G. Gori, O. Le Maître, P.M. Congedo
Debiased Multifidelity Approach to Surrogate Modeling in Aerospace Applications
Journal of Aircraft, published online 22/07/2025
doi:10.2514/1.c037765

A debiased multi-fidelity approach to surrogate modeling in aerospace applications

Giulio Gori *
Department of Aerospace Science and Technology, Politecnico di Milano, Via La Masa 34, 20156, Milano, Italy

Olivier Le Maître†
CNRS/Inria, Centre de Mathématiques Appliquées, École Polytechnique, IPP, Route de Saclay Palaiseau Cedex 91128, France

Pietro M. Congedo‡
Inria, Centre de Mathématiques Appliquées, École Polytechnique, IPP, Route de Saclay Palaiseau Cedex 91128, France

We propose a multi-fidelity formulation for generating co-kriging surrogates of complex physics models. First, we show that the standard autoregressive recursive approach may be subject to substantial limitations due to possible Modeler's biases/errors. These are inherent to the process of establishing a nested hierarchy concerning the alleged fidelity of the available models. The formulation we propose mitigates this issue. At each hierarchy level, the predictor consists of a linear combination of all previous levels instead of just the underlying one. The methodology implies a slightly higher training cost for the surrogate. However, the higher training cost is acceptable considering the effort typically required to generate data in aerospace applications. Few artificial tests, including the optimization of a 2D airfoil, illustrate strengths and weaknesses of the approach.

Nomenclature

C	=	prediction variance from a $\mathcal{GP}$
K	=	correlation kernel
$\mathcal{K}$	=	augmented correlation kernel
L	=	level of highest fidelity
$\mathcal{L}$	=	negative marginal log likelihood
l	=	fidelity level
M	=	true model
$\mathcal{M}$	=	surrogate model
μ	=	mean prediction from a $\mathcal{GP}$
$\mathcal{N}$	=	Gaussian random variable
X	=	set of input points
Y	=	set of output data
ρ	=	regression coefficient
σ	=	observation noise
Θ	=	vector of $\mathcal{GP}$ hyperparameters
θ_i	=	$\mathcal{GP}$ hyperparameter
$\mathcal{U}$	=	uniform random variable

*Assistant Professor of Fluid Dynamics, Department of Aerospace Science and Technology, Politecnico di Milano, Via La Masa 34, 20156, Milano, Italy; giulio.gori@polimi.it (Corresponding Author)

†CNRS Research Director, PLATON Team, CNRS/Inria, Centre de Mathématiques Appliquées, École Polytechnique, IPP, Route de Saclay Palaiseau Cedex 91128, France

‡Inria Research Director, PLATON Team, Centre de Mathématiques Appliquées, École Polytechnique, IPP, Route de Saclay Palaiseau Cedex 91128, France

I. Introduction

Multi-fidelity methods leverage on the concatenation of data sets that present enormous diversity in terms of information, size, and behavior. Pieces of information of diverse fidelity and complexity complement each other, leading to improved estimate accuracy and to a minimization of the cost associated with parametrization. In recent decades, they have been successfully employed in many applications e.g., aerospace design and optimization [1–3]. Multi-fidelity regression models bring clear advantages to the preliminary design phases of engineering products, where data of different cost and quality are exploited to define the best feasible configuration. At this stage, prediction errors, defined as the difference between a predicted parameter and its actual value after product completion and testing, should ideally not exceed a few percent [4]. Not only do multi-fidelity approaches benefit decision-making processes, they also support the discovery of physics laws governing complex systems, and accelerate outer-loop applications e.g., optimization.

In a multi-fidelity setting, it is fundamental to establish the correct hierarchy in terms of fidelity with respect to target applications. Unfortunately, this can vary significantly along the spectrum between low and high. In particular, the complexity characterizing engineering applications usually makes direct estimation of data credibility difficult, if not intractable. Furthermore, the proper implementation of a standard method concerning the Credibility Assessment Scale has long since debated [5]. This leaves ground for modeling biases and, ultimately, poses a limit to multi-fidelity strategies.

Experiments are critical since they utterly support the model validation process [6]. Unfortunately, experiments often consist of a simplified, limited, and partial imitation of reality. Sometimes, computational models provide a more accurate representation. A resounding example is the NASA Common Research Model (CRM) [7]. The CRM test cases were devised for the purpose of validating specific applications of Computational Fluid Dynamics (CFD). Geometry features were specifically designed to promote the occurrence of phenomena of interest for research and development, e.g., flow separation. However, the wind tunnel model underwent significant deformations during the experimental campaign [8], thus frustrating the design choices. In the CRM test case, computational simulations outperform wind-tunnel tests since the goal is to predict the aerodynamic performance of the unbent design.

It follows that the process of establishing any specific fidelity hierarchy between data sets, either from computer models or experiments, is at least questionable. Although advanced Uncertainty Quantification (UQ) techniques can be exploited, the deductive Scientific Modeling approach typically adopted by the Modeler i.e., the person who devises a new model, is strongly hypotheses-driven and hence inherently biased [9].

The current state-of-the-art offers a plethora of multi-fidelity techniques [1, 2]. Kriging [10, 11] is a powerful technique and very well serves applications entailing the enrichment of a data set e.g., sequential design [12]. A pioneering example of multi-fidelity Kriging (the so-called co-Kriging models) can be found in [13]. In the work, the authors construct an autoregressive formulation to surrogate complex computer codes that can be run at different levels of sophistication and cost. The combination of predictions of variable fidelity improves the efficiency and accuracy of the resulting metamodel. Several methods have been developed on top of this work e.g., see [14–17]. Above all, we mention Refs. [18, 19] proposing a recursive formulation endowed with higher computational efficiency and stability.

In this paper, we focus on co-kriging and propose an extension to the formulation from [18]. The goal is to develop a multi-fidelity framework capable of mitigating modeling biases introduced by the Modeler. In particular, a formulation robust to biases affecting the alleged hierarchy between available data sets. In doing so, we will make no distinction between data generated by computer codes or experiments.

This paper is organized as follows. In Section II we recall the standard formulation for multi-fidelity co-kriging and present the details of the extension proposed to overcome the highlighted barriers. In Sec. III, we expose the key weakness of the standard model and assess the performance of our debiased approach. Eventually, in Sec. IV we summarize the findings and provide future perspectives.

II. Multi-fidelity Gaussian Process regression

We refer to the physical process under investigation with the appellative *reality of interest*. Being $l = 0, \dots, L$, lets assume we have access to L models of different cost and fidelity to predict our reality of interest. All $M^{(l)}$ models share the same input space Ω and map it to a certain scalar Quantity of Interest (QoI). Namely, $M^{(l)} : \Omega \subset \mathbb{R}^d \mapsto \mathbb{R}$. Let $X^l = \{\mathbf{x}_1^l, \dots, \mathbf{x}_{N^l}^l\}$ be the set of N^l training points s.t. $\mathbf{x} \in \Omega \subset \mathbb{R}^d$. Each training point $\mathbf{x}_n^l$ is associated with a noisy observation $y_n^l = M^{(l)}\left(\mathbf{x}_n^l\right) + \epsilon_n^l$ from a certain model $M^{(l)}\left(\mathbf{x}_n^l\right)$, assuming i.i.d. Gaussian random variables $\epsilon_n^l \sim \mathcal{N}\left(0, \sigma_l^2\right)$. The set of available observations is included in $Y^{(l)}$.

The classical autoregressive model [13] requires sorting data sets according to their alleged fidelity. For $l = 0$, the

response of the model $M^{(0)}$ is substituted by its approximation $\mathcal{M}^{(0)}$ via Gaussian process to obtain

$$\mathcal{M}^{(0)}(\mathbf{x}) = \delta_0(\mathbf{x}). \tag{1}$$

For $l > 0$, the response of each model $M^{(l)}$ is substituted by the (scaled) lower predictor plus a correction term i.e., a Gaussian Process, modeling the residual between $Y^{(l)}$ data and $\mathcal{M}^{(l-1)}$ predictions

$$\begin{cases} \mathcal{M}^{(l)}(\mathbf{x}) = \rho_{(l-1)}\mathcal{M}^{(l-1)}(\mathbf{x}) + \delta_l(\mathbf{x}), \\ \mathcal{M}^{(l-1)}(\mathbf{x}) \perp \delta_l(\mathbf{x}), \end{cases} \tag{2}$$

where, for every level l,

$$\delta_l(\mathbf{x}) \sim \mathcal{GP}\left(\mathbf{f}_l^T(\mathbf{x})\boldsymbol{\beta}_l, K\left(\mathbf{x}, \mathbf{x}'; \Theta^{(l)}\right)\right). \tag{3}$$

The superscript ^{T} indicates the matrix transpose, $\perp$ the independence relationship, $\mathcal{GP}$ is used to identify Gaussian Processes, whereas K is a correlation function (or kernel) depending on a set of hyperparameters θ_i^l included in the $\boldsymbol{\Theta}^l$ vector. The vector $\mathbf{f}_l(\mathbf{x})$ includes the explicit basis functions in case a nonzero mean is specified for the $\mathcal{GP}$, being $\boldsymbol{\beta}_l$ a vector of scaling coefficients. Here, we choose to rely on zero mean priors for the $\mathcal{GP}$.

In the described framework, $\rho_{(l-1)} \in \mathbb{R}$ can be interpreted as the correlation factor between two consecutive models,

$$\rho_{(l-1)} = \frac{\text{cov}\left(\mathcal{M}^{(l)}(\mathbf{x}), \mathcal{M}^{(l-1)}(\mathbf{x})\right)}{\text{var}\left(\mathcal{M}^{(l-1)}(\mathbf{x})\right)}. \tag{4}$$

In other words, $\mathcal{M}^{(l)}(\mathbf{x})$ is given by the sum of $\mathcal{M}^{(l-1)}(\mathbf{x})$ predictions, scaled by $\rho_{(l-1)}(\mathbf{x})$, plus a correction term δ^l modeling the discrepancy between predictions $\rho_{(l-1)}\mathcal{M}^{(l-1)}(X^{(l)})$ and observations $Y^{(l)}$ at training points $X^{(l)}$.

A. Standard recursive co-kriging model

The recursive formulation proposed in [19] builds surrogates sequentially, from the lowest to the highest fidelity level. The authors take advantage of a different expression for $\mathcal{GP}$ and prove that building a set of L independent Kriging models is formally equal to building a L-level co-Kriging surrogate i.e., the two formulations are equivalent and deliver identical predictive mean and variance. The process of training the multi-fidelity surrogate benefits significantly from this proof as the sequential inversion of L sub-matrices (one for each level) is performed in place of inverting a unique (large) matrix. In their work, the authors also propose to extend the original formulation by introducing a spatial dependency of the adjustment parameters $\rho_{(l)}$. With no loss of generality, in this paper we neglect this spatial dependency to simplify the treatment and lighten the notation. The correlation among models is therefore assimilated to an affine transformation with scaling factor $\rho_{(l-1)}$, see [13]. According to [19], the predictive mean $\mu_{(l)}$ and covariance C_l read

$$\mu^l(\mathbf{x}) = \rho_{(l-1)}\mu_{(l-1)}(\mathbf{x}) + \mathcal{K}^l\left(\mathbf{x}, X^l|\rho_{(l-1)}, \boldsymbol{\Theta}^l\right)\left[\mathcal{K}^l\left(X^l, X^l|\rho_{(l-1)}, \boldsymbol{\Theta}^l\right) + \sigma_l^2\mathrm{I}\right]^{-1}\left(Y^l - \rho_{(l-1)}\mu^{(l-1)}(\mathbf{x})\right), \tag{5}$$

and

$$C^l(\mathbf{x}, \mathbf{x}') = \mathcal{K}^l\left(\mathbf{x}, \mathbf{x}'|\rho_{(l-1)}, \boldsymbol{\Theta}^l\right) - \left[\mathcal{K}^l\left(\mathbf{x}, X^l|\rho_{(l-1)}, \boldsymbol{\Theta}^l\right)\right]\left[\mathcal{K}^l\left(X^l, X^l|\rho_{(l-1)}, \boldsymbol{\Theta}^l\right) + \sigma_l^2\mathrm{I}\right]^{-1}\left[\mathcal{K}^l(X^l, \mathbf{x}'|\rho_{(l-1)}, \boldsymbol{\Theta}^l)\right], \tag{6}$$

whereas $\mathcal{K}$ reads

$$\mathcal{K}^l\left(A, B|\,\rho_{(l-1)}, \boldsymbol{\Theta}^l\right) = \rho_{(l-1)}^2 C^{(l-1)}(A, B) + K^l\left(A, B|\boldsymbol{\Theta}^l\right). \tag{7}$$

Note that the observation noise σ_l^2 is explicitly accounted within the μ and C expressions.

B. Increasingly recursive co-kriging model

In the autoregressive procedure presented in Sec. II.A, if $\mathcal{M}^{(l-1)}(X^l)$ predictions are completely off Y^l i.e., the data set $Y^{(l-1)}$ has little or nothing to do with Y^l, then the regression parameter $\rho_{(l-1)}$ will be very small and close to zero. In other words, if $\mathcal{M}^{(l-1)}(X^l)$ were a bad approximation of $\mathcal{M}^{(l)}(X^l)$, little information would be exploitable from it. As information from $\mathcal{M}^{(l-1)}(X^l)$ are discarded, then information from all lower levels $(l-2, \ldots, 0)$ are also ignored and this poses significant limits to the multi-fidelity approach in case the Modeler mistakenly order the data sets.

We propose an increasingly including strategy. That is, we now seek an extension of [19] where the lower level predictor consists in a linear combination of all previous levels, with coefficients $\rho^l_{l'<l}$. For convenience, we write

$$\mu^{<l}(\mathbf{x}|\boldsymbol{\rho}^l) \equiv \sum_{l'<l} \boldsymbol{\rho}^l_{l'} \mu^{l'}(\mathbf{x}), \quad C^{<l}(\mathbf{x},\mathbf{x}'|\boldsymbol{\rho}^l) \equiv \sum_{l'<l} (\boldsymbol{\rho}^l_{l'})^2 C^{l'}(\mathbf{x},\mathbf{x}'). \tag{8}$$

With this notation, $\boldsymbol{\rho}^l = \{\rho^l_0, \ldots, \rho^l_{l-1}\}^T$ is a vector including a set of regression coefficients. The predictive mean and covariance then read

$$\mu^l(\mathbf{x}) = \mu^{<l}(\mathbf{x}|\boldsymbol{\rho}^l) + \mathcal{K}^l\left(\mathbf{x}, X^l|\boldsymbol{\rho}^l, \boldsymbol{\Theta}^l\right)\left[\mathcal{K}^l\left(X^l, X^l|\boldsymbol{\rho}^l, \boldsymbol{\Theta}^l\right) + \sigma_l^2 \mathrm{I}\right]^{-1}\left(Y^l - \mu^{<l}(\mathbf{x}|\boldsymbol{\rho}^l)\right), \tag{9}$$

$$C^l(\mathbf{x},\mathbf{x}') = \mathcal{K}^l\left(\mathbf{x},\mathbf{x}'|\boldsymbol{\rho}^l, \boldsymbol{\Theta}^l\right) - \left[\mathcal{K}^l\left(\mathbf{x}, X^l|\boldsymbol{\rho}^l, \boldsymbol{\Theta}^l\right)\right]\left[\mathcal{K}^l\left(X^l, X^l|\boldsymbol{\rho}^l, \boldsymbol{\Theta}^l\right) + \sigma_l^2 \mathrm{I}\right]^{-1}\left[\mathcal{K}^l(X^l, \mathbf{x}'|\boldsymbol{\rho}^l, \boldsymbol{\Theta}^l)\right], \tag{10}$$

where $\mathcal{K}$

$$\mathcal{K}^l\left(A, B|\,\boldsymbol{\rho}^l, \boldsymbol{\Theta}^l\right) = C^{<l}(A, B|\boldsymbol{\rho}^l) + K^l\left(A, B|\boldsymbol{\Theta}^l\right). \tag{11}$$

A graphical comparison of the information flow that occurs in the standard [19] (left-hand side) and in the proposed formulation (right-hand side) is presented in Fig. 1. In the standard formulation, imposing a wrong hierarchy may

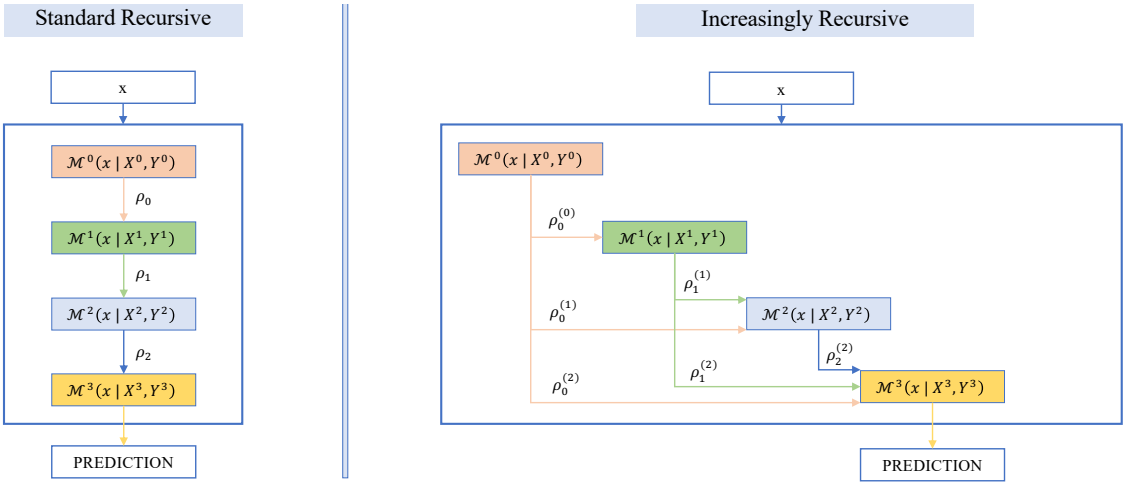


Fig. 1 Graphical comparison of the information flow occurring in the standard recursive formulation [19] (left) and in the proposed formulation (right).

frustrate the effort of relying on a multi-fidelity approach. To set an extreme example, imagine that the Y^1 set is corrupted, that is, data are a bad approximation of reality, whereas the remaining sets are fair. The training process will learn a very poor correlation of level 1 with level 0 and 2 i.e., both the ρ_0 and ρ_1 coefficients will be close to zero, if not just zero. As a consequence, all information brought by Y^1 will be discarded together with the information encoded in Y^0. Instead, with the proposed formulation, the information can still reach all levels above the corrupted one. Being the correlations to lower levels learned simultaneously, the framework is able to automatically exploit multiple models at once, selecting only the useful ones. It follows a multi-fidelity framework more robust to biases. Moreover, the correlation of each level with others is explicitly available through the $\boldsymbol{\rho}^l$ vectors, providing additional information for interpreting the model.

One may argue that, in the proposed setting, the contribution from low levels is accounted multiple times. Namely, it is accounted explicitly in $\mu^{<l}(\mathbf{x}|\boldsymbol{\rho}^l) = \sum_{l'<l} \boldsymbol{\rho}^l_{l'} \mu^{l'}(\mathbf{x})$ and implicitly in the $\mu^{l'}(\mathbf{x})$ terms because of the recursive structure. In opposing this argument, we point out that each $\mathcal{M}^{(l)}$ is trained to fit Y^l. Consequently, the $l' < l$ data sets contribute to improving the approximation of $M^{(l)}$, not the approximation of reality of interest $M^{(L)}$. The proposed formulation opens the path to inferring how the available models correlate with the reality of interest. Ultimately, it can be exploited to obtain physics insights about complex phenomena. Naturally, the proposed formulation yields an increased complexity since it requires the estimation of the $\boldsymbol{\rho}^l$ components for $l = 1, \ldots, L$. This may render the approach demanding in case the spatial dependence of the correlation components proposed in [19] is retained.

C. Parameters estimation

The regression problem requires the estimation of the $\boldsymbol{\rho}^l$ (or $\rho_{(l-1)}$ for [19]), $\boldsymbol{\Theta}^l$ and σ_l parameters at each level l. Thanks to the recursive formulation, the estimation of these parameters can be done sequentially, from the lowest to the highest level. This task can be achieved following diverse approaches. Here, we apply a type II maximum likelihood method. We seek a combination of parameters maximizing the negative marginal log likelihood $\mathcal{L}$ which, for the proposed strategy, takes the form

$$\begin{aligned}\mathcal{L}\left(Y^l|X^l,\Theta^l,\rho^l,\sigma_l\right) = -\frac{1}{2}\left(Y^l-\mu^{<l}(X^l|\rho^l)\right)\left[\mathcal{K}^l\left(X^l,X^l|\rho^l,\Theta^l\right)+\sigma_l^2\mathrm{I}\right]^{-1}\\ \times\left(Y^l-\mu^{<l}(X^l|\rho^l)\right)-\frac{1}{2}\log\left|\mathcal{K}^l\left(X^l,X^l|\rho^l,\Theta^l\right)+\sigma_l^2\mathrm{I}\right|,\end{aligned} \tag{12}$$

being $\mathcal{K}$ defined by Eq. (11). For the standard formulation [19], the expression for $\mathcal{L}$ is analogous except that $\mathcal{K}$ is given by Eq. (7), $\mu^{(l-1)}(X^l)$ is used in place of $\mu^{<l}(X^l|\rho^l)$, and the scalar $\rho_{(l-1)}$ substitutes the vector ρ^l.

III. Applications

We present three exemplary tests to expose the strengths and weaknesses of the proposed formulation. We label the multi-fidelity model obtained according to the standard recursive strategy from [19] as SR (Standard Recursive), the surrogate obtained with the proposed approach as IR (Increasingly Recursive), and the single-fidelity $\mathcal{GP}$ surrogate i.e., trained on the Y^L data set only, as SF (Single-Fidelity). In all applications hereinafter, the kernel function is defined as

$$K^l = K\left(\mathbf{x},\mathbf{x}'|\boldsymbol{\Theta}^l\right) \equiv \theta_1^l \exp\left(\frac{-(\mathbf{x}-\mathbf{x}')^2}{2\theta_2^l}\right)+\sigma_l^2\delta(\mathbf{x}),$$

with $\boldsymbol{\Theta}^l = \{\theta_1^l,\theta_2^l\}$. The parameter θ_1^l is the signal variance whereas θ_2^l is the correlation length. The σ_l^2 term is an unknown homoscedastic gaussian noise. Both $\boldsymbol{\Theta}^l$ and σ_l^2 are inferred recursively from the available data, considering uniform priors. At each level, the training i.e., the maximization of the negative marginal log-likelihood is executed using the Limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS-B) algorithm [20, 21]. The procedure is repeated 10 times from random initialization points.

To evaluate and score the performances of the different modeling approaches, we employ independent data sets and rely on the so-called coefficient of determination

$$R^2 = \left(1-\frac{\int_\Omega\left(T(\mathbf{x})-\mu^L(\mathbf{x})\right)^2}{\int_\Omega\left(T(\mathbf{x})-\frac{1}{\Omega}\int_\Omega T(\mathbf{x})\right)^2}\right), \tag{13}$$

where $\mu^L(\mathbf{x})$ indicates predictions from the top level surrogate, $T(\mathbf{x})$ is the truth i.e., the reality of interest the surrogate tries to mimic, and Ω is the input space. In the best case, the surrogate predicts the reality of interest exactly and $R^2 = 1$. Worse models have $R^2 < 1$. For test cases A and B, to reduce the risk of drawing general conclusions based on peculiar data sets endowed with special informative characteristics, we randomize the training sets by resampling them $K = 100$ times. Then, we score the quality of a multi-fidelity surrogate in terms of averaged R^2 and its standard deviation

$$\mathrm{score}_{\mathrm{AVG}} = \frac{1}{K}\sum_k \mathrm{score}_k \quad \text{and} \quad \mathrm{score}_{\mathrm{STD}} = \frac{1}{K}\sum_k\left(\mathrm{score}_k - \mathrm{score}_{\mathrm{AVG}}\right). \tag{14}$$

Training sets are generated randomly and independently. Whenever analytic expressions are available, training points are drawn uniformly from the input space Ω and synthetic observations are obtained. Whenever real observations are available, the training sets correspond to subsets of the original (larger) database.

A. Analytic function

We assume that the reality of interest $T : \Omega \subset \mathbb{R}^1 \mapsto \mathbb{R}$ one-dimensional

$$T(x) = x\sin(8\pi x)+x, \quad \text{with} \quad x \in [0.0, 1.0], \tag{15}$$

and therefore require $\mathcal{M}^L(x) \doteq T(x)$. We then assume that four models of different fidelity are available to us, to approximate $T(x)$

$$\begin{aligned} M1(\mathbf{x}) &= x, \\ M2(\mathbf{x}) &= 0.7x\sin(8\pi x), \\ M3(\mathbf{x}) &= x\sin(8.2\pi x) + x, \\ M4(\mathbf{x}) &= -5x + 1. \end{aligned} \tag{16}$$

Note that the model list in (16) is random and items are not ordered in any particular manner. Note also that the following approximation holds $T \approx M1 + 1.429M2$.

We consider two arbitrarily hierarchies for building surrogates, namely $S_A = \{M1, M2, M3, M4, T\}$ and $S_B = \{M1, M2, M4, M3, T\}$, noting the swapping of the third and fourth elements. For all $M1$, $M2$, $M3$, $M4$ and T, a finite set of independent observations is available. The cardinality of the training sets is defined according to the following rule $\mathfrak{c}(Y^T) = N$, $\mathfrak{c}(Y^{M4}) = 2 \times N$, $\mathfrak{c}(Y^{M3}) = 3 \times N$, $\mathfrak{c}(Y^{M2}) = 4 \times N$ and $\mathfrak{c}(Y^{M1}) = 5 \times N$. The different cardinality simulates data sets from computations/experiments of varying cost, according to the criterion that cheap data are usually plentiful. We train the surrogates and score them considering a test set of 1000 evenly distributed points in Ω.

We first consider training sets with cardinality $\mathfrak{c}(Y^{M1}) = 30$, $\mathfrak{c}(Y^{M2}) = 24$, $\mathfrak{c}(Y^{M3}) = 18$, $\mathfrak{c}(Y^{M4}) = 12$ and $\mathfrak{c}(Y^T) = 6$. We report the averaged performance analysis in Table 1, for the IR, the SR and the SF approaches. Table 1 also reports the averaged regression coefficients plus/minus their standard deviation. In particular, the IR

Table 1 Test case A: performance comparison of surrogates trained according to a different recursive order.

		$\rho_{\text{AVG}} \pm \rho_{\text{STD}}$				$\text{score}_{\text{AVG}} \pm \text{score}_{\text{STD}}$
		M1	M2	M3	M4	
S_A	IR	0.882 ± 0.185	1.292 ± 0.281	0.110 ± 0.233	0.001 ± 0.035	0.961 ± 0.120
	SR	-	-	-	-0.246 ± 0.138	-0.355 ± 0.665
S_B	IR	0.884 ± 0.186	1.292 ± 0.282	0.110 ± 0.234	0.001 ± 0.035	0.961 ± 0.122
	SR	-	-	0.875 ± 0.309	-	0.552 ± 0.437
	SF	-	-	-	-	-0.264 ± 0.481

formulation leads to performances that, on average, are similar despite the different ordering dictated by S_A and S_B. The $\text{score}_{\text{AVG}}$ reads about 0.96 considering both sequences, also with a similar standard deviation. The averaged regression coefficients indicate a strong correlation of T with $M1$ and $M2$. In particular, the scaling values of the reality of interest ($T \approx M1 + 1.429M2$) are well included within the plus/minus one standard deviation, indicating that the IR method is capable of selecting bits of information that contribute to improving the accuracy of $\mathcal{M}^L$. In addition, the IR provides physics insights about the reality under investigation whose governing laws can be reconstructed as a combination of simpler model units.

The SR formulation reveals the limit associated with establishing an arbitrary hierarchy between models. Since $M4$ is a poor approximation of the truth, its position in the hierarchy is of the utmost importance. In S_A, $M4$ is mistakenly considered at the level $L-1$. The training procedure correctly discards $M4$ by assigning a low correlation coefficient, i.e., on average -0.246. Therefore, the framework is encouraged to learn the target model exclusively from the Y^L data. As a consequence, on average surrogates have a poor score (negative) associated to a quite large standard deviation. Possibly, the large ρ_{STD} indicates that the surrogate score strongly depends on the particular Y_k^L realization. The surrogate behavior changes when we instead consider S_B. Since $M3$ is a quite close approximation of the truth, the SR now learns a high correlation coefficient (about 0.875) between $\mathcal{M}^L$ and $\mathcal{M}^{L-1}$. The averaged score is now positive and significantly higher, whereas the standard deviation is smaller.

Lastly, Table 1 also reports the performance of the classical single-fidelity approach SF. Since there is no recursion, there are no regression coefficients. The average score is about -0.264, slightly better than the SR based on S_A, but significantly worse than the SR based on S_B. This may appear counterintuitive since, despite the ordering, one may expect SR to have an intermediate score between a lower bound (established by the SF) and the perfect fit. However, the SR approach based on S_A does not completely neglect the information from $M4$. In fact, it learns an average scaling

factor of about -1/5, which is consistent with learning the linear contribution to T associated to the $+x$ term. As a consequence, in many data set realizations, the SR framework models the deviation of the $M4$ predictions (scaled by -1/5) from the Y^L data as noise, thus significantly lowering the score. Possibly, this behavior is fostered by the limited cardinality of Y^L considered for this test case. In this regard, it has to be noted that different training strategies e.g., the Leave-One-Out cross-validation, may help mitigating this dependency.

In Fig. 2, we report $\mathcal{M}^{(L)}$ predictions over the whole domain of interest. The figure is relative to one random training set among the K available realizations. Figure 2 develops per row: the first row is relative to S_A whereas the second row

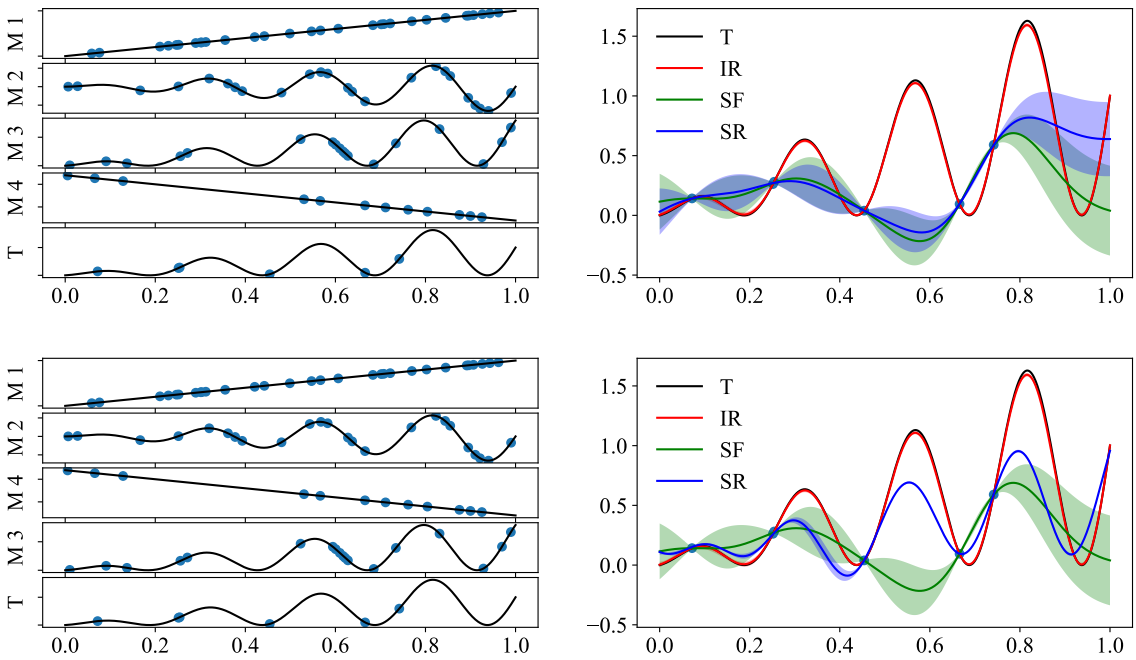


Fig. 2 Test case A, $N = 6$. Top row: S_A. Bottom row: S_B. Left column: analytic curves and the training points. Right column: prediction comparison.

concerns S_B. In each row of Fig. 2, the left hand-side picture plots the five available models M^l, together with the available Y^l observations. The right-hand side compares the prediction $\mu^{(L)}(\mathbf{x})$ from the IR, SF, and SR surrogates, complemented by the envelope of one-standard deviation. Regardless of the model hierarchy, the IR formulation leads to similar predictions. On the other hand, for S_A the SF, and the SR, formulations produce quite similar results. In fact, the low regression coefficient assigned to $M4$ causes SR to almost degenerate into the single-fidelity approach. However, the regression coefficient is not exactly zero, and therefore some differences are noticeable for $x \gtrsim 0.8$. Taking into account S_B, we observe a quite different scenario in which SR predictions have a qualitatively improved behavior, significantly different from SF, due to a strong correlation of $M3$ with T.

We now investigate the hyperparameters of the $\mathcal{GP}$ kernel of the L level in the three models considered. Table 2 reports the averaged kernel hyperparameters as resulting from training on the K data sets. For IR, the reported values confirm that the correction term $\delta^L(x)$ is of little relevance. The variance of the signal (θ_1) and the noise level (σ^L) approach zero on average. At the same time, the correlation length of the corrective $\mathcal{GP}$ is one order of magnitude larger than that of the input domain $\Omega \in [0, 1]$. These considerations hold for S_A and S_B. In contrast, the $\mathcal{GP}$ in SR has a significant dependence on the imposed hierarchy. Kernel hyperparameters vary according to the amount of correction needed i.e., less correction is needed if the sequence is favorable.

We build the same surrogate considering enriched training sets namely, $\mathfrak{c}(Y^{M1}) = 75$, $\mathfrak{c}(Y^{M2}) = 60$, $\mathfrak{c}(Y^{M3}) = 30$, $\mathfrak{c}(Y^{M4}) = 45$ and $\mathfrak{c}(Y^T) = 15$ and report quantitative results in Table 3. Not surprisingly, increasing data leads to general improvements. All approaches return higher $\text{score}_{\text{AVG}}$ and lower $\text{score}_{\text{STD}}$. In particular, the IR consistently scores a value close to 1. The SR approach is also endowed with high performance, but the strong dependency on the prescribed hierarchy is still predominant. Note that SF returns $\text{score}_{\text{AVG}}$ of about 0.803 which is again intermediate w.r.t. SR performances obtained with S_A and S_B.

Again, Fig. 3 reports the $\mathcal{M}^{(L)}$ predictions over the input domain for one random training data set. Despite the

Table 2 Test case A: comparison of kernel hyperparameters averages considering $N = 6$.

		$\theta^L_{1_{\text{AVG}}} \pm \theta^L_{1_{\text{STD}}}$	$\theta^T_{2_{\text{AVG}}} \pm \theta^L_{2_{\text{STD}}}$	$\sigma^L_{\text{AVG}} \pm \sigma^L_{\text{STD}}$
S_A	IR	0.010 ± 0.000	9.903 ± 0.967	0.000 ± 0.000
	SR	0.151 ± 0.254	2.384 ± 4.163	0.061 ± 0.081
S_B	IR	0.010 ± 0.000	9.903 ± 0.967	0.000 ± 0.000
	SR	0.066 ± 0.193	2.296 ± 3.901	0.010 ± 0.019
	SF	0.494 ± 0.0.454	1.417 ± 3.039	0.066 ± 0.104

Table 3 Test case A: performance comparison of surrogates trained according to a different recursive order.

		$\rho_{\text{AVG}} \pm \rho_{\text{STD}}$				$\text{score}_{\text{AVG}} \pm \text{score}_{\text{STD}}$
		M1	M2	M3	M4	
S_A	IR	0.968 ± 0.008	1.426 ± 0.004	0.002 ± 0.003	-0.006 ± 0.002	0.999 ± 0.000
	SR	-	-	-	-0.273 ± 0.111	0.735 ± 0.515
S_B	IR	0.968 ± 0.008	1.426 ± 0.004	0.002 ± 0.003	-0.006 ± 0.002	0.999 ± 0.000
	SR	-	-	0.846 ± 0.058	-	0.928 ± 0.097
	SF	-	-	-	-	0.803 ± 0.243

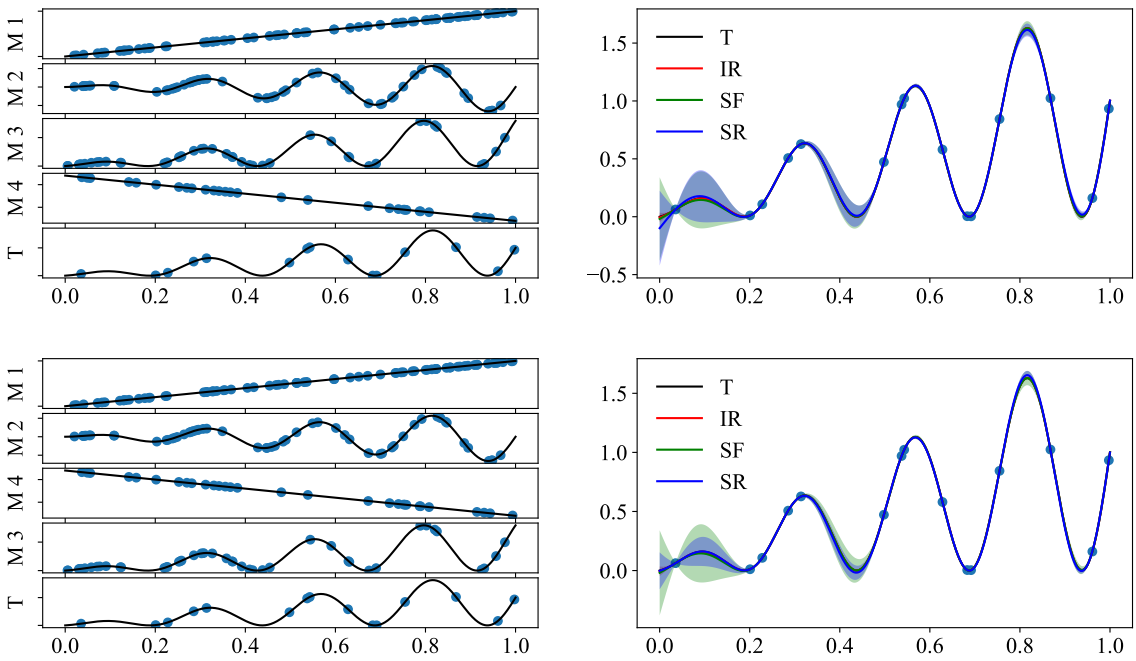


Fig. 3 Test case A, $N = 15$. Top row: S_A. Bottom row: S_B. Left column: analytic curves and the training points. Right column: prediction comparison.

different ordering sequences, qualitatively no particular difference is appreciable. A slight reduction in the prediction variance is notable when comparing SR and SF, for $x = 0.1$ and $x = 0.4$, if S_B is employed.

Lastly, in Fig. 4(a-b) we report the averaged score for S_A and S_B as resulting by increasing the size of the data sets N. For the sequence S_A, the SF approach is consistently endowed with an averaged score higher than the one related to

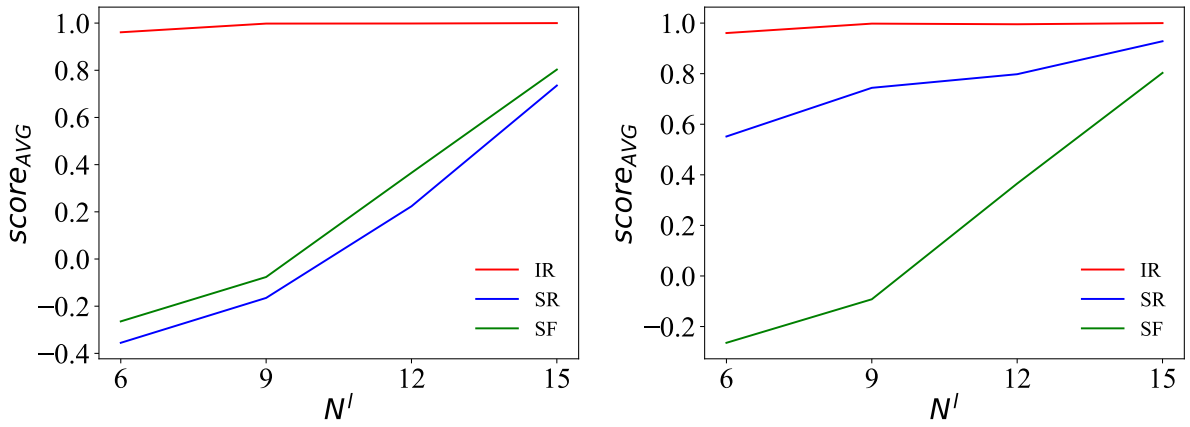


Fig. 4 Test case A. Comparison of the $\text{score}_{\text{AVG}}$ associated to the different approaches w.r.t. an increasing size of the training data sets. a) S_A. b) S_B.

SR. On average, this result is systematic regardless of the dimension of the data sets considered. For the sequence S_B, SR clearly outperforms SF. At the same time, the IR strategy delivers the same performance regardless of the imposed hierarchy of models. Similar considerations apply to the analysis of $\text{score}_{\text{STD}}$, reported in Figure 5(a-b).

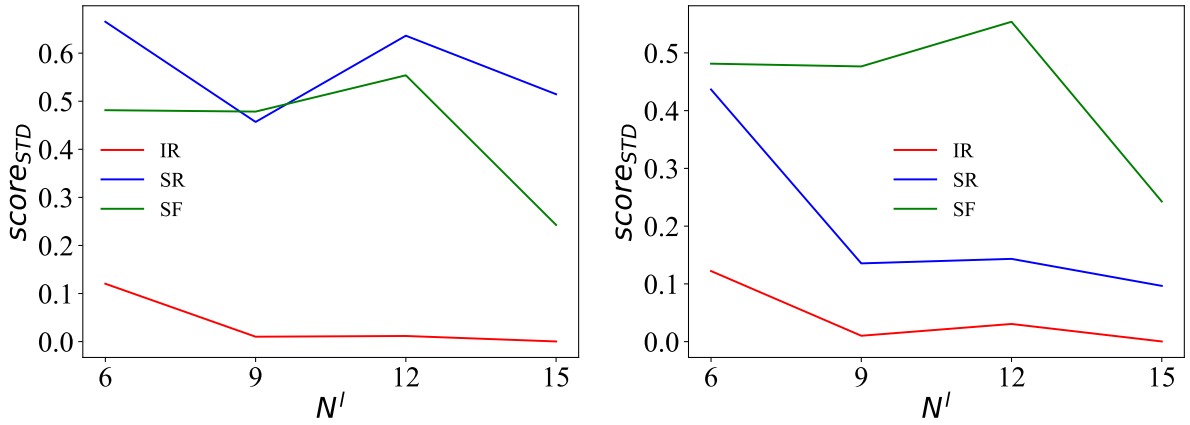


Fig. 5 Test case A. Comparison of the $\text{score}_{\text{STD}}$ associated to the different approaches w.r.t. an increasing size of the training data sets. a) S_A. b) S_B.

Lastly, in Table 4 we report the computational time required to train the K surrogates. The aim is to provide a rough indication of the training time increase due to the different formulations, based on popular choice kernels. The analysis is carried out using a personal laptop and is meant to stress that training the IR surrogate is more demanding than training the SR one. The task is not dramatic, showing a training time ratio of about 1.5. However, both approaches are significantly more expensive than SF. In any case, the training time required to build 100 surrogates is on the order of minutes, that is, tens of seconds to train the single surrogate. This amount of time is acceptable considering the effort typically required to generate data in aerospace applications i.e., run a CFD simulation or execute an experimental campaign. Moreover, in this paper we used a Python implementation of the IR, SR, SF approaches. For instance, a C++ implementation would result in a faster framework, rendering the surrogate training phase even more negligible in

absolute terms.

Table 4 Test case A: computational time required for training the K surrogates, using the three different formulations and for different $\mathfrak{c}(Y^T)$.

	$\mathfrak{c}(Y^T)$			
	6	9	12	15
IR	3148 [s]	3084 [s]	3388 [s]	3694 [s]
SR	1898 [s]	1948 [s]	2350 [s]	2623 [s]
SF	147 [s]	160 [s]	163 [s]	166 [s]

B. Wall-Mounted 2-D Hump

We investigate test n. 83 from the ERCOFTAC Classic Collection Database [22–24]. The application case consists of a subsonic air flow over a wall-mounted Glauert-Goldschmidt type body. The pressure coefficient is measured experimentally. Figure 6 depicts the geometry and reports the operating conditions. Details of the experiment

P_∞	=	101325	[Pa]
T_∞	=	298	[K]
ρ_∞	=	1.185	[Kg/m^3]
μ_∞	=	$18.4 \cdot 10^{-6}$	[Kg/ms]
U_∞	=	34.6	[m/s]
Re_∞	=	$2.23 \cdot 10^6$	[-]
M_∞	=	0.1	[-]
L_{ref}	=	0.42	[m]

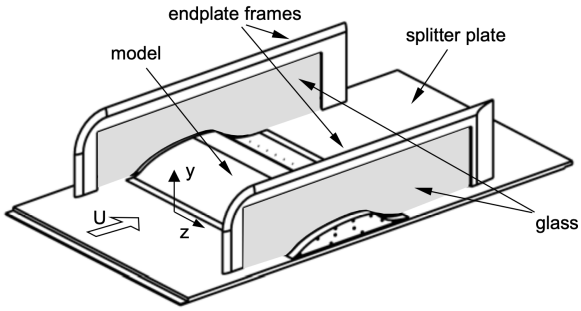


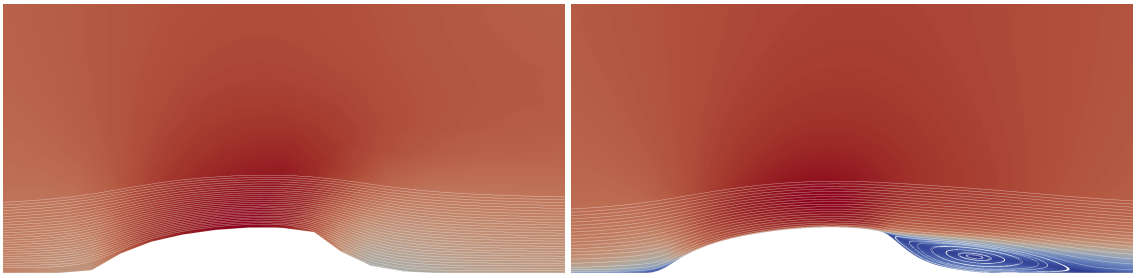
Fig. 6 Test case B. The geometry of the experiment (image from Ref. [22]) alongside with the test conditions.

implementation can be found in the referenced paper. We just recall that the test rig includes a flap deflector and a blowing/suction device to control the air flow. Here, we consider the uncontrolled configuration, namely, fixed flap and no blowing/suction. In general, this experiment is approximated as two-dimensional.

The reality of interest $T : \Omega \subset \mathbb{R}^1 \mapsto \mathbb{R}$ is the experiment itself. The available data consist of observations of the pressure coefficient $C_P(x)$ at discrete stations along the flow center line (x axis). We simulate the experiments using computational models of varying complexity and cost, based either on Euler or Reynolds-Averaged Navier-Stokes (RANS) equations. Specifically, $M1$ solves the Euler equations on a coarse grid (37k elements), whereas $M2$ solves the Euler equations on a fine grid (83k elements). The models $M3$ and $M4$ solve the RANS equations on a grid of, respectively, 100k and 131k elements (differences are limited to the resolution of the boundary layer). The Spalart-Allmaras turbulence model [25] is used.

The open-source SU2 CFD solver [26] was used to generate data from $M1$, $M2$, $M3$, and $M4$. Numerical fluxes are computed by a generalized Approximate Riemann solver of Roe type, with a Monotone Upstream-centered Scheme for Conservation Laws (MUSCL) [27] with the Venkatakrishnan flux limiter. The convergence criterion monitors the density residual, requiring a reduction of 7 orders of magnitude. In any case, simulations are ended after 10000 iterations. Note that this loose arresting criterion is applied to allow the collection of data from simulations that did not converge. This is done on purpose, to maintain an agnostic perspective about the reliability of model predictions.

Figure 7 (a-b) report, respectively, the Mach field predicted by the Euler ($M1$) and RANS ($M3$) models. Streamlines are superimposed to highlight flow features. The $M1$ Euler model, Fig. 7(a), predicts a fully attached flow downstream the bump. On the other hand, a recirculation bubble is clearly visible in the solution from the RANS model $M3$. In



(a) (b)

Fig. 7 Test case B. The Mach field and the flow streamlines computed using the SU2 CFD solver. a) Euler model $M1$. b) RANS model $M3$.

the Euler model, the inviscid flow assumption prevents flow separation. However, CFD relies on the discrete form of the governing equations and a certain amount of numerical dissipation is always present. A conscious exploitation of numerical dissipation can help enforcing a local separation and render the Euler model reasonable even for a separated flow. Typically, the amount of numerical dissipation decreases with the resolution of the grid. Therefore, for the test considered, we expect the $M2$ data to be less reliable than the data from $M1$. Anyway, the widespread bias higher-cost-equals-higher-accuracy may very well mislead a Modeler unfamiliar with CFD. Concerning the RANS model, the extent of the predicted recirculation bubble strongly depends on the fluid transport properties and the turbulence closure. In particular, turbulence models are devised with different applications in mind. Therefore, establishing a fidelity hierarchy between closures is not straightforward.

Figure 8 compares the pressure coefficient along the test section as predicted from the CFD models. The comparison is complemented by experimental data taken from [22–24]. Qualitatively, $M1$ and $M2$ return reasonable predictions for $x \lesssim 0.6$. A slight overestimation of the pressure peak for $x \lesssim 0.2$. The stream expands over the hump, reaching a quite sharp edge in the aft part at $x \sim 0.7$, leading to the prediction of an excessively low C_p. At this edge, the lack of viscosity in the Euler model prevents separation, and the flow follows the wall profile, suffering rapid compression downstream. Since the recirculation bubble does not develop, the compression results in an overshoot of the experimental observation. The deviation is more evident for $M2$, which employs a finer grid. In contrast, the RANS model predictions are closer to the experiments and include the recirculation bubble. However, $M3$ and $M4$ predict a fairly different behavior between $x = 0.8$ and $x = 1.4$. As mentioned, this discrepancy is possibly due either to the diverse grid resolution or to not fully converged simulation.

According to this analysis, the fidelity hierarchy between the available models is not straightforward. We expect the IR formulation to be capable of mitigating the consequences of erroneous assumptions concerning the hierarchy. We then consider the $M1$-$M4$ models and produce $K = 100$ randomized training data sets to investigate the averaged performances. For each realization, the cardinality of the training sets is defined according to $\mathfrak{c}(Y^T) = N$, $\mathfrak{c}(Y^{M4}) = 2\times N$, $\mathfrak{c}(Y^{M3}) = 3 \times N$, $\mathfrak{c}(Y^{M2}) = 4 \times N$ and $\mathfrak{c}(Y^{M1}) = 5 \times N$. In addition, we also randomize the model hierarchy for each realization. We perform the same analysis for $N = \{6, 9, 12, 15, 20, 30\}$.

Table 5 reports the performance assessment. Since the four models predict a qualitatively similar behavior, the surrogates have similar performances. Between IR and SR, the difference in terms of averaged performances is limited and surely not dramatic. Both approaches perform consistently better than the single-fidelity one. Not surprisingly, all methods perform similarly as the dimension of the data sets increases; see Fig. 9(a) and (b). It is interesting to point out that the IR suffers from greater variability in the case of poorly populated data sets. This is due to the larger number of degrees of freedom available, which makes the inference process more difficult and more dependent on the particular data set realization. In SR we are forcing a hierarchy. That is, we are introducing some prior knowledge into the training with the ultimate effect of regularizing the model.

C. Multi-fidelity optimization

We present an application of the proposed IR approach to the optimization of a 2D airfoil. We consider the unconstrained optimization

$$\mathbf{x} = \arg\max_{\mathbf{x}\in\mathbb{R}^d} J(\mathbf{x}), \tag{17}$$

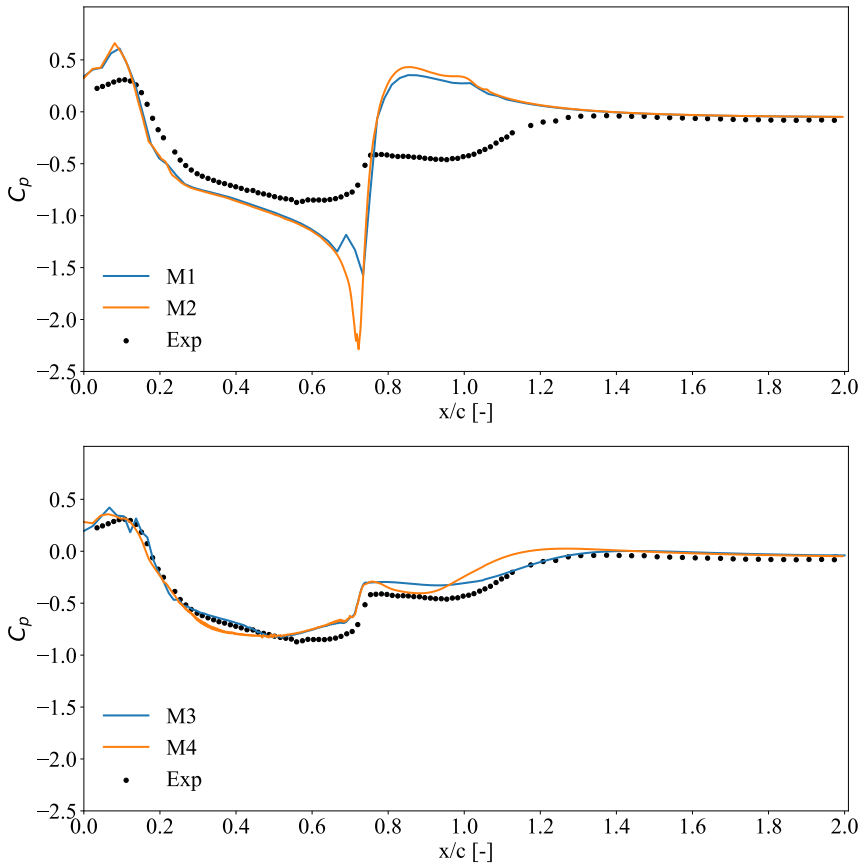


Fig. 8 Test case B. Pressure coefficient CFD predictions and experimental observations.

Table 5 Test case B: performance comparison of surrogates trained according to different sized data sets.

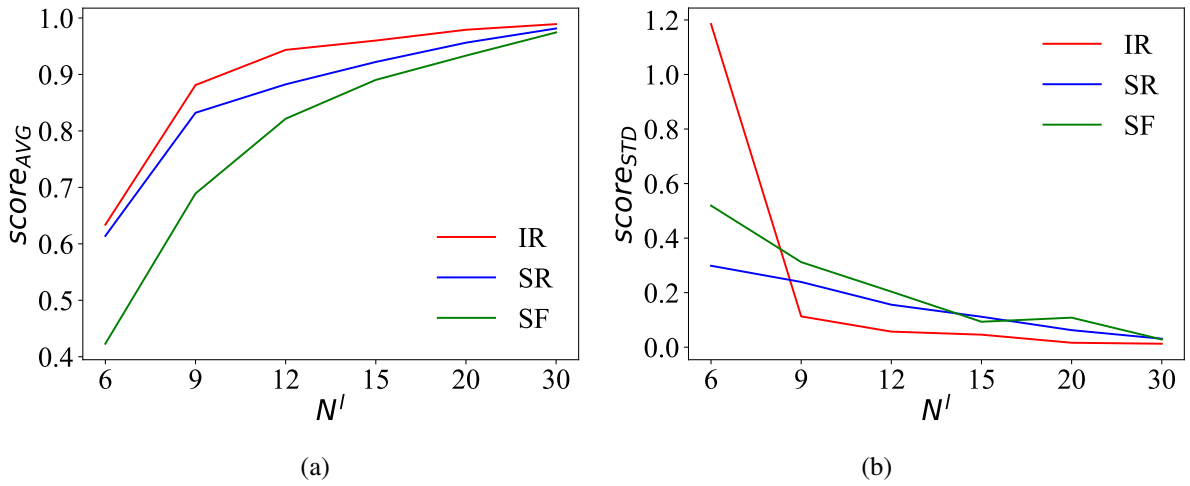
	$\text{score}_{\text{AVG}} \pm \text{score}_{\text{STD}}$					
$\mathfrak{c}(Y^T)$	6	9	12	15	20	30
IR	0.634 ± 1.185	0.881 ± 0.113	0.944 ± 0.057	0.960 ± 0.046	0.979 ± 0.016	0.989 ± 0.013
SR	0.614 ± 0.299	0.832 ± 0.139	0.883 ± 0.156	0.922 ± 0.112	0.956 ± 0.063	0.982 ± 0.031
SF	0.423 ± 0.519	0.689 ± 0.312	0.822 ± 0.204	0.890 ± 0.093	0.933 ± 0.108	0.975 ± 0.029

where the objective is the maximization of the lift coefficient c_l.

The baseline geometry is provided by NASA TMR (Turbulence Modeling Resource) [28, 29], specifically the 2DN00 test case. We consider the NACA 0012 airfoil with unit chord and the following operating conditions: Mach $M_\infty = 0.15$, static temperature $T_\infty = 300$ K and Reynolds number $Re = 6 \cdot 10^6$. More details can be found at [28, 29].

1. Computational models

We take advantage of the XFOIL [30] solver with default viscous correction. The inviscid formulation implemented within XFOIL entails a simple linear-vorticity stream function panel method with an explicit Kutta condition. A Kármán-Tsien compressibility correction is incorporated to improve predictions in the subsonic flow regime. The airfoil is discretized using 170 nodes, all the other settings are left as default. To emulate a multi-fidelity framework, we generate three different computational models based on a different airfoil discretization. Namely, $M1$ (180 elements), $M2$ (260 elements), and $M3$ (320 elements). We will arbitrarily refer to $M3$ as the reality of interest T for this test case.



(a) (b)

Fig. 9 Test case B. Comparison of the performances delivered by the different approaches w.r.t. an increasing size of the training data sets. a) $\text{score}_{\text{AVG}}$. b) $\text{score}_{\text{STD}}$.

To simulate the incorrect specification of one model, we pollute the $M2$ output with Gaussian noise. For c_l we consider $\mathcal{N}(0, 0.2)$, whereas $\mathcal{N}(0, 0.02)$ affects c_d.

2. The optimization algorithm

We implement the Efficient Global Optimization (EGO) technique from the class of Bayesian methods; see [31]. We rely on the *Expected Improvement* (EI) combined with a co-kriging model based on the formulation presented in Sec. II. The parameter ξ controlling the trade-off between exploration and exploration is set to 0.01.

The design space is sampled randomly, considering a uniform probability distribution, and independently for each co-kriging level, to produce the initial training data sets $\{X^{M1}, Y^{M1}\}$, $\{X^{M2}, Y^{M2}\}$, and $\{X^T, Y^T\}$. We thus obtain a multi-fidelity surface approximating $J(\mathbf{x})^T$. Once initialized, the optimization process explores the design space collecting data from the L model only based on the expected improvement. Except for initialization, the $l = 0, \ldots, l-1$ models are never sampled and, therefore, intermediate surrogates are never updated. In other words, we limit the optimization procedure to sampling the L level only, exploiting less accurate data to enrich knowledge in regions poorly populated by high-fidelity information. Future works may be devoted to establishing criteria for selecting the level to be sampled e.g., see [32], exploiting the additional degrees of freedom provided by the IR formulation. Based on predictions at $\mathcal{M}^L$ level, the EI is maximized using the Limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS-B) method [33] to select the most promising design point $\mathbf{x}_n$. Once a new promising design point is identified, a simulation with model $M3$ is automatically performed to generate new data and update the co-kriging surrogate. The optimizer searches the design space until the maximum budget of 100 evaluations is exhausted or until the stopping criterion $\sum_{i=it-5}^{it} |J(\mathbf{x}_i) - J^{BEST}|/|J^{BEST}| < 0.01$ is fulfilled. Note that this latter stopping criterion is defined considering the last five samples of the optimization procedure.

3. Design Parametrization

To better test the IR strategy, we perform the optimization considering a design parametrization of increasing complexity. We rely on a combination of the NACA 4 digit series codification with Class Shape Transformation (CST) [34]. Namely, we first specify the airfoil based on the three NACA X-X-XX parameters (each "X" represents a digit in the codification). Note that the last two digits concur to form a single design parameter. To extend the range of attainable designs, each parameter is cast to a floating-point number in place of an integer. Bounds are enforced to avoid degenerate profiles, respectively, $[0.0, 8.0]$ (percentage of the chord) for the maximum camber, $[4.5, 8.0]$ (tenths of the chord) for the distance of the maximum camber point to the leading edge, and $[1.0, 2.0]$ (percent of the chord) for the maximum airfoil thickness. The resulting NACA 4 digit airfoil is then perturbed using CST to define an arbitrarily complex parametrization based on recursive polynomials. We perturb the upper (U subscript) and lower (L subscript) sides independently, using Chebyshev polynomials Ψ^n of degree n. The perturbation function ϕ is defined along the

nondimensional coordinate $s = x/c$, being c the chord of the airfoil, and it reads

$$y(s)_U = y_U^{\text{NACA}} + \sqrt{s}(1-s)\Psi_U^n(\ \ s)/10, \tag{18}$$

$$y(s)_L = y_L^{\text{NACA}} - \sqrt{s}(1-s)\Psi_L^n(-s)/10. \tag{19}$$

The design parameters consists of the coefficients of the Chebyshev series, bounded within $[0.0, 1.0]$. Furthermore, we consider the Angle of Attack (AoA) $\alpha \in [0.0°, 25.0°]$ as an additional design parameter.

The design vector $\mathbf{x} \in \Omega \subset \mathbb{R}^d$ has dimension d corresponding to α plus the 3 NACA parameters plus the coefficients of the Chebyshev polynomial $n + n$ (n for each side) . Note that, for $n = 0$, the parametrization corresponds to the NACA 4 digit convention plus α. Exemplary (random) design realizations are reported in Fig. 10.

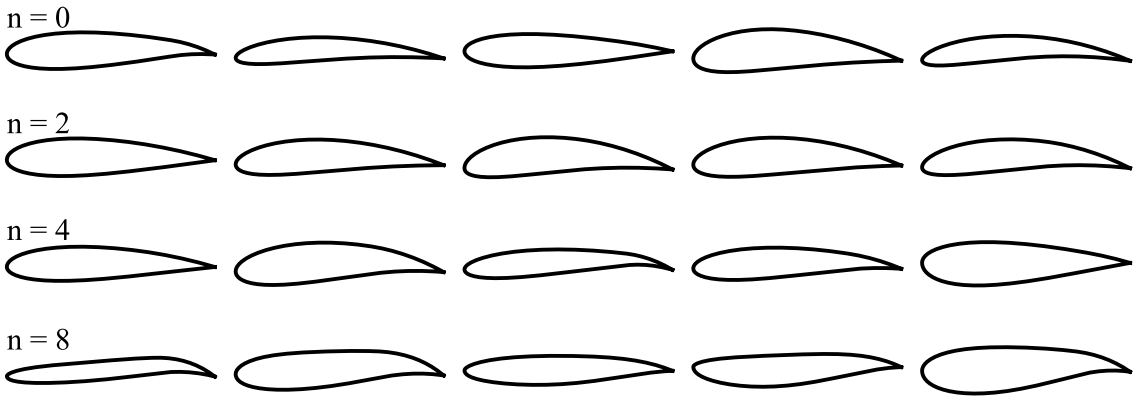


Fig. 10 Test case C. Random realizations of the airfoil, for a varying order of the Chebyshev polynomial.

4. Optimal designs

We present the optimizations performed using the three different metamodels i.e., SF, SR, and IR. According to our parametrization, we consider the following cases: $n = 0 \therefore d = 4$, $n = 2 \therefore d = 6$, $n = 4 \therefore d = 8$, and $n = 8 \therefore d = 12$. All optimization processes are initialized using the very same database of size $\mathfrak{c}(Y_0^T) = 5$, $\mathfrak{c}(Y_0^{M2}) = 40$, and $\mathfrak{c}(Y_0^{M1}) = 500$ respectively.

The optimization history of the four cases is reported in the left column of Fig. 11, whereas the optimal designs are reported on the right side (note that each row corresponds to the different values of n, in an increasing order from top to bottom). Note that the optimization histories report predictions from the L level only, the gray-shaded area identifies points in the initial database. For $n = 0$, all optimizations based on SF, SR, and IR-, converge to the very same solution. The IR methodology requires only 5 iterations to satisfy the convergence criterion, whereas the SR methodology requires more evaluations than the single-fidelity approach. For $n = 2$, IR/SR methods have a very similar behavior and both converge in 5 iterations. The SF approach instead requires a larger number of steps. The three resulting profiles are similar, but not exactly identical. The delivered performance is also slightly different. For $n = 4$, the IR-based optimization achieves convergence in just 5 steps, the other methods requiring a bit more. Notably, the designs provided by the SR and SF methods are suboptimal and deliver lower performance. Although the IR and SR optima present slightly different maximum performances, the resulting airfoil shapes are quite similar. For $n = 8$, the SF approach converges to a very poor solution using more than 20 iterations. Instead, the IR and the SR approaches return the very same optimal solution, with significant differences in the optimization history. The IR converges in 11 steps, whereas the SR requires twice the number of evaluations.

As we are interested in the implications of a (possibly) biased fidelity hierarchy, we investigate how the regression parameters evolve with new samples collected during optimization. We recall here that only the model at level L is sampled. Figure 12 reports the evolution of the regression coefficients for SR (on the left) and IR (on the right) based optimizations. For $n = 0$, we observe differences between SR and IR. Considering the SR case, initially the value of the inferred coefficient ρ_{L-1}^L is close to one. A strong correlation is inferred, but this result is fortuitous since the initial data

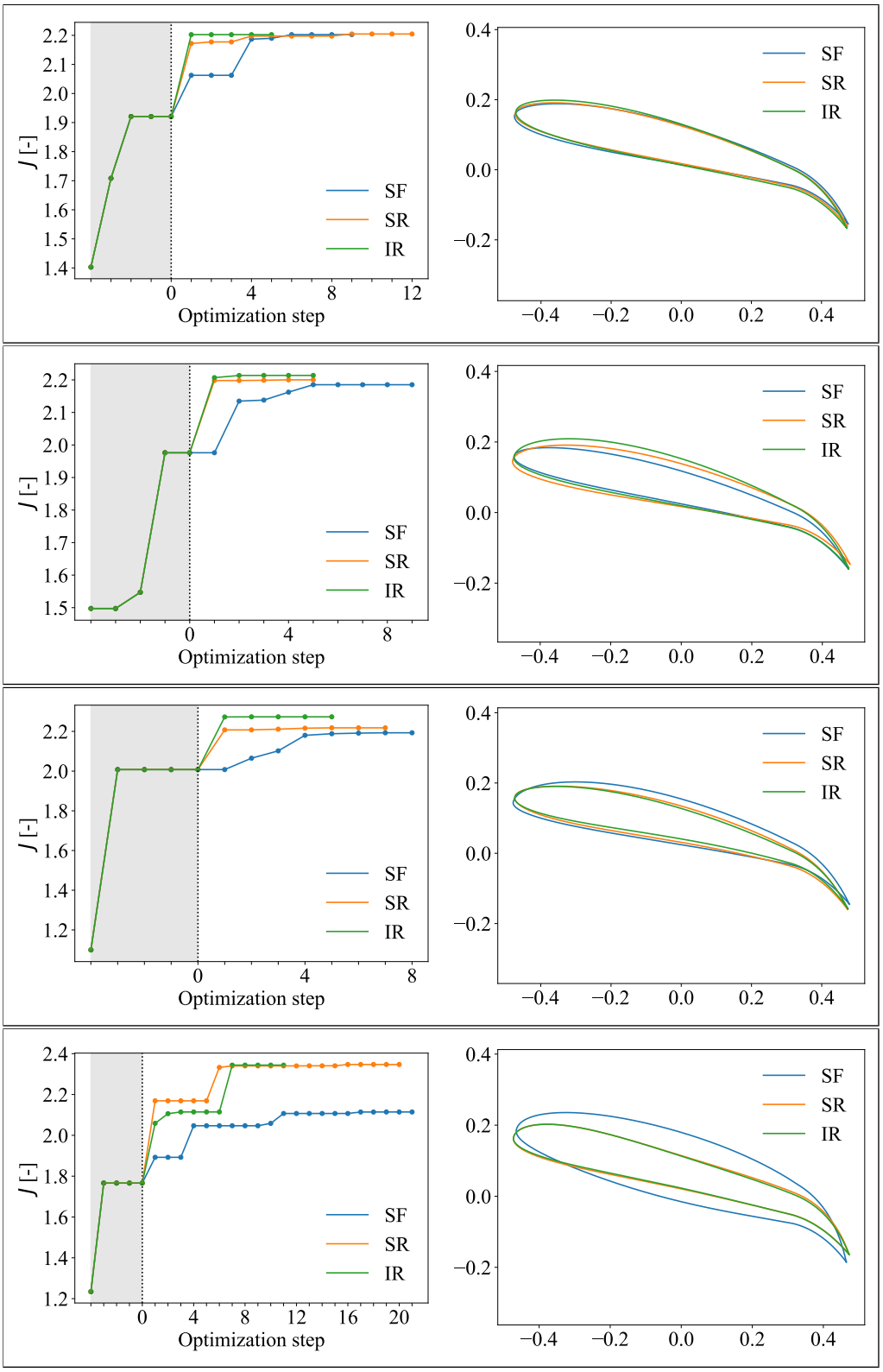


Fig. 11 Test case C, small database. Optimization history and resulting designs. From top to bottom: $n = \{0, 2, 4, 8\}$**.**

sets are random. As the optimizer samples the design space, it realizes a poor correlation between L and $L-1$ due to noise that corrupts the lift coefficient predictions. This fact provides an explanation for why the SR formulation behaves worse than the single-fidelity one. The abrupt decrease of ρ^L_{L-1} corresponds to an abrupt variation in the optimal values for the kernel hyperparameters. At some point, the information brought about by the new samples produces an abrupt decrease in ρ^L_{L-1}. The SR no longer correlates the L data with the $L-1$, and the whole machinery is downgraded to a single-fidelity $\mathcal{GP}$. The IR formulation recognizes a strong correlation between the L and $l=0$ data sets i.e., $\rho^L_0 \approx 1$, instead discarding $l=1$ by learning $\rho^L_1 \approx 0$. In other words, the IR formulation is capable of neglecting the more expensive but less accurate information from the model $M2$. As we consider $n=2$ and $n=4$, we observe a qualitative similar behavior, with only quantitative differences. For the SR, ρ^L_{L-1} is initially close to one and decreases as the design space is sampled. Since the design space is larger, the regression coefficient does not go to zero and instead approaches a non-negligible value of about 0.75 and 0.85. This is possibly due to the amount of data available, which becomes poor in relation to the larger input space. In both cases, the corrective term $\delta_l(\mathbf{x}) \sim \mathcal{GP}$ interprets the discrepancy between $\mathcal{M}^{L-1}$ and $\mathcal{M}^L$ as pure noise which, in principle, is not wrong. However, the noise component hinders the overall optimization process, as it introduces uncertainty in the predictions of the co-kriging model at the level L. The optimizer then needs to acquire more samples to reduce the prediction variance i.e., more exploration is performed. Instead, the IR formulation still infers strong and poor correlations of level L w.r.t. $L-2$ and $L-1$, respectively, discarding the noisy data set. The same conclusions apply to the $n=8$ case. However, the amount of data in relation to the dimensionality of the problem is very poor, and both the SR and IR approaches struggle to produce an interpretable behavior.

To provide a confirmation of the above discussion, we also repeat the very same optimization procedure considering richer initial data sets. Consequently, $\mathfrak{c}(Y^T_0) = 20$, $\mathfrak{c}(Y^{M2}_0) = 100$, and $\mathfrak{c}(Y^{M1}_0) = 600$. The optimization histories are reported in Fig. 13. Again, the IR formulation seems to provide the best solution in the least number of iterations quite consistently. The SR and the SF surrogates lead either to suboptimal solutions or require more search steps to find the optimum, especially as the degrees of freedom for the optimization problem increase. The evolution of the regression coefficient is reported in Fig. 14. Conclusions similar to those drawn for the smaller database. That is, the SR formulation infers a poor correlation between L and $L-1$ since the very beginning, basically acting as a single-fidelity approach. On the contrary, in the IR case the ρ^L_0 and the ρ^L_1 coefficients do not vary with the sampling of new points, showing that the corrupted information brought to the process by $M2$ are discarded from the very beginning.

Overall, the IR is promising and, at least for this exemplary case, is more robust to corrupted data from intermediate levels. It must be stressed that we do not claim a general superiority of the proposed formulation w.r.t. standard optimization multi-fidelity approaches. Our goal is to focus attention on possible biases introduced in the establishment of the fidelity hierarchy between models. A thorough investigation of the implications of using the proposed formulation within an optimization procedure is required.

IV. Conclusions

We propose a modification of the standard autoregressive recursive formulation for building multi-fidelity co-kriging surrogates. The mathematical formulation represents a straightforward extension of the existing approach. The complexity of the implementation into a computer code is limited, whereas the computational cost required to training the surrogate remains acceptable.

The proposed multi-fidelity formulation is capable of overcoming the possible limitations inherent in the process of establishing an (arbitrary) fidelity hierarchy between available models. That is, the formulation mitigates the role played by the Modeler's prior knowledge in selecting the appropriate ordering and is more resilient to subjective biases. The framework recognizes the coherence of predictions of different fidelity w.r.t. the reality of interest, selecting the appropriate models while neglecting inaccurate, or wrong, information. This should not be intended as a framework capable of preventing the Modeler from committing errors in establishing a fidelity hierarchy between models, but rather as a framework capable of mitigating the consequences of such errors.

The proposed formulation can also improve the interpretability of the co-kriging surrogate. The a posteriori analysis of the regression parameters helps identify direct correlations between available models. Instead, in the standard recursive autoregressive formulation the correlation was available only for consecutive models.

Moreover, the proposed formulation can be employed straightforwardly in aerodynamic shape optimization and in design problems in general. In this regard, it may be exploited to support the early design stages Multi-Disciplinary Analysis and Optimization in aerospace applications.

However, future works should focus on further investigating the implications of arbitrary multi-fidelity sequences. Although the proposed approach is capable of mitigating modeling biases, a not quantified relevance of the ordering

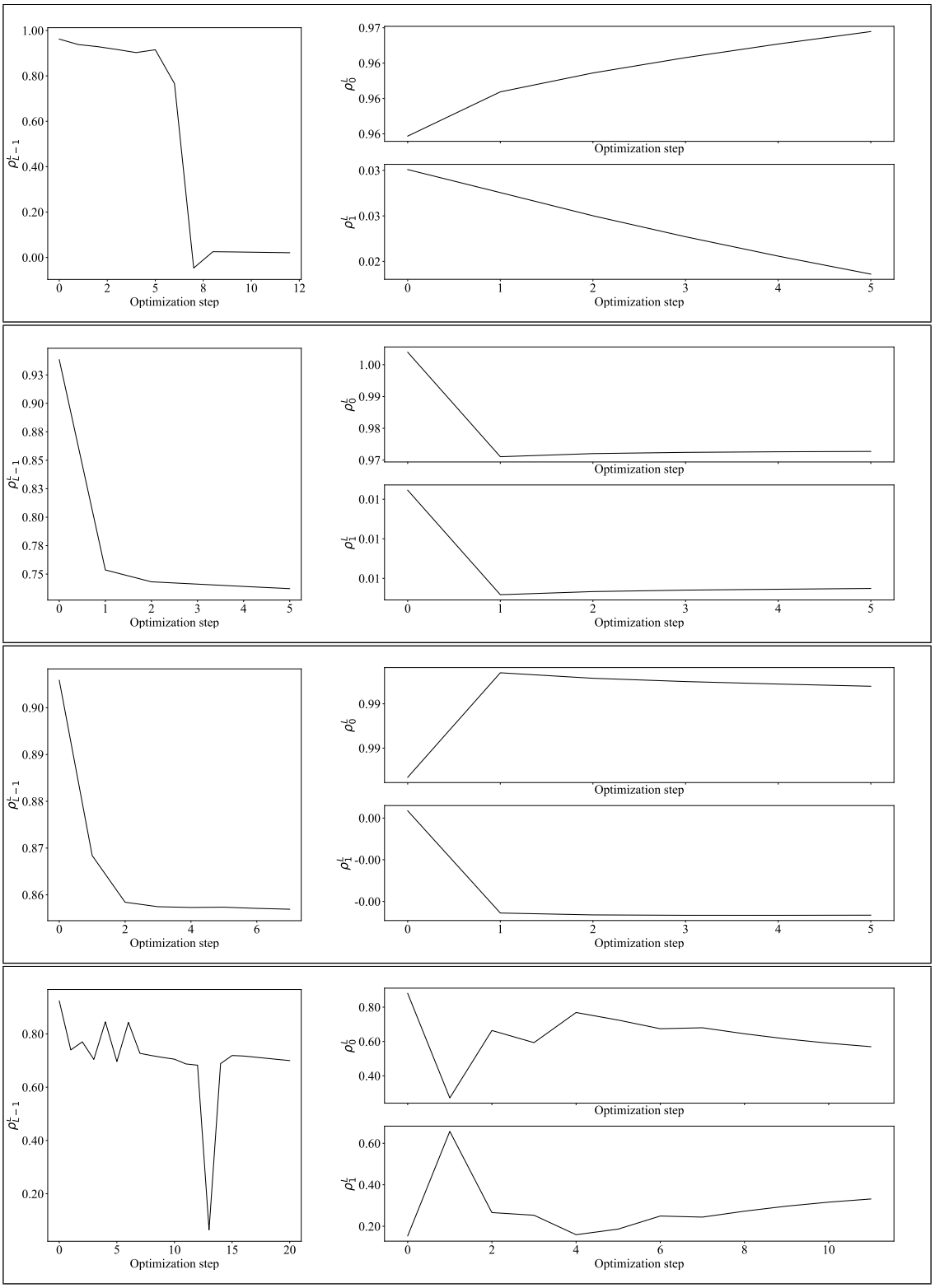


Fig. 12 Test case C, small database. Regression parameters for level L SR (left) and IR (right). From top to bottom: $n = \{0, 2, 4, 8\}$.

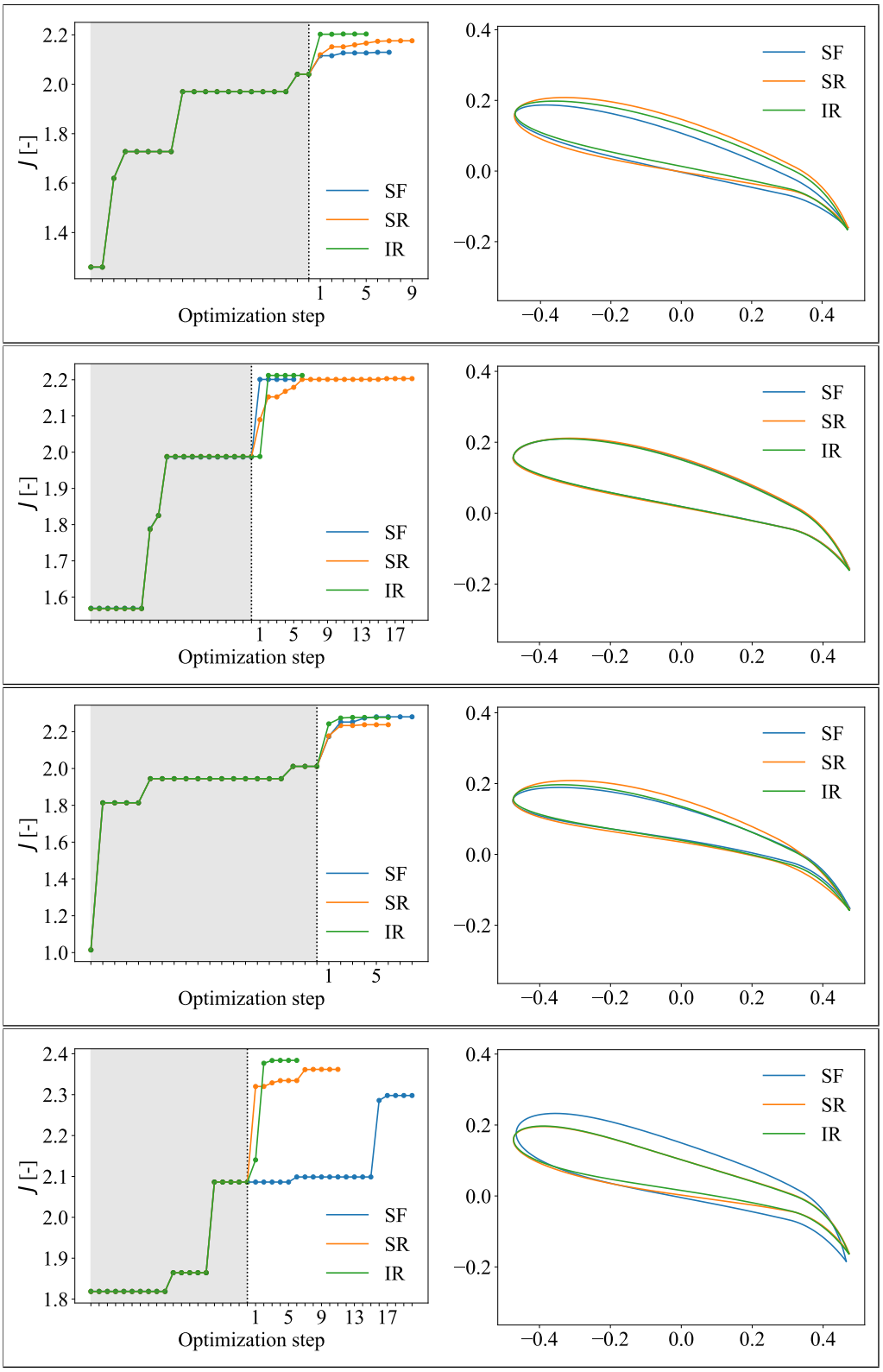


Fig. 13 Test case C, large database. Optimization history and resulting designs. From top to bottom: $n = \{0, 2, 4, 8\}$**.**

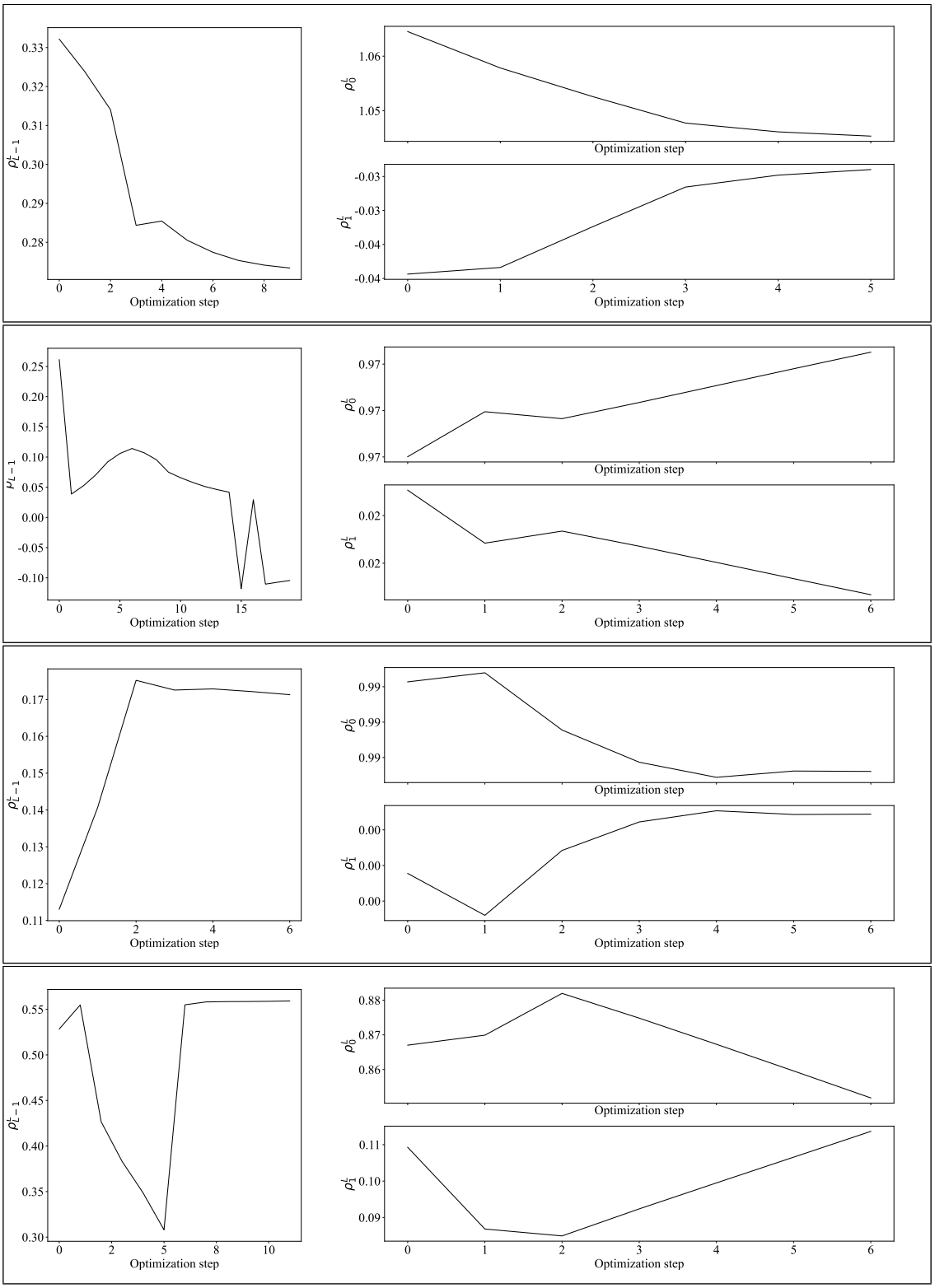


Fig. 14 Test case C, large database. Regression parameters for level L SR (left) and IR (right). From top to bottom: $n = \{0, 2, 4, 8\}$.

sequence is still retained and deserves thorough attention. This lies in the recursive structure of the co-kriging model. At each level, the training fits the corresponding data set Y^l. Changing the hierarchy may change the quality of the surrogate obtained at each level and, since the formulation is autoregressive recursive, also the quality of the surrogate at the level L. Surely, a possible future development is introducing an automatic ordering rule for the available models, to achieve an optimal sequence w.r.t. the specific data set realization. Another key aspect is to consider the possible spatial dependency of the regression coefficients to shed light on the implications. Eventually, database infill strategies may be developed based on the proposed formulation.

Funding Sources

G. Gori would like to acknowledge that this work was partially funded by the European Union (Project 101059320 - UN-BIASED). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the European Research Executive Agency can be held responsible for them.

References

[1] Giselle Fernández-Godino, M., Park, C., Kim, N. H., and Haftka, R. T., "Issues in Deciding Whether to Use Multifidelity Surrogates," *AIAA Journal*, Vol. 57, No. 5, 2019, pp. 2039–2054. https://doi.org/10.2514/1.J057750.

[2] Peherstorfer, B., Willcox, K., and Gunzburger, M., "Survey of Multifidelity Methods in Uncertainty Propagation, Inference, and Optimization," *SIAM Review*, Vol. 60, No. 3, 2018, pp. 550–591. https://doi.org/10.1137/16M1082469.

[3] Mukhopadhaya, J., Whitehead, B. T., Quindlen, J. F., Alonso, J. J., and Cary, A. W., "Multi-fidelity Modeling of Probabilistic Aerodynamic Databases for Use in Aerospace Engineering," *International Journal for Uncertainty Quantification*, Vol. 10, No. 5, 2020, pp. 425–447. https://doi.org/10.1615/Int.J.UncertaintyQuantification.2020032841.

[4] Torenbeek, E., *Advanced Aircraft Design: Conceptual Design, Analysis and Optimization of Subsonic Civil Airplanes*, John Wiley and Sons, Ltd, 2013. https://doi.org/10.1002/9781118568101.

[5] Bertch, W. J., Babula, M., Green, L., Hale, J., Moser, G., Steele, M., Sylvester, A., and Woods, J., "NASA standard for models and simulations : credibility assessment scale," , 2008. https://doi.org/2014/41622, URL https://hdl.handle.net/2014/41622.

[6] Kieweg, S., and Witkowski, W. R., "Experimental Credibility and its Role in Model Validation and Decision Making," 2018. https://doi.org/10.1007/978-3-319-74793-4_5.

[7] Vassberg, J., Dehaan, M., Rivers, M., and Wahls, R., *Development of a Common Research Model for Applied CFD Validation Studies*, 2008. https://doi.org/10.2514/6.2008-6919.

[8] Levy, D. W., Laflin, K. R., Tinoco, E. N., Vassberg, J. C., Mani, M., Rider, B., Rumsey, C. L., Wahls, R. A., Morrison, J. H., Brodersen, O. P., Crippa, S., Mavriplis, D. J., and Murayama, M., "Summary of Data from the Fifth Computational Fluid Dynamics Drag Prediction Workshop," *Journal of Aircraft*, Vol. 51, No. 4, 2014, pp. 1194–1213. https://doi.org/10.2514/1.C032389.

[9] Montáns, F. J., Chinesta, F., Gómez-Bombarelli, R., and Kutz, J. N., "Data-driven modeling and learning in science and engineering," *Comptes Rendus Mécanique*, Vol. 347, No. 11, 2019, pp. 845–855. https://doi.org/https://doi.org/10.1016/j.crme.2019.11.009, data-Based Engineering Science and Technology.

[10] Rasmussen, C. E., and Williams, C. K. I., *Gaussian Processes for Machine Learning*, The MIT Press, 2005. https://doi.org/10.7551/mitpress/3206.001.0001.

[11] Stein, M. L., *Interpolation of spatial data*, Springer Series in Statistics, 1999. https://doi.org/https://doi.org/10.1007/978-1-4612-1494-6.

[12] Jones, D., Schonlau, M., and Welch, W., "Efficient Global Optimization of Expensive Black-Box Functions," *Journal of Global Optimization*, Vol. 13, No. 4, 1998, pp. 455–492. https://doi.org/10.1023/A:1008306431147.

[13] Kennedy, M. C., and O'Hagan, A., "Predicting the Output from a Complex Computer Code When Fast Approximations Are Available," *Biometrika*, Vol. 87, No. 1, 2000, pp. 1–13. https://doi.org/https://www.jstor.org/stable/2673557.

[14] Kennedy, M. C., and O'Hagan, A., "Bayesian calibration of computer models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Vol. 63, No. 3, 2001, pp. 425–464. https://doi.org/https://doi.org/10.1111/1467-9868.00294.

[15] Higdon, D., Kennedy, M., Cavendish, J. C., Cafeo, J. A., and Ryne, R. D., "Combining Field Data and Computer Simulations for Calibration and Prediction," *SIAM Journal on Scientific Computing*, Vol. 26, No. 2, 2004, pp. 448–466. https://doi.org/10.1137/S1064827503426693.

[16] Reese, C. S., Wilson, A. G., Hamada, M., Martz, H. F., and Ryan, K. J., "Integrated Analysis of Computer and Physical Experiments," *Technometrics*, Vol. 46, No. 2, 2004, pp. 153–164. URL http://www.jstor.org/stable/25470801.

[17] Qian, P. Z. G., and Wu, C. F. J., "Bayesian Hierarchical Modeling for Integrating Low-Accuracy and High-Accuracy Experiments," *Technometrics*, Vol. 50, No. 2, 2008, pp. 192–204. URL http://www.jstor.org/stable/25471459.

[18] Le Gratiet, L., "Multi-fidelity Gaussian process regression for computer experiments," Theses, Université Paris-Diderot - Paris VII, Oct. 2013. URL https://tel.archives-ouvertes.fr/tel-00866770.

[19] Le Gratiet, L., and Garnier, J., "Recursive Co-Kriging Model for Design of Computer Experiments with Multiple Levels of Fidelity," *International Journal for Uncertainty Quantification*, Vol. 4, No. 5, 2014, pp. 365–386. https://doi.org/10.1615/Int.J.UncertaintyQuantification.2014006914.

[20] Malouf, R., "A Comparison of Algorithms for Maximum Entropy Parameter Estimation," *Proceedings of the 6th Conference on Natural Language Learning - Volume 20*, Association for Computational Linguistics, USA, 2002, p. 1–7. https://doi.org/10.3115/1118853.1118871.

[21] Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C., "A Limited Memory Algorithm for Bound Constrained Optimization," *SIAM Journal on Scientific Computing*, Vol. 16, No. 5, 1995, pp. 1190–1208. https://doi.org/10.1137/0916069.

[22] Greenblatt, D., Paschal, K., Yao, C., Harris, J., Schaeffler, N., and Washburn, A., *A Separation Control CFD Validation Test Case. Part 1: Baseline & Steady Suction*, 2004. https://doi.org/10.2514/6.2004-2220.

[23] Greenblatt, D., Paschal, K., Yao, C.-S., and Harris, J., *A Separation Control CFD Validation Test Case Part 2. Zero-Efflux Oscillatory Blowing*, 2005. https://doi.org/10.2514/6.2005-485.

[24] Naughton, J., Viken, S., and Greenblatt, D., *Wall Shear Stress Measurements on the NASA Hump Model for CFD Validation*, 2004. https://doi.org/10.2514/6.2004-2607.

[25] Spalart, P., and Allmaras, S., *A one-equation turbulence model for aerodynamic flows*, 1992. https://doi.org/10.2514/6.1992-439.

[26] Palacios, F., Alonso, J., Duraisamy, K., Colonno, M., Hicken, J., Aranake, A., Campos, A., Copeland, S., Economon, T., Lonkar, A., Lukaczyk, T., and Taylor, T., "Stanford University Unstructured (SU²): An open-source integrated computational environment for multi-physics simulation and design," *AIAA Paper 2013-0287*, 2013. https://doi.org/10.2514/6.2013-287.

[27] van Leer, B., "Towards the ultimate conservative difference scheme. V. A second-order sequel to Godunov's method," *Journal of Computational Physics*, Vol. 32, No. 1, 1979, pp. 101–136. https://doi.org/https://doi.org/10.1016/0021-9991(79)90145-1.

[28] "NASA Langley Turbulence Modeling Resource website," , 2023. URL http://turbmodels.larc.nasa.gov.

[29] Jespersen, D., Pulliam, T., and Childs, M., "OVERFLOW: Turbulence Modeling Resource Validation Results," Tech. Rep. NASA-2016-01, NASA Ames Research Center, Moffett Field, CA, 2010.

[30] Drela, M., "XFOIL: An Analysis and Design System for Low Reynolds Number Airfoils," *Low Reynolds Number Aerodynamics*, edited by T. J. Mueller, Springer Berlin Heidelberg, Berlin, Heidelberg, 1989, pp. 1–12. https://doi.org/https://doi.org/10.1007/978-3-642-84010-4_1.

[31] Brochu, E., Cora, V. M., and de Freitas, N., "A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning," Tech. rep., 2010. https://doi.org/https://doi.org/10.48550/arXiv.1012.2599.

[32] Sacher, M., Maître, O. L., Duvigneau, R., Hauville, F., Durand, M., and Lothodé, C., "A Non-Nested Infilling Strategy for Multifidelity Based Efficient Global Optimization," *International Journal for Uncertainty Quantification*, Vol. 11, No. 1, 2021, pp. 1–30. https://doi.org/10.1615/Int.J.UncertaintyQuantification.2020032982.

[33] Virtanen, P., Gommers, R., and Oliphant, T. e. a., "SciPy 1.0–Fundamental Algorithms for Scientific Computing in Python," *Nat Methods*, Vol. 17, 2020, p. 261–272. https://doi.org/https://doi.org/10.1038/s41592-019-0686-2.

[34] Kulfan, B., and Bussoletti, J., *"Fundamental" Parameteric Geometry Representations for Aircraft Component Shapes*, 2006. https://doi.org/10.2514/6.2006-6948.