ONERA

THE FRENCH AEROSPACE LAB

Data fusion with multi-fidelity Gaussian processes for aerodynamic experimental and numerical databases

SIAM Conference on Computational Science and Engineering (CSE)

Andrés López-Lopera[1], Nathalie Bartoli[1], Thierry Lefebvre[1] and Sylvain Mouton[2]

March 5, 2021

ONERA – The French Aerospace Lab, France
[1]Information Processing and Systems Department (DTIS)
[2]Le Fauga-Mauzac Wind Tunnels Department (DSFM)

∗ In collaboration with S. Defoort, J. Gauvrit-Ledogar and M. Méheut (BWB database)

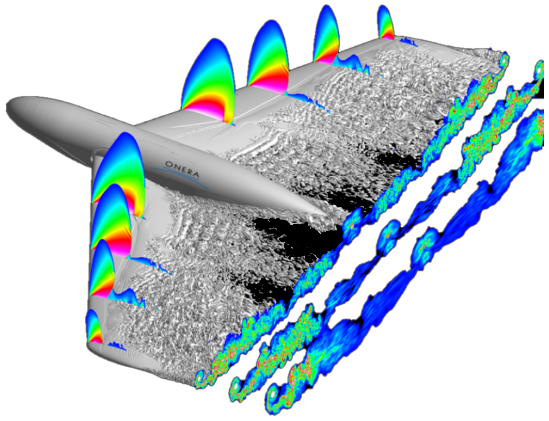
Data fusion in aerodynamics

Data fusion in aerodynamics

· In aerodynamics, data may come from different types of sources:

Flight tests

Wind-tunnel experiments

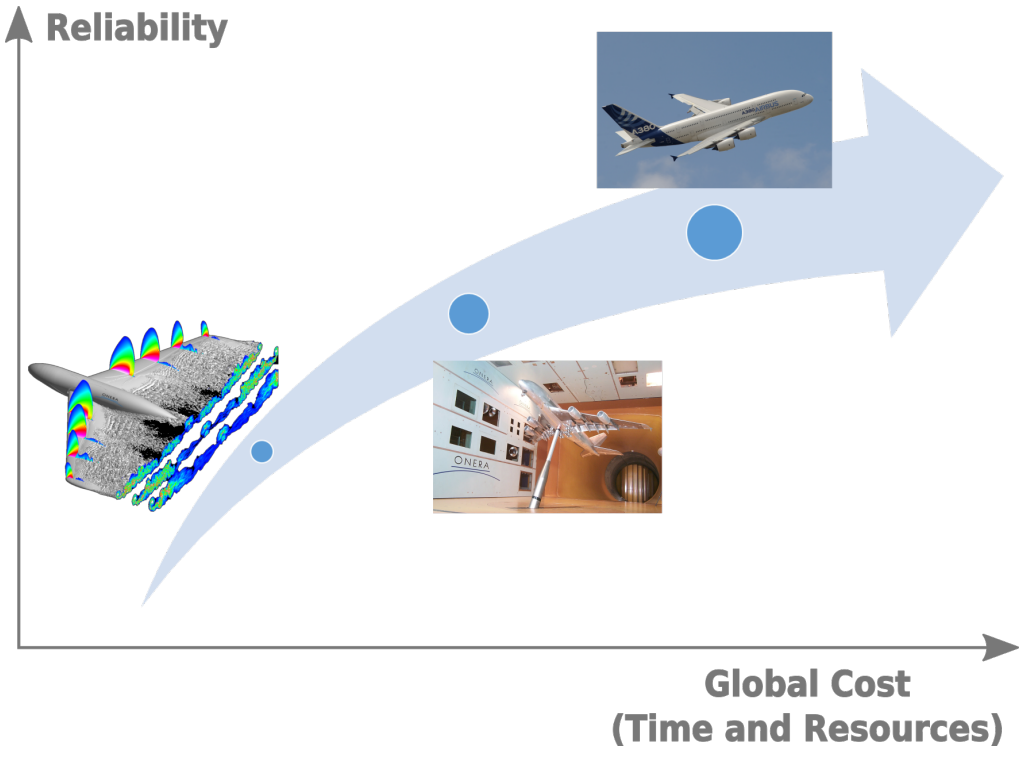


Aerodynamic numerical simulators

- Flight and wind-tunnel tests lead to more reliable assessments
- Simulators can evaluate situations that cannot be performed in practice

Data fusion in aerodynamics

· Different data acquisition schemes may lead to different levels of fidelity



Acquisition Scheme	Level of Fidelity	Data Availability	Cost
flight tests	high	very low	expensive
wind-tunnel tests	upper-intermediate	intermediate	moderate
simulators	low or intermediate	high	cheap or moderate

Data fusion in aerodynamics

· In practice,

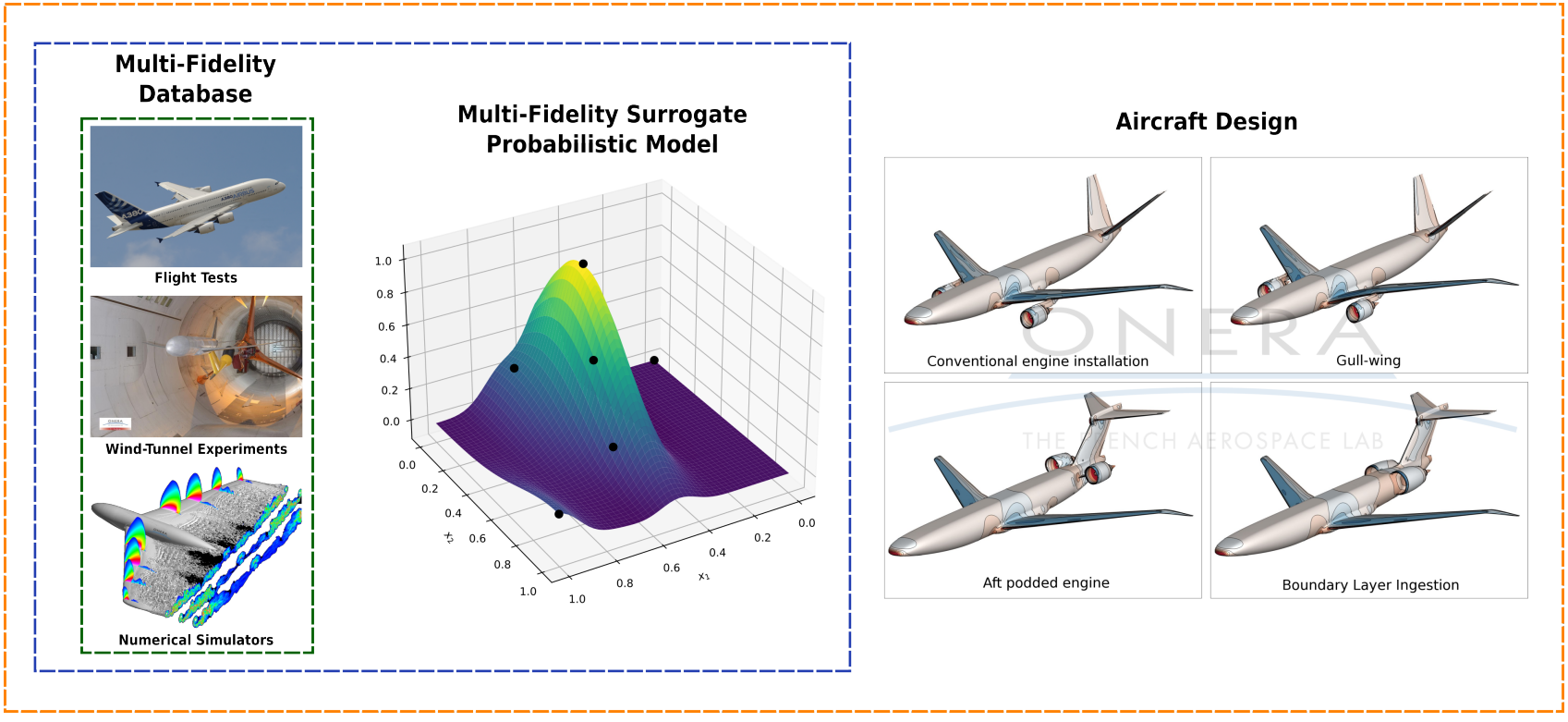
- An independent model is considered for each data acquisition scheme
- Aggregation steps are performed to reduce discrepancy between models
- Decisions are taken using the (empirically) aggregated framework

· The aggregation step is not always straightforward in many cases

- Wrong aggregations can lead to misleading results!

Data fusion in aerodynamics

· Data fusion (DF)-based frameworks aim at jointly treating data acquisition schemes while accounting for their corresponding levels of fidelity



Multi-fidelity DF-based architecture

Data fusion in aerodynamics

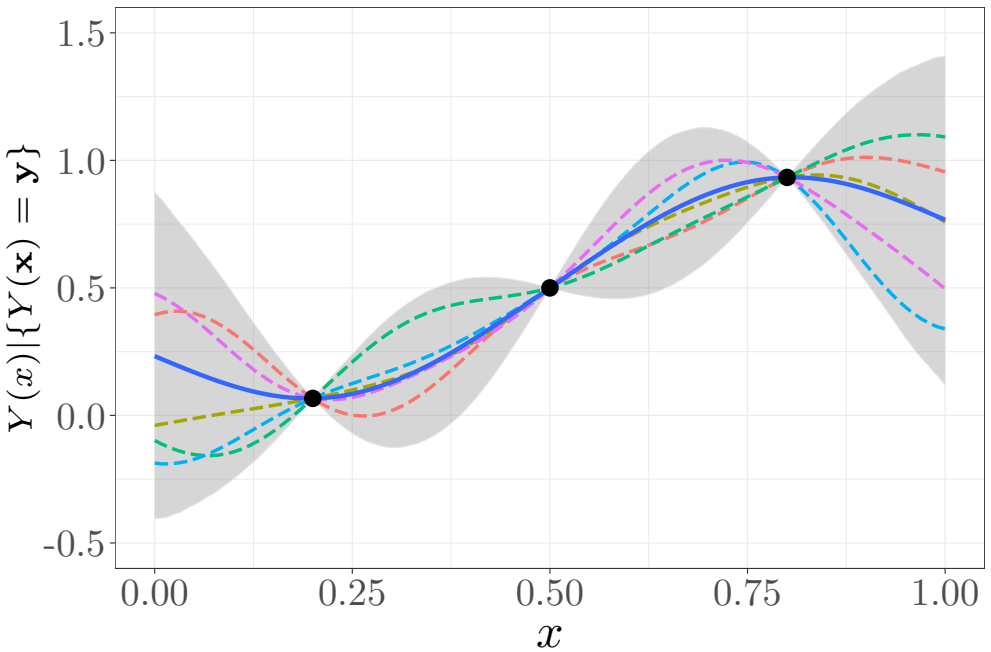
· In this talk, we focus on

- Multi-fidelity DF-based modelling based on Gaussian processes
- An open-source Python-based surrogate modelling toolbox (SMT): `https://github.com/SMTorg/smt`

Multi-fidelity Gaussian processes

Gaussian processes (GPs)

· GP modelling assumes Gaussian priors on (aerodynamic) functions:



- mean function
- prediction intervals
- · · · random realisations

- aerodynamic data

$$(x_i, y_i)_{i=1}^{n} = (\boldsymbol{X}_n, \boldsymbol{y}_n)$$

Gaussian processes (GPs)

· Let $\{Y(\boldsymbol{x}), \boldsymbol{x} \in \mathcal{D}\}$ be a GP prior in $\mathbb{R}$ with compact input space $\mathcal{D} \subset \mathbb{R}^d$:

$$Y \sim \mathcal{GP}(m, k),$$

with mean function m and covariance function (kernel) k.

· GP priors let to tractable computations of the conditional distribution:

$$Y(\boldsymbol{x}) | \{Y(\boldsymbol{X}_n) = \boldsymbol{y}_n\} = \mathcal{GP}(\mu, c), \tag{1}$$

where

conditional mean: $\mu(\boldsymbol{x}) = m(\boldsymbol{x}) + \boldsymbol{k}(\boldsymbol{x}, \boldsymbol{X}_n)^\top \boldsymbol{K}^{-1}(\boldsymbol{X}_n, \boldsymbol{X}_n)[\boldsymbol{y}_n - \boldsymbol{m}(\boldsymbol{X}_n)]$

conditional covariance: $c(\boldsymbol{x}, \boldsymbol{x}') = k(\boldsymbol{x}, \boldsymbol{x}') - \boldsymbol{k}(\boldsymbol{x}, \boldsymbol{X}_n)\boldsymbol{K}^{-1}(\boldsymbol{X}_n, \boldsymbol{X}_n)\boldsymbol{k}(\boldsymbol{X}_n, \boldsymbol{x}')$

predictions: $\mu(\boldsymbol{x}) \pm 2\sqrt{v(\boldsymbol{x})}$, with $v(\boldsymbol{x}) = c(\boldsymbol{x}, \boldsymbol{x})$

Gaussian processes (GPs)

· Benefits of GPs [Rasmussen and Williams, 2005]:

- They form a **flexible prior** over functions while providing a **well-founded probabilistic framework**
- They are commonly used for modelling **expensive black-box functions**
- **Regularity assumptions** can be encoded into kernels
- They have been successfully applied in a wide range of applications

Multi-fidelity model based on Gaussian processes

· We can consider the autoregressive model [see, e.g., Le Gratiet, 2013]:

$$Y_\ell(\boldsymbol{x}) = \rho_\ell(\boldsymbol{x}) Y_{\ell-1}(\boldsymbol{x}) + \delta_\ell(\boldsymbol{x}), \quad \text{for } \ell = 1, \ldots, L \tag{2}$$

where

- $L+1$ is number of fidelity levels
- $Y_0 \sim \mathcal{GP}(m_0, k_0)$ and $\delta_\ell(\boldsymbol{x}) \sim \mathcal{GP}(m_\ell, k_\ell)$
- $\rho_\ell : \mathcal{D} \to \mathbb{R}$ is a scale factor between Z_ℓ and $Z_{\ell-1}$
- $\delta_\ell : \mathcal{D} \to \mathbb{R}$ is the discrepancy function tasked with capturing the differences between Z_ℓ and $Z_{\ell-1}$ beyond scaling

· If $Y_0, \delta_1, \ldots, \delta_L$ are independent GPs, then $Y_1, \ldots, Y_L$ are also GP-distributed

Multi-fidelity model based on Gaussian processes

· We can consider the autoregressive model [see, e.g., Le Gratiet, 2013]:

$$Y_\ell(\boldsymbol{x}) = \rho_\ell(\boldsymbol{x}) Y_{\ell-1}(\boldsymbol{x}) + \delta_\ell(\boldsymbol{x}), \quad \text{for } \ell = 1, \ldots, L \tag{2}$$

where

- $L+1$ is number of fidelity levels
- $Y_0 \sim \mathcal{GP}(m_0, k_0)$ and $\delta_\ell(\boldsymbol{x}) \sim \mathcal{GP}(m_\ell, k_\ell)$
- $\rho_\ell : \mathcal{D} \to \mathbb{R}$ is a scale factor between Z_ℓ and $Z_{\ell-1}$
- $\delta_\ell : \mathcal{D} \to \mathbb{R}$ is the discrepancy function tasked with capturing the differences between Z_ℓ and $Z_{\ell-1}$ beyond scaling

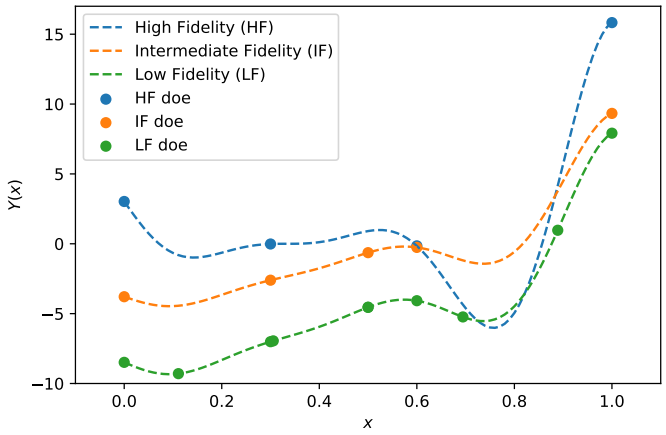
· If $Y_0, \delta_1, \ldots, \delta_L$ are independent GPs, then $Y_1, \ldots, Y_L$ are also GP-distributed

∗ As shown by [Le Gratiet, 2013], efficient implementations are obtained using nested designs of experiments (DoEs):

$$D_\ell \subseteq D_{\ell-1}, \tag{3}$$

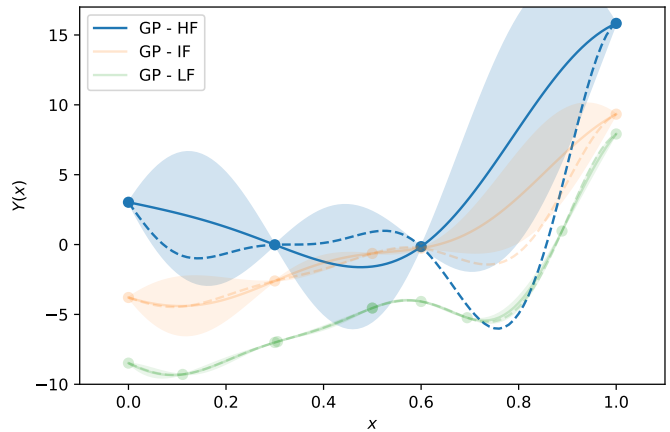
with $D_\ell = (\boldsymbol{x}_{\ell,i}, y_{\ell,i})_{0 \leq \ell \leq L-1, 1 \leq i \leq n_\ell}$.

Multi-fidelity model based on Gaussian processes

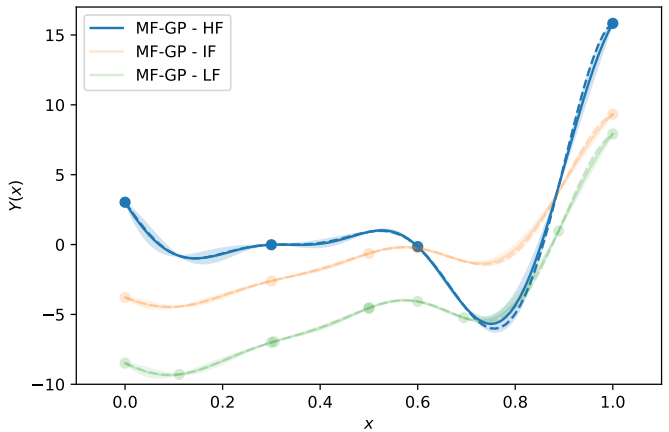


(a) True functions and design points

- The (nested) dataset contains:
  - 4 high-fidelity (HF) design points
  - 5 intermediate-fidelity (IF) design points
  - 8 low-fidelity (LF) design points



(b) Independent GP models

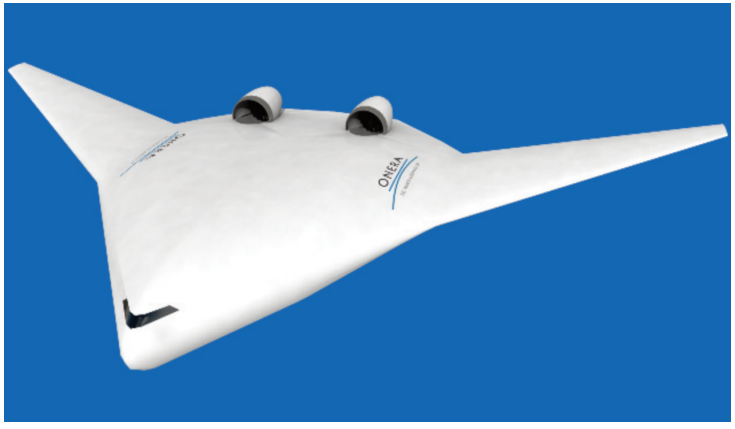


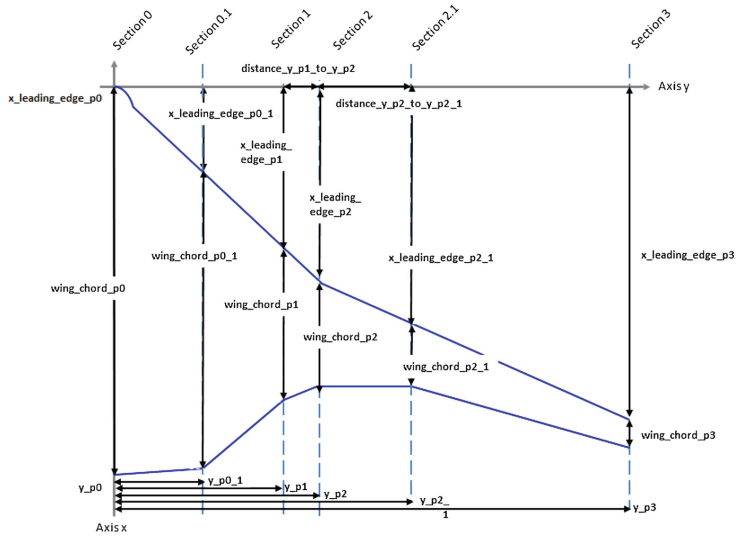
(c) Multi-fidelity GP model

1D multi-fidelity regression example with 3 levels of fidelity

ONERA
THE FRENCH AEROSPACE LAB

Blended wing body (BWB) database [Gauvrit-Ledogar et al., 2020]





Example of the configuration of a BWB studied in Europe [Gauvrit-Ledogar et al., 2020]

Inputs x: 7 design variables including,

- wing chord P2
- wing span
- wing thickness ratio P0, P2
- taper ratio P3
- wing sweep P0, P2

Output y: Drag coefficient (CD) for a fixed lift coefficient (CL).

· The nested database contains 3 levels of fidelity, each with 241 observations:

- lowest- and medium-fidelity: empirical models
- high-fidelity: CFD based on Euler equations

ONERA
THE FRENCH AEROSPACE LAB

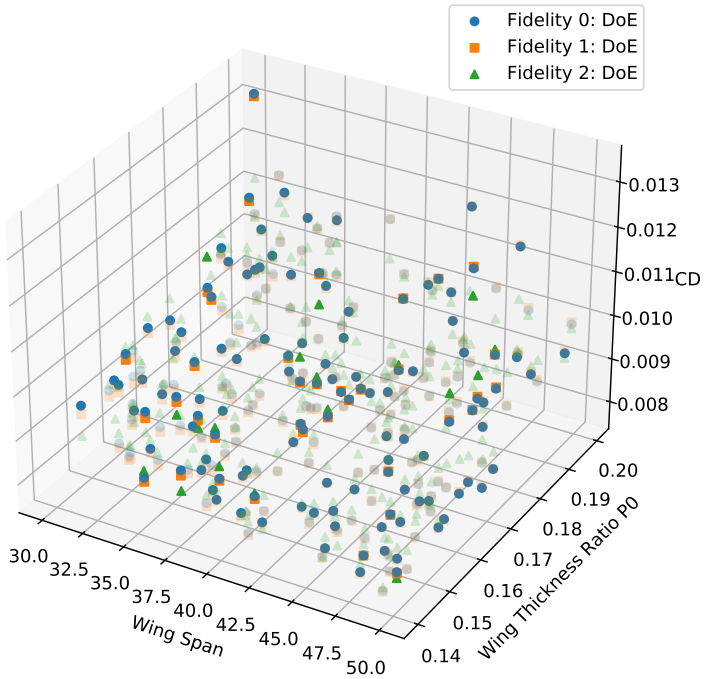
Blended wing body (BWB) database [Gauvrit-Ledogar et al., 2020]

· For illustration purposes, we consider as inputs:

$$x = (\text{wing span, wing thickness ratio P0})$$

· Missing inputs are modelled as a noise term in the GP

· The multi-fidelity GP is trained with:

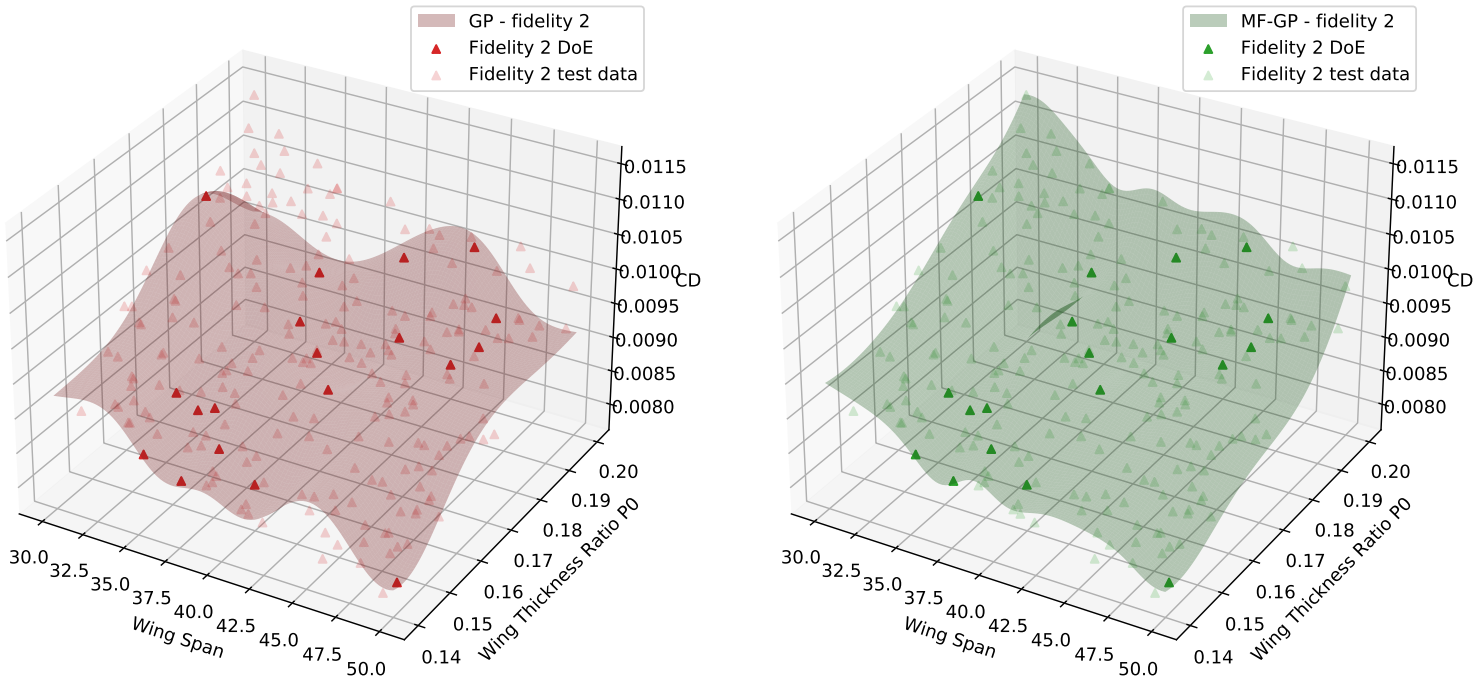


2D dataset used for training the model

- 50% of the lowest fidelity database:
$$(x_{0,i}, y_{0,i})_{1 \leq i \leq 121}$$
- 15% of the intermediate fidelity database:
$$(x_{1,i}, y_{1,i})_{1 \leq i \leq 37}$$
- 7.5% of the highest fidelity database:
$$(x_{2,i}, y_{2,i})_{1 \leq i \leq 19}$$

ONERA
THE FRENCH AEROSPACE LAB

Blended wing body (BWB) database [Gauvrit-Ledogar et al., 2020]

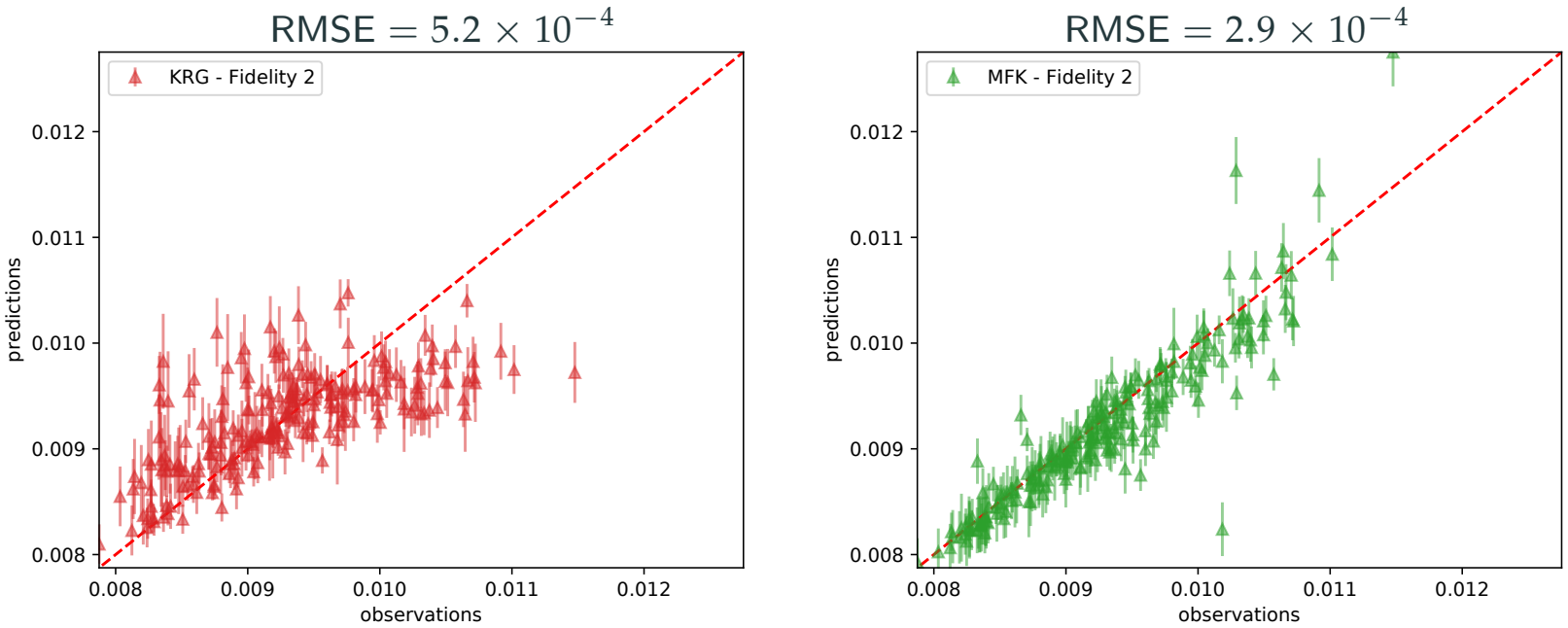


(a) Classical GP at $\ell = 2$ (b) Multi-fidelity GP at $\ell = 2$

Predictions on the 2D BWB database

Blended wing body (BWB) database [Gauvrit-Ledogar et al., 2020]

By considering the 7 inputs with the same DoE used in the 2D example:



(a) Classical GP at $\ell = 2$ (b) Multi-fidelity GP at $\ell = 2$

ground truth vs predictions ($\pm$ 1 standard-deviation confidence intervals [68%])

Predictions on the 7D BWB database

ONERA
THE FRENCH AEROSPACE LAB

Surrogate modelling Python toolbox

Surrogate modelling Python toolbox [Bouhlel et al., 2019]

SMT: Surrogate Modeling Toolbox

The surrogate modeling toolbox (SMT) is an open-source Python package consisting of libraries of surrogate modeling methods (e.g., radial basis functions, kriging), sampling methods, and benchmarking problems. SMT is designed to make it easy for developers to implement new surrogate models in a well-tested and well-document platform, and for users to have a library of surrogate modeling methods with which to use and compare methods.

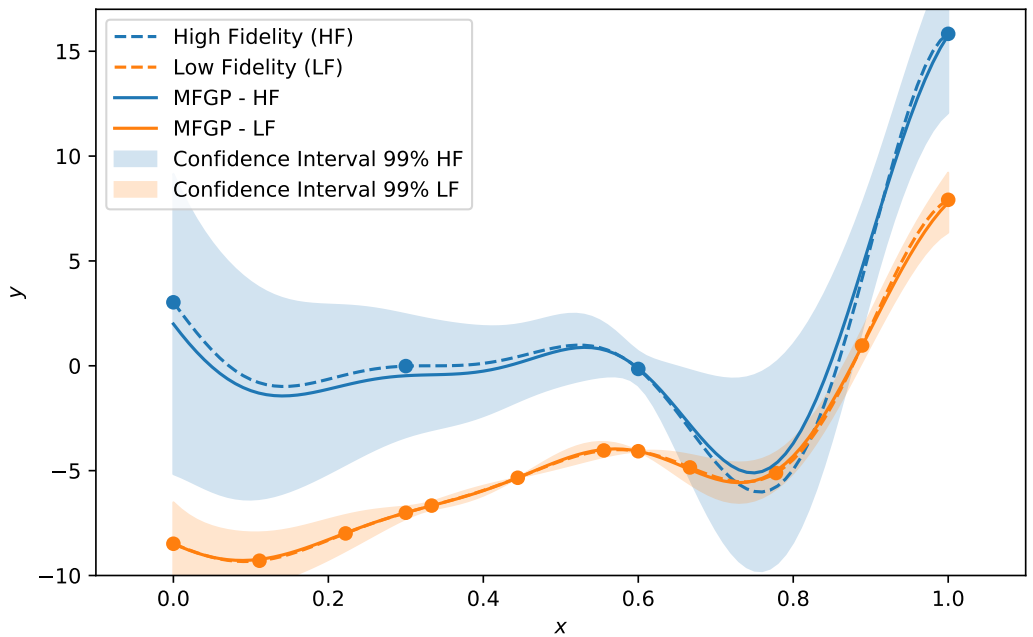
- Surrogate modeling methods
  - Radial basis functions
  - Inverse-distance weighting
  - Regularized minimal-energy tensor-product splines
  - Least-squares approximation
  - Second-order polynomial approximation
  - Kriging
  - KPLS
  - KPLSK
  - GEKPLS
  - GENN
  - Marginal Gaussian Process (MGP)
- Benchmarking problems
  - Sphere function
  - Branin function
  - Lp norm function
  - Rosenbrock function
  - Tensor-product function
  - Cantilever beam function
  - Robot arm function
  - Torsion vibration function
  - Water flow function
  - Welded beam function
  - Wing weight function
- Applications
  - Mixture of experts (MOE)
  - Variable-fidelity modeling (VFM)
  - Multi-Fidelity Kriging (MFK)
  - Multi-Fidelity Kriging KPLS (MFKPLS)
  - Multi-Fidelity Kriging KPLSK (MFKPLSK)
  - Efficient Global Optimization (EGO)
  - Mixed-Integer Sampling and Surrogate (Continuous Relaxation)

Link: `https://github.com/SMTorg/smt`

Heteroscedastic multi-fidelity GP models

$$Y_\ell(\boldsymbol{x}_{\ell,i}) + \varepsilon_{\ell,i} = y_{\ell,i}, \quad \varepsilon_{\ell,i} \sim \mathcal{N}\left(0, \tau^2_{\ell,i}\right) \tag{4}$$

with noise variance parameters $\tau^2_{\ell,i} \in \mathbb{R}^+$, for $\ell = 0, \ldots, L$ and $i = 1, \ldots, n$



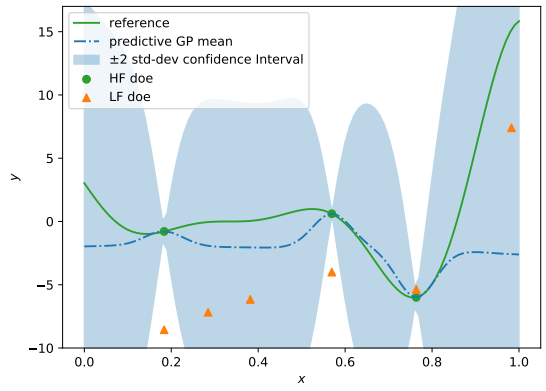
- 4 HF design points
- 12 LF design points

Heteroscedastic multi-fidelity GP regression example

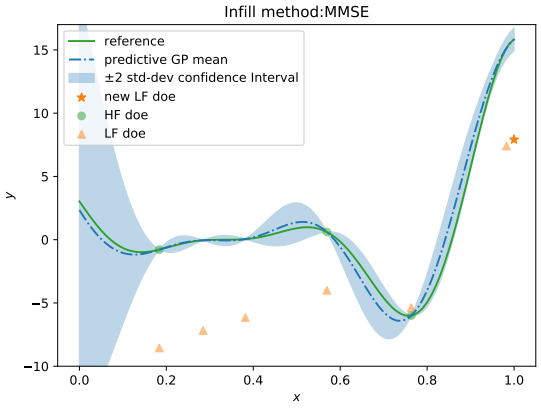
· τ_i^2 can be fixed or estimated if considering repetitions of observations

ONERA
THE FRENCH AEROSPACE LAB

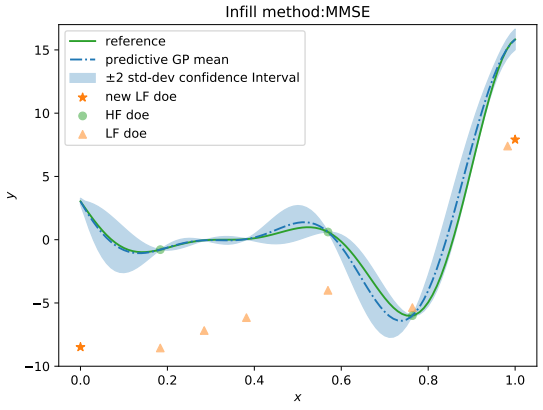
Adaptive design of experiments (DoEs) for multi-fidelity GPs



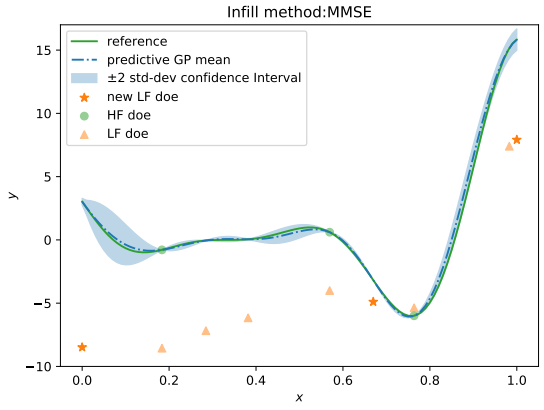
(a) Iter. 0



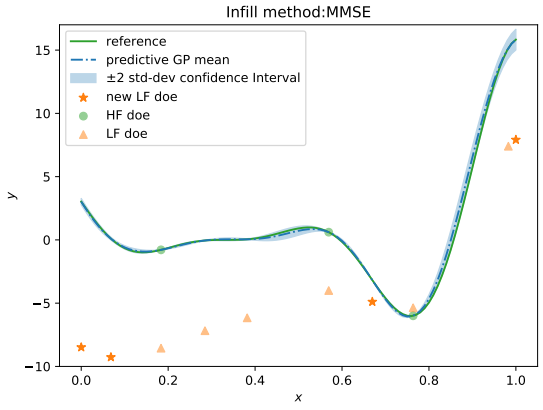
(b) Iter. 1: new LF point



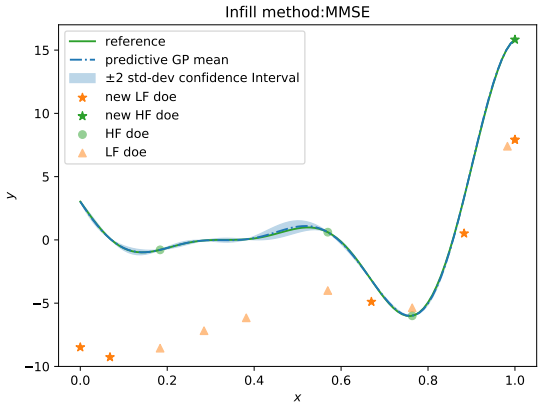
(c) Iter. 2: new LF point



(d) Iter. 3: new LF point



(e) Iter. 4: new LF point



(f) Iter. 6: new HF point

Initial DoE: 3 HF points 6 LF points
Updated DoE: 4 HF points 12 LF points

ONERA
THE FRENCH AEROSPACE LAB

SMT: latest implementations

· For adaptive DoEs, we aim at reducing the uncertainty at the L-th GP:

$$Y_L(\boldsymbol{x}) = \rho_L(\boldsymbol{x})Y_{L-1}(\boldsymbol{x}) + \delta_L(\boldsymbol{x})$$

Step 1. Optimise an acquisition criterion for placing a new point, e.g.:

$$\boldsymbol{x}_* = \text{argmax}_{\boldsymbol{x}\in\mathcal{D}}\, v_{L,n}(\boldsymbol{x}), \tag{5}$$

where $v_{L,n}$ is the variance of $Y_L(\boldsymbol{x})|\{Y_0(\boldsymbol{X}_{0,n_0}) = \boldsymbol{y}_{0,n_0}, \ldots, Y_L(\boldsymbol{X}_{L,n_L}) = \boldsymbol{y}_{L,n_L}\}$

Step 2. Select the number of levels to be enriched, e.g. [Meliani et al., 2019]:

$$\ell_* = \text{argmax}_{\ell\in\{0,\ldots,L\}} \frac{\sum_{\kappa=0}^{\ell} \left[v_{\delta,\kappa,n}(\boldsymbol{x}_*) \prod_{j=\kappa}^{L-1} \rho_j^2(\boldsymbol{x}_*)\right]}{\left[\sum_{\kappa=0}^{\ell} c_\kappa\right]^2} \tag{6}$$

where

- $v_{\delta,\kappa,n}$ is the variance of $\delta_\kappa(\boldsymbol{x})|\{Y(\boldsymbol{X}_{0,n_0}) = \boldsymbol{y}_{0,n_0}, \ldots, Y(\boldsymbol{X}_{\kappa,n_\kappa}) = \boldsymbol{y}_{\kappa,n_\kappa}\}$
- c_κ is a user-defined cost of enrichment for the κ-th fidelity level

Step 3. Enrich the DoE $(\boldsymbol{X}_{\ell,n_\ell+1}, \boldsymbol{y}_{\ell,n_\ell+1}) = ((\boldsymbol{X}_{\ell,n_\ell}, \boldsymbol{y}_{\ell,n_\ell}), (\boldsymbol{x}_*, y_{\ell,*}))$ for $\ell = 1, \ldots, \ell_*$ and repeat **steps 1** and **2**

ONERA
THE FRENCH AEROSPACE LAB

Conclusions

Conclusions

· Data fusion can be performed via multi-fidelity Gaussian processes (GPs)

- Aerodynamic data aggregation is considered in a unified framework
- The resulting model leads to more accurate predictions

· The GP-based framework can account for:

- uncertainty quantification
- noisy observations
- adaptive designs of experiments (DoEs)
- Bayesian global optimisation

· Developments are freely available in the Python SMT toolbox

Challenges:

- High number of observations: complexity $\mathcal{O}(n^3)$ and storage $\mathcal{O}(n^2)$
- Solving high-order integrals in adaptive DoEs based on:

$$\boldsymbol{x}_* = \text{argmin}_{x \in \mathcal{D}} \int_{\mathcal{D}} v_{n,\boldsymbol{\nu}}(\boldsymbol{x}) d\boldsymbol{x}$$

- Building proper nested aerodynamic DoEs

References

M. Binois, R. B. Gramacy, and M. Ludkovski. Practical heteroscedastic Gaussian process modeling for large simulation experiments. *Journal of Computational and Graphical Statistics*, (4), 2018.

M. A. Bouhlel, J. T. Hwang, N. Bartoli, R. Lafage, J. Morlier, and J. R. R. A. Martins. A Python surrogate modeling framework with derivatives. *Advances in Engineering Software*, 2019.

Julie Gauvrit-Ledogar, Arnault Tremolet, and Loïc Brevault. *Blended Wing Body Design*. Springer International Publishing, Cham, 2020.

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

M. Meliani, N. Bartoli, T. Lefebvre, M. A. Bouhlel, J. R. R. A. Martins, and J. Morlier. Multi-fidelity efficient global optimization: Methodology and application to airfoil shape design. In *AIAA Aviation Forum*. 2019.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, Cambridge, MA, 2005.