved, depending on the correlation structure between the variables of interest A and C, and the control variates B and D, as seen in Figure 1.

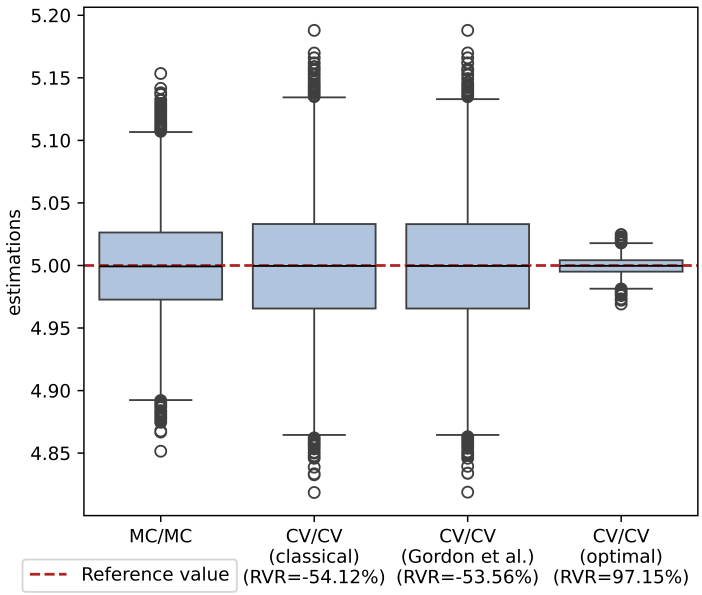


Figure 1: Boxplots of ratio estimations on highly-correlated simulated Gaussian data (n=100). From left to right: Monte Carlo estimator, CV estimator with single-mean coefficients [1], CV estimator with ratio coefficients [4], and CV estimator with the proposed optimal coefficients [2].

To illustrate the approach, we applied the proposed optimized CV estimators on three realistic problems on multi-fidelity datasets: strut mass fraction estimation, conditional Value-at-Risk (cVaR) estimation, and Extreme Value Index (EVI) estimation [3].

Short biography (PhD student)

Louison Bocquet-Nouaille is a PhD student at ONERA and ISAE-SUPAERO. She holds an MSc in applied mathematics from INSA Toulouse. Her research focuses on statistical transfer learning for extreme value analysis, aiming to improve inference in data-scarce settings. The PhD is funded at 50% by ONERA and at 50% by a MESRI scholarship.

References

[1] Søren Asmussen and Peter W. Glynn. Variance-Reduction Methods. In *Stochastic Modelling and Applied Probability*, pages 126–157. Springer New York, New York, NY, 2007.

[2] Louison Bocquet-Nouaille, Jérôme Morio, and Benjamin Bobbia. Control variates for variance-reduced ratio of means estimators. *arXiv preprint arXiv:2510.13504*, 2025.

[3] Louison Bocquet-Nouaille, Jérôme Morio, and Benjamin Bobbia. Variance-reduced extreme value index estimators using control variates in a semi-supervised setting. *arXiv preprint arXiv:2511.15561*, 2025.

[4] Heather L. Gordon, Stuart M. Rothstein, and Timothy R. Proctor. Efficient variance-reduction transformations for the simulation of a ratio of two means: application to quantum Monte Carlo simulations. *Journal of Computational Physics*, 47:375–386, 1982.

[5] Alex A. Gorodetsky, Gianluca Geraci, Michael S. Eldred, and John D. Jakeman. A generalized approximate control variate framework for multifidelity uncertainty quantification. *Journal of Computational Physics*, 408, 2020.

Bayesian calibration for computer models with functional outputs using elastic partial matching

Paul Castéras[†,1,2,3], Julien Bect[§,1], Josselin Garnier[§,2], Gwenaël Salin[§,3]

[†] PhD student (presenting author). [§] PhD supervisor

PhD expected duration: Dec. 2024 – Nov. 2027

[1] Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des signaux et systèmes, France

[2] CMAP, Ecole polytechnique, Institut Polytechnique de Paris, France

[3] CEA / DAM / DIF, France

paul.casteras@cea.fr

Abstract

To study complex phenomena, computer models are often used for simulation. These models typically depend on unknown parameters that must be estimated prior to their use. This task, known as model calibration, becomes particularly challenging when the outputs are high-dimensional. In this work, we focus on calibration problems where the outputs are time series.

To calibrate the model parameters β, one has to compare model outputs with experimental data for different experimental settings x. Since there are multiple ways of comparing time-dependent outputs, [2] suggests modeling error terms to account for time deformation using the framework of [5]. This approach, referred to as elastic calibration, offers two main advantages. First, an elastic transformation may simplify the representation of simulation outputs, thereby improving the performance of surrogate models. Second, elastic calibration makes it possible to introduce a temporal discrepancy term to explicitly account for time deformations.

However, because time series are measured on preselected time intervals, they are mecessarily truncated after an arbitrary final time. As a result, [1] argues that time deformations should be used that allow for this additional degree of freedom through a framework called elastic partial matching. In this work, we propose to extend the approach of [2] with elastic partial matching. This extension makes it possible to apply elastic calibration to a broader range of problems. Within the elastic framework, a function is decomposed into two components: an amplitude and a phase. A phase is a time warping function that aligns a function with a template. The amplitude is the aligned version of a function. For all experiments $y_{exp}(x,t)$ and simulation runs $y_{sim}(x,\beta,t)$, we partially match the functions to a template to obtain amplitudes, denoted by $f_{exp}(x,t)$ and $f_{sim}(x,\beta,t)$, and phases, denoted by $\gamma_{exp}(x,t)$ and $\gamma_{sim}(x,\beta,t)$. Figure 1 shows the result of the partial matching procedure for one dataset.

We then want to apply a Bayesian calibration framework, as introduced by [3], in the amplitude-phase space. However, the phase space is not a vector space and it is difficult to compare phases in this space. Thus, following [2], we first transform each phase $\gamma(t)$ into a shooting vector function $v(t)$ and a truncation parameter t_f.

To replace computationally expensive computer models, we can first run the computer model for a set of model parameters $(\beta_j)_{j=1,\dots,N_\beta}$ for each experimental setting. Then, based on these data, we build time-dependent surrogate models in the amplitude and shooting-vector spaces,

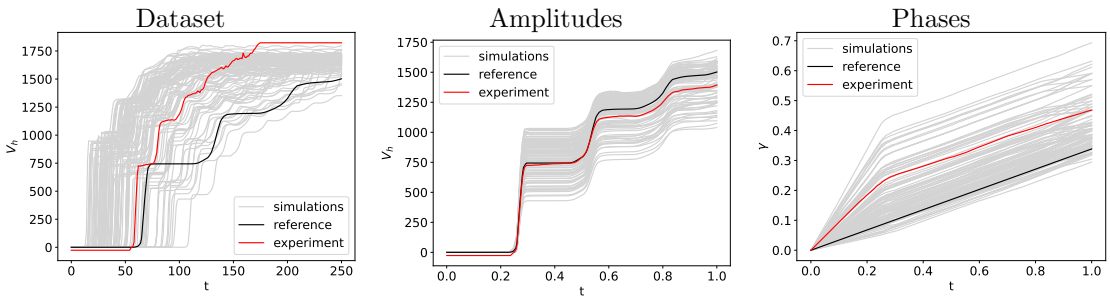


Figure 1: Example of decompostion into amplitudes and phases

and a scalar surrogate model for t_f. We denote these surrogates by $\tilde{f}^{\text{sim}}(x,t,\beta)$, $\tilde{v}^{\text{sim}}(x,t,\beta)$, and $\tilde{t}_f^{\text{sim}}(x,\beta)$. They are constructed using Gaussian process regression combined with PCA, as explained in [4].

Relations between the experiments and simulations can be written in the new spaces:

$$f^{\text{exp}}(x,t) = \tilde{f}^{\text{sim}}(x,t,\beta) + \varepsilon_f(x,t),$$
$$v^{\text{exp}}(x,t) = \tilde{v}^{\text{sim}}(x,t,\beta) + \varepsilon_v(x,t),$$
$$t_f^{\text{exp}}(x) = \tilde{t}_f^{\text{sim}}(x,\beta) + \varepsilon_t(x).$$

To estimate the calibration parameters and their uncertainties, after modeling errors and choosing prior distributions, one can use the experimental data $y^{\text{exp}}(x_i,t)_{i=1,\dots,N_x}$ measured at different experimental settings $(x_i)_{i=1,\dots,N_x}$ to compute and sample the posterior distributions of β and the hyperparameters of the errors, denoted by θ_f, θ_v and θ_t :

$$p(\beta,\theta_f,\theta_v,\theta_t|y^{\text{exp}}) \propto p(f^{\text{exp}}|\beta,\theta_f)p(v^{\text{exp}}|\beta,\theta_v)p(t_f^{\text{exp}}|\beta,\theta_t)\pi(\beta)\pi(\theta_f)\pi(\theta_v)\pi(\theta_t).$$

We demonstrate the benefits of this method on the calibration of an equation of state, which links thermodynamic variables such as pressure, temperature, volume, and internal energy.

Short biography

Before my PhD, I studied at the engineering school CentraleSupélec and completed the MVA master's program. I did my end-of-study internship at CEA and then started my PhD thesis, funded by CEA. I am now a second-year PhD student working on Bayesian calibration with time-series outputs, focusing on misalignment issues that arise in real CEA data.

References

[1] D. Bryner and A. Srivastava. Shape analysis of functional data with elastic partial matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9589–9602, 2021.

[2] D. Francom, J. D. Tucker, G. Huerta, K. Shuler, and D. Ries. Elastic Bayesian model calibration. *SIAM/ASA Journal on Uncertainty Quantification*, 13(1):195–227, 2025.

[3] M. C. Kennedy and A. O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464, 2001.

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

[5] A. Srivastava and E. P. Klassen. *Functional and Shape Data Analysis*, volume 1. Springer, 2016.

Model order reduction with randomized methods for parametrized evolution problems

U. Desgroppes[†,2,3], M. Billaud-Friess[§,1], A. Nouy[§,2], K. Smetana[§,3]

[†] PhD student (presenting author). [§] PhD supervisors

PhD expected duration: Sep. 2024 – Sep. 2028

[1] Centrale Méditerranée, Institut de Mathématiques de Marseille UMR 7373
[2] Centrale Nantes, Nantes Université, LMJL UMR 6629
[3] Stevens Institute of Technology, Department of Mathematical Sciences, Hoboken, NJ USA

Abstract

When dealing with parameter-dependent partial differential equations (PDEs), classical numerical methods lead to costly numerical simulations. One usually relies on model order reduction to approximate the original problem by a reduced one which we can solve more efficiently. It becomes possible to fastly compute a solution for a given input (parameters, initial or boundary conditions) in real time, or to compute a bunch of solutions in a many-query scenario or uncertainty quantification.

Different reduction methods have been developed for time-dependent problems such as the Proper Orthogonal Decomposition (POD) [1, 3]. The goal is to construct a reduced basis of an approximation space that captures the essential dynamic of the problem. However, POD needs to precompute the global solution trajectory in order to build such a reduced space. This means having to solve sequentially in time (at least for some part of the global time interval) before projecting the dynamic onto a reduced space.

An alternative has been developed in [4] where it has been proposed to construct the reduced basis in a parallel manner in time. It consists in solving the full dimensional problem on small chunks of the global time interval, as in Figure 1.

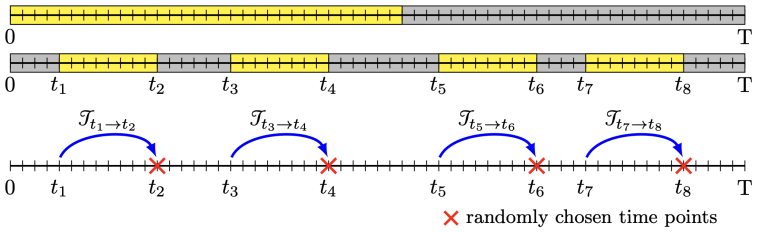


Figure 1: From [4]. Comparison of the computational budget in time, POD (top row) versus split in time approach (middle row). Representation of the local transfer operators at randomly chosen time points (bottow row).

The objective is to construct approximation spaces capturing the action of the transfer operator mapping initial condition to final state on a time interval. For linear PDE operators, learning the range of the affine transfer operators can be split into two tasks, the approximation of solutions on subintervals with zero initial conditions, or with zero source term. The latter task is related to the approximation of the range of a linear operator, for which we rely on randomized numerical linear algebra methods (RNLA) (range finder or random probes [2]), that translates into the resolution of the equation with random initial conditions. Using random initial conditions on each sub-interval, the local trajectories are completely independent from each other, hence allowing parallelization.

The main interest of such an approach is to capture local in time features of the problem and construct relevant reduced spaces that can be used afterwards for an efficient resolution over the whole time interval.

In this work, we extend the approach from [4] to the resolution of evolution problems in tensor spaces (e.g. resulting from the discretization of high-dimensional PDEs or stochastic PDEs) and rely on RNLA for tensors and dynamical low-rank methods for the resolution of dynamical systems in tensor spaces.

We will present numerical examples for parabolic PDEs.

Short biography (PhD student)

I started my PhD after finishing my Master in applied mathematics at Centrale Nantes. Having studied PDEs as well as stochastic numerical methods, this thesis was a smooth following to my time at Centrale Nantes. My PhD supervisors are Marie Billaud-Friess, Kathrin Smetana and Anthony Nouy. During this PhD, I will spend my time equally between Centrale Nantes and the Stevens Institute of Technology in Hoboken, NJ.

References

[1] G. Berkooz, P. Holmes, and J. L. Lumley. The proper orthogonal decomposition in the analysis of turbulent flows. *Annual Review of Fluid Mechanics*, 25:539–575, 1993.

[2] N. Halko, P-G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions, 2010.

[3] K. Kunisch and S. Volkwein. Galerkin proper orthogonal decomposition methods for parabolic problems. *Numerische Mathematik*, 90(1):117–148, 2001.

[4] J. Schleuß, K. Smetana, and L. ter Maat. Randomized quasi-optimal local approximation spaces in time. *SIAM J. Sci. Comput.*, 45(3):1066–1096, 2023.

Multifidelity Gaussian process regression for solving nonlinear partial differential equations

Fatima-Zahrae El-Boukkouri[†,1], Josselin Garnier[§,2], Olivier Roustant[§,1]

[†] PhD student (presenting author). [§] PhD supervisor

PhD expected duration: Oct. 2024 – Sep. 2027

[1] Institut de Mathématiques de Toulouse, Université de Toulouse, INSA, Toulouse
{el-boukkouri,roustant}@insa-toulouse.fr

[2] CMAP, CNRS, Ecole polytechnique, Institut Polytechnique de Paris, Palaiseau
josselin.garnier@polytechnique.edu

Abstract

Kernel-based methods provide a principled alternative to classical numerical solvers for nonlinear partial differential equations (PDEs), especially when mesh-free formulations, regularization, and uncertainty quantification are of interest. Traditional discretization techniques such as finite differences or finite elements [2] often become computationally demanding for nonlinear or multiscale problems and require careful mesh design near singularities. In contrast, the variational framework introduced by Owhadi et al. [1] formulates PDE solving as a constrained minimization problem in a reproducing kernel Hilbert space (RKHS), providing both a functional-analytic interpretation and a probabilistic Gaussian-process (GP) viewpoint.

Let $\Omega \subset \mathbb{R}^p$ be a bounded domain and consider the nonlinear PDE

$$\begin{cases} \mathcal{P}(u^\star)(x) = f(x), & x \in \Omega, \\ \mathcal{B}(u^\star)(x) = g(x), & x \in \partial\Omega, \end{cases}$$

where $\mathcal{P}$ is a nonlinear differential operator and $\mathcal{B}$ a boundary operator. Given an RKHS $\mathcal{H}(K)$ with sufficiently smooth kernel K, and collocation points $\{x_m\}_{m=1}^M \subset \overline{\Omega}$, the framework proposed in [1] defines the numerical solution as

$$\begin{cases} \min_{u \in \mathcal{H}(K)} \|u\|_{\mathcal{H}(K)} \\ \text{s.t. } \mathcal{P}(u)(x_m) = f(x_m), \quad 1 \le m \le M_\Omega, \quad \text{and} \quad \mathcal{B}(u)(x_m) = g(x_m), \quad M_\Omega + 1 \le m. \end{cases}$$

This formulation admits a probabilistic interpretation: if $\varepsilon \sim \mathcal{GP}(0, K)$, then the solution coincides with the maximum a posteriori (MAP) estimator of ε conditioned on the PDE constraints [1]. As a consequence, the choice of kernel K and its hyperparameters directly determines the admissible solution space, its regularity, and the numerical stability of the solver. Recent studies have demonstrated a strong sensitivity of kernel-based PDE solvers to hyperparameter selection, and proposed bilevel and Gauss–Newton strategies to mitigate this issue [4].

In this work, we introduce a multifidelity, physics-informed methodology for constructing RKHSs adapted to PDE resolution, exploiting hierarchical simulation data rather than relying on ad hoc kernel choices. We assume access to an ensemble of M low-fidelity simulations $\{y_L^i\}_{i=1}^M$ defined on a grid X_L, together with high-fidelity observations y_H available on a nested grid $X_H \subseteq X_L$. From the low-fidelity ensemble, we estimate the empirical mean and covariance k_L. Since k_L is

only defined on $X_L \times X_L$ and may lack smoothness, we approximate it by a differentiable kernel k_{opt} belonging to a smooth kernel family S by solving

$$k_{\text{opt}} = \arg\min_{k \in S} d(k|_{X_L \times X_L},\, k_L)\,,$$

where d is a matrix distance, chosen in practice as the Frobenius norm for numerical robustness.

To incorporate high-fidelity information, we adopt the autoregressive cokriging model of Kennedy and O'Hagan [3, 5], which assumes

$$Y_H(x) = \rho\, Y_L(x) + Y_d(x), \qquad Y_d \sim \mathcal{GP}(\mu_d, k_d),$$

with k_d a stationary kernel modeling the discrepancy between fidelity levels. This leads to the high-fidelity kernel

$$k_H^\star(x, x') = \rho^2 k_{\text{opt}}(x, x') + k_d(x, x'),$$

whose parameters are learned by maximizing the marginal likelihood of the high-fidelity data. The RKHS $\mathcal{H}(k_H^\star)$ is then directly used in the PDE-constrained minimization problem above, yielding a multifidelity-informed kernel that combines large-scale regularity from low-fidelity ensembles with fine-scale corrections inferred from high-fidelity observations.

Beyond kernel construction, we extend the framework to non-centered Gaussian processes, using a cokriging-informed mean as a physically motivated prior. When sufficient high-fidelity data are available, we also consider learning a smooth kernel directly from high-fidelity residuals by likelihood maximization, bypassing the approximation of k_L and complementing multifidelity constructions.

We validate these methodologies on linearized and fully nonlinear Burgers equations. Numerical results show that multifidelity-informed kernels substantially reduce sensitivity to hyperparameter selection compared with single-fidelity baselines, while preserving the differentiability required by operator-based RKHS solvers.

Short biography (PhD student)

I graduated from École Polytechnique and hold a Master's degree in Mathematics, Vision, and Learning from the Institut Polytechnique de Paris. My PhD focuses on developing physics-informed methods for predicting water-related extreme events, aiming to improve accuracy and reduce computational costs. The research is funded by specific doctoral grants for École Polytechnique graduates and ANITI.

References

[1] Y. Chen, B. Hosseini, H. Owhadi, and A. M. Stuart. Solving and learning nonlinear pdes with Gaussian processes. *Journal of Computational Physics*, 447:110668, 2021.

[2] C. Johnson. *Numerical solution of partial differential equations by the finite element method*. Courier Corporation, 2009.

[3] M. C. Kennedy and A. O'Hagan. Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, 87(1):1–13, 2000.

[4] N. H. Nelsen, H. Owhadi, A. M. Stuart, X. Yang, and Z. Zou. Bilevel optimization for learning hyperparameters: Application to solving pdes and inverse problems with gaussian processes. *arXiv preprint arXiv:2510.05568*, 2025.

[5] X. Yang, X. Zhu, and J. Li. When bifidelity meets cokriging: An efficient physics-informed multifidelity method. *SIAM Journal on Scientific Computing*, 42(1):A220–A249, 2020.

Generalized Functional ANOVA for Categorical Inputs : Closed-Form Decomposition and Fast Computations

B. Ferrere[⋆,1,2], N. Bousquet[1], F. Gamboa[2], JM. Loubes[2,3], J. Muré[1]

[⋆] PhD student (presenting author).

PhD expected duration: Jan. 2025 – Dec. 2027

[1] EDF Lab Chatou, Chatou, France
{baptiste.ferrere , nicolas.bousquet , joseph.mure}@edf.fr

[2] IMT, Toulouse, France
{fabrice.gamboa , loubes}@math.univ-toulouse.fr

[3] INRIA, Regalia Team, Toulouse, France

Abstract

Functional ANOVA (or Hoeffding Decomposition) [6] provides a principled framework to model interpretability by decomposing a prediction function $f(X)$ into main effects and higher-order interactions. However, beyond the independent feature setting, practitioners typically lack an explicit decomposition basis and must resort to sampling based approximations that are computationally expensive and can be unreliable under strong dependence. In this work, we address this limitation for categorical inputs X by proposing an explicit family of basis functions onto which $f(X)$ can be projected with guarantees. This family is fully explicit and satisfies the defining ANOVA properties originally formalized by [2] through a constrained optimization viewpoint, thereby providing a direct and exact decomposition mechanism.

By combining tools from functional analysis with Boolean function analysis [5], we obtain a fully explicit, closed-form expression for the Functional ANOVA components under *arbitrary* dependence structures. The resulting decomposition is strictly exact: it coincides with the classical orthogonal ANOVA when features are independent, yet it also remains valid in regimes that are typically problematic for existing methods, including highly sparse support and strong—even functional—dependencies between features.

Our results also align with, and refine, the generalization hypotheses put forward in [1] and later in [3]. In particular, these works provide necessary conditions for the *existence* and *uniqueness* of the generalized ANOVA decomposition; the conclusions induced by our explicit decomposition family, especially regarding uniqueness, recover the previously established boundary cases identified in these foundational contributions.

Beyond this theoretical scope, the proposed closed-form expressions lead to extremely fast computation: experiments indicate that the decomposition can be obtained *ultra fast* while accurately recovering main effects and interactions in challenging settings, including very sparse support and high number of features.

Finally, we highlight an implication for attribution methods. In the independent case, SHAP [4] can be viewed as arising from the orthogonal Functional ANOVA decomposition; our decomposition therefore suggests a method to extend SHAP-style attributions to correlated categorical variables by grounding them in the generalized decomposition derived here.

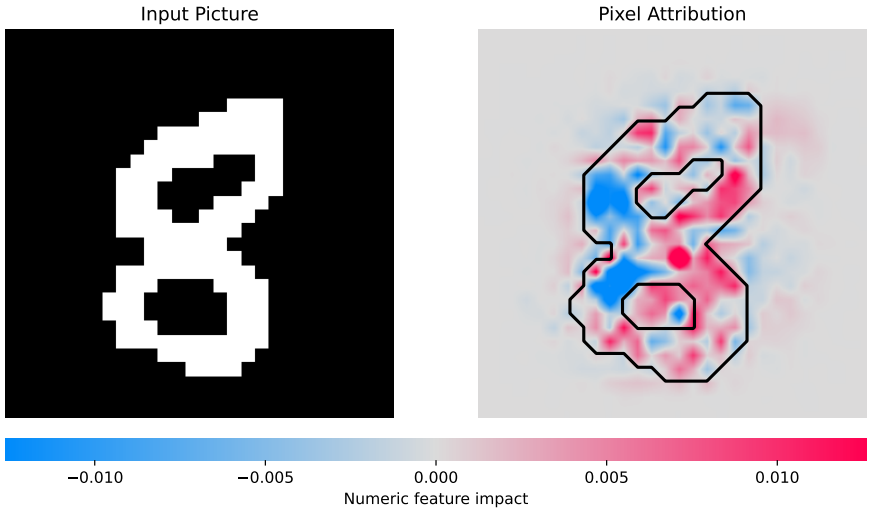


Figure 1: Example of attribution on the binarized MNIST dataset. Here, we analyze 784 correlated binary features where $f(X) = Pr\,(\text{predicted class} = 3)$. Red pixels increase this probability, while blue pixels decrease it.

Short biography (PhD student)

I started my PhD in January 2025, focusing on Explainable AI (XAI) and model interpretability. My research aims to bridge the gap between classical statistics and modern Machine Learning by applying principles from functional analysis to ML frameworks. Building on my experience at EDF R&D, where I worked on sensitivity analysis via Functional ANOVA and Hoeffding decompositions, I am currently investigating how to generalize and improve popular XAI methods (such as SHAP, Orthogonal ANOVA, and other decomposition-based techniques). My goal is to address the theoretical limitations of these methods while developing robust additive decomposition properties for real-world applications.

References

[1] Gaëlle Chastaing, Fabrice Gamboa, and Clémentine Prieur. Generalized Hoeffding-Sobol Decomposition for Dependent Variables – Application to Sensitivity Analysis. *Electronic Journal of Statistics*, 6:2420–2448, March 2012.

[2] Giles Hooker. Generalized functional anova diagnostics for high-dimensional functions of dependent variables. *Journal of Computational and Graphical Statistics*, 16(3):709–732, 2007.

[3] Marouane Il Idrissi, Nicolas Bousquet, Fabrice Gamboa, Bertrand Iooss, and Jean-Michel Loubes. Hoeffding decomposition of functions of random dependent variables. *Journal of Multivariate Analysis*, page 105444, 2025.

[4] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[5] Ryan O'Donnell. *Analysis of boolean functions*. Cambridge University Press, 2014.

[6] Charles J Stone. The use of polynomial splines and their tensor products in multivariate function estimation. *The annals of statistics*, pages 118–171, 1994.

Simulation of extreme functionals for design of experiments of coastal flood numerical models

N. Gorse[†,1], O. Roustant[§,1], J. Rohmer[§,2], D. Idier[§,2]

[†] PhD student (presenting author). [§] PhD supervisor

PhD expected duration: Nov. 2024 – Oct. 2027

[1] UMR CNRS 5219, Institut de Mathématiques de Toulouse, INSA, Université de Toulouse, France
{gorse,roustant}@insa-toulouse.fr

[2] BRGM, F-45060 Orléans, France
{rohmer,idier}@brgm.fr

Abstract

Context and objective The role of meteoceanic conditions is crucial in coastal flooding but is also difficult to capture with traditional statistic methods. In this context, our objective is to enhance the modelling of coastal flooding [5] by a design of experiments (DoE) where the inputs of numerical hydrodynamic models are time series. The difficulty is that the scarcity of recorded extreme meteoceanic conditions would prevent a DoE from efficiently exploring the most relevant input regions to coastal flooding. To address this limitation, we propose a method for simulating extreme time series which have the same behaviour as the observed conditions but are extrapolated towards high values.
This work is guided by the application presented in [3], focusing on the site of Gâvres in French Brittany, the amplitude of coastal flooding induced by storms such as Johanna is estimated using numerical models. Since such results are computationally intensive, a surrogate model can be trained using selected forcing conditions from the computational domain and outputs from numerical models.
We analyse the dynamics of the forcing conditions within this domain and use a database based on the paper of [3]. This town being located in a macro-tidal area, we focus on meteoceanic conditions occurring (+/-)3h around the high tide (with a fixed time step of 10 minutes).

Methodology We use the notation X_M^t to describe the value obtained at time t for the M^{th} tidal cycle. Our observations do not meet the standard assumptions of independence and regular variations.
Construction of a probabilistic model The first step of our method detailed in [2] thus consists in a *"whitening"* pre-processing of detrended winter time series $\tilde{X}_M^t$. We impose a minimal duration Δ between each event and introduce an autoregressive model with residuals ε_M^t to account for the temporal dependence *between* tidal cycles while preserving the dependence *within* each one of them. In a second step, after applying marginal transformations to the residuals, we construct a probabilistic model within the framework of [1], which relies on a polar coordinate representation [4].
Usage of the model for simulation We simulate extreme time series ε_{sim} by first sampling from this representation and then applying inverse transformations to recover time series in the original space. The final inversion of the autoregressive model depends on an initial time series $\tilde{X}_{M-\Delta}$, which can be selected to tune the desired level of extremes. In our case, because we aim to generate realistic extreme simulations, $\tilde{X}_{M-\Delta}$ is selected according to the level of $\ell(\varepsilon_{\text{sim}})$.

Numerical results We apply our method to the surge data and assess the quality of our method by checking that simulations and extreme observations share common behaviours. First of all, they should have shape similarities (Figure 1).

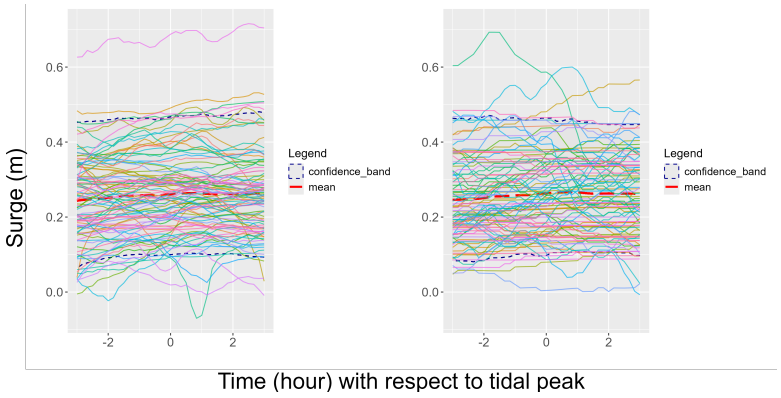


Figure 1: Comparing simulated time series and extreme observations. Left panel: simulated sample of 100 extreme time series, right panel: sample of 100 extreme observations. The dotted lines represent the 95% confidence bands

To further analyze the simulations, we apply principal component analysis (PCA) to the normalized data $\tilde{X}_M/\ell(\tilde{X}_M)$, as the simulations extrapolate the L^2 norm of the observations. The simulated time series exhibit behavior similar to that of the observed extremes, as their corresponding PCA coordinates are close. Finally, we apply two-sample classification tests, to assess whether a classifier can distinguish between observed and simulated time series. We show that the simulations are consistent with the observations, as the value 50% often lies within the confidence intervals for several classifiers.

Short biography (PhD student)

Former ENSAI student, I began to work on extreme functionals during an internship before working on it as a PhD student. The objective is to enlarge the toolbox of simulation methods in extreme time series and to use it in metamodelling for coastal flooding. The thesis is co-funded by AI Interdisciplinary Institute ANITI and BRGM.

References

[1] C. Dombry and M. Ribatet. "Functional regular variations, Pareto processes and peaks over threshold". In: *Statistics and its Interface* 8.1 (2015), pp. 9–17.

[2] N. Gorse et al. "Simulation of extreme functionals in meteoceanic data: Application to surge evolution over tidal cycles". In: *arXiv preprint arXiv:2508.13687* (2025).

[3] D. Idier et al. "Coastal flood: a composite method for past events characterisation providing insights in past, present and future hazards—joining historical, statistical and modelling approaches". In: *Natural Hazards* 101.2 (2020), pp. 465–501.

[4] P. Kokoszka, S. Stoev, and Q. Xiong. "Principal components analysis of regularly varying functions". In: *Bernoulli* 25.4B (2019), pp. 3864–3882.

[5] J. Rohmer et al. "Partitioning the contributions of dependent offshore forcing conditions in the probabilistic assessment of future coastal flooding". In: *Natural Hazards and Earth System Sciences* 22.10 (2022), pp. 3167–3182.

Maximin Designs on a Reproducing Kernel Hilbert Space

L.Calzolari[†,1,3,4], C. Helbert[§,2], M. Munoz Zuniga[§,3], D. Sinoquet[§,3], C. Prieur[§,1,4]

[†] PhD student (presenting author). [§] PhD supervisor

PhD expected duration: Nov.2024 – Nov. 2027

[1] Univ. Grenoble Alpes, LJK,

{lorenzo.calzolari,clementine.prieur}@univ-grenoble-alpes.fr

[2] École Centrale de Lyon

celine.helbert@ec-lyon.fr

[3] IFP Energies Nouvelles

{lorenzo.calzolari,miguel.munoz-zuniga,delphine.sinoquet}@ifpen.fr

[4] Inria

{lorenzo.calzolari,clementine.prieur}@inria.fr

Abstract

Numerous advances have been made in the field of computer simulations that replace real life experiments or time-consuming numerical models [1]. These methods involve learning a function $f : \mathcal{X} \to \mathbb{R}$ that mimics the behaviour of a real-world phenomenon or a numerical model, using an input/output dataset. This involves the choice of the input points at which to observe the real-world phenomenon or evaluate the numerical model. This selection process is referred to as the search of an *Experimental Design*. To the best of our knowledge, Design of Experiments (DoEs) with functional inputs rely on a truncated basis expansion, which projects the entries over a fixed subspace, as in [3]. In this framework, constructing an initial DoE consists in selecting a set of coefficient vectors. As observed in [4] this introduces a bias in the choice of the basis and a loss of information with the truncation. We propose two methods for finding Space Filling Designs over a Reproducing Kernel Hilbert Space (RKHS) defined on a compact subset of $\mathbb{R}$, without any prior dimension reduction. These functional spaces are often used in the resolution of PDEs or in approximation theory. Since this task amounts to covering an infinite dimensional space with a finite amount of elements, we will focus on the case of subsets where the norm of the functions in the DoE can be controlled, e.g. the unitary ball of the RKHS.

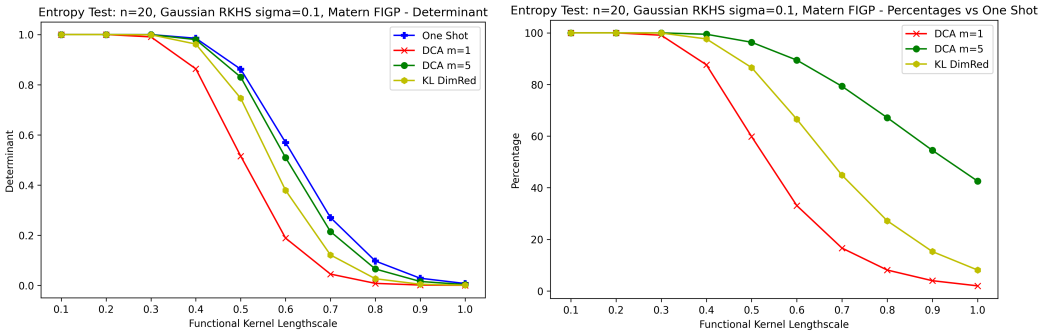
A distance based criterion Φ_p^{func} is proposed to ensure the functions are spread out over the domain. This is based on a generalization of the Morris Criterion , often called maximin criterion, introduced in [2]. Its optimization over the unitary ball of a RKHS $\mathcal{H}_k$ reads

$$\begin{aligned} \text{minimize} \quad & \Phi_p^{func} := \left(\sum_{i<i'} d(f_i, f_{i'})^{-p} \right)^{\frac{1}{p}} \\ \text{subject to} \quad & \|f_i\|_{\mathcal{H}_k} - 1 \leq 0 \quad i = 1, ..., n. \end{aligned} \tag{P}$$

The uniqueness result by Moore-Aronszajn ensures that any function in $\mathcal{H}_k$ can be written as a linear combination of kernel evaluations. The explicit dependence of this formulation on the centers and on the coefficients is exploited for the resolution of (P). To this end the Dynamic Cloud Algorithm (DCA) is proposed for a RKHS with a stationary and smooth kernel, allowing the use of gradient-based optimization procedures.

An analytical study of (P) is performed, from which necessary and sufficient conditions for optimality over the unitary ball of $\mathcal{H}_k$ are derived. These amount to a system of equalities on the norms and on the inner products among the functions of a DoE. These conditions are exploited to create a quick-to-evaluate algorithm, in which no optimization routines are needed.

As in [4], stationary kernels can be extended to functional spaces like RKHSs, provided the distance used is induced by the associated inner product. This allows the construction of Functional Input Gaussian Processes (FIGPs). Experimental designs can then be studied through the entropy of the trained FIGP. As shown in [1], in Gaussian process regression, the dedicated DoEs of maximal entropy are the ones that maximise the determinant of the correlation matrix. The entropies of FIPGs trained on 4 different functional DoE are compared: the first is produced by our fast algorithm solving (P), the second and third are produced by DCA for functions of $m = 1, 5$ kernel evaluations and the fourth corresponds to a Karhunen-Loeve inspired dimension reduction. With this measure, we show empirically as in Figure 1 that the one shot algorithm outperforms both the optimization based method and the dimension reduction approach.



(a) Values of the determinant.

(b) Percentages against the One Shot Algorithm.

Figure 1: Analysis of the determinant of the correlation function of a FIGP with Matern 5/2 covariance function, with inputs defined on a RKHS associated to the Gaussian kernel with $\sigma = 0.1$, and a DoE of $n = 20$ functions over $[0, 1]$.

Short biography (PhD student)

I am a second years PhD student enrolled at Universite Grenoble-Alpes, after graduating in Applied Mathematics at the University of Bologna. My Thesis focuses on the development of meta-models over functional spaces, and their application to optimization. This project is a collaboration between IFPEN (which provides the fundings) and the AIRSEA research team from Inria Grenoble, where I am currently working at the LJK laboratory.

References

[1] Robert B Gramacy. *Surrogates: Gaussian process modeling, design, and optimization for the applied sciences*. Chapman and Hall/CRC, 2020.

[2] Max D Morris and Toby J Mitchell. Exploratory designs for computational experiments. 43(3):381–402, 1995. Publisher: Elsevier.

[3] Thomas Muehlenstaedt, Jana Fruth, and Olivier Roustant. Computer experiments with functional inputs and scalar outputs by a norm-based approach. 27:1083–1097, 2017. Publisher: Springer.

[4] Chih-Li Sung, Wenjia Wang, Fioralba Cakoni, Isaac Harris, and Ying Hung. Functional-input gaussian processes with applications to inverse scattering problems. 2022.

Prediction of physical fields under linear constraints: uncertainty quantification for RANS simulation of the Buice-Eaton diffuser

N. Mahamat Hamdan[†,1,3], C. Gauchy[1], P-E. Angeli [2], S. Da Veiga[§,3],

[†] PhD student (presenting author). [§] PhD supervisor

PhD expected duration: Nov. 2024 – Nov. 2027

[1] Université Paris-Saclay, CEA, Service de Génie Logiciel pour la Simulation, France
[2] Université Paris-Saclay, CEA, Service de Thermohydraulique et de Mécaniques des Fluides, France
{mahamat-hamdan.nassouradine, clement.gauchy, pierre-emmanuel.angeli}@cea.fr

[3] ENSAI, CREST, F-35000 Rennes, France
sebastien.da-veiga@ensai.fr

Abstract

In the context of industrial computational fluid dynamics (CFD), Uncertainty Quantification (UQ) and optimization processes rely heavily on Reynolds-Averaged Navier-Stokes (RANS) simulations. These simulations often involve numerous physical parameters determined experimentally, leading to a classical uncertainty propagation problem. However, the computational cost associated with high-fidelity CFD codes makes Monte Carlo-based approaches prohibitive due to the sheer number of required evaluations. Consequently, the construction of surrogates, or metamodels, has become an essential strategy to emulate the input-output relationship of the solvers. Modeling full physical fields presents specific challenges: primarily the high dimensionality of the discretization mesh (often millions of degrees of freedom) and the necessity to adhere to fundamental physical constraints (e.g., conservation laws, boundary conditions). In this work, we address the problem of constructing metamodels for vector-valued physical fields subject to linear constraints, formulated as:

$$\sum_{k=1}^{Q} \alpha_k(\mathbf{x}) f_k(\mathbf{x}) = 0 \tag{1}$$

where $\mathbf{x}$ denotes the input parameters, and $\{f_k\}_{k=1}^{Q}$ are the Q components of the physical, each mapping $\mathbb{R}^D$ to a high-dimensional space $\mathbb{R}^S$ corresponding to the mesh resolution S. While the literature often combines Principal Component Analysis (PCA) for dimensionality reduction with Gaussian Process (GP) regression [3], the optimal strategy for handling multiple coupled fields under constraints remains an open question.

Dimensionality reduction strategies for vector fields: A core contribution of this work is a detailed analysis of dimensionality reduction strategies for multi-component fields. When dealing with Q physical fields, the application of PCA is not trivial [4]. We explore three distinct architectures:

- Field-wise PCA: This approach computes a specific reduced basis for each physical component independently. While it minimizes the reconstruction error for each field locally, it generates distinct latent representations for components that are physically coupled. This

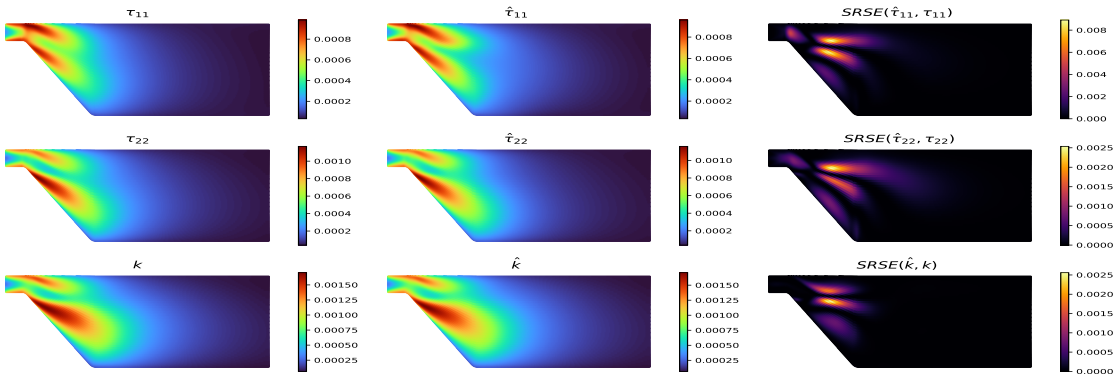


Figure 1: CFD test case. An exemple of fields predicted by our model (left) in comparison to simulated fields (middle) and the Spatial Relative Squared Error (SRSE) (right). From top to bottom, the first and second component of Reynolds stress tensor (τ_{11}, τ_{22}) , and the turbulent kinetic energy k.

results in a disjointed latent space where the physical correlation between fields is difficult to capture.

- Column-wise PCA: This strategy enforces a single common latent representation (coefficients) across all fields. Its optimality is strictly conditional on the correlation between the latent coefficients of the fields when projected onto their own eigenbases.
- Row-wise PCA: Here, we construct a common basis by concatenating the fields. We demonstrate that this approach offers a critical advantage for our problem: it allows for the preservation of linear constraints within the latent space itself. Although it may theoretically yield a sub-optimal reconstruction error compared to the field-wise approach (as it seeks a compromise basis for all fields), it solves the interpretability issue by projecting all physical fields in an unique basis. We show that the optimality of the Row-wise approach is dependent on the correlation of the eigenvectors across all fields.

Constraint handling and proposed framework : Ensuring physical validity in predictions is crucial. A naive approach to handling summation constraints—when data is noise-free—is to project the data onto the constraint hyperplane. For instance, one could model $Q-1$ fields and deduce the final component via the constraint equation. However, we argue that this introduces a significant selection bias depending on which field is chosen as the dependent variable. To overcome this limitation, we propose a bias-free framework that uses all available data. Our method combines:

- Row-wise PCA for dimensionality reduction, which, as discussed, provides an interpretable common basis that respects the linear structure of the constraints.
- Constrained Multi-Output Gaussian Process (MOGP) Regression where instead of projecting data, we enforce the constraint probabilistically by embedding the linear operator directly into the multi-output GP prior via a parameterized covariance kernel [2]. This guarantees that posterior samples satisfy the linear constraint (1) exactly, while capturing inter-output correlations beyond the constraint.

We validate this framework on a case study involving the prediction of the turbulent Reynolds stress tensor for an incompressible fluid flow in a diffuser [1]. The physical validity of the predicted tensor field is governed by the incompressibility condition (trace constraint), expressed as:

$$\tau_{11} + \tau_{22} - k = 0$$

where k represents the turbulent kinetic energy. We benchmark our proposed Row-wise PCA + Constrained MOGP approach against standard projection-based methods and unconstrained baselines. The results demonstrate that our method not only strictly satisfies the physical constraints but also eliminates selection bias, yielding robust predictions with physically meaningful uncertainty quantification.

Short biography (PhD student)

Nassouradine Mahamat Hamdan is a PhD Student in CEA Saclay and is directed by Prof. Sébastien Da Veiga at ENSAI CREST. He obtained a Master's degree in Applied Mathematics from the University of Reims-Champagne Ardenne, specializing in scientific computing. The PhD topic consists of proposing an uncertainty quantification framework for the Reynolds stress tensor modeling.

This work has been half funded by the *Agence Nationale de la Recherche* Exa-MA project, under the France 2030 initiative, with reference ANR-22-EXNU-0002.

References

[1] Carl U. Buice and John K. Eaton. Experimental investigation of flow through an asymmetric plane diffuser: (data bank contribution)1. *Journal of Fluids Engineering*, 122(2):433–435, 01 2000.

[2] A. WIlls et Thomas B.Shon C. Jidling, N. Wahlstrom. Linearly constrained Gaussian processes. *31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.*, pages 2–5, 2017.

[3] Dave Higdon, James Gattiker, Brian Williams, and Maria Rightley. Computer model calibration using high-dimensional output. *Journal of the American Statistical Association*, 103(482):570–583, 2008.

[4] R. W. Preisendorfer and C. D. Mobley. *Principal Component Analysis in Meteorology and Oceanography*. Elsevier, 1988.

Multitask Gaussian processes with functional covariates

Razak C. Sabi Gninkou[†,1], Andrés F. López-Lopera[§,2], Franck Massa[3], Rodolphe Le Riche[§,4]

[†] PhD student (presenting author). [§] PhD supervisors

PhD expected duration: Dec. 2024 – Nov. 2027

[1] CERAMATHS, INSA Hauts-de-France, Univ. Polytechnique Hauts-de-France
RazakChristophe.SabiGninkou@uphf.fr

[2] IMAG, Univ. de Montpellier, Inria LEMON, Montpellier, CNRS

[3] LAMIH, INSA Hauts-de-France, Univ. Polytechnique Hauts-de-France, CNRS

[4] LIMOS, Univ. Clermont Auvergne, CNRS

Abstract

Many numerical models in mechanics and engineering involve inputs that are described as curves or fields, such as loading histories, operating profiles, or spatially distributed boundary conditions (see an example in Figure 1 for a force-displacement driver considered in a riveting process [6]). Within the functional data analysis framework, such quantities are modeled as elements of infinite-dimensional function spaces, which allows their intrinsic structure and variability to be preserved rather than reduced to finite-dimensional vectors [4]. Gaussian process (GPs) regression then provides a natural probabilistic surrogate modeling framework for such settings, enabling flexible nonparametric prediction together with principled uncertainty quantification [5].

Recent studies have demonstrated the effectiveness of GP based surrogates for models driven by functional covariates [2, 3, 7]. However, most existing approaches focus on single-task settings and do not explicitly account for dependencies between multiple responses. Empirical correlation analyses between tasks may reveal strong dependencies, including both positive and negative correlations, which will be inherently ignored by independent GPs. This limitation motivates the construction of multitask formulations designed to jointly model correlated responses [1].

In this contribution, we propose a multitask GP surrogate model tailored to functional covariates, based on a separable covariance structure across three complementary dimensions: the functional description of the inputs, scalar variables, and tasks. This construction enables inter-task correlations to be explicitly modeled while retaining an interpretable structure. From a computational perspective, the resulting covariance operator admits a Kronecker product representation for tensor-structured data, which can be efficiently exploited through structured tensor algebra for exact inference and prediction, even in the presence of high-dimensional functional data.

Our framework is assessed on a realistic riveting process application involving inputs and outputs representing force-displacement curves. For the latter application, Figure 1 shows the predictive envelopes obtained by our framework, and the estimated inter-task correlation matrix. According to our experiments, the multitask formulation consistently improves predictive accuracy and uncertainty calibration compared to independent GPs models, highlighting the benefits of jointly modeling correlated responses in complex numerical simulations.

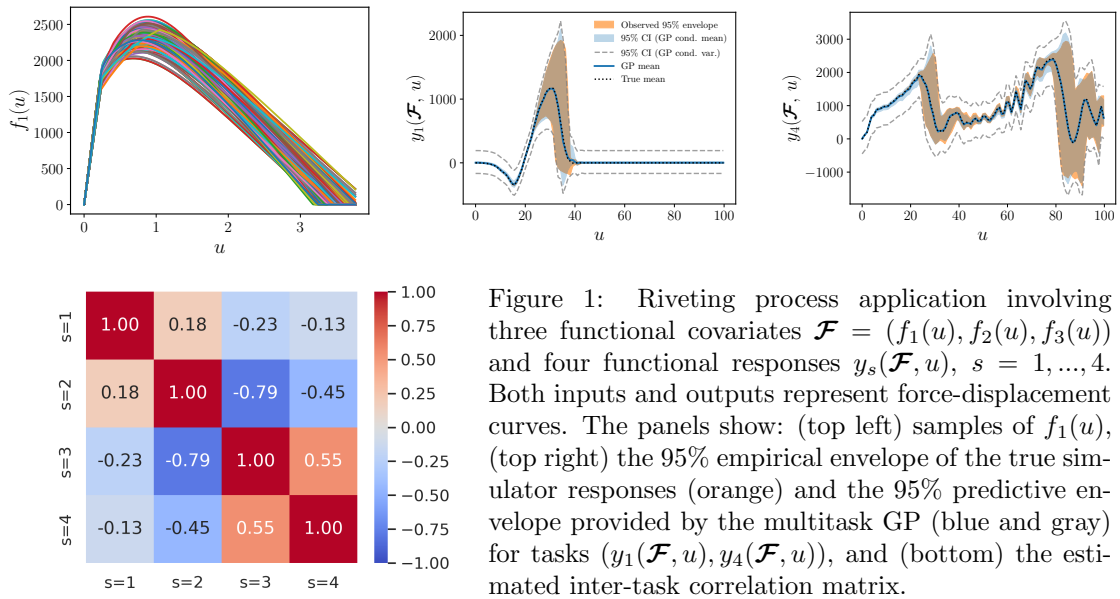


Figure 1: Riveting process application involving three functional covariates $\boldsymbol{\mathcal{F}} = (f_1(u), f_2(u), f_3(u))$ and four functional responses $y_s(\boldsymbol{\mathcal{F}}, u)$, $s = 1, ..., 4$. Both inputs and outputs represent force-displacement curves. The panels show: (top left) samples of $f_1(u)$, (top right) the 95% empirical envelope of the true simulator responses (orange) and the 95% predictive envelope provided by the multitask GP (blue and gray) for tasks $(y_1(\boldsymbol{\mathcal{F}}, u), y_4(\boldsymbol{\mathcal{F}}, u))$, and (bottom) the estimated inter-task correlation matrix.

References

[1] E. V. Bonilla, A. K. M. Chai, and C. K. I. Williams. Multitask Gaussian process prediction. *Advances in Neural Information Processing Systems*, 20:153–160, 2008.

[2] A. F. López-Lopera, F. Bachoc, J. Idier, R. Pedreros, and J. Rohmer. Multi-output Gaussian processes with functional data: A study on coastal flood hazard assessment. *Reliability Engineering & System Safety*, 219:108227, 2022.

[3] A. F. López-Lopera, F. Massa, I. Turpin, and N. Leconte. Modeling complex mechanical computer codes with functional input via Gaussian process. In *Proceedings of the XLIII Ibero-Latin American Congress on Computational Methods in Engineering (CILAMCE)*, 2022.

[4] J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer, 2 edition, 2005.

[5] C. E. Rasmussen and C. K. I. Williams. *Gaussian processes for Machine Learning*. MIT Press, Cambridge, MA, 2006.

[6] R. C. Sabi Gninkou, A. F. López-Lopera, F. Massa, and R. Le Riche. Scalable multitask Gaussian processes for complex mechanical systems with functional covariates. *Work to be submitted in Computer Methods in Applied Mechanics and Engineering*, 2026.

[7] C. L. Sung, W. Wang, F. Cakoni, I. Harris, and Y. Hung. Functional-input Gaussian processes with applications to inverse scattering problems. *Statistica Sinica*, pages SS–2022–0180, 2022.

Short biography (PhD student)

Razak C. Sabi Gninkou received the Master's degree in Statistics and Machine Learning from Sorbonne University (Paris). He is currently a PhD student in Applied Mathematics at Univ. Polytechnique Hauts-de-France, in collaboration with École des Mines de Saint-Étienne and Univ de Montpellier. His research focuses on the metamodeling of complex systems with functional covariates. His PhD is conducted within the ANR JCJC GAME project.

Statistical test to detect misspecification of a covariance kernel in Gaussian process regression

Théo Sylvestre[†,1,3], Guillaume Damblin[§,1], Raksmy Nop[§,1], Amandine Marrel[¶,2,3]

[†] PhD student [¶] PhD supervisor [§] PhD co-supervisor

PhD expected duration: Nov. 2024 – Oct. 2027

[1] CEA Saclay, 91190 Gif-sur-Yvette, France,
{theo.sylvestre,guillaume.damblin,raksmy.nop}@cea.fr

[2] CEA, DES, IRESNE, DER, Cadarache, France,
amandine.marrel@cea.fr

[3] LMA Université d'Avignon, EA 2151, 84029 Avignon,

Abstract

In computer experiments, numerical codes $\mathcal{M}$ aim to simulate physical phenomena as accurately as possible. However, most of the time, those high-fidelity simulation tools are time consuming, preventing their intensive use for an in-depth understanding of the simulated phenomena. To overcome this, the simulator is replaced by a surrogate model, also called a metamodel. A popular method for modeling expensive computer experiment is the Gaussian Process Regression ($\mathcal{GPR}$). As emphasised in [5] and [2], this non-parametric method is particularly well suited thanks to its flexibility and the natural uncertainty quantification in its predictions, which is essential for decision-making in safety-critical domains such as nuclear engineering or environmental risk assessment as explained in [7] and [3]. A $\mathcal{GPR}$ model is fully specified by its prior mean function—taken as zero here—and its covariance kernel $k_{\boldsymbol{\theta}}(\cdot,\cdot)$, whose hyperparameters (HPs) $\boldsymbol{\theta}$ are estimated from the data, usually using Maximum Likelihood (ML) estimation. In most applications, $k_{\boldsymbol{\theta}}(\cdot,\cdot)$ is assumed to be stationary, meaning that it only depends on the difference between two input locations, not on their absolute positions. Assuming such stationary covariance may be inappropriate when the numerical model exhibits distinct behaviours across different regions of the input space. This situation arises in some thermal-hydraulic transients studied in reactor safety. One notable example is reactivity insertion accidents occurring in pool-type research reactors, where non-stationarity is suspected. Appropriate covariance kernels, whose structure depends on absolute input locations rather than solely on relative differences, should then be used, leading to complex non-stationary $\mathcal{GPR}$ models, as emphasised in [6].

Before resorting to such advanced methods, it is important to assess whether non-stationarity is actually present. Its detection may rely on expert judgement or be suggested by descriptive statistics of the output variable. More commonly, it is inferred from diagnostics applied to an initially fitted stationary $\mathcal{GPR}$ model, following the approach described in [1], which helps to assess whether the stationary assumption is adequate. In this framework, our contribution is a new statistical test designed to assess whether the assumed covariance structure is adequate for a given $\mathcal{GPR}$ problem, in a setting where the hyperparameters are neither known nor fixed. Only a parametric family for the kernel is specified, and the test explicitly accounts for the uncertainty induced by hyperparameter estimation. This provides a framework for $\mathcal{GPR}$ model

validation and new insights for detecting non-stationarity. In particular, if a broad class of stationary kernels is rejected, this strongly suggests the presence of non-stationarity in the data.

For a given $\mathcal{GPR}$ problem with learning sample $B_\mathrm{l} = (\mathbf{X}_\mathrm{l}, \mathbf{Y}_\mathrm{l})$, where $\mathbf{X}_\mathrm{l} \in \mathcal{M}_{n_\mathrm{l},d}(\mathbb{R})$ is the experimental design matrix and $\mathbf{Y}_\mathrm{l} \in \mathbb{R}^{n_\mathrm{l}}$ contains the corresponding scalar outputs of $\mathcal{M}$, the competing hypotheses about the data are:

$$H_0 : \exists \boldsymbol{\theta},\ \mathcal{M} \mid B_\mathrm{l} \ \sim\ \mathcal{GP}(\bar{m}_{\boldsymbol{\theta}}(\cdot), \bar{k}_{\boldsymbol{\theta}}(\cdot,\cdot)) \quad \text{vs} \quad H_1 : \forall \boldsymbol{\theta},\ \mathcal{M} \mid B_\mathrm{l} \not\sim \mathcal{GP}(\bar{m}_{\boldsymbol{\theta}}(\cdot), \bar{k}_{\boldsymbol{\theta}}(\cdot,\cdot))$$

with $\bar{m}_{\boldsymbol{\theta}}(\cdot)$ the posterior mean and $\bar{k}_{\boldsymbol{\theta}}(\cdot,\cdot)$ the posterior covariance.

We consider two test statistics T that are representative of H_0. The predictivity coefficient Q^2, computed from the posterior mean, quantifies the proportion of variance in the observed data that is explained by the model. The coverage criterion IAE, computed from the whole predictive distribution, measures the overall average L^1 error of the predictive interval coverage rates. Q^2 and IAE are evaluated on a validation set $B_\mathrm{v} = (\mathbf{X}_\mathrm{v}, \mathbf{Y}_\mathrm{v})$ of size n_v, distinct from B_l. Both T-based tests are designed to assess whether the observed T is consistent with the values it would take under the tested model. The bilateral Q^2-based test assesses whether the model's predictive performance is significantly better or worse than expected, while the right-sided IAE-based test evaluates whether the predictive variance provides as reliable confidence intervals as expected. For a significance level $\alpha \in [0,1]$ and each test statistic T (either Q^2 or IAE), the following Algorithm 1 is applied to a learning sample B_l, a validation sample B_v, and a given parametric covariance $k_\theta(\cdot,\cdot)$:

Algorithm 1 New procedure to detect misspecification of a covariance kernel

Step 1: Model training on B_l using ML estimation

1. Estimate $\hat{\boldsymbol{\theta}}$ by maximising the likelihood induced by the prior $\mathcal{GP}(0, k_\theta(\cdot,\cdot))$ on the learning sample B_l
2. Condition the $\mathcal{GP}(0, k_{\widehat{\boldsymbol{\theta}}}(\cdot,\cdot))$ on B_l using $\widehat{\boldsymbol{\theta}}$ to obtain the predictive distribution
3. Compute the observed test statistic T_obs on the validation set B_v

Step 2: Parametric bootstrap to approximate the distribution of $\widehat{\boldsymbol{\theta}}$ under H_0

1. Generate $n_{\boldsymbol{\theta}}$ virtual observations sets $\left(\mathbf{Y}^{(i)}_{\mathrm{l}_\mathrm{virt}}\right)_{i\in[\![1,n_{\boldsymbol{\theta}}]\!]}$ at $\mathbf{X}_\mathrm{l}$ from the prior $\mathcal{GP}\big(0,\ k_{\widehat{\boldsymbol{\theta}}}(\cdot,\cdot)\big)$
2. For each i, estimate HPs $\widehat{\boldsymbol{\theta}}_{i,\mathrm{virt}}$ with ML on virtual learning set $B^{(i)}_{\mathrm{l}_\mathrm{virt}} = (\mathbf{X}_\mathrm{l}, \mathbf{Y}^{(i)}_{\mathrm{l}_\mathrm{virt}})$

Step 3: Simulation-based approximation of the H_0 distribution of T

1. **For each** $i \in [\![1, n_{\boldsymbol{\theta}}]\!]$:
   1) Generate new virtual observations $(\mathbf{Y}_{\mathrm{l}_\mathrm{virt}}, \mathbf{Y}_{\mathrm{v}_\mathrm{virt}})$ at $(\mathbf{X}_\mathrm{l}, \mathbf{X}_\mathrm{v})$ from the prior $\mathcal{GP}(0, k_{\widehat{\boldsymbol{\theta}}_{i,\mathrm{virt}}}(\cdot,\cdot))$
   2) Estimate $\widetilde{\boldsymbol{\theta}}_{i,\mathrm{virt}}$ with ML on virtual learning sample $B_{\mathrm{l}_\mathrm{virt}} = (\mathbf{X}_\mathrm{l}, \mathbf{Y}_{\mathrm{l}_\mathrm{virt}})$
   3) Condition the $\mathcal{GP}(0, k_{\widetilde{\boldsymbol{\theta}}_{i,\mathrm{virt}}}(\cdot,\cdot))$ on $B_{\mathrm{l}_\mathrm{virt}} = (\mathbf{X}_\mathrm{l}, \mathbf{Y}_{\mathrm{v}_\mathrm{virt}})$ using $\widetilde{\boldsymbol{\theta}}_{i,\mathrm{virt}}$ to obtain the predictive distribution
   4) Compute T on $B_{\mathrm{v}_\mathrm{virt}} = (\mathbf{X}_\mathrm{v}, \mathbf{Y}_{\mathrm{v}_\mathrm{virt}})$
2. Compute the rejection region at level α from the empirical distribution of T (using empirical quantiles)

Step 4: Final decision of the test

Compare the T_obs with the level-α rejection zone

Across all significance levels α, the bilateral Q^2-based test appears poorly calibrated: it fails to reject H_0 as often as the predefined level α. In contrast, the right-sided IAE-based test shows correct calibration across all significance levels, with rejection frequencies consistent with the expected behaviour under H_0. Nevertheless, to remain consistent with the usual framework of $\mathcal{GPR}$ model validation based on both the posterior mean and variance, those criteria are combined into a multiple testing procedure using a Bonferroni correction [4], allowing the control of the final significance level α by an upper bound.

As this procedure yielded encouraging results to detect the misspecification of the covariance kernel on various numerical examples, it is applied to a reactivity insertion use case involving 3 uncertain inputs, with $n_{\mathrm{l}} = 275$ and $n_{\mathrm{v}} = 45$. Resulting p-values, for each T-based test, together with corresponding final test decisions for $\alpha = 5\%$, are reported in Table 1:

Kernel	p-value Q^2	p-value IAE	Final decision of the Bonferroni correction
$\frac{1}{2}$-Matérn	0.955	0.000	Rejected
$\frac{3}{2}$-Matérn	0.052	0.006	Rejected
$\frac{5}{2}$-Matérn	0.000	0.100	Rejected

Table 1: Statistical tests results for the reactivity insertion use case

The proposed test procedure rejected all three stationary kernels, providing strong evidence of non-stationarity in the data. This result is consistent with prior expert assessments and highlights the need to fit an appropriate non-stationary surrogate model.

Short biography (PhD student)

Theo Sylvestre is a PhD student at CEA Saclay supervised by Amandine Marrel. He completed a Master's degree in Statistics at ENSAI before starting this PhD. This project aims to develop a novel Gaussian process regression model to emulate non-stationary phenomena, intended to improve the analysis of reactivity insertion accidents.

References

[1] Leonardo S. Bastos and Anthony O'Hagan. Diagnostics for gaussian process emulators for estimating piecewise continuous regression functions. Technometrics, 51(4):425–438, 2009.

[2] Robert B. Gramacy. *Surrogates: Gaussian Process Modeling, Design, and Optimization for the Applied Sciences*. Chapman and Hall/CRC, 1st edition, 2020.

[3] Amandine Marrel and Bertrand Iooss. Probabilistic surrogate modeling by gaussian process: A review on recent insights in estimation and validation. Reliability Engineering and System Safety, 2024.

[4] J. Neyman and E. S. Pearson. On the use and interpretation of certain test criteria for purposes of statistical inference part i. *Biometrika*, 20A(1-2):175–240, 12 1928.

[5] Jerome Sacks, William J. Welch, Toby J. Mitchell, and Henry P. Wynn. Design and Analysis of Computer Experiments. *Statistical Science*, 4(4):409 – 423, 1989.

[6] Annie Sauer, Andrew Cooper, and Robert B. Gramacy. Non stationary gaussian process surrogates. Reliability Engineering and System Safety, 2024.

[7] Kush R. Varshney. *Trustworthy Machine Learning*. Independently Published, Chappaqua, NY, USA, 2022.

Reliable criterion for decision-making using risk measures

M. Temple-Boyer[†,1,2], G. Perrin[§,2], V. Chabridon[1], J. Pelamatti[1], E. Remy[1]

[†] PhD student (presenting author). [§] PhD supervisor

PhD expected duration: Sep. 2024 – Aug. 2027

[1] EDF R&D, 6 quai Watier, 78401 Chatou, France
{marie.temple-boyer,vincent.chabridon,julien.pelamatti,emmanuel.remy}@edf.fr

[2] Université Gustave Eiffel, COSYS, 14-20 Boulevard Newton, 77447 Marne-la-Vallée, France
{marie.temple-boyer2, guillaume.perrin}@univ-eiffel.fr

Abstract

The reliability assessment of a critical industrial system (such as electricity power plant) is based on an analysis of uncertainties and, ultimately, on the estimation of a *risk measure* characterizing the risk of a system failure. This estimate can be obtained using a numerical simulator that models the behavior of the system in its environment. The risk measures generally used correspond to failure probabilities or high-order quantiles of the ouput Y of the simulator. Nevertheless, a wider range of risk measures may be considered in practice. For example, the superquantile [3] and the buffered failure probability [1] are quantities used in risk quantification in finance, and may also be relevant in engineering, partly due to their many interesting theoretical properties.

The risk measures can be classified into two different categories: Y-homogeneous risk measures, that have the same unit of measure as the output Y, and $[0,1]$-homogeneous risk measures, that are homogeneous to a probability, i.e., valued in $[0,1]$. The quantile and the superquantile are Y-homogeneous risk measures, whereas the failure and buffered failure probabilities are $[0,1]$-homogenous risk measures. A connection can be made between these two categories: the quantile is associated to the failure probability, and the superquantile to the buffered failure probability. More precisely, guaranteeing that a Y-homogeneous risk measure is below some threshold is equivalent to guaranteeing that the associated $[0,1]$-homogeneous risk measure is below some acceptable risk [4].

A general formalism for decision-making based on risk measure estimation is proposed in [4]. This formalism allows to guarantee, with some confidence level, that a risk measure is below a given threshold. When used on the failure probability, this framework is similar as the one developped in [2]. It gives a criterion to guarantee that the failure probability is below some acceptable risk. It is equivalent to a decision criterion made on the Wilks estimator of the quantile [5]. If we further assume that output Y is bounded, then the framework can be used on the superquantile, giving the guarantee that the superquantile is below a given safety threshold, or equivalently that the buffered failure probability is small enough.

The main interest of such a framework is the non-asymptotical statistical guarantee. However, there is a critical sample size that must be reached to obtain a relevant criterion. For classical Monte-Carlo estimators and when the acceptable risk is close to zero, this critical sample size may be prohibitive. The usage of more advanced estimators, with reduced variance in the rare event case, is therefore studied in this decision-making framework.

Short biography (PhD student)

Marie Temple-Boyer graduated from *École Nationale des Ponts et Chaussées* in June 2023 and was awarded the "Agrégation de Mathématiques" in June 2024. She is now involved in a PhD program financed by a CIFRE grant between EDF R&D and *Université Gustave Eiffel* (under the supervision of Guillaume Perrin). Her thesis aims to propose certifiable approaches for the simulation and estimation of rare events.

References

[1] A. Mafusalov and S. Uryasev. Buffered Probability of Exceedance: Mathematical Properties and Optimization. *SIAM Journal on Optimization*, 28(2):1077–1103, 2018.

[2] G. Perrin, J. Reygner, and V. Chabridon. Enhancing reliability analysis with limited observations: A statistical framework for system safety margins. *Structural Safety*, page 102670, 2025.

[3] R.T. Rockafellar and J.O. Royset. Engineering Decisions under Risk Averseness. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, 1(2):04015003, 2015.

[4] M. Temple-Boyer, E. Cussenot, G. Perrin, V. Chabridon, J. Pelamatti, E. Remy, and B. Iooss. Risk measures in engineering reliability: comparison, duality and decision-making. 2025.

[5] S. S. Wilks. Determination of Sample Sizes for Setting Tolerance Limits. *The Annals of Mathematical Statistics*, 12(1):91–96, 1941.

Uncertainty quantification for nonlinear terms approximation in reduced order modeling

R. Tiphaigne[†,1], V.Resseguier[§,1], G.Stabile[§,2], D.Heitz[§,1]

[†] PhD student (presenting author). [§] PhD supervisor

PhD expected duration: Nov. 2024 – Oct. 2027

[1] INRAE, UR OPAALE, Rennes, France
{romain.tiphaigne, valentin.resseguier, dominique.heitz}@inrae.fr

[2] Sant'Anna School of Advanced Studies, The Biorobotics Institute, Pontedera, Pisa, Italy
giovanni.stabile@santannapisa.it

Abstract

For real-time monitoring, such as digital twins, we want to predict a system behavior on the fly. We therefore need a low complexity model as accurate as possible. Assuming availability of spatially sparse measurements, we can enhance these observations with a priori knowledge on the system. It leads to a two step approach involving a stochastic Reduced-Order Model (sROM) [4] and data assimilation [3]. The sROM learns a priori knowledge from high-fidelity simulations and physics equations through a low dimensional model. One can see it either as a generative model or as a predictor including uncertainty quantification. The data assimilation step combines the sROM prediction with the accessible measurements.

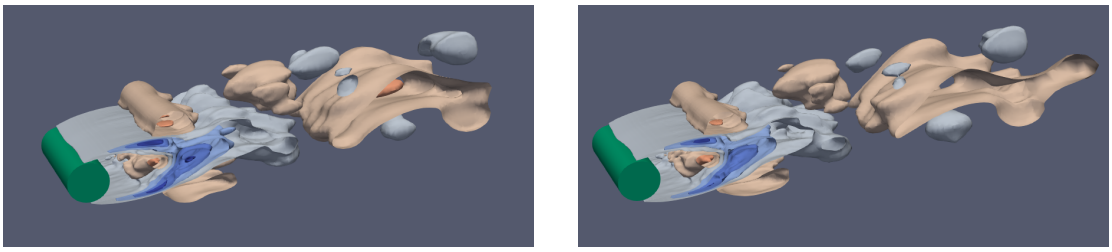
In this work, we focus on the hyperreduction part of the sROM. The core of the sROM is the dimensionality reduction that we perform with POD-Galerkin, where POD stands for Proper Orthogonal Decomposition. It mainly consists of projecting linearly the physics partial differential equation onto a linear reduced basis built through PCA (POD). However, there is no complexity reduction when applying this method directly to nonlinear terms of the original physical equation. Hyperreduction comes into play to preserve dimensionality reduction when projecting nonlinear terms.

There are two main types of hyperreduction method, referred to as the Discrete Empirical Interpolation Method (DEIM) [1] and Empirical Cubature [2]. Both methods are approximations, so neither gives an exact projection. The modeling of their errors to enhance the uncertainty quantification in the sROM is the main point of the presentation. We focus on DEIM, which interpolates the nonlinear terms from few local evaluations before projecting onto the PCA modes.

We show that the DEIM method is similar to Gaussian process regression (kriging) with a covariance built from an additional PCA learned on a training set of the nonlinear terms. This PCA suits well reduced models, also built from a training data set of high fidelity simulations. Similitude between DEIM and kriging is particularly helpful to quantify the hyperreduction uncertainty. It includes contributions from the PCA truncation and Bayesian quadrature. This second contribution results from the projection of the kriging approximation onto the PCA modes.

We apply this methodology in the context of 3D unsteady turbulent fluid dynamics. The fields are statistically non-stationary in both space and time. Results focus on time extrapolation of

the nonlinear terms inside the sROM. Quality of the deterministic hyperreduction prediction and its associated confidence interval are assessed.



(a) Reference high dimensional field

(b) Kriging conditional mean approximation with 16 PCA modes and 16 observation points

Figure 1: Iso-surfaces of the target nonlinear term at a given time step.

Short biography (PhD student)

My PhD thesis is funded by the RedLUM ANR project. The project's objectives are to develop tools for real-time estimation and short-term prediction of 3D unsteady fluid flows, using limited computational resources. It is a collaboration with researchers from INRAE Rennes, Sant'Anna School of Pisa and Inria Bordeaux, as well as the companies Weather Measures and SCALIAN DS. I previously graduated from the Applied Mathematics section of INSA Toulouse.

References

[1] Saifon Chaturantabut and Danny C. Sorensen. Nonlinear Model Reduction via Discrete Empirical Interpolation. *SIAM Journal on Scientific Computing*, 32(5):2737–2764, January 2010. Publisher: Society for Industrial and Applied Mathematics.

[2] J. A. Hernández, M. A. Caicedo, and A. Ferrer. Dimensional hyper-reduction of nonlinear finite element models via empirical cubature. *Computer Methods in Applied Mechanics and Engineering*, 313:687–722, January 2017.

[3] Valentin Resseguier, Matheus Ladvig, and Dominique Heitz. Real-time estimation and prediction of unsteady flows using reduced-order models coupled with few measurements. *Journal of Computational Physics*, 471:111631, 2022.

[4] Valentin Resseguier, Agustin M. Picard, Etienne Memin, and Bertrand Chapron. Quantifying Truncation-Related Uncertainties in Unsteady Fluid Dynamics Reduced Order Models. *SIAM/ASA Journal on Uncertainty Quantification*, 9(3):1152–1183, January 2021.

Importance Sampling for Sobol' Indices: Variance Minimization and Bayesian Connections - poster

Haythem Boucharif[†,1,2], Jérôme Morio[§,1,2], Paul Rochet[§,2]

[†] PhD student (presenting author). [§] PhD supervisor

PhD expected duration: Oct. 2024 – Sep. 2027

[1] ONERA/DTIS, Université de Toulouse, Toulouse, 31055, France

[2] Fédération ENAC ISAE-SUPAERO ONERA, Université de Toulouse, Toulouse, France

Abstract

We propose a preferential sampling framework for the efficient estimation and analysis of Sobol' indices [6, 5]. Let $X = (X_1, \dots, X_d)$ be a random input vector with independent components and density p, and let $f : \mathbb{R}^d \to \mathbb{R}$ be a model such that $f(X) \in L^2$. In the classical setting, first-order Sobol' indices satisfy

$$S_i = \frac{\mathrm{Var}(\mathbb{E}[f(X) \mid X_i])}{\mathrm{Var}(f(X))} = \frac{\eta_i - \mathbb{E}[f(X)]^2}{\mathrm{Var}(f(X))}, \qquad \eta_i := \mathbb{E}\big[m_i(X_i)^2\big], \quad m_i(x_i) = \mathbb{E}\big[f(X) \mid X_i = x_i\big],$$

so that the conditional second-moment η_i is the key quantity to estimate and typically the main source of statistical difficulty. Efficient estimators and asymptotic efficiency results for Sobol' indices motivate focusing on η_i [7, 2, 4].

We show that η_i, originally defined under the reference distribution p, can be estimated from samples drawn under an auxiliary density q by reweighting model outputs, in the spirit of importance sampling [3]. For a factorized proposal $q(x) = q_i(x_i)\, q_{-i}(x_{-i} \mid x_i)$, define

$$Z_i = \sqrt{\frac{p_i(X_i)}{q_i(X_i)}}\, \frac{p_{-i}(X_{-i})}{q_{-i}(X_{-i} \mid X_i)}\, f(X), \qquad X \sim q,$$

which yields the identity

$$\eta_i = \mathbb{E}_{q_i}\Big[\big(\mathbb{E}_{q_{-i}(\cdot \mid X_i)}[Z_i \mid X_i]\big)^2\Big].$$

This representation provides importance sampling estimators of η_i and enables a direct analysis of their asymptotic variance. We derive an optimal joint sampling distribution minimizing the asymptotic variance of efficient estimators and show that, when $f \geq 0$, there exists $q^\star$ such that the optimal variance vanishes [1], with

$$q^\star(x) \propto p(x)\, f(x)\, m_i(x_i).$$

This characterization clarifies how preferential sampling should concentrate on influential regions of the input space.

Pseudo-Bayesian viewpoint and harmonic-mean estimation. When $f \geq 0$, the optimal marginal density admits the pseudo-Bayesian form

$$\underbrace{q_i^\star(x_i)}_{\text{posterior}} \propto \underbrace{p_i(x_i)}_{\text{prior}} \times \underbrace{m_i(x_i)^2}_{\text{pseudo-likelihood}}, \qquad \underbrace{\eta_i}_{\text{evidence}} = \int p_i(x_i)\, m_i(x_i)^2\, dx_i.$$

Hence, the quantity η_i appears as a normalizing constant (marginal likelihood). For any density φ such that $\int \varphi(x_i)\, dx_i = 1$,

$$\rho_i = \mathbb{E}_{q_i^\star}\left[\frac{\varphi(X_i)}{p_i(X_i)\, m_i(X_i)^2}\right] = \frac{1}{\eta_i}.$$

Moreover,

$$\mathrm{Var}_{q_i^\star}\left(\frac{\varphi(X_i)}{p_i(X_i) m_i(X_i)^2}\right) = 0 \quad \text{iff} \quad \varphi = q_i^\star,$$

highlighting the optimality of the target density itself.

This interpretation turns the estimation of η_i into a marginal-likelihood problem and enables the use of standard Bayesian evidence techniques. Crucially, once an approximation of $q_i^\star$ is learned (e.g. via MCMC or related methods), η_i can be estimated directly from the generated samples without requiring additional evaluations of the expensive model f, in contrast to classical importance sampling schemes that necessitate new runs under $q^\star$. This significantly reduces the overall computational cost while preserving variance reduction properties.

Short biography (PhD student)

Haythem Boucharif is a second-year PhD student at ONERA and ENAC, supervised by Jérôme Morio and Paul Rochet. He holds Master's degrees in Econometrics and Statistics (ISFA) and in Statistical Engineering (ISUP–Sorbonne Université). His research focuses on variance reduction and Monte Carlo methods for global sensitivity analysis. The PhD is funded by the Fédération ENAC ISAE-SUPAERO ONERA.

References

[1] H. Boucharif, J. Morio, and P. Rochet. Importance Sampling for Sobol' Indices Estimation. *arXiv preprint arXiv:2507.05958 [math.ST]*, 2025.

[2] Alexandre Janon, Thierry Klein, Agnès Lagnoux, Maëlle Nodet, and Clémentine Prieur. Asymptotic normality and efficiency of two Sobol index estimators. *ESAIM: Probability and Statistics*, 18:342–364, 2014.

[3] H. Kahn and T. E. Harris. Estimation of particle transmission by random sampling. *National Bureau of Standards Applied Mathematics Series*, 12:27–30, 1951.

[4] T. Klein and P. Rochet. Efficiency of the averaged rank-based estimator for first-order sobol' index inference. *Statistics & Probability Letters*, 207, 2024.

[5] A. Saltelli, S. Tarantola, F. Campolongo, and M. Ratto. *Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models*. John Wiley Sons, 2004.

[6] I. M. Sobol. Sensitivity analysis for non-linear mathematical models. *Mathematical Modelling and Computational Experiment*, 1(4-5):407–414, 1993.

[7] S. Da Veiga and F. Gamboa. Efficient Estimation of Sensitivity Indices. *Journal of Nonparametric Statistics*, 25(3):573–595, 2013.

Sampling on Riemannian manifolds with stochastic interpolants

Thibaut Pellerin[†,1], Alain Durmus[§,1], Eric Barat[§,2], Elinor Berger[§,2]

[†] PhD student (presenting author). [§] PhD supervisor

PhD expected duration: Nov. 2024 – Nov. 2027

[1] CMAP, CNRS, Ecole polytechnique, Institut Polytechnique de Paris, 91120 Palaiseau, France
{thibaut.pellerin,alain.durmus}@polytechnique.edu

[2] Université Paris-Saclay, CEA, List, F-91120 Palaiseau, France
{thibaut.pellerin,eric.barat,elinor.berger}@cea.fr

Abstract

Sampling from an un-normalized density is a fundamental problem in computational statistics, with applications in Bayesian inference and statistical physics. Such problems are usually tackled with Markov Chain Monte Carlo (MCMC) methods, which, to some extent, can be extended to Riemannian manifolds [4]. There are several reasons to consider sampling on Riemannian manifolds. The state-space might naturally have a manifold structure, e.g. the Grassmann manifold in shape analysis. Constraints might also naturally be expressed as a manifold, e.g., the spherical Ising model. Finally, when sampling a Bayesian posterior, the Fischer metric provides a natural geometry to the parameter space.

Regardless of the latent geometry, MCMC methods often struggle with multi-modal distributions, as mode-trapping results in poor mixing properties. Recent works proposes adapting stochastic interpolants - at the core of several gnerative models such as diffusion and flow-matching - to these sampling problems [5]. The core idea is to replace the initial problem with several intermediary, easier sampling sub-problems, for which MCMC methods are efficient.

In my first PhD year I worked on a generalization of this apporach to Riemannian manifolds [3]. Instead of the diffusion-based interpolations used in the Euclidean case, we used deterministic interpolations, inspired by recent works on Riemannian flow-matching [2].

The setting is the following: given the target distribution π_1, the interpolant is

$$X_t = \mathrm{Exp}_{X_0}\left((1-t)\,\mathrm{Log}_{X_1}(X_0)\right), \quad t \in [0,1], \tag{1}$$

the geodesic interpolation between $X_0 \sim \pi_0$ and $X_1 \sim \pi_1$, where π_0 is a simple distribution, e.g., uniform if M is compact. We define a Markov projection, solution of an ODE driven by the vector field $\mathbf{u}_t(X_t) = \mathbb{E}[\dot{X}_t|X_t]$, whose time-marginal densities $(p_t)_{t\in[0,1]}$ are the same as the intractable interpolant X. This process can be simulated using sequential Monte-Carlo estimation of $\mathbf{u}$. This requires (a) an explicit expression of the conditional densities $p_{1|t}$, which we derived for a large class of manifolds, and (b) efficient sampling of these densities.

We have illustrated this approach for several bi-modal distributions defined on the d-sphere and the Grassmann manifold, where our algorithm outperforms a naive method when using the same overall number of MCMC steps. The main drawback of this approach is the high dependency on the hyperparameter t_0, the starting time of the interpolation. Numerical results suggest the

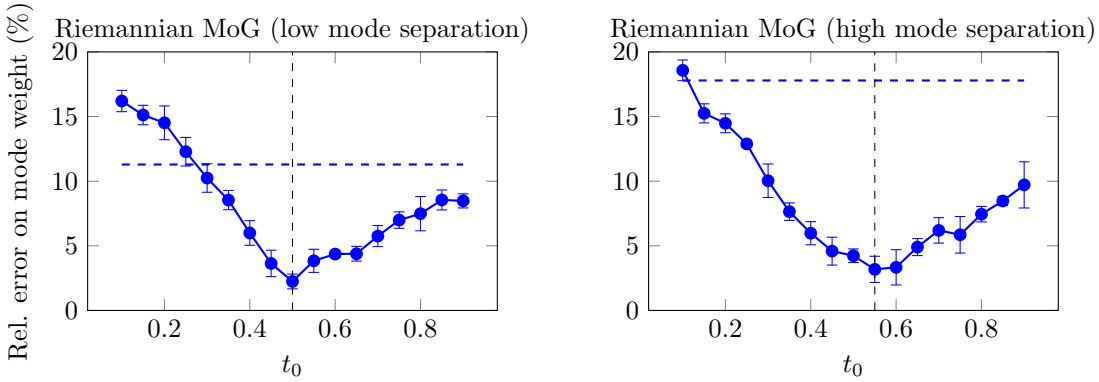


Figure 1: Performances of our algorithm (solid blue line) and MALA (dashed blue line) instances for a bi-modal Gaussian targets on the Grassmann manifold $\mathrm{Gr}(8, 3)$, as a function of t_0, in *low mode separation* regime (left) and *high mode separation* regime (right).

existence of an optimal t_0, for which high-quality samples are obtained. However, tuning this hyperparameter is still an open problem, partially adressed in [5, 3].

Since February 2026, I have started working on a concrete sampling problem. *Positron emission tomography* is a medical imaging technique which involves a difficult Poisson inverse problem, with a non-trivial Fisher metric. Most existing methods for this problem rely on optimisation, and typically do not include any uncertainty quantification. If sampling methods proved efficient, it would open the door to uncertainty quantification in this context, which can be crucial for medical applications. Riemannian sampling methods that leverage the Fisher metric are a natural approach. First, one needs a prior distribution, which is not straightforward in this context, as PET samples are not alike natural images. A prior could be learned from data using Poisson diffusion models [1], that could fit with the non-Gaussian nature of the distribution.

Short biography (PhD student)

I have started my PhD at CEA List and CMAP in november 2024, after graduating from CentraleSupélec. My PhD is funded by CEA's PRIMaL project. I work on sampling methods and their applications to Bayesian inverse problems, e.g., in the context of medical imaging.

References

[1] Sagnik Bhattacharya, Abhiram Gorle, Ahsan Bilal, Connor Ding, Amit Kumar Singh Yadav, and Tsachy Weissman. Itdpdm: Information-theoretic discrete poisson diffusion model, 2026.

[2] Ricky T. Q. Chen and Yaron Lipman. Flow matching on general geometries, 2024.

[3] Alain Durmus, Maxence Noble, and Thibaut Pellerin. Sampling from multi-modal distributions on riemannian manifolds with training-free stochastic interpolants, 2026.

[4] Mark Girolami, Ben Calderhead, and Siu A Chin. Riemann Manifold Langevin and Hamiltonian Monte Carlo.

[5] Louis Grenioux, Maxence Noble, Marylou Gabrié, and Alain Oliviero Durmus. Stochastic localization via iterative posterior sampling, 2024.

Gradient-enhanced global sensitivity analysis with Poincaré chaos expansions

D. Heredia[†,1], O. Roustant[§,1], N. Lüthen[2], B. Sudret[2]

[†] PhD student (presenting author). [§] PhD supervisor

PhD expected duration: Oct. 2023 – Sep. 2026

[1] UMR CNRS 5219, Institut de Mathématiques de Toulouse, INSA, Université de Toulouse, France

[2] Chair of Risk, Safety and Uncertainty Quantification, ETH Zürich, 8093 Zürich, Switzerland

Abstract

The analysis of complex input/output systems has received growing attention in the last decades. Two key challenges are the construction of surrogate models to approximate expensive computational codes, and the computation of Sobol indices to quantify the influence of input variables. Among the various approaches to address these tasks, sparse regression polynomial chaos expansions are a well-established and cost-efficient tool. They rely on representing the model in bases of polynomials which are orthogonal with respect to the distribution of the input parameters.

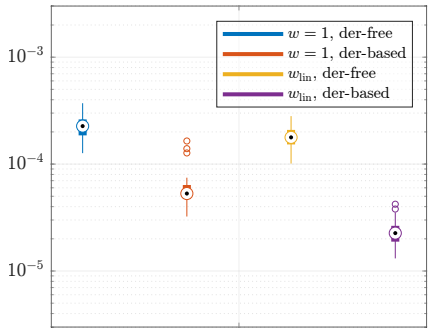
In this work we propose to use weighted Poincaré chaos expansions, where polynomial bases are replaced by the so-called Poincaré bases. Each one of these bases consists of eigenfunctions of a differential operator involving a non-negative function w, that we call a weight, as diffusion constant. A key advantage of using Poincaré bases is that they are uniquely characterized by the property that differentiating the basis functions yield another orthogonal basis, as we show by generalizing the classical case (with $w \equiv 1$) addressed in [3]. This structural property allows one to use the derivatives of the model, when available, as additional information to improve the accuracy of surrogate models and Sobol indices estimators.

Our work presents two significant contributions with respect to the existing literature [3, 4]: it incorporates gradient information into the construction of surrogate models and extends the classical Poincaré chaos framework (with $w \equiv 1$) to non-constant weights w. The introduction of a weight provides an additional degree of freedom to enhance the accuracy of the surrogate models and Sobol indices estimators. In [2] we have develop guidelines for choosing it. In our analysis, the weight w is chosen to ensure that the first non-null eigenfunction e_1, which is inherently monotonic, coincides with any prescribed monotonic function. In particular this allows us to:

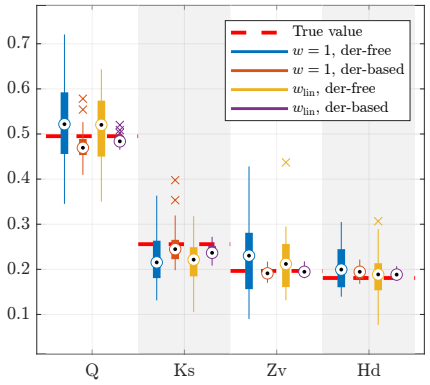
- consider the classical weight w_{lin} for which the eigenfunction e_1 is linear (also called the Stein kernel), which is therefore well suited to models with linear trends. This choice yields as eigenfunctions the three classical families of orthogonal polynomials: Hermite, Laguerre and Jacobi, associated to the normal, gamma and beta distributions, respectively. As such, gradient-enhanced surrogate modelling using these bases has already been developed for models involving these canonical measures (see e.g. [1]). Our approach makes it possible to deal with more general probability distributions;

- consider data-driven weights constructed from monotonic estimators of the so-called main effects, which capture the model's behavior with respect to each individual variable. The performance of such weights was tested in the context of Poincaré-inequalitiy-based upper bounds of total Sobol indices, a method essentially relying only on the eigenfunction e_1. By using chaos decomposition we involve the whole family of eigenfunctions.

We apply the proposed methodology to a simplified flood-dyke model, with non-standard input distributions such as truncated Gumbel, truncated normal, triangular. We present numerical results obtained using the classical Poincaré chaos, where $w \equiv 1$, as well as those obtained using the Stein kernel w_{lin}. These are shown in Figure 1, illustrating the L^2 error between the true model and the constructed surrogate models, and estimations of total Sobol indices. All the results are computed using either model evaluations only (der-free) or a combination of model and gradient evaluations (der-based). We observe that the best performance is achieved when gradient information is included and when we use the bases associated to the weight w_{lin}.



(a) L^2 error



(b) Total Sobol indices of the four most influential variables Q, K_s, Z_v and H_d.

Figure 1: Results for the annual maintenance cost of a dyke.

Short biography (PhD student)

After completing my undergraduate studies in my home country, Ecuador, I continued my studies at the Université Paris-Saclay with the support of the FMJH. I then obtained a ministerial doctoral contract at the Institut de Mathématiques de Toulouse, where I am conducting my PhD under the supervision of Aldéric Joulin and Olivier Roustant. My thesis lies at the interface between global sensitivity analysis and functional inequalities.

References

[1] Ben Adcock and Yi Sui. Compressive Hermite interpolation: Sparse, high-dimensional approximation from gradient-augmented measurements. *Constructive Approximation*, 50, 08 2019.

[2] David Heredia, Aldéric Joulin, and Olivier Roustant. On one dimensional weighted poincaré inequalities for global sensitivity analysis. *Journal of Mathematical Analysis and Applications*, 554(2), 2026.

[3] N. Lüthen, O. Roustant, F. Gamboa, B. Iooss, S. Marelli, and B. Sudret. Global sensitivity analysis using derivative-based sparse Poincaré chaos expansions. *Int. J. Uncertain. Quantif.*, 13:57–82, 2023.

[4] O. Roustant, F. Gamboa, and B. Iooss. Parseval inequalities and lower bounds for variance-based sensitivity indices. *Electron. J. Stat.*, 14:386–412, 2020.

On the construction of predictors under monotonicity and fairness constraints

M. Deronzier[†,1], F. Bachoc[§,2], O. Roustant[§,1,3], A.F. López-Lopera[§,4], L. De Lara[⋆,5], E. Odin[⋆,1] F. Gamboa[⋆,1]

[†] PhD student (presenting author). [§] PhD supervisor. [⋆] Coauthor.

PhD expected duration: Apr. 2023 – June. 2026

[1] Institut de Mathématiques de Toulouse (IMT), Université Paul Sabatier, CNRS, 5219,

[2] Laboratoire Paul Painlevé, Université de Lille, CNRS, 8524

[3] Institut National des Sciences Appliquées (INSA) de Toulouse, CNRS, 5219

[4] IMAG, Université Montpellier, CNRS, Inria LEMON, Montpellier, 34090

[5] Centre Hospitalier Universitaire de Toulouse

Abstract

In this presentation, we will address the following question:

How can we construct a predictor under functional constraints?

Classical methodologies for constructing predictors from observational data are well established; however, they generally neglect prior structural information about the underlying function. Our goal is to incorporate such knowledge by requiring that the predictor $\widehat{f}$ belongs to some constraint functional space $\mathcal{F}$.

More precisely, when $\mathcal{X}$ denotes the input space and $\mathbb{R}$ the output space, we focus on two specific families of constrained functions:

1. **Monotone functions.** Given a partially ordered space $(\mathcal{X}, \preceq)$, a function $f : \mathcal{X} \to \mathbb{R}$ is called *monotone* whenever
$$a \preceq b \;\Rightarrow\; f(a) \leq f(b).$$

2. **Equalizing maps.** Given two probability measures P and Q on $\mathcal{X}$, a map f is said to *equalize* P and Q if their pushforward by f coincide: $f_\sharp P = f_\sharp Q$, where for any probability measure μ, the pushforward of μ by f denoted $f_\sharp \mu$ is defined by $f_\sharp \mu(B) = \mu\big(f^{-1}(B)\big)$ for every Borel set $B \subset \mathbb{R}$.

These constraints arise in different contexts. Monotonicity constraints are typically motivated by physical considerations, where the output is known to evolve monotonically with respect to certain variables. Equalization constraints, instead, are rooted in fairness concerns in machine learning. Given a binary protected attribute $S \in \{0, 1\}$ (such as gender or race) and random variables X and S defined on a probability space $(\Omega, \Sigma, \mathbb{P})$, one may require statistical parity, that is, asking the outup to be independent of the protected attribute: $\widehat{f}(X, S) \perp\!\!\!\perp S$ that can be framed as $\widehat{f}$ equalizing $\mathcal{L}(X \mid S = 0)$ and $\mathcal{L}(X \mid S = 1)$.

Our presentation will discuss two tools we used to integrate structural information in the construction of a predictor. First relies on the theory of Gaussian Processes while the second on convex analysis methods.

Gaussian Process (GP) Predictors. To construct predictors $\widehat{f}$ that satisfy functional constraints, we rely on GP models and incorporate the constraint directly into the conditioning of the process. Given observational data $\{(x_1, y_1), \dots, (x_n, y_n)\} \subset \mathcal{X} \times \mathbb{R}$, a functional constraint $\mathcal{F}$, and a GP $\{W(x) : x \in \mathcal{X}\}$, we aim to construct

$$\widehat{W}_{\mathcal{F}} := \big(W \mid W(x_i) = y_i + \epsilon_i,\ i = 1, \dots, n,\ W \in \mathcal{F}\big),$$

where (ϵ_i) are i.i.d. centered Gaussian variables independent of W. We will detail approximation strategies for these conditional laws and how to derive predictors from them. In the case of monotonicity constraints, we present the work of [3], extending and generalizing approaches in [1]. For statistical parity constraints, we rely on a recent characterization [2], noting that this constraint is fundamentally non-convex, unlike the monotone case. To the best of our knowledge, we propose the first construction of "GP predictors" satisfying statistical parity constraint.

Convex-Analytic View on Monotone Predictors. Let $\mathcal{F}$ denote the space of continuous monotone functions on $\mathcal{X}$. The set $\mathcal{F}$ is a convex cone in $\mathcal{C}(\mathcal{X})$, the Banach space of continuous real-valued functions on $\mathcal{X}$. Denote $I := \{f \in \mathcal{C}(\mathcal{X}) : f(x_i) = y_i,\ i = 1, \dots, n\}$ the affine subspace of interpolating functions, and let Ω be an integral convex regularizer [4]. A valid predictor $\widehat{f}_n$ that incorporates both observational constraints monotonicity constraint is defined as the solution of the convex optimization problem

$$\min_{f \in \mathcal{C}(\mathcal{X})} \ \Omega(f) + \chi_{\mathcal{F} \cap I}(f), \tag{$\star$}$$

where χ_A denotes the indicator function of a set A, equal to 0 on A and $+\infty$ otherwise.
($\star$) can be interpreted as the minimization of Ω over the feasible set $\mathcal{F} \cap I$. The Rockafellar–Fenchel duality Theorem, together with characterization of the subdifferential of the Legendre-Fenchel transform of Ω [4], allows to construct the minimizer of ($\star$) from the maximizer of the associated dual problem. This dual formulation highlights the importance of characterizing the dual of the cone of monotone functions:

$$\mathcal{F}^\star := \{\mu \in \mathcal{R}(\mathcal{X}) : \int_{\mathcal{X}} f \, d\mu \leq 0, \text{ for all } f \in \mathcal{C}(\mathcal{X})\}.$$

The result we will briefly demonstrate is the following equivalence:

$$\mu \in \mathcal{F}^\star \iff \mu(S) \leq 0 \quad \text{for all monotone sets } S \subset \mathcal{X},$$

where a set S is called *monotone* whenever it satisfies $x \in S$ and $x \preceq y$ imply $y \in S$.

Short biography Mathis Deronzier did his master degree at Mines de Saint-Étienne and achieved the agrégation in 2022. He started his thesis at the IMT april 11 2023 thanks to the ANR GAP grant and will defend his PhD thesis between april and june 2026.

References

[1] François Bachoc, Andrés F. López-Lopera, and Olivier Roustant. Sequential Construction and Dimension Reduction of Gaussian Processes Under Inequality Constraints. *SIAM Journal on Mathematics of Data Science*, 4(2):772–800, June 2022.

[2] Lucas De Lara, Mathis Deronzier, Alberto González-Sanz, and Virgile Foy. On the Non-convexity of Push-Forward Constraints and its Consequences in Machine Learning. *SIAM Journal on Mathematics of Data Science*, 7(2):597–620, June 2025.

[3] Mathis Deronzier, Andrés F. López-Lopera, François Bachoc, Olivier Roustant, and Jérémy Rohmer. Block-Additive Gaussian Processes under Monotonicity Constraints. *Statistics and Computing*, 36(1):42, December 2025.

[4] Ralph Rockafellar. Integrals which are convex functionals. II. *Pacific journal of mathematics*, 39(2):439–469, 1971.

An Efficient Framework for A- and B-Basis Value Estimation under Epistemic Uncertainty using Importance Sampling and Control Variates

E. Donfack-Siewe[†,1,2], J. Morio[§,2], S. Dubreuil[§,2], J.-P. Navarro[§,1], C. Fagiano[§,3]

[†] PhD student. [§] PhD supervisor

PhD expected duration: jan. 2024 – Dec. 2026

[1] AIRBUS OPERATIONS SAS, 316 route de Bayonne, Toulouse, 31060, France
{elton.e.donfack-siewe,jean-philippe.navarro}@airbus.com

[2] ONERA/DTIS, Universite de Toulouse, Toulouse, 31055, France
{jerome.morio, sylvain.dubreuil}@onera.fr

[3] ONERA/DMAS, Universite Paris Saclay, Châtillon, 92320, France
christian.fagiano@onera.fr

Abstract

Estimating conservative quantiles, specifically A- and B-basis values, is essential for robust aerospace certification. These statistical tolerance bounds are defined at a 95% confidence level, where the A-basis represents the 1st percentile ensuring that at least 99% of the population exceeds the threshold, and the B-basis represents the 10th percentile guaranteeing 90% population coverage. While traditionally established through extensive physical testing campaigns [1], the ongoing research now investigates how to estimate these values directly from stochastic numerical models within a mixed aleatory and epistemic uncertainty while being complaint with the current means of compliance (analysis supported by tests). This work estimates these quantiles within a mixed aleatory-epistemic framework mirroring the multi-scale aerospace certification pyramid using stochastic numerical models. We explicitly model epistemic uncertainties at each level: probabilistic model identification from limited test data [2], surrogate modeling approximation error [3], and statistical estimation uncertainty arising from finite computational budgets [2].

We propose a probabilistic framework for estimating these confidence bounds, where all epistemic uncertainties are modeled as random variables. The method leverages variance reduction techniques, specifically importance sampling and control variates [4], to obtain accurate quantile estimates. A key advantage of this strategy is that it uses the surrogate model strictly as a variance reduction tool, thereby preserving the unbiasedness of the quantile estimator regardless of the surrogate's fidelity. Furthermore, the framework enables sensitivity analysis based on Sobol indices at no extra computational cost, identifying whether physical testing or numerical simulation contributes most to the variance of the quantile estimator.

To illustrate the method's industrial applicability, we apply the framework to a critical interface of the **main deck cargo door** of a commercial aircraft, aiming to estimate the B-basis of internal transmitted loads (denoted ϕ) subject to 28 uncertain geometric gaps (in a similar approach than one described in [5]). Figure 1 presents results for a scenario defined by three specific data budgets: $N_{\mathbf{test}} = 600$ **physical test** (used for probabilistic input identification), $N_{\mathbf{DoE}} = 350$ **numerical training points** (used to build the surrogate model), and $N_{\mathbf{sim}} =$

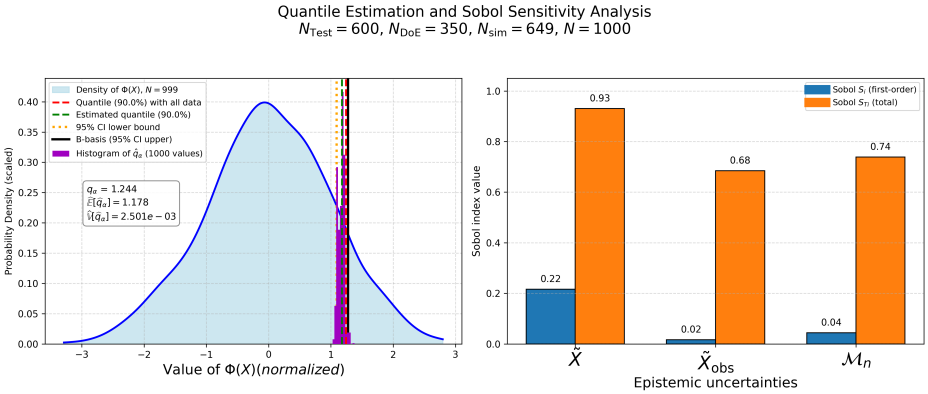


Figure 1: **Left:** histogram of the estimated quantile and the corresponding B-basis; **Right:** first-order and total Sobol indices for all sources of epistemic uncertainty.

649 **simulation samples** (used for the Monte Carlo estimator). The **left panel** displays the probability density of ϕ overlaid with the histogram of the quantile estimator $\widehat{q}_\alpha$ ($N = 1000$ quantile estimations) obtained via our framework. The resulting B-basis value is the upper 95% confidence bound. The **right panel** displays the Sobol sensitivity analysis quantifying the impact of epistemic uncertainties on the variance of $\widehat{q}_\alpha$. The Sobol indices decompose the total variance into contributions from Monte Carlo sampling ($\widetilde{\mathbf{X}}$), input identification uncertainty ($\widetilde{\mathbf{X}}_{\text{obs}}$), and surrogate modeling error ($\mathcal{M}_n$). This diagnosis reveals that, in this high-fidelity scenario, the uncertainty is driven by complex interactions between sampling and model fidelity, thereby guiding the engineer to efficiently balance numerical resources against physical testing.

Short biography (PhD student)

I am an ISAE-SUPAERO aerospace engineer currently pursuing a CIFRE PhD at Airbus Operations and ONERA. My research focuses on the multi-scale reliability of composite aerostructures, specifically through the development of a stochastic framework to propagate uncertainty from the coupon level to full-scale structures. This work supports the transition to 'Certification by Analysis' by substituting empirical safety factors with physics-based uncertainty quantification.

References

[1] U.S. Department of Defense, *Composite Materials Handbook, Volume 3: Polymer Matrix Composites Materials Usage, Design, and Analysis*, MIL-HDBK-17-3F, Washington, D.C., 2002.

[2] C. Surget, S. Dubreuil, J. Morio, C. Mattrand, J.-M. Bourinet, and N. Gayton, "A sensitivity analysis based trade-off between probabilistic model identification and statistical estimation," *Reliability Engineering & System Safety*, 2024.

[3] M. Menz, S. Dubreuil, J. Morio, C. Gogu, N. Bartoli, and M. Chiron, "Variance-based sensitivity analysis for Monte Carlo and importance sampling reliability assessment with Gaussian processes," *Structural Safety*, vol. 89, p. 102048, 2021.

[4] C. Cannamela, J. Garnier, and B. Iooss, "Controlled Stratification for Quantile Estimation," *Annals of Applied Statistics*, vol. 2, no. 4, pp. 1554–1580, 2008.

[5] G. Capasso, C. Gogu, C. Bès, J.-P. Navarro, and M. Kempeneers, "Semi-Probabilistic Codesign Framework for Tolerance bound Optimization complying with Static Strength Requirements," in *Proceedings of the 15th World Congress on Structural and Multidisciplinary Optimization (WCSMO15)*, Cork, Ireland, 2023.

Mixture-Based Generative Modeling for Data Imputation and Synthesis

A. Faul[†,1], D. Ginsbourger[§,1], B. Spycher[§,2],

[†] PhD student (presenting author). [§] PhD supervisor

PhD expected duration: Oct. 2022 – Sep. 2026

[1] Institute of Mathematical Statistics and Actuarial Science, University of Bern, Switzerland
antoine.faul@unibe.ch david.ginsbourger@unibe.ch

[2] Institute of Social and Preventive Medicine, University of Bern, Switzerland
ben.spycher@unibe.ch

Abstract

In medical research, data scarcity and missing information are prevalent, posing significant challenges for statistical analysis. These issues often stem from the high cost of collecting comprehensive patient datasets, the complexity of measuring specific variables, privacy concerns, or varying levels of patient adherence.

This work addresses these challenges by developing statistical methods for data imputation and synthetic data generation using mixtures of distributions. Our motivation originates from a case study at the University Hospital of Bern, for which we have made the data publicly available [3]. The goal was to predict the 10-year risk of cardiovascular disease when some clinical inputs of the risk calculator were systematically missing. Our method involved probabilistically imputing the missing variables and propagating the resulting uncertainty into the risk calculator.

Formally, the aim of systematic imputation is to sample from the conditional distribution of a random vector $\mathbf{Y} \in \mathbb{R}^q$ given observations of $\mathbf{X} \in \mathbb{R}^p$, based on existing samples of $(\mathbf{X}, \mathbf{Y})$. We propose a generative approach which consists in estimating the joint density $f(\mathbf{x}, \mathbf{y})$ using a parametric probability density function and sampling from the conditional distribution $f(\mathbf{y}|\mathbf{x})$ through analytical formulas.

For this purpose, our work [1] introduces families of multivariate distributions stable by conditioning, including multivariate Gaussian, Student t, and skew normal distributions, but excluding, for example, multivariate q-Exponential distributions. We demonstrate that stability by conditioning of a family of trans-dimensional probability distributions can be extended to finite mixtures and marginal transformations.

This extends the applicability of analytical conditioning across a broader range of multivariate distributions. We developed an algorithm for conditional sampling using implicit copulas and latent spaces (see Figure 1).

While our initial study used Gaussian copulas [5], more complex dependence structures like the Gaussian Mixture copula model (GMCM) were employed to capture multi-modalities and tail dependencies in our latest work [1].

These generative approaches, based on copulas, can be applied for synthetic tabular data generation. They produce realistic synthetic data points and are competitive with machine learning

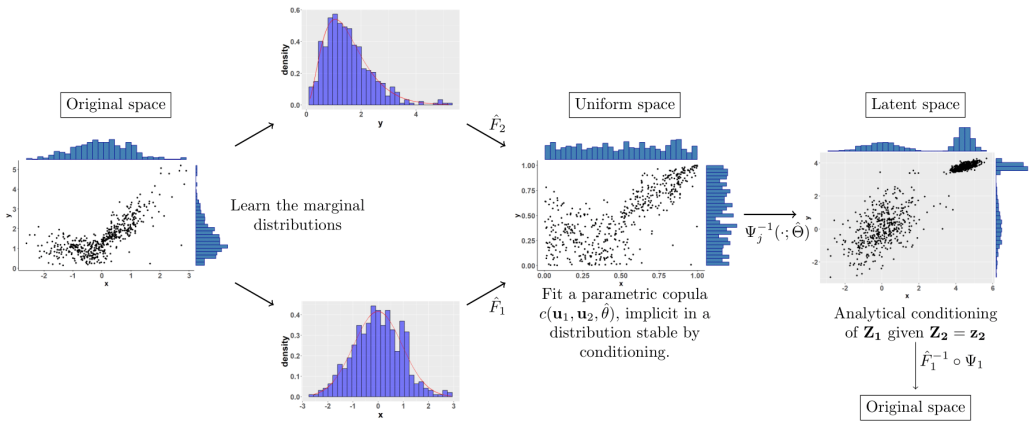


Figure 1: Workflow of the conditioning algorithm on a 2-dimensional example.

methods in moderate dimensions [2]. Evaluating data quality is crucial, yet there's no consensus on the best metrics. Therefore, we utilized a variety of existing metrics and introduced additional ones tailored to assess the statistical utility and privacy specific to our problem.

Finally, addressing complex scenarios where single generators fail to capture multi-modalities, we developed an iterative procedure inspired by the Expectation–Maximization framework to combine mixtures of diverse generators, each specializing in different data space regions [4]. We showed empirically that our approach produces high-quality synthetic data and we provide theoretical guarantees by establishing convergence rates of the mixture distribution.

Short biography (PhD student)

Antoine Faul graduated with a MSc in Engineering from Isae-Supaero and with a MSc in Statistics from Université Paris-Saclay. He started his PhD at the University of Bern in October 2022 on statistics with application to medicine. The thesis is funded by the Multidisciplinary Center for Infectious Diseases (MCID) of the University of Bern.

References

[1] Antoine Faul, David Ginsbourger, and Ben Spycher. Easy conditioning far beyond gaussian. *arXiv preprint arXiv:2409.16003*, 2024.

[2] Antoine Faul, David Ginsbourger, Ben Spycher, and Petra Stute. Copula-based synthetic data generation in medicine. 2026. Work in Progress.

[3] Antoine Faul, Philip Stange, Anja Mülhemann, Manuela Moraru, Suzanne Theis, Lena Friederichsen, Lukas Bütikofer, and Petra Stute. Menobalance: Cardiovascular risk factors from the CIMBOLIC study. *Dataset on Zenodo*, 2024.

[4] Antoine Faul, Xiao Zhou, Ossi Raisa, Mihaela Van der Shaar, and Cem Tekin. Modelling complex tabular datasets with a mixture of diverse generative models. 2026. Submitted to AISTATS 2026.

[5] Anja Mühlemann, Philip Stange, Antoine Faul, Serena Lozza-Fiacco, Rowan Iskandar, Manuela Moraru, Susanne Theis, Petra Stute, Ben D Spycher, and David Ginsbourger. Comparing imputation approaches to handle systematically missing inputs in risk calculators. *PLOS Digital Health*, 4(1):e0000712, 2025.

Combining Kriging with optimal transport to estimate measure-valued data - application to metamodeling in nuclear safety

Florian Gossard[†,1], Jean Baccou[§,2], François Bachoc[§,3],

[†] PhD student (presenting author). [§] PhD supervisor

PhD expected duration: Nov. 2023 – Nov. 2026

[1] Institut de Mathématiques de Toulouse, Univ. de Toulouse
florian.gossard@math.univ-toulouse.fr

[2] Autorité de Sûreté Nucléaire et de Radioprotection (ASNR), PSN-RES, SEMIA, Cadarache, Saint Paul-Lez-Durance 13115, France.
jean.baccou@asnr.fr

[3] Laboratoire Paul Painlevé Univ. Lille, CNRS, UMR 8524
francois.bachoc@univ-lille.fr

Abstract

Probability measure-valued data interpolation plays an important role in many applications. It arises for example in fluid mechanics to analyse the dynamics of complex phenomena or in image processing for histogram interpolation. The key ingredient is a reformulation of the problem in the optimal transport framework [6]. Several works such as [3] have also exploited the connection between interpolation and optimal transport for smooth interpolation of probability distributions. This type of approach is particularly motivated by the analysis of computationally expensive simulation codes, where building accurate surrogate models is essential to enable detailed studies of complex physical systems.

In this context, Kriging [4] is a popular approach to perform smooth interpolation of spatially correlated processes. It is widely employed in various disciplines such as environment monitoring or natural resources evaluation but also in industrial applications involving complex computer code simulations. A specificity of Kriging is to allow estimating the spatial correlation via the so-called correlation function or semivariogram before its integration in the construction of a predictor. However, in its original formulation, Kriging is restricted to real-valued processes. To tackle the prediction of probability measures we propose a Kriging approach in the optimal transport framework [2].

By interpreting the Kriging predictor [4] as a barycenter of the observations, we generalize it to probability measures using the Wasserstein distance, a standard metric from optimal transport theory. In the one-dimensional case, this distance admits an explicit formulation in terms of quantile functions, which allows us to derive a new Kriging estimator acting directly on quantiles.

Beyond this methodological extension, we address the practical issue of sparse observations, which is common when dealing with expensive numerical simulations. We introduce a leave-one-out cross-validation strategy to estimate the spatial dependence structure, adapting the virtual cross-validation formulas from the classical Kriging setting [1] to the case of quantile-based Kriging.

Finally, we illustrate the proposed approach on a nuclear safety application, focusing on the prediction of probability measures associated with simulated temperature in a reactor core during a loss-of-coolant accident. Preliminary results demonstrate a clear improvement in predictive accuracy compared to more traditional methods based on temperature maps, particularly for high quantiles, which are critical quantities of interest in nuclear safety analysis.

This work is presented in detail in the following preprint [5].

Short biography (PhD student)

I'm a 3rd year PhD student at Institut de Mathématiques de Toulouse (IMT). This PhD thesis is carried out within the QUTHY project, funded by the NEEDS program, which involves several partners including EDF R&D, CEA, ASNR, IMT, and the Institut de Mathematiques de Marseille. The project focuses on uncertainty quantification in thermohydraulics, with applications related to nuclear safety.

As part of this work, I actively participated in several scientific events related to rt-uq, including the SAMO conference, the ETICS thematic school, and various specialized workshops.

References

[1] François Bachoc. Cross validation and maximum likelihood estimations of hyper-parameters of Gaussian processes with model misspecification. *Computational Statistics & Data Analysis*, 66:55–69, 2013.

[2] Antonio Balzanella and Antonio Irpino. Spatial prediction and spatial dependence monitoring on georeferenced data streams. *Statistical Methods and Applications*, 29:101–128, 3 2020.

[3] Sinho Chewi, Julien Clancy, Thibaut Le Gouic, Philippe Rigollet, George Stepaniants, and Austin Stromme. Fast and smooth interpolation on Wasserstein space. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pages 3061–3069, 2021.

[4] Noel A.C. Cressie. *Statistics for spatial data revised edition.* Wiley, 4 2015.

[5] Florian Gossard, François Bachoc, Jean Baccou, Thibaut Le Gouic, Jacques Liandrat, and Tony Glantz. Kriging measure-valued data with sparse observations: application to nuclear safety studies. working paper or preprint, October 2025.

[6] Filippo Santambrogio. Optimal transport for applied mathematicians. *Springer International Publishing*, 87, 2015.

Gradient-based active learning for Global Sensitivity Analysis

Guerlain Lambert[†,1,2], Céline Helbert[§,1], Claire Lauvernet[§,2]

[†] PhD student [§] PhD supervisor

PhD expected duration: Oct. 2023 – Sep. 2026

[1] Institut Camille Jordan, CNRS UMR 5208, École Centrale de Lyon, Écully, France
{guerlain.lambert,celine.helbert}@ec-lyon.fr

[2] INRAE, RiverLy, 69625 Villeurbanne, France
claire.lauvernet@inrae.fr

Abstract

Global Sensitivity Analysis (GSA) is widely used to understand and improve physical models in industry and environmental sciences by quantifying how uncertain inputs drive the variability of a quantity of interest. In the classical setting, one considers a computational model $Y = f(X)$ where $X = (X_1, \dots, X_d)$ are the inputs, and aims at ranking the components of X according to their influence on the output variability. Variance-based sensitivity measures, in particular Sobol' indices, are popular because they provide an interpretable variance attribution; for instance, the first-order Sobol' index of X_i can be written as $S_i = \frac{\mathrm{Var}(\mathbb{E}[Y|X_i])}{\mathrm{Var}(Y)}$, while the total-effect index $S_i^{\top} = 1 - S_{\sim i}$ accounts for all effects involving X_i, including interaction terms. Such indices are standard tools for GSA [2]. Despite their interpretability, accurate Sobol' index estimation often requires numerous calls to the expensive computer code f, which is prohibitive when a single run may take several dozen hours. A standard approach is to replace f by a surrogate model $\hat{f}$ (e.g., a Gaussian-process, GP) and then to compute Sobol' indices from $\hat{f}$ using Monte Carlo.

Nevertheless, constructing a reliable surrogate still requires carefully chosen observations of f. This motivates adaptive designs of experiments (DoE) that allocate expensive runs where they most improve the learning objective. Recent work [1, 3] suggests using derivative-based global sensitivity measures (DGSM) to drive such sequential designs. DGSM rely on gradient information and indicate which directions in the input space are locally influential. Under a GP prior with standard kernels (RBF, Matérn), gradients of the GP posterior mean and covariance can be computed analytically, enabling acquisition functions tailored to sensitivity learning. Figure 1 shows, on a 2D toy function, the spatial distribution of active learning points selected by Sobol-based random sampling (left) and by two DGSM-driven acquisition functions, highlighting how gradient-based criteria exploit local sensitivity information.

In this talk, we present an active learning strategy adapted to DGSM in the context of GP metamodeling [3], focusing on acquisition functions designed to reduce uncertainty on sensitivity-related quantities. Figure 2 reports the RMSE of DGSM estimates versus active learning steps on scalar-valued toy functions, demonstrating the effectiveness of the proposed approach.

We then discuss how to incorporate input dependencies into these acquisition criteria, an important requirement in realistic applications where uncertain inputs may not be independent. Finally, we examine how to extend the framework to complex functional inputs such as time series, where the effective dimension is high and adaptive designs must exploit temporal structure.

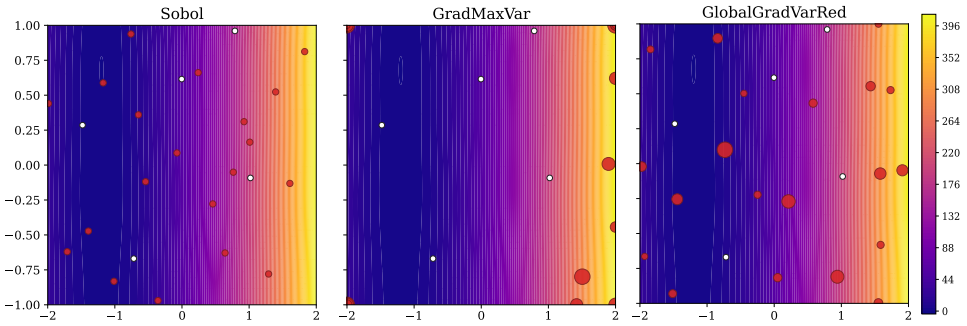


Figure 1: Spatial distribution of active learning points (red) on a two-dimensional toy function for Sobol' random sampling (left) and for DGSM-driven acquisition functions: `LocalGradVarRed` (middle) and `GlobalGradVarRed` (right). White points are initial DoE.

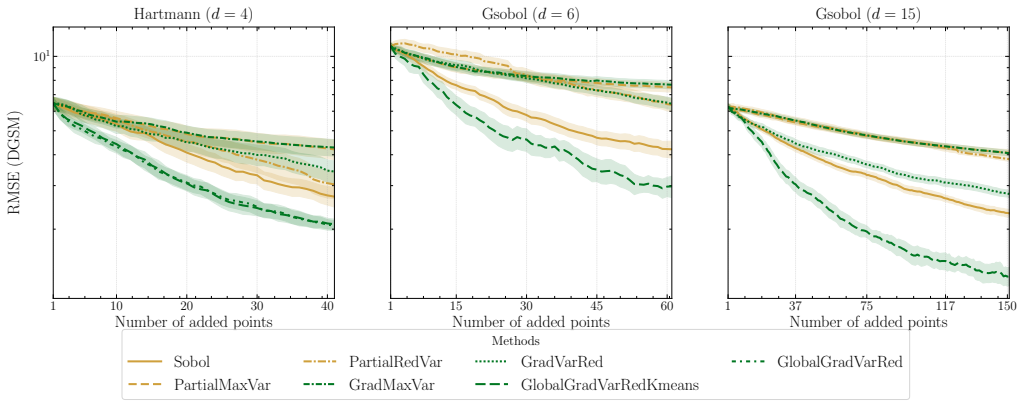


Figure 2: RMSE of DGSM and Sobol total indices estimates versus active learning steps on three toy functions.

The developed methodologies are illustrated on toy cases and are applied to a 3D spatio-temporal model of groundwater pollutant management.

Short biography (PhD student)

I am a third-year PhD student at Centrale Lyon and INRAE Lyon, supervised by Céline Helbert and Claire Lauvernet. My thesis develops active learning methods for complex input data, targeting applications in sophisticated environmental models. The work is part of Water4All's AQUIGROW project, which seeks to boost groundwater service resilience amid rising drought risk. The thesis is also supported by the CIROQUO consortium.

References

[1] S. Belakaria, B. Letham, J. Doppa, B. Engelhardt, S. Ermon, and E. Bakshy. Active learning for derivative-based global sensitivity analysis with gaussian processes. In *Advances in Neural Information Processing Systems*, 2024.

[2] B. Iooss and P. Lemaître. *A Review on Global Sensitivity Analysis Methods*, pages 101–122. Springer US, Boston, MA, 2015.

[3] G. Lambert, C. Helbert, and C. Lauvernet. Gradient-based active learning with gaussian processes for global sensitivity analysis. 2025.

Approximation and learning with compositional tensor trains

Martin Eigel[§,1], Charles Miranda[†,2], Anthony Nouy[§,2], David Sommer[1]

[†] PhD student (presenting author). [§] PhD supervisor

PhD expected duration: Nov. 2023 – Oct. 2026

[1] Weierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany
{eigel,sommer}@wias-berlin.de

[2] Centrale Nantes, Nantes Université, Laboratoire de Mathématiques Jean Leray UMR CNRS 6629, France
{charles.miranda,anthony.nouy}@ec-nantes.fr

Abstract

We introduce compositional tensor trains (CTTs) for the approximation of multivariate functions, a class of models obtained by composing low-rank functions in the tensor-train format. This format can encode standard approximation tools, such as (sparse) polynomials, deep neural networks (DNNs) with fixed width, or tensor networks with arbitrary permutation of the inputs, or more general affine coordinate transformations, with similar complexities.

Formally, a CTT u is defined by linear operators $\mathfrak{L} : \mathbb{R}^d \to \mathbb{R}^p$ and $\mathfrak{R} : \mathbb{R}^p \to \mathbb{R}^{d_o}$ called respectively *lift* and *retraction*, and a univariate basis $\Phi = \{\phi_j : \mathbb{R} \to \mathbb{R}\}_{j=1}^n$, such that

$$u(x) = \mathfrak{R} \circ (\mathrm{Id} + \psi_L) \circ \cdots \circ (\mathrm{Id} + \psi_1) \circ \mathfrak{L}(x),$$

where ψ_k are tensors in the *Tensor-Train format* [6].

This format can be viewed as a DNN with width exponential in the input dimension and structured weights matrices. Compared to DNNs, this format enables controlled compression at the layer level using efficient tensor algebra.

On the optimization side, we derive a layerwise algorithm inspired by natural gradient descent [1], allowing to exploit efficient low-rank tensor algebra. The natural gradient descent tries to mimic the update in the functional space by an update in the parameter space. In the case of L^2 functions, this update simplifies to

$$\theta_{k+1} = \theta_k - \alpha_k G(\theta_k)^\dagger \nabla_\theta L(\theta_k),$$

where $G(\theta)_{ij} := \langle \partial_{\theta_i} u_\theta, \partial_{\theta_j} u_\theta \rangle$ is the *Gram matrix* and $L : \Theta \to \mathbb{R}$ is a *loss function* e.g. $L(\theta) = \frac{1}{2}\|u_\theta - v\|_{L^2}^2$.

In the case of CTT, the Gram matrix G can be stored efficiently due to the *Tensor-Train format* and its inherit low-rank format. Computing the update direction can be done using algorithms such as *alternating linear scheme (ALS)* [3]. However, the Gram matrix associated to each layer ℓ may have a bad condition number, so that ALS without preconditioning may show a slow convergence and yield a highly suboptimal low-rank approximation of the update direction.

A well-established approach to mitigate this issue is to approximate G_ℓ with a low-rank surrogate. In particular, the randomized Nyström method [5, 7, 2, 4] achieves this by projecting

G_ℓ onto a randomly generated, low-dimensional subspace. In the context of this work, the Gram matrix G_ℓ is a linear operator acting on tensor spaces, and so we can compute random projections using a tensor-structured sketch efficiently. The key advantage of this Gaussian sketching approach is that, with high probability, the span of the sketch captures the dominant eigenspace of G_ℓ Viewing the format as a discrete dynamical system, we also derive an optimization algorithm inspired by numerical methods in optimal control.

Numerical experiments on regression tasks demonstrate the expressivity of the new format and the relevance of the proposed optimization algorithms.

The Figure 1 shows the performance of the optimizer for a recovery problem where the TT ranks have provably high. Moreover, by computing layerwise updates, the optimizer is faster than the state of the art solvers.

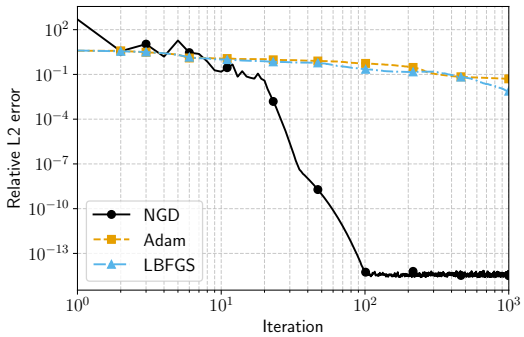


Figure 1: Convergence plot for the optimizers Adam, NGD and L-BFGS for a recovery problem, in log-log scale, for dimensions $d = 4$.

We also studied the *effective condition number* $\kappa_\ell(\theta) := \|G_\ell(\theta)\|_{2\to 2}\|G_\ell(\theta)^\dagger\|_{2\to 2}$ during the optimization. Experiments show that the Gram matrices become highly ill-conditioned during optimization, with condition numbers κ_ℓ ranging from 10^6 up to 10^{14}. In fact, the condition number increases rapidly, reaching values around 10^{13} after approximately 20 iterations. Initially, directions associated with small eigenvalues play a useful role by guiding the optimizer toward a good configuration. However, as the solution approaches optimality, these directions contribute progressively less to the reduction of the loss.

Finally, we applied the randomized method to the recovery problem and studying the convergence behavior for various sketching sizes. We have observed that retaining a rank-30 approximation of the Gram matrix is sufficient to achieve convergence to an optimal solution, which is less than half of the total eigendirections.

Overall, CTTs combine the expressivity of compositional models with the algorithmic efficiency of tensor algebra, offering a scalable alternative to standard deep neural networks.

Short biography (PhD student)

I am a third year PhD student working on *Approximation and learning with compositional functions networks*, supervised by Anthony NOUY (École Centrale de Nantes, Nantes, France), and Martin EIGEL (Weierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany). My research focuses on the approximation capabilities of compositional functions networks, and the design of optimization algorithms for these specific model classes. My research is funded by DFG-ANR Cofnet, and DR Centrale Nantes.

References

[1] Shun-ichi Amari. Natural Gradient Works Efficiently in Learning. *Neural Computation*, 10(2):251–276, February 1998.

[2] Alex Gittens and Michael Mahoney. Revisiting the Nyström method for improved large-scale machine learning. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 567–575, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

[3] Sebastian Holtz, Thorsten Rohwedder, and Reinhold Schneider. The Alternating Linear Scheme for Tensor Optimization in the Tensor Train Format. *SIAM Journal on Scientific Computing*, 34(2):A683–A713, January 2012.

[4] Per-Gunnar Martinsson and Joel A. Tropp. Randomized numerical linear algebra: Foundations and algorithms. *Acta Numerica*, 29:403–572, May 2020.

[5] E. J. Nyström. Über die praktische Auflösung von Integralgleichungen mit Anwendungen auf Randwertaufgaben. *Acta Mathematica*, 54(0):185–204, 1930.

[6] I. V. Oseledets. Tensor-train decomposition. *SIAM Journal on Scientific Computing*, 33(5):2295–2317, January 2011.

[7] Christopher Williams and Matthias Seeger. Using the Nyström method to speed up kernel machines. *Advances in neural information processing systems*, 13, 2000.

Physics-informed, boundary-constrained Gaussian processes with applications in fluid dynamics

Adrian Padilla-Segarra[†,1,2], Pascal Noble[§,2], Olivier Roustant[§,2], Éric Savin[§,3]

[†] PhD student (presenting author) [§] PhD supervision team

PhD expected duration: Nov. 2023 – Oct. 2026

[1] Department of Information Treatment and Systems (DTIS), ONERA, Toulouse, 31400, France
{adrian.padilla_segarra, eric.savin}@onera.fr

[2] Institute of Mathematics of Toulouse (IMT), INSA Toulouse, Toulouse, 31400, France
{pascal.noble, olivier roustant}@insa-toulouse.fr

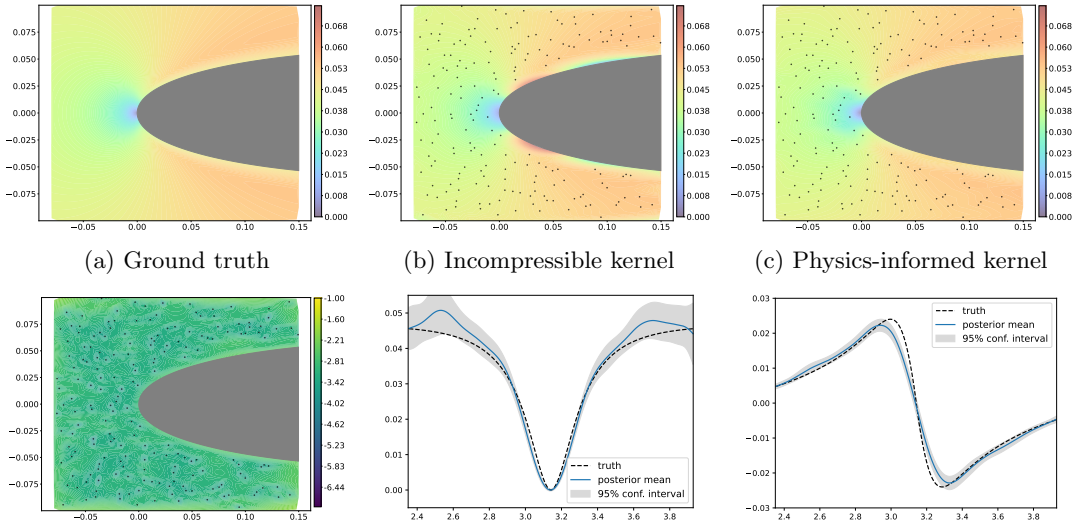
Abstract

In this communication we present a framework for surrogate modelling of incompressible fluid flows using Gaussian processes (GP) constrained by physical conditions and including uncertainty quantification (UQ). We develop a general method to continuously enforce a prescribed GP on an arbitrary compact subset of its domain. Other methods that explore boundary conditions on GPs are based on defining particular structures of the covariance kernels or are not necessarily adapted to consider derivatives of the GP. The proposed strategy uses a spectral expansion of the prescribed GP on the compact set, which in practice corresponds to the boundary of an aerodynamic profile (*e.g.* cylinder, NACA airfoil), to enforce boundary conditions. This enables us to consider observations of derivatives of the GP directly while satisfying other physical laws simultaneously. This framework can be thus flexibly merged within data assimilation schemes based on GP regression for the reconstruction of fluid flows.

Consider a two-dimensional incompressible flow with velocity field $\boldsymbol{u}$ over the computational domain Ω. This field must satisfy scalar boundary conditions $\mathcal{B}(\boldsymbol{u}) = 0$ over a compact set $\Gamma \subset \Omega$, where $\mathcal{B}$ is a scalar-valued linear operator on $\boldsymbol{u}$. In this setting the velocity field can be explained through a scalar stream function ψ such that $\boldsymbol{u} = \mathbf{curl}\,\psi = (-\partial_{x_2}\psi, \partial_{x_1}\psi)^\mathsf{T}$. Our strategy consist in modelling the stream function as a GP prior Z, so that the velocity field can be modelled as $\mathbf{curl}\,Z$, verifying thus the divergence-free condition (from incompressibility) $\partial_{x_1}u_1 + \partial_{x_2}u_2 = 0$ everywhere in Ω. Furthermore, a power-law structure for the energy decay of velocity increments can be accounted for in the definition of the covariance kernel $\boldsymbol{K}$ of $\mathbf{curl}\,Z$, as illustrated in [2, 3]. Lastly, boundary information on the profile delimited by Γ can be included by using our general spectral method [3] based on [1]. In this sense we obtain a modified version $\boldsymbol{K}_0$ of $\boldsymbol{K}$ satisfying also the homogeneous boundary condition everywhere and not only at a discrete set of observations.

Now, given a set of velocity measurements $\boldsymbol{V}(t)$ at positions $\boldsymbol{X}(t)$ (*e.g.* Lagrangian data in particle tracking velocimetry), this physics-informed GP prior can be used for reconstruction of the field as $\boldsymbol{u}^\star(\boldsymbol{x}, t) = \boldsymbol{K}_0(\boldsymbol{x}, \boldsymbol{X}(t))\boldsymbol{K}_0(\boldsymbol{X}(t), \boldsymbol{X}(t))^{-1}\boldsymbol{V}(t)$. UQ counterparts follow in a similar manner, as well as posterior estimates for the stream function ψ and vorticity $\omega = \mathbf{curl}^\mathsf{T}\boldsymbol{u}$. Main results are depicted in Figure 1. In [3], we compare the reconstructions with respect to three choices of the covariance kernel of Z: an incompressible radial basis function (RBF), a multi-scale additive RBF, and the physics-informed boundary-constrained kernel as described

above. The calibration of hyperparameters is performed by cross-validation using UQ coverage indicators of the estimates.



(a) Ground truth (b) Incompressible kernel (c) Physics-informed kernel

(d) Uncertainty quantification (e) Horizontal velocity (f) Vertical velocity

Figure 1: **Reconstruction of the velocity field u of an incompressible flow around the leading edge of a NACA 0412 airfoil [3].** Ground truth field is depicted in (a), with colormap as velocity norm in m/s. In (b), the reconstruction of the field is done from 184 velocity measurements $\boldsymbol{V}$ (black dots •) using GP regression with a one-scale incompressible kernel. In (c), the kernel is further informed with the profile boundary condition (slip) and an energy decay law [3]. This last physics-informed estimate $\boldsymbol{u}^\star$ comes with the UQ counterpart depicted in (d), as the (logarithmic) total standard deviation. Notice that even if there are no discrete measurements at the airfoil boundary, the physics-informed kernel $\boldsymbol{K}_0$ used in (c) is able to capture the velocity field, as displayed in (e) and (f) for each spatial component (horizontal axis is domain of airfoil boundary Γ parameterization).

We further explore the use of this framework to define a data assimilation scheme using the numerical method on the Navier-Stokes equations for vorticity from [2]. Preliminary tests indicate that this scheme can accurately propagate the initial information through subsequent assimilation steps, incorporating Lagrangian velocity and vorticity measurements when available. Moreover, it is competitive in comparison to its pure data-driven version in scarce data regimes. We are also interested in optimal sensor placement to improve the reconstructions using UQ of the estimates.

Short biography (PhD student)

This research project is at the intersection of probabilities, data assimilation and numerical analysis of differential equations. Adrian Padilla–Segarra holds a MSc from Paris–Saclay University (track Analysis, Modelling, Simulation at Orsay and ENSTA, under a FMJH scholarship) and a BSc in Mathematics from Yachay Tech (Ecuador). The PhD project is jointly financed by ONERA and INSA Toulouse (SHOM project "Machine Learning Methods in Oceanography" no-20CP07) and is attached to the IMT. A collaboration with prof. Houman Owhadi from Caltech (Pasadena, United States) is ongoing.

References

[1] Bertrand Gauthier and Xavier Bay. Spectral approach for kernel-based interpolation. *Annales de la Faculté des Sciences de Toulouse: Mathématiques*, 21(3):439–479, 2012.

[2] Houman Owhadi. Gaussian process hydrodynamics. *Applied Mathematics and Mechanics*, 44(7):1175–1198, 2023.

[3] Adrian Padilla-Segarra, Pascal Noble, Olivier Roustant, and Éric Savin. Physics-informed, boundary-constrained Gaussian process regression for the reconstruction of fluid flow fields. *arXiv 2507.17582*, 2025.

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

P. Aurélien[†,1,2], E. Vazquez[§,2]

[†] PhD student. [§] PhD supervisor

PhD expected duration: June 2023 – June 2026

[1] Transvalor S.A., Biot, France
aurelien.pion@centralesupelec.fr

[2] Université Paris-Saclay, CNRS, CentraleSupélec, L2S, Gif-sur-Yvette, France

Abstract

Gaussian processes (GPs) are Bayesian models widely used to interpolate an unknown deterministic function $f : \mathcal{X} \subset \mathbb{R}^d \to \mathbb{R}$ from observed data. A GP prior on f is written $\xi \sim \mathrm{GP}(m, k)$, with mean function $m : \mathcal{X} \to \mathbb{R}$ and positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Conditioning on evaluations of f yields a posterior predictive distribution that provides both point predictions and uncertainty quantification.

In practice, GP predictive distributions are often miscalibrated, so nominal confidence intervals can miss their target coverage when assessed under a design-marginal measure on $\mathcal{X}$. This mismatch is critical in downstream tasks that rely on calibrated uncertainty, since it can bias sequential design decisions. In [3], we formalize this setting via *μ-calibration*, which evaluates calibration under a sampling measure μ that describes how observation locations are generated.

Conformal prediction (CP) provides finite-sample, distribution-free marginal coverage under exchangeability [4] and has been adapted to GPs for post-hoc correction of prediction intervals [2, 1]. Conformal predictive systems (CPS) extend this idea by outputting a full predictive cumulative distribution function (CDF) at each test point [5].

We propose two post-hoc calibration methods for GP interpolation that target μ-calibration. The first, CPS–GP, adapts CPS to kernel methods and returns conformal predictive distributions. The second, BCR–GP, keeps the GP posterior mean and models normalized errors with a generalized normal distribution whose parameters are selected by a Bayesian rule inspired by tolerance intervals.

We compare BCR–GP and CPS–GP to the standard GP, an oracle model selected on a test set, and J$^+$–GP/FCP. Across coverage levels, both methods improve μ-calibration, while BCR–GP yields smooth predictive CDFs that remain usable in optimization and excursion procedures.

Short biography (PhD student)

I graduated from l'Ecole des Ponts in 2023, and in 2022 I had the master "Mathématiques, Vision, Apprentissage" from l'ENS Paris-Saclay. During the MVA, I did a 4 month internship at the L2S/CentraleSupelec with Emmanuel to prepare the thesis. Finally, in 2023 I started the thesis at the L2S and in collaboration with Transvalor S.A., which is funding the thesis under a CIFRE grant.

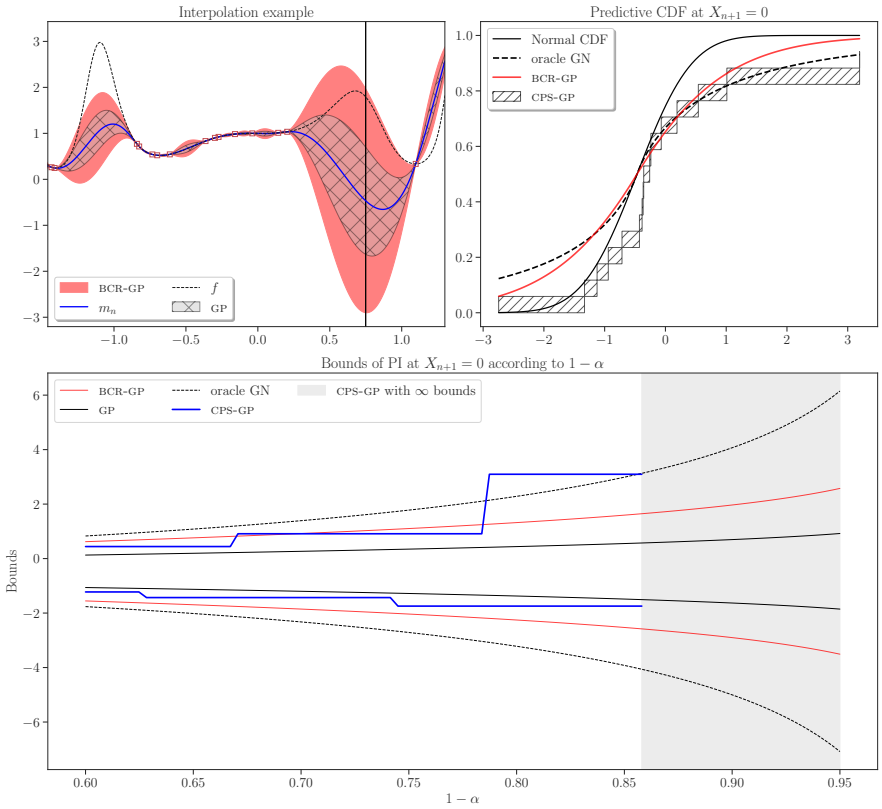


Figure 1: **Top left:** Prediction intervals from the GP posterior and from BCR–GP at $1-\alpha = 0.9$. BCR–GP selects parameters using the generalized normal variance, with $\delta = 0.1$. **Top right:** Predictive CDFs at $x = 0.75$ for the GP posterior, BCR–GP (red), CPS–GP (stepwise, black hatches), and an oracle CDF (black) obtained by fitting a generalized normal model on a test grid with $n_{\text{test}} = 2000$. The GP posterior is strongly miscalibrated on this dataset. **Bottom:** Interval bounds as a function of $1 - \alpha$. The GP posterior underestimates uncertainty; BCR–GP and CPS–GP produce wider intervals. CPS–GP yields unbounded widths for $1 - \alpha \gtrsim 0.85$.

References

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

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

[3] Aurélien Pion and Emmanuel Vazquez. Design-marginal calibration of gaussian process predictive distributions: Bayesian and conformal approaches, 2025.

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

[5] Vladimir Vovk, Jieli Shen, Valery Manokhin, and Min-ge Xie. Nonparametric predictive distributions based on conformal prediction. In Alex Gammerman, Vladimir Vovk, Zhiyuan Luo, and Harris Papadopoulos, editors, *Proceedings of the Sixth Workshop on Conformal and Probabilistic Prediction and Applications*, volume 60 of *Proceedings of Machine Learning Research*, pages 82–102. PMLR, 13–16 Jun 2017.

Probabilistic Neural Networks for Heteroscedastic Count Regression of a Radioactivity Measurement

A. Roblin[†,1,3], J. Baccou[§,2], G. Dougniaux[§,3], S. Velasco-Forero[§,1]

[†] PhD student (presenting author). [§] PhD supervisor

PhD expected duration: Oct. 2023 – Oct. 2026

[1] Center for Mathematical Morphology (CMM), Mines Paris PSL, Fontainebleau, 77300, France
{arthur.roblin,santiago.velasco-forero}@minesparis.psl.eu

[2] Autorité de Sûreté Nucléaire et de Radioprotection (ASNR), PSN-RES/SEMIA/LSMA, Saint-Paul-lez-Durance, F-13115, France

[3] Autorité de Sûreté Nucléaire et de Radioprotection (ASNR), PSN-RES/SCA/LPMA, Saclay, F-91400, France

Abstract

In nuclear facilities, dedicated instruments are used to monitor airborne radioactivity in real-time. The nuclear measurement analysis is performed by rudimentary statistical algorithms that showed serious limitations in atypical situations [2]. In order to improve analysis, we implement a deep neural network for a better interpretation of the raw measurement. In the long term, the results may lead to an effective implementation in nuclear instrumentation, if, and only if satisfying operational and safety guaranties are achieved.

The final objective of the model is a binary classification task: the monitor has to decide whether or not some transuranic activity is detected in the local atmosphere. However, in opposition to our previous work [5], we opt for a preliminary regression framework rather than a direct binary classification by the neural network. The goal is to predict directly the number of transuranic events in the spectrum, i.e. the number of alpha particles emissions measured that are due to a transuranic radioisotope. We motivate this choice with multiple reasons, mainly related to interpretability requirements.

In such a sensitive use case as worker radioprotection, the need of a reliable quantification of prediction uncertainty is clear. The use of distributional prediction is common for continuous regression, but less explored for discrete count regression. The idea is to model our heteroscedastic target count y using a parametric distribution, allowing an uncertainty-aware prediction. Given its obvious limitations concerning variance modelisation, Poisson distribution is insufficient. We thus used the Double Poisson (DP) distribution [1], as well two parameterizations of the Negative Binomial (NB) [4]. Using their negative log-likelihood as loss functions allows to retrieve a satisfying estimation of the aleatoric uncertainty associated to each input [8]. An innovative heuristic optimization process of the DP and NB neural networks is proposed, Stop-Gradient Scheduler (SGS), mitigating a well-known convergence issue and leading to a significant performances improvement.

Even if the uncertainty estimates can reach satisfying levels of empirical coverages, this approach does not achieve any theoretical guaranty, and the confidence intervals may not be always relevant.

Therefore, we exploit the conformal prediction framework to address this limitation. The idea is to correct the predicted confidence intervals on an dedicated portion of the dataset for a designed confidence level, using the normalised residual ordination method [6, 7]. This leads to locally adaptive confidence intervals that ensure an empirical marginal coverage consistent with any confidence level chosen.

Finally, from this transuranic count estimate, the binary decision may be taken from traditional means. In this case, a natural choice is the sigma-factor inequation, used in many monitors [3]. The decision is therefore interpretable in itself, and the deep learning model task is restricted to a purely objective analysis of the measurement.

The comparison with the traditional classification done by the 4-ROI algorithm [3] on airborne radioactive contamination data, detailed Table 1, shows a significant improvement of the performances of the classification results with our new model. It is particularly noticeable in the case of dust-laden atmospheric conditions (inducing background noise in the measure), which is a situation encountered during dismantling operations. The performances are consistent with the network trained for direct binary classification with the binary cross-entropy, but the precision-recall trade-off is more controllable. The precision levels are consistent with the required high conservatism standard.

Table 1: Classification performances obtained with the different algorithms.

Algorithm	Precision	Recall
Moderate background noise		
4-ROI	97.98 %	83.35 %
Direct classification network	98.16 %	96.97 %
DP count network	**1.00** %	97.78 %
NBI count network	**1.00** %	**98.18** %
NBII count network	**1.00** %	97.58 %
High background noise		
4-ROI	61.55 %	67.85 %
Direct classification network	96.06 %	**91.02** %
DP count network	**99.85** %	82.43 %
NBI count network	98.47 %	86.95 %
NBII count network	99.64 %	84.07 %

Short biography (PhD student)

In 2020, I joined ENSAI, specializing in statistical engineering. I completed my final year internship at IRSN under the supervision of Jean Baccou, and continued at ASNR for this PhD thesis focusing on a deep learning application, with an emphasis on model interpretability. This PhD is funded by ASNR, and directed by Santiago Velasco from Mines Paris PSL.

References

[1] Bradley Efron. Double Exponential Families and Their Use in Generalized Linear Regression. *Journal of the American Statistical Association*, 81(395), 1986.

[2] G. Hoarau, Gregoire Dougniaux, Francois Gensdarmes, B. Dhieux Lestaevel, J. Laurent, and Philippe Cassette. Impact de la masse de particules sur le comportement d'un moniteur de mesure de la contamination atmosphérique (cam). 2020.

[3] Alan Justus. Technical Details of the Sigma Factor Alarm Method within Alpha CAMs. *Health Physics*, 120(4):442–453, 2021.

[4] Robert Rigby. Distributions For Modeling Location Scale and Shape Using GAMLSS in R 1st Edition, 2019.

[5] Arthur Roblin, Jean Baccou, Grégoire Dougniaux, and Santiago Velasco-Forero. Deep learning approach for airborne alpha radioactivity monitoring in atypical atmospheric conditions. *Journal of Aerosol Science*, 187:106573, 2025.

[6] Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized Quantile Regression, 2019.

[7] Ryan Tibshirani. Conformal prediction under distribution shift. *Notes for Advanced Topics in Statistical Learning, Spring*, 2023.

[8] Spencer Young, Porter Jenkins, Lonchao Da, Jeff Dotson, and Hua Wei. Flexible Heteroscedastic Count Regression with Deep Double Poisson Networks, 2024.

Nonlinear model reduction based on Compositional Polynomial Networks for Parametric PDEs

Joel Soffo [†,1,2], Antoine Bensalah [§,1], Anthony Nouy [§,2]

[†] PhD student (presenting author). [§] PhD supervisors

PhD expected duration: July 2023 – July 2026

[1] Airbus

[2] Nantes Université, Centrale Nantes, Laboratoire de Mathématiques Jean Leray, CNRS UMR 6629, France

{joel.soffo, anthony.nouy}@ec-nantes.fr {joel-pascal.soffo-wambo, antoine.bensalah}@airbus.com

Abstract

We are interested in this work in developing a data-driven framework based on dimensionality reduction, for solving parametric partial differential equations (Parametric PDEs). Model reduction methods are used to approximate a manifold of functions M from a (high-dimensional) Hilbert space X by a low-dimensional space or manifold. We have recently proposed in [1] a novel approach, which aims at constructing a low-dimensional approximation manifold $M_n := \{\mathcal{L}(a) + \mathcal{N}(a) : a \in \mathbb{R}^n\}$ using samples from M. $\mathcal{L}$ is a linear map onto a n-dimensional space X_n and $\mathcal{N}$ is a nonlinear map (composition of polynomials) onto a complementary space. The method relies on an adaptive strategy to construct the maps $\mathcal{L}$ and $\mathcal{N}$ based on a prescribed relative precision and Lipschitz constant. The first part of this presentation is dedicated in the construction of M_n.

In the second part, we use the manifold approximation technique above in an operator learning context.

Let X, Y be two Hilbert spaces and ρ a measure supported on X. Here, we consider the problem of approximating a (possibly nonlinear) map $\mathcal{T} : X \to Y$, using samples $\{f_j, u_j\}_{j=1}^m$ where the f_j are samples drawn from ρ and $u_j = \mathcal{T} f_j$. We are interested in constructing an approximation T_θ of $\mathcal{T}$ by solving the minimization problem

$$\min_{\theta \in \Theta} \|\mathcal{T} - T_\theta\|_2 := [\mathbb{E}_{f\sim\rho}(\|\mathcal{T}(f) - T_\theta(f))\|_Y^2]^{1/2}$$

In the work proposed in [2], the authors rely on model reduction to construct T_θ. This is done by first applying Principal Component Analysis (PCA) on the input and output spaces, then building a map φ between the resulting finite-dimensional coordinates in the bases of principal components (see Figure 1).

Following this idea, we will propose a new method for constructing T_θ. The approach exploits the offline nonlinear approach above, to approximate the output space by a nonlinear manifold M_n. Given a prescribed precision ϵ, the method aims at constructing

$$T_\theta := D_n^{\text{out}} \circ \varphi \circ E_m^{\text{in}}$$

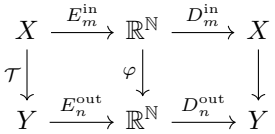


Figure 1: Operator learning using dimension reduction.

which relies on encoder-decoder pairs $(E_m^{\text{in}}, D_m^{\text{in}})$ and $(E_n^{\text{out}}, D_n^{\text{out}})$, such that

$$\|\mathcal{T} - T_\theta\|_2 \leq \epsilon \|\mathcal{T}\|_2.$$

Short biography (PhD student)

I am a PhD student in applied mathematics at Airbus and Nantes Université/Centrale Nantes, Laboratoire de Mathématiques Jean Leray. My background is mainly on statistics, machine learning and numerical analysis. My work focuses on developing new mathematical and numerical data-driven methods, for solving partial differential equations, especially the ones describing acoustic waves propagation. My PhD is funded by Airbus Central Research and Technology.

References

[1] Antoine Bensalah, Anthony Nouy, and Joel Soffo. Nonlinear manifold approximation using compositional polynomial networks, 2025.

[2] Kaushik Bhattacharya, Bamdad Hosseini, Nikola B. Kovachki, and Andrew M. Stuart. Model reduction and neural networks for parametric pdes, 2021.

Leveraging historical effects for a precise inversion of time-varying systems

M. Yachouti[†,1], G. Perrin[§,2], J. Garnier[§,1]

[†] PhD student (presenting author). [§] PhD supervisor

PhD expected duration: Jan. 2024 – Jan. 2027

[1] CMAP, Ecole polytechnique, Institut Polytechnique de Paris
{mouad.yachouti, josselin.garnier}@polytechnique.edu

[2] COSYS, Université Gustave Eiffel
guillaume.perrin@univ-eiffel.fr

Abstract

Dynamical systems are used in real-world applications to model complex time-varying systems. Despite being broadly used, exploiting the historical effects carried by the data for inversion purposes is an understudied question. Traditional methods often assume simplified settings (linear or instantaneous models for instance) and do not fully exploit history while more sophisticated ones often lack explicit Bayesian formalism [2].

In this work, we are interested in reconstructing the time evolution of latent environment variables whose effects are reflected with a certain inertia on a system's output (ie. the latter is a function that rely not only on the present or instantaneous input but also on its past values). More precisely, we propose a simple and yet effective formalism for a precise Bayesian inversion for time-series with historical outputs, and investigate three inversion set-ups. We first consider the scenario where the input-output relationship is approximated based on a standard non-linear model relying solely on the instantaneous input. In such a set-up, the model captures a part of historical effects via the temporal correlation of the input process [3] and although it can achieve good prediction results, we highlight that its usability becomes limited for inversion tasks. Then, we consider another class of models based on linear history (also called distributed lag models [1]) that are often exclusively used for prediction in the literature. We demonstrate to what extent they are useful to handle the inversion problematic being studied and are sometimes sufficient to tackle it. Finally, for an enhanced accuracy and more versatility, we consider a combination of the two aforementioned set-ups where the linear history model is complemented with instantaneous non-linearities.

We evaluate this framework on synthetic data and demonstrate its effectiveness in providing better inversion results in different application cases inspired from models describing the dynamics of biological systems and/or engineering systems.

Short biography (PhD student)

To complete his Master's in applied mathematics from Université Paris Cité, Mouad did an end-of-studies internship on explainable artificial intelligence at Michelin. He then joined the Centre Borelli as a research assistant in Machine Learning, before starting his PhD thesis within the framework of the BNP Paribas stress test chair at CMAP - Ecole polytechnique.

References

[1] A. Gasparrini, B. Armstrong, and M. G. Kenward. Distributed lag non-linear models. *Statistics in Medicine*, 29(21):2224–2234, 2010.

[2] Zhengming Kang, Xiangyu Yang, Haojie Qin, Zhuangzhuang Kang, and Chensheng Wu. An efficient inversion method of array induction logging using lstm neural network. *Journal of Geophysics and Engineering*, 22(4):1059–1073, 05 2025.

[3] Mouad Yachouti, Guillaume Perrin, and Josselin Garnier. Towards History-aware Sensitivity Analysis For Time Series. April 2025. working paper or preprint.

Local synchronisation of unsteady flow features using sequential data assimilation

L. Villanueva[†,1,2], K. Truffin[§,3], J. Borée[§,1], M. Meldi[§,4]

[†] PhD student (presenting author). [§] PhD supervisor

PhD duration: Oct. 2021 – Dec. 2024

[1] Institut Pprime, CNRS - ISAE-ENSMA - Université de Poitiers, France
[2] Present Affiliation : CECI, Université de Toulouse, CNRS, CERFACS, IRD, France
villanueva@cerfacs.fr
[3] Institut Carnot IFPEN Transports Energie, IFP Energies nouvelles, France
[4] Arts et Métiers ParisTech, LMFL - Kampé de Feriet, France

Abstract

In the field of Computational Fluid Dynamics (CFD), a growing number of studies focus on the use of scale-resolved simulations. Less expensive approaches such as averaging methods (Reynolds Averaged Navier Stokes – RANS) are popular but lack the accuracy required to account for the naturally unsteady behaviour of real flows. Scale-resolving methods such as Large-Eddy Simulation (LES) allow for the representation of unsteady features at the expense of higher computational costs. By nature, these approaches are able to compute extreme events, identified by highly non-linear features, local in space and time.

However, these features are often difficult to reproduce without the specific input information leading to such flows. Extreme events can lead to critical issues. For example, Internal Combustion Engines (ICEs) are subjected to strong variability between cycles that may degrade the efficiency or harm the system. The integration of realistic high-fidelity information in the simulation is one of the key elements allowing for a better prediction and representation of such events. This is specifically the goal of Data Assimilation (DA), which are methods designed to combine a numerical model with a high-fidelity sparse source of information in order to enhance model prediction. A simplified engine geometry is chosen. A sequential data assimilation approach known as the Ensemble Kalman Filter (EnKF) [2] is used to infer a simulation of this geometry. The high-fidelity sparse information, known as the *observation*, is sampled on the instantaneous velocity field of a high-fidelity LES (LES-HF) of the system. Fig. 1 shows the geometry alongside the LES-HF velocity field.

The EnKF algorithm is used to infer a low-fidelity LES (LES-LF) of the system with respect to the sampled high-fidelity observation. Observation samples are measured downstream the valve, in the recirculation zones visible in Fig. 1. Two objectives are targeted. First, the *parameter estimation* offered by the algorithm is used to successfully calibrate the inlet description of the simulation with respect to the unsteady velocity information of the observation. Second, *state estimation* is used to synchronise the time evolution of local features of the LES-LF flow with the sparse observation measured on the LES-HF. To that purpose, new developments of the algorithm are introduced. The *hyper-localised* Ensemble Kalman Filter (HLEnKF) [4] allows to treat each observation sensor as a local EnKF by taking into account the correlation distance of flow features. Apart from improving the overall robustness of the procedure, the HLEnKF is able to provide a local synchronisation of the inferred velocity fields with the observed data.

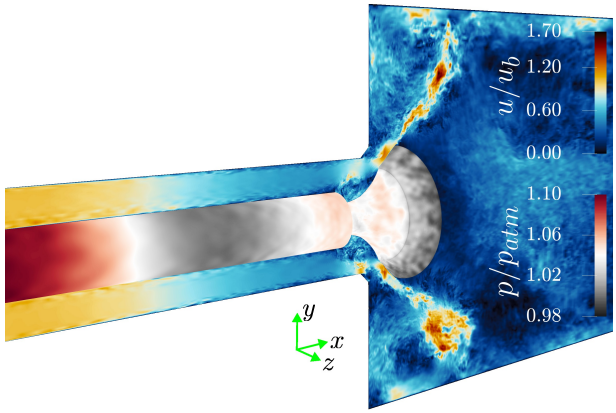


Figure 1: The *intake flow rig* is used by engine manufacturers in the early stages of development of ICEs [1, 3]. It is made of one single steady valve inserted in an intake duct and ejecting the fluid into a cylindrical chamber. The normalised velocity field u/u_b of the high-fidelity reference simulation is visualised on a 2D plane ($z = 0$). The normalised pressure field p/p_{atm} is also visible on the surface of the single guide/valve.

Furthermore, improvements of the flow topology and modal energy distribution are noticed far from the sensors used for the DA procedure, in the flow recirculation region. Overall, sequential data assimilation shows promising features for the inference of scale-resolved simulations, opening perspectives for the study of complex realistic phenomena such as extreme events.

This work was supported by the ANR-20-CE05-0007 ALEKCIA project and was granted access to HPC resources of TGCC under allocation no. A0162B10763.

Short biography (PhD student)

During my engineering studies in Orléans and later Montreal, my interest was focused on CFD. I then began my PhD at the Pprime Laboratory in Poitiers, where I studied the inference of CFD systems with DA means. I successfully defended at the end of 2024. I recently joined the CECI-CNRS team at Cerfacs, Toulouse, as a post-doc to infer wildfire simulations with DA based on a coupled LES atmosphere-fire model.

References

[1] Al Hassan Afailal, Jérémy Galpin, Anthony Velghe, and Rémi Manceau. Development and validation of a hybrid temporal LES model in the perspective of applications to internal combustion engines. *Oil & Gas Science and Technology – Revue d'IFP Energies nouvelles*, 74:56, 2019.

[2] Mark Asch, Marc Bocquet, and Maëlle Nodet. *Data Assimilation: Methods, Algorithms, and Applications*. Society for Industrial and Applied Mathematics, Philadelphia, PA, December 2016.

[3] L. Thobois, G. Rymer, T. Soulères, and T. Poinsot. Large-Eddy Simulation in IC Engine Geometries. In *SAE*, pages 2004–01–1854, June 2004.

[4] Lucas Villanueva, Karine Truffin, Jacques Borée, and Marcello Meldi. Enhancement of large-eddy simulations for the prediction of an intake flow rig using sequential data assimilation. *Journal of Fluid Mechanics*, 1023:A47, November 2025.