Computers & Fluids 84 (2013) 327–350

Contents lists available at SciVerse ScienceDirect

Computers & Fluids

journal homepage: www.elsevier.com/locate/compfluid

# Proper Orthogonal Decomposition, surrogate modelling and evolutionary optimization in aerodynamic design

Emiliano Iuliano [*,1], Domenico Quagliarella [1]

*CIRA, Italian Aerospace Research Center, Via Maiorise, 81043 Capua, Caserta, Italy*

ARTICLE INFO

*Article history:*
Received 3 August 2012
Received in revised form 7 June 2013
Accepted 11 June 2013
Available online 27 June 2013

*Keywords:*
Proper Orthogonal Decomposition
Reduced order modelling
Surrogate modelling
Aerodynamic design
Optimization
Design of experiments

ABSTRACT

A computational methodology is proposed for CFD-based aerodynamic design to exploit a reduced order model as surrogate evaluator. The model is based on the Proper Orthogonal Decomposition of an ensemble of CFD solutions. A zonal approach is presented to better solve the shock wave region and improve the surrogate prediction in transonic flow. Model validation and in-fill criteria are shown as valid tools to examine the accuracy of the surrogate and, therefore, to feed the model back with "intelligent" information. The reduced order model is integrated in an evolutionary optimization framework and used as fitness evaluator to improve the aerodynamic performances of a two-dimensional airfoil. Finally, the performances of the surrogate-based shape optimization are compared to the efficiency of a meta-model assisted optimization and to the accuracy of a plain optimization, where, instead, each aerodynamic evaluation is performed with the high-fidelity model.

© 2013 Elsevier Ltd. All rights reserved.

## 1. Introduction

The intrinsic design complexity of modern aircraft relies more and more on the development and assessment of new theoretical methodologies capable of reducing or even replacing the experimental load. Moreover, theoretical methods are often used to further explore the trade-offs and alternatives when a decision about the design path (e.g., a down selection) has to be faced. Highest-fidelity at lowest possible cost is essential in aircraft design and analysis. High-fidelity is related to the capability of the theoretical method to reproduce "real-life" phenomena with a significant degree of accuracy (e.g., flow transition and separation, aerodynamic stall prediction). The growth in technical and commercial requirements call for the application of advanced theoretical/numerical methods even in the early stages of the design process. Here, as the design space is still huge, semi-empirical tools and rules, derived from classical configurations data, have been traditionally applied thanks to their computational efficiency. However, they exhibit serious validity and flexibility limits when designing novel concepts due to the lack of accuracy on unconventional and not yet assessed configurations. In the recent past, a strong effort has been done to introduce potentially highly accurate design analysis methods both in geometric representation and physical modelling, but the main drawback is that they are computationally expensive. For example, the solution of the Navier–Stokes equations around complex aerodynamic configurations requires a huge amount of computational resources even on modern state-of-art computing platforms. This turns out to be an even bigger issue when hundreds or thousands of analysis evaluations, like in parametric or optimization studies, have to be performed.

In order to speed up the computation while keeping a high level of fidelity, the scientific community is increasingly focusing on surrogate methodologies like meta-models or reduced order models. They can provide a compact, accurate and computationally efficient representation of the aircraft design performance index, usually referred to as objective or fitness function in an optimization context. Both data-fits and reduced order models have their own mathematical machinery and related parameters. Finding the set of parameters which best fit the model to the available data is usually known as the *training* phase. The training data set is usually obtained by sampling the design space (DoE, Design of Experiments) and performing expensive high-fidelity computations on the selected points. Depending on the adopted surrogate technique, design objectives and constraints or vector/scalar fields of interests are used to feed the surrogate model. The strategy to properly and optimally choose the DoE sampling data set is of paramount importance to achieve a satisfactory accuracy of the surrogate model. Of course, this strategy is heavily dependent on the type of the surrogate model and on its target. Indeed, two basic features are associated to a given DoE sampling and related to

[*] Corresponding author. Tel.: +39 0823623913.
*E-mail addresses:* e.iuliano@cira.it (E. Iuliano), d.quagliarella@cira.it (D. Quagliarella).
[1] Fluid Dynamics Department, Air Vehicles Division, Italy.

0045-7930/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compfluid.2013.06.007

the building of a surrogate: the exploratory and predictive capabilities. The first helps to unveil promising regions of the design space where the global/local minima might reside; the second provides for improving the model accuracy in the vicinity of possible minima. For the sake of clarity, an evenly-spaced dense sampling could be classified as purely exploratory, while a properly clustered sampling could fall in the highly predictive class. As a consequence, a trade-off exist between the exploration of the design solutions and the improvement of the surrogate accuracy near predicted minima. Hereinafter, it will be shown that a proper balance between these two concepts is highly effective especially when large design spaces and multi-modal objective functions have to be faced. The aerodynamic shape optimization problems, illustrated later on, will provide a significant evidence of the aforementioned concepts.

## 2. Literature review

Surrogate-based optimization (SBO) has been introduced to tackle the number of function evaluations in many engineering optimization problems. This represents a special challenge in the field of global optimization as state-of-the-art methods often requires more function evaluations than can be comfortably affordable. A well-established approach consists in fitting some kind of response functions to basic data obtained by evaluating the objectives and constraints at a few points. The resulting surfaces, affordable at low cost, can provide fast answers in terms of trade-off analysis and optimization as well as just an intuitive sketch behaviour by means of simple visualization. The basic process consists of the following phases:

- Design space sampling – once the design variables have been chosen, a sampling plan is defined and some initial sample designs are analyzed with an accurate solver.
- Surrogate model construction – a surrogate model type is selected and used to build a meta-model of the underlying problem.
- Model validation – the model is checked according to some statistical metrics and, if not enough accurate, a search is carried out using the model to identify new design points for analysis.
- Model updating – the new results are added to those already available and a new meta-model is built (repeating the last three steps.
- Optimization – the refined surrogate is used to provide objective/constraint functions.

As SBO covers so many topics, the literature on the subject is huge. A plenty of ideas have been proposed in the last twenty years, classified for design space dimensions, surrogate methods, search algorithms, updating algorithms, application areas. Hence, an exhaustive survey of all the possible ideas for each topic and all the possible combination of them would go beyond the scope of the present paper. Here we take a more in depth look at the optimization assisted with surrogates and the exploitation of Proper Orthogonal Decomposition as a technique to derive reduced order approximations.

### *2.1. Surrogate-based optimization*

Jones et al. [1], among the first, proposed a response surface methodology based on modelling the objective and constraint functions with stochastic processes (Kriging). The so-called Design and Analysis of Computed Experiments (DACE) stochastic process model was built as a sum of regression terms and normally distributed error terms. The main conceptual assumption was that the lack of fit due only to the regression terms can be considered as entirely due to modelling error, not measurement error or noise, because the training data are derived from a deterministic simulation. Hence, by assuming that the errors at different points in the design space are not independent and the correlation between them is related to the distance between the computed points, the authors came up with an interpolating surrogate model able to provide not only the prediction of objectives/constraints at a desired sample point, but also an estimation of the approximation error. After the construction of such a surrogate model, this last powerful property is exploited to build an Efficient Global Optimization (EGO), which can be considered as the progenitor of a long and still in development chain of SBO methods. Indeed, they found a proper balancing between the need to exploit the approximation surface (by sampling where it is minimized) with the need to improve the approximation (by sampling where prediction error may be high). This was done by introducing the Expected Improvement (EI) concept, already proposed by Schonlau et al. [2], that is an auxiliary function to be maximized instead of the original objective. Sampling at a point where this auxiliary function is maximized improves both the local (exploitation) and global (exploration) search.

In a further work, Jones [3] proposed a taxonomy of global SBO methods. Seven methods were identified and classified on whether they were interpolating (cubic splines, thin-plate splines, multiquadrics, kriging) or not (quadratic polynomials), whether they provided statistical information (kriging) or not (splines) and whether the method for selecting search points (updating the model by adding new sample points) was two-stage (probability/expected improvement) or one-stage (goal-seeking, credibility function). As discussed by Alexandrov et al. [4], a comfortable way to improve such methods and to ensure convergence to a critical point is to force the gradient of the surface to match the gradient of the function whenever the search stagnates. Alexandrov showed that this additional condition is not sufficient: by using also a trust region approach, a locally convergent method was developed. In this context, it is worth to mention the work from Booker et al. [5], who showed how response surfaces can be used to accelerate a derivative-free method of local optimization.

An overview of SBO techniques was presented also by Queipo et al. [6] and Simpson et al. [7]. They covered some of the most popular methods in design space sampling, surrogate model construction, model selection and validation, sensitivity analysis, and surrogate-based optimization. Forrester and Keane [8] recently proposed a review of some advances in surrogate-based optimization. An important lesson learned is that only calling the true function can confirm the results coming from the surrogate model. Indeed, the path towards the global optimum is made of iterative steps where, even exploiting some surrogate model, only the best results coming from the true function evaluations are taken as optimal or sub-optimal design. The true function evaluation has to be also invoked to improve the surrogate model. With the term “in-fill criteria” it is usually meant some principles which allow to intelligently place new points (in-fill points) at which the true function should be called. The selection of infill points, also referred to as adaptive sampling or model updating, represent the core of a surrogate-based optimization method and helps to improve the surrogate prediction in promising areas of the objective space.

The right choice of the number of points which the initial sampling plan would comprise and the ratio between initial/in-fill points has been the focus of several recent studies. However, it must be underlined that no universal rules exist, as each choice should be carefully evaluated according to the design problem (e.g., number of variables, computational budget, type of surrogate). Forrester and Keane assumed that there is a maximum budget of function evaluations, so as to define the number of points as a fraction of this budget. They identified three main cases according to the aim of the surrogate construction: pure visualization and design space comprehension, model exploitation and balanced exploration/exploitation. In the first case, the sampling plan should contain all of budgeted points as no further refinement of the model is foreseen. In the exploitation case, the surrogate can be used as the basis for an in-fill criterion, that means some computational budget must be saved for adding points to improve the model. They also proposed to reserve less than one half points to the exploitation phase as a small amount of surrogate enhancement is possible during the in-fill process. In the third case, that is two-stage balanced exploitation/exploration in-fill criterion, as also shown by Sóbester et al. [9], they suggested to employ one third of the points in the initial sample while saving the remaining for the in-fill stage. Indeed, such balanced methods rely less on the initial prediction and so fewer points are required. Concerning the choice of the surrogate, the authors observed that it should depend on the problem size, i.e. the dimensionality of the design space, the expected complexity, the cost of the true analyses and the in-fill strategy to be adopted.

In general, a global search would require a surrogate model able to provide an estimate of the error it commits when predicting. Thus, the authors suggested to use Gaussian process based methods like Kriging, although citing the work of Gutmann [10] as an example of one-stage goal seeking approach employing various Radial Basis Functions. Finally, some interesting suitable convergence criterion to stop the surrogate in-fill process were proposed. However, the authors also observed that discussing on convergence criterion may be interesting and fruitful, but "in many real engineering problems we actually stop when we run out of available time or resources, dictated by design cycle scheduling or costs". This is what typically happens in aerodynamic design, where the high-dimensionality of the design space and expensive computer simulations often do not allow to reach the global optimum of the design problem but suggest to consider even a premature, sub-optimal solution as a converged point.

Goel et al. [11] proposed a weighted average of an ensemble of surrogates. For example, a better model can be achieved by combining Kriging, which might accurately predict the non-linear aspects of a function, and polynomials to better capture the regression trends.

### 2.2. POD modelling

Besides interpolating response functions, Proper Orthogonal Decomposition can be also exploited as a model reduction technique. A wide and comprehensive review of POD-based applications can be found in Mifsud [12]. The technique was proposed by several authors at different times, in different fields and under a variety of names [13]. The Proper Orthogonal Decomposition (also known as Principal Component Analysis) and the Singular Value Decomposition are generally treated as the same thing, however it must be underlined that the second technique is just a method of solution of the orthogonal basis, hence they are not strictly the same. Restricting the overview to the fluid dynamics applications, the method has been originally used in stochastic turbulence problems [14], where the POD eigenfunctions were related to the characteristic eddies of the turbulence field. The method has also been used in steady aerodynamic analysis such as the design of inviscid aerofoils by LeGresley and Alonso [15] and parametric studies by Epureanu [16] and Bui-Thank et al. [17,18].

Concerning reduced order models derivation, Sirovich [19] introduced the method of snapshots as a way for efficiently determining the POD basis functions or modes for large problems. Holmes et al. [20] extended and applied the method by Sirovich to fluid dynamics problems. Combined with CFD and unsteady aerodynamics, the method of snapshots has been widely used as in Dowell et al. [21] and Hall et al. [22]. The general approach is to first compute a set of instantaneous flow solutions or snapshots and then apply the POD process to extract an optimal set of basis functions, where optimal means that the error between the originally computed and the reconstructed data is minimized. Once built the optimal orthogonal basis, reduced-order models can be derived by projecting the model onto the reduced space spanned by the POD modes. Therefore, the original problem, formulated in terms of non-linear partial differential equations as Navier–Stokes model, can be converted into a small system of ordinary differential equations which can be solved efficiently.

Everson and Sirovich [23] presented a variation of the basic POD method to handle incomplete data sets (also known as "gappy POD"). The method relies on a least square approximation, built on known data, to reconstruct an incomplete snapshot. Indeed, once computed the POD modes from the known data, an incomplete data vector can be reconstructed accurately by imposing the optimal conditions and solving the resulting linear system of equations. Another approach is also proposed, i.e. when the snapshots themselves are damaged or incomplete. In this case, an iterative method can be used to derive the POD basis, which is in turn used to reconstruct the incomplete data. This method has been successfully applied for the reconstruction of human face images, from partial data with 25% of the data missing.

The "gappy POD" method has been also exploited by Bui-Thanh et al. [17,18] in transonic flow analysis and optimization. The case considered is a NACA 0012 aerofoil at a free-stream Mach number of 0.8. With just this limited surface pressure data available, the complete pressure field was determined accurately with only six POD modes.

From technical literature analysis, the use of POD methods for capturing the time variation in unsteady fluid dynamics problems is widespread, while few papers focusing on its application to parametric variation or even shape modification problems can be found. In the same works by Bui-Thanh et al. [17,18], the authors applied the POD technique to steady transonic external aerodynamic problem. In both works, all snapshots were computed by an inviscid steady-state CFD code which uses a finite volume formulation. In Bui-Thanh et al. [17] the POD technique was coupled with a cubic-spline interpolation method in order to develop low-order models that capture the variation in parameters. The problem considered in this work is steady flow about the NACA 0012 aerofoil with varying angle of attack and Mach number. Results showed that the POD method combined with interpolation allows models to be derived that accurately predict steady-state pressure fields over a range of parameter values. However, it is emphasized that in order for the interpolated result to be reliable, the properties of interest must vary smoothly with the parameters under consideration. It has been stated that "the approach can be extended to the case where more than two parameters vary and which may include geometrical properties in order to apply the models in an optimization context".

The POD combined with a response surface method was employed by Tang et al. [24] for the reconstruction and prediction of aerodynamic and aero-thermal solutions of an X-34 configuration. It was reported that this module proved to be not only computationally more efficient than the low-level engineering methods, but also as accurate as the high-level CFD methods, making it valid for MDO and real-time applications.

In the field of aerodynamic shape optimization, the work by LeGresley and Alonso [15,25] and Bui-Thanh et al. [18] demonstrated that the POD method could be used as a low-cost, low-order approximation to enhance the design process. In LeGresley and Alonso [15,25], a set of pressure field distributions corresponding to different aerofoil profiles were computed using an Euler solver. Different aerofoil profiles were created by perturbing the design variables of the base shape. The POD basis was then computed and used to construct a reduced-order model for Euler equations to estimate new, approximate solutions for any arbitrary aerofoil at significantly lower computational costs.

However, one of the most significant challenges is the use of POD-based reduced order models in high-speed flows with parametric variation. Indeed, as the shape/boundary condition parameter changes, the shock waves moves and classical POD/ROM techniques, which work well for subsonic flows, no longer provide reliable and accurate predictions. Lucia et al. [26] and Lucia [27] proposed and used a technique to exploit POD for accurately treating moving shock waves. A zonal approach was proposed to isolate the shock wave and two reduced-order models were developed independently in the inner and outer domains. The application to a one-dimensional quasi-steady nozzle flow-field demonstrated the suitability of the approach. Moreover, LeGresley and Alonso [25] applied this technique for the shape optimization of a two-dimensional aerofoil. The attained results were satisfactory though some discrepancies could still be detected between the high-fidelity solution and the POD/ROM with domain decomposition.

Buffoni et al. [28] discuss three possible methods to adapt domain decomposition to transonic flows with shocks. The first is based on a Schur iteration where the solution of the low-order model is obtained by a projection step in the space spanned by the POD modes. The second is in the same spirit but instead of a Dirichlet–Neumann iteration they employ a Dirichlet–Dirichlet iteration in the frame of a classical Schwartz method. The last approach is of different nature since the solution of the low-order model is not simply based on a projection in the space of the POD modes. It takes into account in a weak sense the governing equations by minimizing the residual norm of the canonical approximation in the space spanned by the POD modes. The main application is about the compressible Euler equations in a nozzle. The authors observe that the obtained results depend to a large extent on the database used for the POD modes. The locality of the approach is recognized, meaning that the approximation error can be large when the reconstructed solution is far from the training ones in the parameter space. They also point out that a major limitation of their method lies in the nonavailability of an efficient method to improve the approximation quality. Indeed, they conclude that, in order to get better results, it is fundamental to increase the approximation accuracy by enriching the functional space in which the solution is sought, based on some objective in-fill criteria.

Toal et al. [29] proposed a POD-based re-parameterization for optimization purposes. This strategy, termed geometric filtration, was found to outperform a traditional kriging-based optimization, producing better designs for a considerable reduction in overall optimization cost. The optimization of a transonic airfoil for minimum drag to lift ratio was used as a test case to compare the geometric filtration strategy to a traditional kriging based optimization and an extensive direct optimization using a genetic algorithm. The traditional kriging strategy achieved 76.3% of the improvement obtained by the genetic algorithm but with only 300 objective function evaluations. However, applying geometric filtration to the same problem, again using 300 objective function evaluations, produced designs achieving 84.1% of the improvement obtained with the genetic algorithm, a substantial improvement over the traditional kriging strategy.

In a more recent paper, Braconnier et al. [30] combined steady compressible RANS equations, a POD reduced-basis method and a leave-one-out adaptive sampling technique. The proposed strategy was tested on an analytic test case and on the two-dimensional turbulent flow around a RAE2822 airfoil. It was shown that the adaptive resampling led to a higher speed of convergence with respect to classical Latin Hypercube *a priori* design of experiments. However, the method applicability is not demonstrated on a real shape optimization problem, but just on a two-parameter Mach-angle of attach design space.

POD has been also investigated in multi-disciplinary analysis and design. Lieu and Farhat [31] and Lieu et al. [32] applied POD-based ROMs for an aero-elastic analysis of a complete F-16 aircraft configuration for varying Mach number and low angles of attack. However, changes in the Mach number or the angle of attack often require the reconstruction of the ROM in order to maintain accuracy. Consequently, this destroys computational efficiency. In that work it was shown that "straightforward approaches for ROM adaptation lead to inaccurate POD bases in the transonic flight regime". Thus, a new ROM adaptation scheme is proposed and evaluated for varying Mach number and angle of attack. The new approach is reported to have a significant potential for accurate, real-time, aero-elastic predictions.

## 3. POD-based surrogates

In this section a review of the mathematical core of Proper Orthogonal Decomposition is presented. POD is a mathematical procedure that allows to perform a modal decomposition of a large set of multi-dimensional data so as to derive a dimensionality reduction and describe the original system with much less number of unknowns. The mathematical development of POD for fluid flow applications, in particular, is described in some detail in Lumley [14]. Details of the mathematical machinery related to POD theory can be found in Holmes et al. [20] and Reed and Simon [33]. Here, the main aspects related to the construction of a reduced order model through singular value decomposition are presented and the use of this technique for steady-state problems is mainly addressed.

### 3.1. Model order reduction

Physics-based approximation concepts require a deep understanding of the governing equations and the numerical methods employed for their solution. The substantial difference between a reduced order model and data fit model consists in retaining an explicit dependency between state variables, related to the governing equations, and design parameters. In other words, the reduced order models operate on the dimensionality of the discretization of the state equations rather than on the design space. Thus, such models are partially independent of a notable increase in the number of design variables. A reduced-order model, in fact, mimics the basic structure of the problem and not just a functional relationship between input and output parameters. Hence, the main advantage of using reduced order models lies in being mostly insensitive to the curse of dimensionality.

To illustrate the model order reduction concept, consider the discrete mathematical model (e.g., RANS equations) of a physical system written in the form

$$\mathbf{R}(\mathbf{w},\mathbf{x}(\mathbf{w}))=0 \qquad (1)$$

where $\mathbf{w}\in\mathbb{R}^t$ is the vector of design variables and $\mathbf{x}\in\mathbb{R}^q$ the discretized vector of state (or field) variables (velocity, energy, density). Note that $\mathbf{x}$ is an implicit vector function of the design variable vector. Unlike classical data fit methods (e.g., Kriging, RBF) which work on local or integral values of the state variables, reduced order methods, instead, provide an approximation of the state vector in the form

$$\hat{\mathbf{x}}=c_1\boldsymbol{\phi}_1+\cdots+c_M\boldsymbol{\phi}_M=\boldsymbol{\Phi}\mathbf{c} \qquad (2)$$

where $\boldsymbol{\Phi} = \{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_M\} \in \mathbb{R}^{q \times M}$ is a matrix of known basis vectors and $\mathbf{c} = \{c_1, \ldots, c_M\} \in \mathbb{R}^M$ is a vector of unknown coefficients. The underlying approximation is that the state vector lies in the subspace spanned by a set of basis vectors. Obviously, this is not true for each state vector, but later on it will be shown how a proper choice of an orthonormal basis leads to the minimization of the approximation error in a least square sense. This is how a Proper Orthogonal Decomposition is derived. Following this approach, the problem of representing a state vector with $q$ unknowns can be recast into a problem with $M$ unknowns and, as usually $q \gg M$, it is possible to get an approximation of $\mathbf{x}$ very efficiently. The estimation of the vector $\mathbf{c}$ can be obtained with different techniques, classified as intrusive and non-intrusive: the first introduce the approximation in the governing equations and find the coefficients by minimization of the residual norm; the second employ data fit techniques trained on a set of known coefficients.

The basis vectors can be computed starting from state solutions of the discrete governing equations which correspond to $M$ different values of the parameters $\mathbf{w}$. As a consequence, the matrix $\boldsymbol{\Phi}$ contains the basis vectors of the subspace:

$$\boldsymbol{\Phi} = \mathrm{span}\{\mathbf{x}(\mathbf{w_1}), \mathbf{x}(\mathbf{w_2}), \ldots, \mathbf{x}(\mathbf{w_M})\} \in \mathbb{R}^{q \times M} \tag{3}$$

For instance, the state solutions $\mathbf{x}_i = \mathbf{x}(\mathbf{w}_i)$ are obtained by solving the RANS equations on $M$ different configurations generated by applying the employed parameterization method on $M$ design vectors $\mathbf{w}_i$. The definition of the $M$ design sites where to compute the solutions is not a trivial issue: generally speaking, standard design of experiments techniques are used to sample the design space with good coverage properties, but, as it will be discussed in next sections, this approach may lead to erroneous results when facing highly multi-modal, highly non-linear problems. Indeed, the quality of the approximation strongly depends on the location of training data in the design space.

### 3.2. POD solution

The so-called snapshot method introduced by Sirovich [34] allows to find the POD basis by solving an eigenvalue problem whose dimension is equal to the number of snapshots, which is much lower than the dimension of the eigenvalue problem. This is a big advantage in aerodynamic design, where usually the number of mesh points, and hence the snapshot size, is huge in comparison to the number of snapshots. In this section, the singular value decomposition (SVD) solution of the POD basis vectors and coefficients is described for steady-state problems. This matrix solution method is normally preferred to the eigenvalue/eigenvector solution as it is faster and easier to implement. The discussion will unfold with specific reference to compressible aerodynamic problems, hence the space domain will be the discretized volume occupied by the flowing air and the snapshot vectors will be defined from computed flow fields.

#### 3.2.1. Snapshots collection

POD fields are usually provided and predicted on fixed meshes. Indeed, as the presented methodology is mainly aimed at shape design and optimization, this assumption is hardly satisfied as the design parameters act on the modification of the geometry at hand and, consequently, the resulting shape will be different for each design vector. The physical domain varies with the design variables, so that the set of discretized spatial points are not the same for each design candidate. In other words, the snapshots collection in turn would be affected by the discrepancy in the physical domain mapping. Indeed, the POD modes would no longer remain at fixed places within the computational domain and consequently a shift error would be introduced as, being the POD model a space-index transformation, the correspondence between space location and vector index would be lost. An interesting solution has been proposed by Pettit and Beran [35] with the definition of a common domain, shared by all the design candidates, and transpiration boundary conditions to take into account for the domain change. LeGresley [36] defined the scalar product as a weighted sum of vector component products where the weight is the cell volume, so that any deformation of the computational cell due to geometry modification is taken into account. Here, quite a different approach is used. The snapshot structure is conceived as a combination of mesh coordinates (i.e. spatial locations) and flow field variables (i.e. state vectors). The idea is to let the POD basis catch the coupling effects between space location and state field. Hence, once the surrogate model is built, not only a flow field can be computed, but also an approximation of the volume mesh. Such a surrogate model would be able to catch, although in a reduced order form, the cross effects of geometry modification and aerodynamic flow change.

With respect to the previous section, here a change in notation is introduced to better fit with the usual CFD terminology. In particular, the discretized spatial locations are indicated explicitly with the three spatial components $(\xi, \upsilon, \zeta)$ and the general snapshot vector is expressed as $\mathbf{s}$. Let $\{\mathbf{w}_j\}$ be a set of design vectors (e.g., sampled from the design space with a DoE technique) and $\{\mathbf{s}_j\}$ the corresponding snapshots, i.e. column vectors containing the volume grid and flow variables as obtained from a CFD solution:

$$\begin{aligned}
&\mathbf{s} = (\mathbf{s}_{\mathrm{grid}}, \mathbf{s}_{\mathrm{flow}})^T \\
&\mathbf{s}_{\mathrm{grid}} = (\xi_1, \ldots, \xi_q, \upsilon_1, \ldots, \upsilon_q, \zeta_1, \ldots, \zeta_q) \\
&\mathbf{s}_{\mathrm{flow}} = (\rho_1, \ldots, \rho_q, \rho\xi'_1, \ldots, \rho\xi'_q, \rho\upsilon'_1, \ldots, \rho\upsilon'_q, \rho\zeta'_1, \ldots, \rho\zeta'_q, p_1, \ldots, p_q)
\end{aligned}$$

where $q$ is the number of mesh nodes involved in the POD computation, $(\xi, \upsilon, \zeta)$ are the nodes coordinates in a Cartesian reference system, $\rho$ is the flow density, $(\xi', \upsilon', \zeta')$ are the three Cartesian velocity components and $p$ is the static pressure. Being $s = 8$, the global size of the snapshot is $N = 8 \times q$. Each snapshot is constructed by placing in order the solution at each grid point for the whole grid. This order can be determined arbitrarily, but it must be consistent throughout the whole set of snapshots. Moreover, the POD modes are sensitive to the scaling of the flow variables as the dataset is made of heterogeneous variables. State variables are usually expressed with different units and have different range of variation, so that this can represent a big issue. Consequently, proper scaling factors have to be applied for each state variable in order to get uniformity along the snapshot.

#### 3.2.2. SVD solution

Starting from the vectors $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_M$ obtained by the CFD expensive computations for a representative set of design sites $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M$, finding a Proper Orthogonal Decomposition means to compute a linear basis of vectors to express any other $\mathbf{s}_j \in \mathbb{R}^N$ with the condition that this basis is optimal in some sense. To compute the optimal basis, we first define the snapshot deviation matrix

$$\mathbf{P} = \begin{pmatrix} \mathbf{s}_1 - \bar{\mathbf{s}} & \mathbf{s}_2 - \bar{\mathbf{s}} & \cdots & \mathbf{s}_{\mathrm{M}} - \bar{\mathbf{s}} \end{pmatrix}$$

where the ensemble mean vector is computed as

$$\bar{\mathbf{s}} = \frac{1}{M}\sum_{j=1}^{M}\mathbf{s}_j$$

Then, we search for a set of orthonormal vectors $\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_M$ such that

$$\mathbf{s}_j = \bar{\mathbf{s}} + \sum_{i=1}^{M}\alpha_i(\mathbf{w}_j)\boldsymbol{\phi}_i = \bar{\mathbf{s}} + \sum_{i=1}^{\widehat{M}}\alpha_i(\mathbf{w}_j)\boldsymbol{\phi}_i + \boldsymbol{\epsilon}^j_{\widehat{M}} = \bar{\mathbf{s}} + \sum_{i=1}^{\widehat{M}}(\mathbf{s}_j, \boldsymbol{\phi}_i)\boldsymbol{\phi}_i + \boldsymbol{\epsilon}^j_{\widehat{M}}$$

with $\widehat{M} \leqslant M$ and the error $\epsilon$ is the smallest possible. Indeed, for any set of orthonormal vectors $\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \ldots, \boldsymbol{\psi}_M$ the optimality condition is expressed as:

$\|a\|$

$$\epsilon_{\widehat{M}} = \sum_{j=1}^{M}\|\mathbf{s}_j - \bar{\mathbf{s}}\| - \sum_{i=1}^{\widehat{M}}(\mathbf{s}_j, \boldsymbol{\phi}_i)\boldsymbol{\phi}_i \leqslant \sum_{j=1}^{M}\|\mathbf{s}_j - \bar{\mathbf{s}} - \sum_{i=1}^{\widehat{M}}(\mathbf{s}_j, \boldsymbol{\psi}_i)\boldsymbol{\psi}_i\|$$

To this aim, we take the singular value decomposition (SVD) of $\mathbf{P}$

$$\mathbf{P} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T = \mathbf{U}\begin{pmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_M \\ 0 & \cdots & 0 \end{pmatrix}\mathbf{V}^T \tag{4}$$

with $\mathbf{U} \in \mathbb{R}^{N\times N}$, $\mathbf{V} \in \mathbb{R}^{M\times M}, \boldsymbol{\Sigma} \in \mathbb{R}^{N\times M}$ and the singular values $\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_M \geqslant 0$. The POD basis vectors, also called POD modes, are the first $M$ column vectors of the matrix $\mathbf{U}$, while the POD coefficients $\alpha_i(\mathbf{w}_j)$ are obtained by projecting the snapshots onto the POD modes:

$$\alpha_i(\mathbf{w}_j) = (\mathbf{s}_j - \bar{\mathbf{s}}, \boldsymbol{\phi}_i) \tag{5}$$

If a fluid dynamics problem is approximated with a suitable number of snapshots from which a rich set of basis vectors is available, the singular values become small rapidly and a small number of basis vectors are adequate to reconstruct and approximate the snapshots as they preserve the most significant ensemble energy contribution. In this way, POD provides an efficient means of capturing the dominant features of a multi-degree of freedom system and representing it to the desired precision by using the relevant set of modes. The reduced order model is derived by projecting the CFD model onto a reduced space spanned by only some of the proper orthogonal modes or POD eigenfunctions. This process realizes a kind of lossy data compression through the following approximation

$$\mathbf{s}_j \simeq \bar{\mathbf{s}} + \sum_{i=1}^{\widehat{M}}\alpha_i(\mathbf{w}_j)\boldsymbol{\phi}_i \tag{6}$$

where

$$\widehat{M} \leqslant M \Rightarrow \frac{\sum_{i=1}^{\widehat{M}}\sigma_i^2}{\sum_{i=1}^{M}\sigma_i^2} \geqslant \epsilon \tag{7}$$

and $\epsilon$ is a pre-defined energy level. In fact, the truncated singular values fulfils the relation

$$\sum_{i=\widehat{M}+1}^{M}\sigma_i^2 = \epsilon_{\widehat{M}}$$

If the energy threshold is high, say over 99% of the total energy, then $\widehat{M}$ modes are adequate to capture the principal features and approximately reconstruct the dataset. Thus, a reduced subspace is formed which is only spanned by $\widehat{M}$ modes.

### 3.2.3. Pseudo-continuous global representation

Eq. (6) allows to get a POD approximation of any snapshot $\mathbf{s}_j$ belonging to the ensemble set. Indeed, the model does not provide an approximation of the state vector at design sites which are not included in the original training dataset. In other words, the POD model by itself does not have a predictive feature globally, i.e. over the whole design space. Among the possible options to accomplish this task, here it is shown how it can be done by establishing a functional relation between the $\alpha_i$ coefficients, which represent the projection of a generic CFD flow field onto the set of POD basis vectors, and the design variables.

It is well known that regression techniques are particularly suitable to fit experimental data, as they filter the random noise out from the data. This behaviour is less desirable when working with computer simulations based on determinism. In this case, one asks the data fit model to exactly reproduce the sample data used for training and to consistently catch the local data trends. A Radial Basis Function (RBF) network answers to these criteria.

A Radial Basis Function is a real-valued function whose value depends on the Euclidean distance from a point called centre. A RBF network uses a linear combination of radial functions. A RBF model can be expressed as

$$\alpha(\mathbf{w}) = \sum_{i=1}^{M} f_i r(\|\mathbf{w} - \mathbf{w}_i\|, \theta_i) \tag{8}$$

where the approximating function is represented by a sum of $M$ RBFs $r$, each associated with a different centre $\mathbf{w}_i$, weighted by a real-valued weight $f_i$ and characterized by a width parameter $\theta_i$. Hence, an RBF network can be defined as a weighted sum of translations of radially symmetric basis function. Given $d = \|\mathbf{w} - \mathbf{w}_i\|$, typical RBFs $r$ are:

Gaussian – $r(d, \theta) = e^{-\frac{d^2}{2\theta^2}}$
Multi-quadric – $r(d, \theta) = \sqrt{1 + \frac{d^2}{\theta^2}}$
Inverse quadratic – $r(d, \theta) = \frac{1}{1 + \frac{d^2}{\theta^2}}$

Once fixed the types of RBF to be used and the "optimal" width parameters, the RBF network is defined only by the weights $f_i$. More generally, the RBF centres may not coincide with the training points as they can be treated as additional parameters (i.e., to be found "optimally") of the RBF approximation. However, the present approach considers the exact matching between RBF centres and training points. RBF networks are built over known values of the POD coefficients to predict at a generic design site which is not included in the original ensemble [10]. For each of the $\widehat{M}$ preserved modal coefficients, RBF interpolations are trained on the following correspondences:

$$\begin{aligned} &\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M\} \to \{\alpha_1(\mathbf{w}_1), \alpha_1(\mathbf{w}_2), \ldots, \alpha_1(\mathbf{w}_M)\} \\ &\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M\} \to \{\alpha_2(\mathbf{w}_1), \alpha_2(\mathbf{w}_2), \ldots, \alpha_2(\mathbf{w}_M)\} \\ &\vdots \\ &\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M\} \to \{\alpha_{\widehat{M}}(\mathbf{w}_1), \alpha_{\widehat{M}}(\mathbf{w}_2), \ldots, \alpha_{\widehat{M}}(\mathbf{w}_M)\} \end{aligned} \tag{9}$$

In other words, the RBF parameters are found by imposing the interpolation condition on the training set for any modal coefficient $i \leqslant \widehat{M}$, which in turn results in solving $\widehat{M}$ linear systems

$$\begin{pmatrix} r(0, \theta_1) & \ldots & r(\|\mathbf{w}_1 - \mathbf{w}_M\|, \theta_M) \\ r(\|\mathbf{w}_2 - \mathbf{w}_1\|, \theta_1) & \ldots & r(\|\mathbf{w}_2 - \mathbf{w}_M\|, \theta_M) \\ \vdots & \vdots & \vdots \\ r(\|\mathbf{w}_M - \mathbf{w}_1\|, \theta_1) & \ldots & r(0, \theta_M) \end{pmatrix}\begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_M \end{pmatrix} = \begin{pmatrix} \alpha_i(\mathbf{w}_1) \\ \alpha_i(\mathbf{w}_2) \\ \vdots \\ \alpha_i(\mathbf{w}_M) \end{pmatrix} \tag{10}$$

The width parameters have a big influence both on the accuracy of the RBF model and on the conditioning of the solution matrix. In particular, it has been found [10,37] that interpolation errors become high for very small and very large values of $\theta$, while the condition number of the coefficient matrix increases with increasing values of $\theta$. Therefore, they have to be "optimal" in the sense that a tuning of the width parameters is needed to find the right trade-off between interpolation errors and solution stability (Ref. [37] for a discussion about how to properly select the best set of parameters). In the present approach, instead of fixing *a priori* the RBF types and width parameters, the RBF implementation is based on a tuning of RBFs aiming at minimizing the Leave-One-Out error

[38]. Multi-quadrics and Gaussian RBF types are used. The tuning autonomously chooses the RBF type and adjusts the width parameters in order to have an accurate model. Given a combination of RBF type and width parameter, the total cross-validation error (i.e., the sum of the cross-validation error at each available test data) is obtained and the combination with the minimum error is selected. The evaluation of each Leave-One-Out error is not expensive thanks to the usage of the efficient formula by Rippa [39], which states that the only required information to compute the cross-validation residual have been already computed during the construction of the full RBF model (with no sampling left out).

The pseudo-continuous prediction of the flow field at a generic design site $w$ is then expressed as:

$$\mathbf{s}(\mathbf{w}) = \bar{\mathbf{s}} + \sum_{i=1}^{\widehat{M}} \alpha_i(\mathbf{w})\boldsymbol{\phi}_i \tag{11}$$

This provides a useful surrogate model which combines design of experiments for sampling, CFD for training, POD for model reduction and RBF network for global approximation. In conclusion, an explicit, global, low-order and physics-based relation between the design vector and the state vector has been derived thanks to the following steps:

- Design of computer experiments – sample the design space with a DoE technique (with or without auto-adaptation) and define a set of $M$ design sites $(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M)$;
- Snapshots collection – compute a series of CFD simulations in parallel with the full-order CFD model for the selected set of design sites $(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M)$ and fill $M$ snapshot vectors $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_M$ with the state variables values;
- Model order reduction – evaluate the set of POD modes $\{\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_M\}$ and POD coefficients $\{\alpha_1, \alpha_2, \ldots, \alpha_M\}$ as described in Section 3.2.2 and according to Eqs. (4) and (5);
- Basis truncation – store the first $\widehat{M}$ modes according to energy criteria (Ref. Eqs. (6) and (7));
- Model training – train $\widehat{M}$ RBF models for $\widehat{M}$ POD coefficients on $M$ data set according to Eqs. (8)–(10);
- Model prediction – assemble the model response as a snapshot vector $\mathbf{s}$ of state variables at an arbitrary design site $\mathbf{w}$ (Ref. Eq. (11)).

## 4. POD model validation: an application example

In this section the POD-based reduced order model is assessed in transonic conditions. The aerodynamic case is represented by a steady, viscous flow about a scaled RAE2822 airfoil. This case was selected as it is a standard geometry used to validate CFD numerical modelling [40]. The POD snapshots are obtained by perturbing the RAE2822 airfoil by means of the parameterization described later on. A mixed POD/CFD approach (zonal POD) is proposed to increase the accuracy level of the surrogate model in transonic conditions.

### 4.1. Parameterization and design space definition

In the present context, surrogate modelling is aimed at providing a fast and accurate tool to speed up the process in aerodynamic shape design. As a consequence, one of the most important issues is to show its suitability and applicability to the shape optimization problem. Indeed, the definition of the design space through shape modification parameters typically involves a complex, often highly nonlinear relation between the flow field and the design variables. Moreover, modifying an aircraft component (e.g., a wing airfoil) requires several parameters, thus enlarging the dimensions of the design space. It is straightforward, then, that the complexity of the problem increases and approaches a real-world application level. The CST method [41] provides an analytical form to represent various geometry of aeronautical interest and it shows interesting properties of continuity, differentiability and reproducibility of a huge number of test shapes. It allows to specify the airfoil contour as a product of a class function, which in the proposed case defines the rounded leading edge/pointed trailing edge airfoil class, and a shape function obtained as a linear combination of $n$th-order Bernstein polynomials. The parameterization is described by the equation set (12)

$$\begin{aligned} y_{u,l}(\psi) &= C(\psi)S_{u,l}(\psi) \\ C(\psi) &= \psi\sqrt{1-\psi} \\ S_{u,l}(\psi) &= \sum_{i=0}^{n} A_i^{u,l} K_{i,n} \psi^i (1-\psi)^{n-i} \\ K_{i,n} &\equiv \binom{n}{i} = \frac{n!}{i!(n-i)!} \end{aligned} \tag{12}$$

where $y_{u,l}(\psi)$ are respectively the upper and lower airfoil shapes, $\psi$ is the dimensionless abscissa along the airfoil chord direction, $R_{le}$ is the leading edge radius, $c$ is the airfoil chord length, $\beta$ the trailing edge closure angle and the design vector is defined as:

$$\mathbf{w} = \left(A_0^u, A_1^u, \ldots, A_n^u, A_0^l, A_1^l, \ldots, A_n^l\right)$$

The first and last parameters $A_0$, $A_n$ are related respectively to the leading edge radius ($A_0 = \pm\sqrt{2R_{le}/c}$) and to the trailing edge closure angle ($A_n = \tan\beta$), as shown in detail in Ref. Kulfan [41].

The subscript/superscript $^u$ and $^l$ refer respectively to the upper and lower airfoil surface. In the present context, 7th-order Bernstein polynomials are considered, hence each airfoil side is described by eight design variables. The design space $DW$ is then a subset of $\mathbb{R}^{16}$. A scaled 14% thickness ratio RAE 2822 airfoil is selected as the baseline airfoil. The airfoil geometry is shown in Fig. 1. The corresponding design parameters values, which define the RAE 2822 profile according to Eq. (12), and their range of variation, which define the design space, are reported in Table 1.

### 4.2. Design of experiments

Design-of-Experiment theory (DoE) [42] was born to provide experimentalists with a tool to optimally choose the independent variable values for a limited number of experiments. One of the

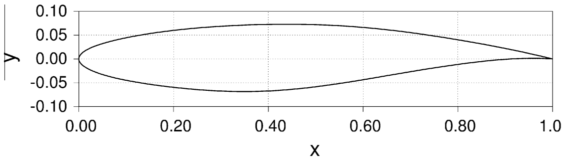


**Fig. 1.** Baseline geometry, RAE 2822 airfoil.

**Table 1**
Design parameters, values and ranges.

| Parameter | Upper values | Lower values | Upper range | Lower range |
|---|---|---|---|---|
| $A_0$ | 0.1293 | −0.1280 | [0.1293, 0.2293] | [−0.2280, −0.1280] |
| $A_1$ | 0.1282 | −0.1483 | [0.1282, 0.2282] | [−0.2483, −0.1483] |
| $A_2$ | 0.1771 | −0.1080 | [0.1771, 0.2771] | [−0.2080, −0.1080] |
| $A_3$ | 0.1219 | −0.2580 | [0.1219, 0.2219] | [−0.3580, −0.2580] |
| $A_4$ | 0.2393 | −0.0918 | [0.2393, 0.3393] | [−0.1918, −0.0918] |
| $A_5$ | 0.1662 | −0.1079 | [0.1662, 0.2662] | [−0.2079, −0.1079] |
| $A_6$ | 0.1976 | −0.0561 | [0.1976, 0.2976] | [−0.1561, −0.0561] |
| $A_7$ | 0.2110 | 0.0638 | [0.2110, 0.3110] | [−0.0362, 0.0638] |

most widely adopted is the Latin Hypercube Sampling (LHS) technique by Mckay et al. [43]. The basic principle of LHS is to bound the randomness of the sample selection. In fact, let $t$ be the number of design variables, each design variable range is divided into $m$ intervals or bins of equal probability. This generates a total of $m \times t$ bins in the whole space. Within each bin only one sample is allocated randomly. This ensures that a one-dimensional projection onto the parameter space will produce one sample in each bin.

LHS is useful for the initialization of POD-based surrogate models, but, as will be detailed later on, it exhibits some major limits which prevent them from being used as a standard sampling technique for optimization purposes. Indeed, LHS is optimal in the sense of design space coverage, but it does not allow for refining the sampling distribution according to enrichment or improvement criteria, e.g. design space exploration or objective function minimization. Here, a random Latin Hypercube sampling has been used, without using any technique to optimize the space filling.

### 4.3. Zonal POD

The POD surrogate model is mainly designed as reduced order model (ROM) within a shape optimization process, where the geometry and, hence, the volume mesh vary with the design site. Moreover, the application is focused on transonic aerodynamics with potential flow separations and shock waves. Therefore, care must be taken about the definition of the snapshot domain and how to extract the integral quantities of interest (e.g., aerodynamic force coefficients) from the snapshot structure. Indeed, as the snapshots are expressed through a linear combination of POD modes, shock waves, flow separations and other non-linearities present in the training ensemble would be captured and replied in the POD modes, so that any prediction of a new snapshot would likely bring the footprint of those flow features with it. This is a desirable behaviour on average for a physics-based approach, but when approaching the optima, which should be featured with shock-less and fully attached flow profile, a POD approximation of this type would hide the potential improvement behind the trace of the original snapshots. This issue is of paramount importance and can be tackled by introducing and combining two concepts: zonal POD and adaptive sampling. The first will allow to reduce the inherent variability of the snapshots by means of a domain partitioning, thus avoiding the POD basis to capture all the physics within the field. The second technique will allow to enrich the POD approximation by sampling at new points which are "optimal" in the sense of exploration/model improvement balance. In this section, the discussion will be focused on the zonal POD approach.

The basic idea, proposed by Iuliano [44], is to use a mixed full order (FOM)/reduced order (ROM) model by splitting the solution domain into two sub-domains: the FOM (i.e. the CFD RANS model) is used only in the vicinity of the surface to accurately solve the near wall boundary layer, non-linearities (e.g., shock waves) and flow separations where they occur; the ROM (i.e. the POD surrogate model) is exploited to reconstruct the flow field far from the solid wall, where a smoother and weakly varying solution is expected.

Fig. 2 shows a sketch of the domain decomposition. The POD based surrogate model is built on the spatial domain defined in the light grey region, hence the size $N$ of each snapshot, as described in Section 3.2, is eight times the number of mesh points in this zone. Once trained the POD model, the surrogate response on the FOM/ROM boundary interface (blue curves in the figure) is extracted and used as boundary conditions to iterate the full order CFD solver in the inner domain (orange). Details about the specific boundary condition formulation across the two domains can be found in Iuliano [44]. An useful advantage of the zonal POD is that any aerodynamic coefficient or surface distribution of interest (e.g., pressure or skin friction distributions) can be directly extracted from the CFD solution in the inner domain. On the other hand, when the full POD model is used (i.e., trained on the whole domain), the surrogate model would provide a prediction of the mesh and state variables, so that a properly designed "condensation" procedure has to be applied to retrieve the integral coefficients like $C_l$, $C_d$, $C_m$. In particular, the general definition of the aerodynamic coefficients is

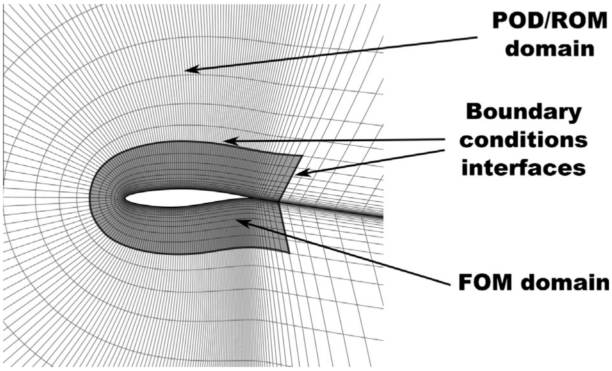


**Fig. 2.** Zonal approach, FOM/ROM domains.

$$C_{l,d,m} = \frac{1}{S}\int_S F_{l,d,m} dS \simeq \frac{1}{S}\sum_{i=0}^{N_S} F^i_{l,d,m}\Delta S_i \tag{13}$$

where $S$ is the integration surface, $F^i_{l,d,m}$ and $\Delta S_i$ are the discretized components of $F_{l,\ d,m}$ and $S$. The function $F_{l,d,m}$ represent the pressure and shear stresses distribution on the solid wall times the scalar product of the unit vector ($\underline{n}$) normal to the wall and the unit vector of the axis along which the aerodynamic force is being projected. For example, the integrand function for computing the lift coefficient is

$$F_l = \frac{2}{\rho_\infty V_\infty^2}[(p - p_\infty)\underline{n}\cdot\underline{j} - (\underline{\underline{\tau}}\cdot\underline{j})\cdot\underline{n}] \tag{14}$$

where $\underline{j}$ is the unit vector normal to the wind direction. In order to be able to predict the aerodynamic coefficients, the discretized functions $(F^i_{l,d,m}\Delta S_i)$, as computed from the training solutions, are appended to the snapshot vectors. Hence, when predicting with the POD surrogate at a new design site, the approximated values of $(F^i_{l,d,m}\Delta S_i)$ would be also available and a simple sum would be needed to get the integral coefficients.

The described zonal approach is generally applicable and it does not depend on the particular flow solver adopted. Anyway, the content of the following sections is related to the implementation of such an approach into the in-house developed ZEN flow solver [45], which is able to solve Euler and Reynolds-Averaged Navier Stokes (RANS) equations on multi-block structured grids.

This section is concluded with a discussion about the comparison of the present zonal approach with two main references that exist in the literature, namely the works by Buffoni et al. [28] and LeGresley and Alonso [25]. With respect to the first one, the proposed approach presents some similarities, especially in the geometrical domain decomposition and in the introduction of an internal boundary condition. However, there are some basic differences mainly due to the fact that the present method is conceived to answer the needs of an optimization search.

Buffoni et al. [28] solve two different sub-problems on two sub-domains in an iterative fashion and update the internal boundary conditions on the interface up to convergence. The present method, instead, solves a problem similar to the original one in a much smaller domain by assuming that the values of the internal boundary conditions are estimated with the POD model from the outer domain. This approach inherently brings a good accuracy potential as selecting the region to be solved with POD model far off the solid wall would embed smaller aerodynamic disturbances. Summing up, here the POD model is simply used to give an estimation of the internal boundary condition. Moreover, the present method has been conceived to handle changing domain boundaries (e.g. airfoil/wing shape modification) whose shape is function of the design parameters. The authors' experience suggests that this leads to a completely different POD energy decomposition with respect to the case of changing only the external boundary condition set. This issue has been handled including the mesh points into the POD snapshots definition. Finally, Buffoni et al. [28] find the POD coefficients at an unknown point by projecting the high-fidelity solution onto the trace of the POD modes on the internal boundary interface or by residual minimization. The present method, instead, uses RBF interpolation techniques.

On the other hand, the work by LeGresley and Alonso [25] proposes a coupling between full order and POD model in order to reduce the estimation error in selected sub-domains. Indeed, the idea is to augment the POD basis with additional top hat functions to improve the approximation where the prediction error is high (typically, in the vicinity of shocks and discontinuities). The number and dimensions of the sub-domains where to compute the estimation error is rather arbitrary, as well as the threshold of the estimation error to decide whether that particular sub-domain should be selected for full order evaluation or not. Their approach is an *a posteriori* correction of the original POD approximation based on mesh partitioning and sophisticated and costly error estimation scheme, while the present methodology can be defined as an *a priori* decomposition. Indeed, the domain decomposition is here selected before constructing the model and its accuracy is checked in a parametric way (i.e., varying the distance of the POD/CFD domain from the solid wall). The underlying idea is that in a transonic viscous flow the non-linear disturbances (boundary layer and shock waves) are confined in a well known region in the vicinity of the body and, hence, the region in which the POD error is maximized can be easily identified. Moreover, the domain decomposition by LeGresley and Alonso [25] requires a multi-block approach or a mesh partitioning ($8 \times 8$ block cells division in the proposed application), while the present method is based on the definition of an interface boundary between two domains. Furthermore, the authors applied the domain decomposition methodology to inviscid compressible flow, while the present work deals with RANS solutions. This difference impacts not only the computational cost, but also the reliability of the shape optimization results due to the well-known interaction between shock wave and boundary layer. This would strongly affect the domain partitioning near the wall and the choice of a suitable basis of the correction term.

### 4.4. Surrogate models training

Before getting into the validation process, the POD/ROM model has to be trained, so that an initial Latin Hypercube sampling is done on the design space made of sixteen variables (Ref. Section 4.1). The size $M$ of the training sampling is chosen very large ($M = 180$) to cover each design variable with a sufficient number of samples. As a rather arbitrary rule of thumb, the sampling size has been assumed to be at least 10 times the design space dimension (i.e. $\geqslant 160$). Anyway, the correctness of this assumption will be investigated in future research. The set of design sites $\{\mathbf{w}_i\}$ are then transformed into the physical representation of the airfoil geometry thanks to Eq. (12). The baseline geometry is a modified RAE 2822 airfoil, scaled to 14% thickness-to-chord ratio to amplify compressibility effects. One hundred eighty calls of the volume mesh generator and CFD solver are launched in parallel at fixed flow conditions to compute the flow field around each airfoil shape. Thanks to a proper selection of the baseline geometry and design weight ranges, a wide and varied distribution of shock waves locations and flow separations is obtained through the training dataset. This is a highly desirable feature to test the predictive capability of such a physics-based surrogate model. The flow conditions are summarized in Table 2. Fully turbulent flow is assumed. For each airfoil shape, a single-block structured volume mesh made of 25,186 points (12,288 cells) is computed by means of an automatic hyperbolic grid generator. Using fixed topology, mesh parameters and sizes, standard quality grids are obtained for each geometry. The first cell at the wall is placed so as to have a unit $y^+$ at the specified flow conditions. Fig. 3 shows a sketch of the volume mesh around the baseline airfoil.

**Table 2**
Design point.

| Mach number | Reynolds number | Angle of attack (°) |
|---|---|---|
| 0.729 | 6.5e+06 | 2.0 |

With reference to Fig. 4, the mesh partitioning is applied to define the FOM/ROM domains, which are required to be non-overlapping and adjoining. This can be easily done when a structured mesh is available as the grid lines can be used as interfaces between domains. To this aim, the $d$ parameter is introduced as the distance of the FOM/ROM interface from the airfoil leading edge. Indeed, different POD-based reduced order models can be defined by varying this distance and, hence, reducing or increasing not only the size but also the inherent variability of the snapshot set. This mechanism has to be carefully considered beside the coexistence of eight heterogeneous variables (spatial coordinates, density, pressure and velocity) in the same snapshots, as it could introduce a bias in the correlation process. For example, the POD reduction could give more importance to the flow features related to the snapshot variables which exhibit the largest absolute values or the widest range of variation. To avoid this, a scaling operator is applied to the snapshot set prior to feed the POD model. The scaling factors are designed so as to map each variable to the interval $[0,1]$ by normalizing as follows:

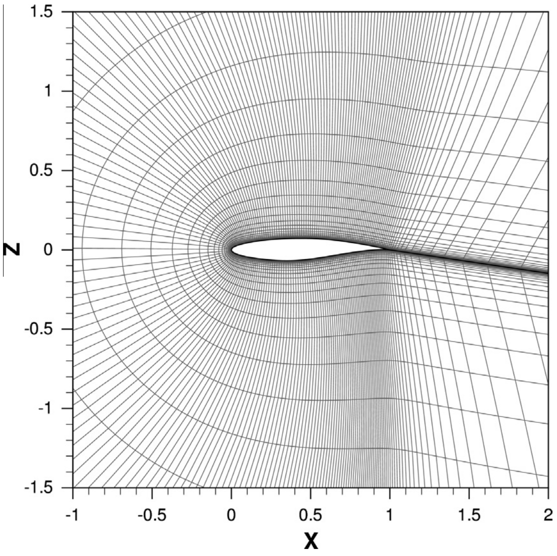


**Fig. 3.** Volume mesh around the airfoil.

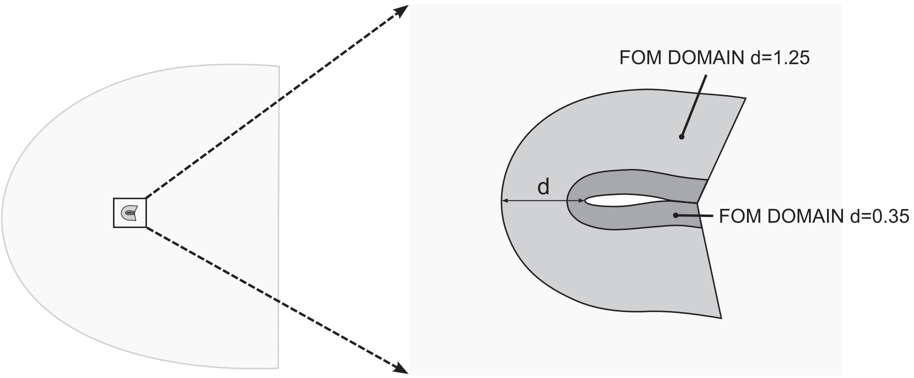


**Fig. 4.** FOM/ROM domains with varying interface.

$$\mathbf{x}_h^* = \frac{\mathbf{x}_h - \min(\mathbf{x}_h)}{\max(\mathbf{x}_h) - \min(\mathbf{x}_h)}$$

where $\mathbf{x}_h$ is the vector containing the $h$th flow variable in the snapshot $\mathbf{s} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_8)^T$ and the minimum and maximum are taken over the vector $\mathbf{x}_h$. In the present investigation the scaling factors are defined once for each variable and kept constant even when varying the FOM/ROM domains (and hence of the snapshot size).

Five surrogate models (mentioned herein with the initials “SM”) are built exploiting the computed CFD data:

SM1 – a POD model with $d = 0$, i.e. the full-order model is not invoked, the POD approximation is used to get the flow field everywhere and no boundary condition is exchanged. The snapshot size $N$ is 201,488;
SM2 – a POD model with $d = 0.35$, i.e. the full-order domain is the blue one in Fig. 4 while the POD approximation is used to get the flow field anywhere else. The snapshot size $N$ is 91,792;
SM3 – a POD surrogate model with $d = 1.25$, similar to the previous one but now the full-order domain is the orange one in Fig. 4. The snapshot size $N$ is 75,232;
SM4 – a Kriging interpolation model with Gaussian correlation using the aerodynamic efficiency $C_l/C_d$ as response function, the implementation from the Dakota package [46] is used;
SM5 – a quadratic polynomial regression model using the aerodynamic efficiency $C_l/C_d$ as response function. Given $t$ the number of design variables, at least $(t+1) \times (t+2)/2$ design sites should be evaluated to train this type of model. In the present case, $(t+1) \times (t+2)/2 = 153$, hence the size of the *a priori* sampling is sufficient.

SM1, SM2 and SM3 are POD-based reduced order models, while SM4 and SM5 are introduced on purpose to compare the presented methodology with standard meta-models. The ensemble energy content threshold $\epsilon$ is reported in Fig. 5a as a function of the number of POD modes for each of the POD-based approximations. It is clearly evident that SM1 requires a big number of modes even to reproduce a relatively low energy level (95%), while SM3 performs considerably better (97%) with just four modes preserved. SM2 requires more modes with respect to SM3 because the corresponding ROM domain embeds part of the supersonic region on the airfoil suction side: Fig. 5b clarify this issue as it reports the FOM/ROM domains (as in Fig. 4) superimposed with the local Mach number contours. The solution is here computed around an airfoil selected within the ensemble database. While SM3 ROM domain is quite far off the supersonic region, the SM2 FOM/ROM interface lies across it, thus introducing a stronger source of variability (and of slight discontinuity due to the shock wave) into the ensemble. Therefore, for each model a given energy level is obtained with different number of modes. In order to make a fair assessment, the models will not be compared using a pre-defined number of modes, but at a fixed energy level (95%). Indeed, the number of preserved modes is ten for SM1 (95%), seven for SM2 (96.4%) and four for SM3 (97%).

### 4.5. Surrogate models validation and error analysis

In order to assess the performance of surrogate models, a validation plan is needed after the training phase. Validation means measuring the goodness of a surrogate model with respect to the highest-fidelity model and, therefore, drawing information to eventually optimize it. The goal is to evaluate the potential of the model to globally approximate the design space. Once trained, classical validation of a surrogate model is carried out by sampling the design space once more, estimating the full and reduced order models on the new sampling set and computing a set of statistics from the obtained data. This approach requires to compute new CFD solutions and, for this reason, is computationally expensive. In order to reduce the number of full order computations, the validation points could be represented by the same set used for training, provided that a cross-validation technique is used (e.g., leave-one-out). Indeed, cross-validation implies the partitioning of a sample of data into complementary subsets: one subset is used for training, the other one for validation or testing. The variability due to the choice of the partitions is usually reduced by performing multiple rounds of cross-validation and averaging over the rounds. Here a classical validation is performed, while cross-validation will be used later on in auto-adaptive sampling, when the estimation of the quality of the POD model basis and coefficients will be required.

A new LHS sampling of size $\overline{M} = 50$ is performed on the 16-dimensional design space and the new design sites are evaluated by means of both the surrogate and the full order model. Then, the aerodynamic efficiency $E = \frac{C_L}{C_d}$ is computed and used to assess the following error measures:

Percentage error

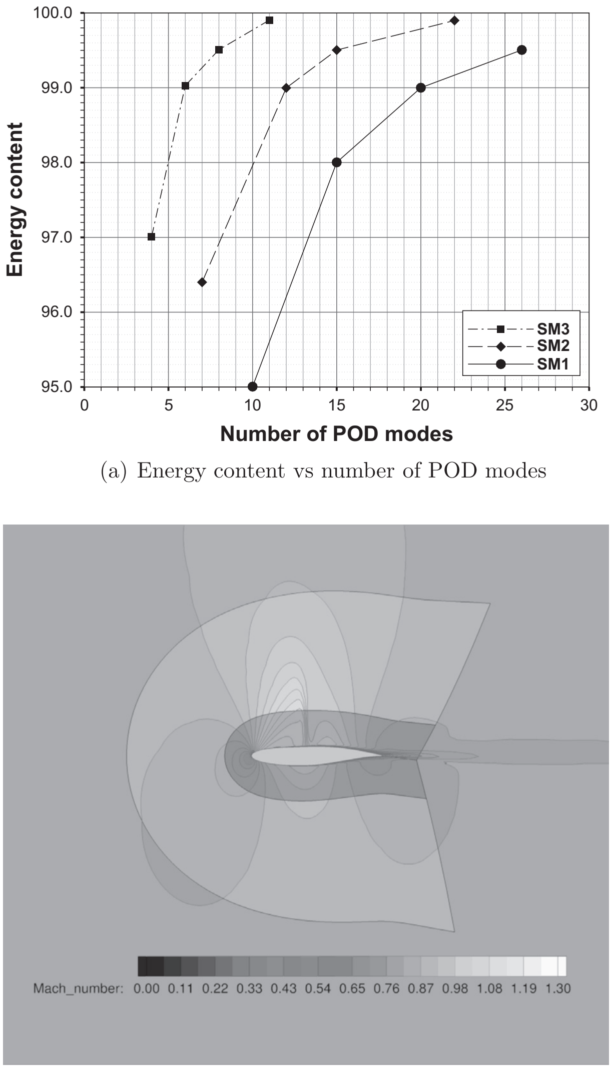
$$\mathrm{PE}_i = \left|\frac{E_i - \widehat{E}_i}{E_i}\right| \times 100, \quad i = 1, \ldots, \overline{M}$$

Mean percentage error

$$\mathrm{MPE} = \frac{1}{\overline{M}} \sum_{i=1}^{\overline{M}} \mathrm{PE}_i \tag{15}$$

Standard deviation of the percentage error

$$\mathrm{SDPE} = \frac{1}{\overline{M} - 1} \sum_{i=1}^{\overline{M}} [\mathrm{PE}_i - \mathrm{MPE}]^2 \tag{16}$$



(a) Energy content vs number of POD modes

(b) FOM/ROM interface cutting and embedding the expansion lobe

**Fig. 5.** Effect of zonal interface on the energy amount captured by POD.

R-squared coefficient of determination

$$R^2 = 1 - \frac{\sum_{i=1}^{\overline{M}}\left(E_i - \widehat{E}_i\right)^2}{\sum_{i=1}^{\overline{M}}\left(E_i - \frac{1}{\overline{M}}\sum_{i=1}^{\overline{M}} E_i\right)^2} \tag{17}$$

where index $i$ denotes the $i$th sample of the DoE validation plan, the hat quantities refer to the surrogate predictions while the hat-less to the full order ones. These type of error measures provide a picture of how the POD model reconstruction error is propagated on a surface integral, as only aerodynamic force coefficients appear. Therefore, it is very useful to understand the suitability of the surrogate model to approximate the fitness function in an aerodynamic optimization process, which usually requires the evaluation of aero-coefficients. However, in order to ensure a more general error analysis, the mean absolute percentage error between the exact CFD computation and the predicted value is introduced at snapshot level:

$$\mathrm{Er}_i = \frac{1}{N}\sum_{j=1}^{N}\left|\frac{s_{i,j} - \hat{s}_{i,j}}{s_{i,j}}\right| \times 100 \tag{18}$$

**Table 3**
Surrogate goodness-of-fit estimation.

| Surrogate | *R*-squared | MPE | SDPE | *G* | Ranking |
|---|---|---|---|---|---|
| SM1 | 0.5876 | 10.33 | 48.14 | 1597.22 | 4 |
| SM2 | 0.8899 | 4.55 | 12.85 | 647.83 | 2 |
| SM3 | 0.9791 | 2.30 | 1.61 | 171.10 | 1 |
| SM4 | 0.8657 | 4.56 | 26.61 | 853.98 | 3 |
| SM5 | 0.06074 | 15.62 | 171.62 | 1761.64 | 5 |

where $N$ is the snapshot size, $s_{i,j}$ ($\hat{s}_{i,j}$) is the $j$th element of the computed (predicted) snapshot vector at the $i$th validation site.

The goodness-of-fit for each model is estimated and results are summarized in Table 3. The surrogate models can be ranked as reported in the rightmost column: SM3 exhibits superior performances for each quality index, while the quadratic polynomial surface is very poor in approximating the objective function. SM3 performs very well even on the SDPE estimate which measures the variation of the percentage prediction error along the validation sampling. Hence, the prediction errors at any validation site are comparable and close to the mean value. This is a very desirable feature for a surrogate model designed for optimization. On the other hand, SM5 shows very poor performances because such a polynomial regression is unable to approximate a multimodal, rapidly changing objective function.

Looking at the figures in Table 3, the models SM2 and SM4 show similar results, even if they differ completely for methodology and construction. This suggests a useful indication when seeking the proper balance between the FOM and ROM domains (i.e. to determine the distance $d$): the POD surrogate accuracy increases by moving the FOM/ROM interface away from the airfoil surface and there exists a peculiar value of the distance $d$ for which its predictive power is very close to standard and efficient interpolation techniques.

Monotonicity is one of the properties a good surrogate should have in an optimization process. Given two "true" data $f(\mathbf{w}_i)$ and $f(\mathbf{w}_j)$ and the corresponding surrogate predictions $\hat{f}(\mathbf{w}_i)$ and $\hat{f}(\mathbf{w}_j)$, a surrogate model is monotonic when:

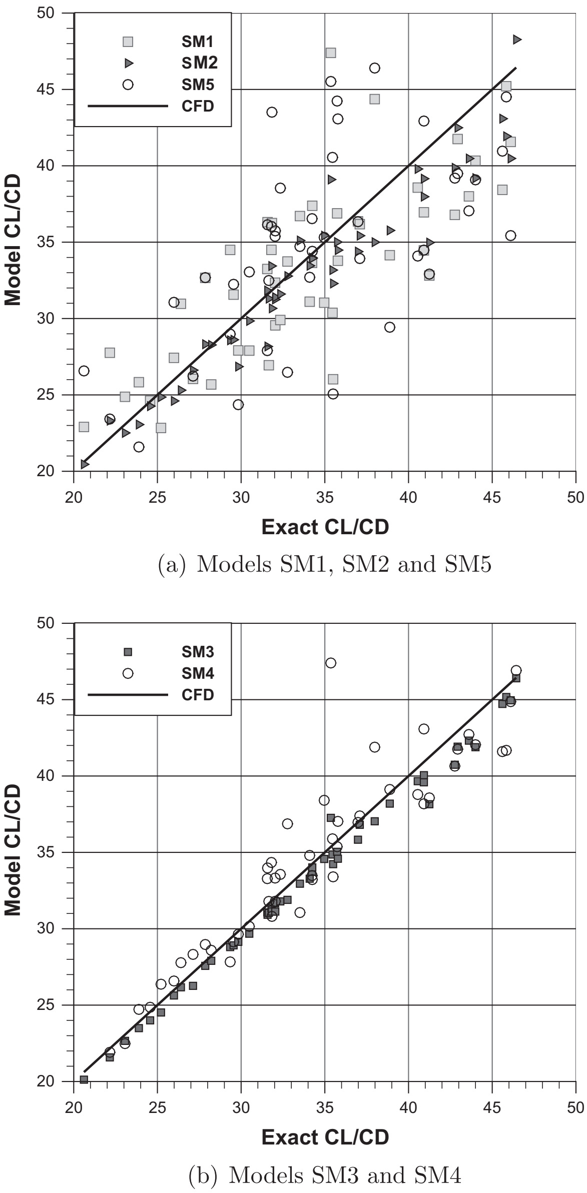
$$f(\mathbf{w}_i) \leqslant f(\mathbf{w}_j) \Rightarrow \hat{f}(\mathbf{w}_i) \leqslant \hat{f}(\mathbf{w}_j)$$

This property can be global (i.e., valid for each $\mathbf{w}_i, \mathbf{w}_j \in DW \subset \mathbb{R}^t$) or local. In order to measure the monotonicity, the following metric is introduced:

$$G = \sum_{i=1}^{\overline{M}}\sum_{j=1}^{i} -\min\left(0, \frac{\Delta\widehat{E}_{ij}}{\Delta E_{ij}}\right)$$

where $\Delta E_{ij} = E_i - E_j$ and $\Delta\widehat{E}_{ij} = \widehat{E}_i - \widehat{E}_j$. The $G$ index can assume any non-negative value, zero value indicates global monotonicity and the higher the magnitude, the more significant is the monotonicity loss. As shown in Table 3, this ranking measure is coherent with the previously introduced indicators and, considering the big difference between SM3 and other models, it provides additional evidence of the quality of this model.

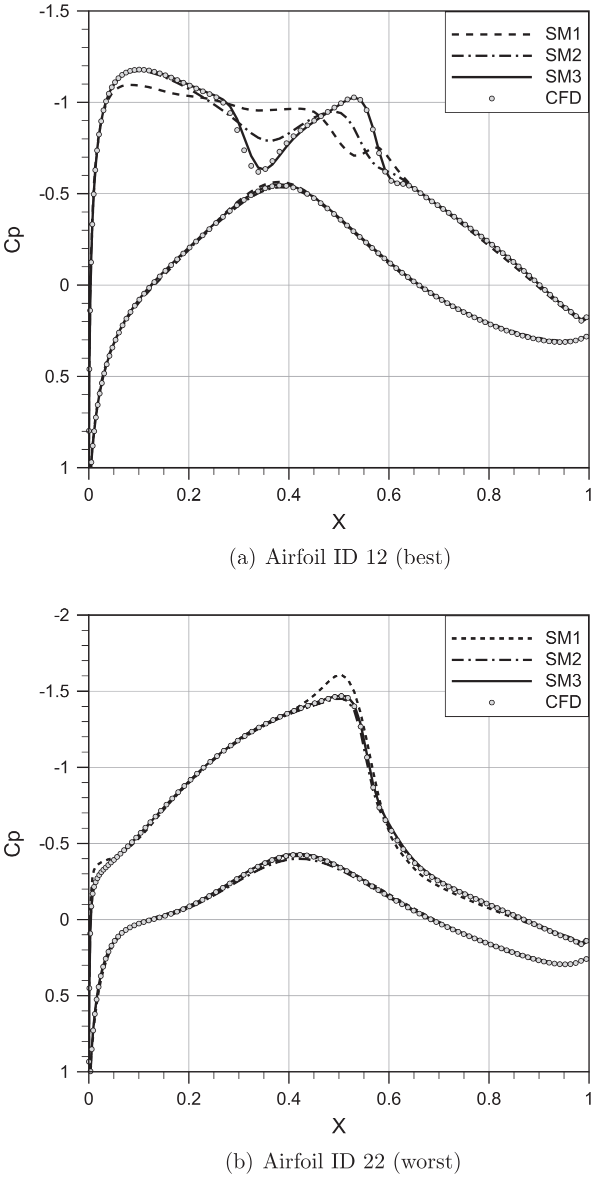
Fig. 6 reports the correlation plot between the models prediction and the "true" CFD data. Again, SM2, SM3 and SM4 are globally closer to the linear trend, resulting in a better fit. The correlation plot highlights another significant feature of SM3 model, as it generally underestimates the aerodynamic efficiency. For further comparisons, Table 4 summarizes the validation set indices where each model predicts the highest and lowest efficiency, the corresponding values of aerodynamic efficiency and the percentage error with respect to the CFD datum. This is useful to evaluate the capability of the model to identify the global extrema of the objective function. It is observed that only SM4



(a) Models SM1, SM2 and SM5

(b) Models SM3 and SM4

**Fig. 6.** Correlation plot of surrogate models prediction.



(a) Airfoil ID 12 (best)

(b) Airfoil ID 22 (worst)

**Fig. 7.** Pressure coefficient comparison.

**Table 4**
Surrogate estimations of aerodynamic efficiency for best and worst validation airfoils.

| Surrogate | ID of max | ID of min | max | min | Δmax (%) | Δmin (%) |
|---|---|---|---|---|---|---|
| CFD | 12 | 22 | 46.43 | 20.61 | 0 | 0 |
| SM1 | 12 | 22 | 52.19 | 22.84 | 12.40 | 10.81 |
| SM2 | 12 | 22 | 48.27 | 20.45 | 3.95 | −0.77 |
| SM3 | 12 | 22 | 46.40 | 20.12 | −0.07 | −2.39 |
| SM4 | 26 | 22 | 47.40 | 19.86 | 2.08 | −3.65 |
| SM5 | 12 | 39 | 54.62 | 16.78 | 17.63 | −18.59 |

leads to a wrong estimation of the position of the "optimal" airfoil while SM5 underestimates the performance of the worst profile.

The last two properties, i.e. the capability to preserve the monotonicity of the dataset and to correctly identify the best/worst candidates, are crucial aspects in surrogate-based optimization (SBO), so that models SM2 and SM3 seem to be more suitable for this purposes. A qualitative comparison is also proposed at airfoil level. Fig. 7a and b show the pressure coefficient distribution as obtained from CFD computation and surrogate models SM1, SM2 and SM3. Two airfoils are compared: the most efficient (airfoil ID 12) and the least efficient (airfoil ID 22). Airfoil ID 12 is featured with a double shock structure on the upper surface which is not captured by SM1, partially captured by SM2 and fully captured by SM3. On the other hand, a strong shock wave and a shock-induced separation feature the aerodynamics of airfoil ID 22, but all models show a pretty good prediction for this case. This outwardly strange behaviour is probably motivated by the fact that the training database contains more than one CFD solutions which present aerodynamic features similar to ID 22, so that even less accurate models provide a satisfying prediction. On the contrary, airfoil ID 12 represents a quite unique sample.

POD models accuracy is also evaluated and compared in terms of the point-to-point snapshot error given by Eq. (18). The computation of the snapshot error includes both the mesh points coordinates and the flow field values. Fig. 8 shows the results for each snapshot belonging to the validation plan (again ranging from 1 to 50). The error index is plotted in logarithmic scale. It comes out that, in strong transonic conditions, training a POD model on

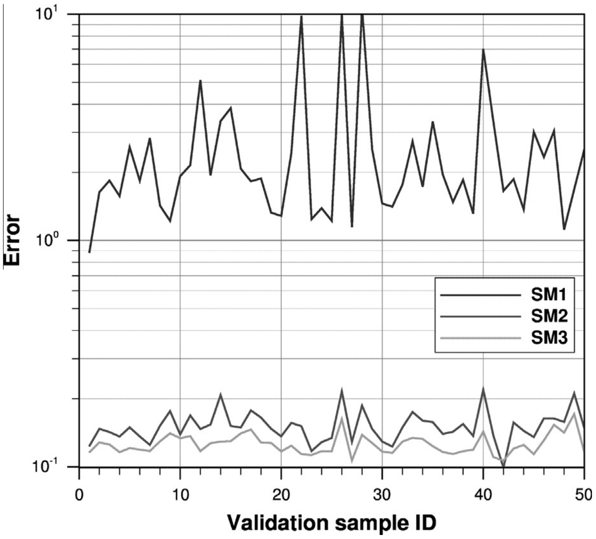


**Fig. 8.** Snapshot error prediction.

the full CFD domain (SM1) would lead to misleading results in the prediction phase, as the model would not be able to catch the highly non linear trends which characterize these kind of flows. Indeed, the high number of POD modes required and the low goodness-of-fit performance suggest that further modelling is needed to optimize the computation of the basis vectors and modal coefficients in transonic aerodynamics. In the following sections, we will introduce some adaptive sampling concepts to globally improve the reduced order models predictions.

A final comparison is possible in terms of POD model accuracy *versus* computational time and cells saving. In particular, the R-squared prediction error can be taken as a measure of the model accuracy, while the time saving index *TS* and the cells saved index *CS* are defined as

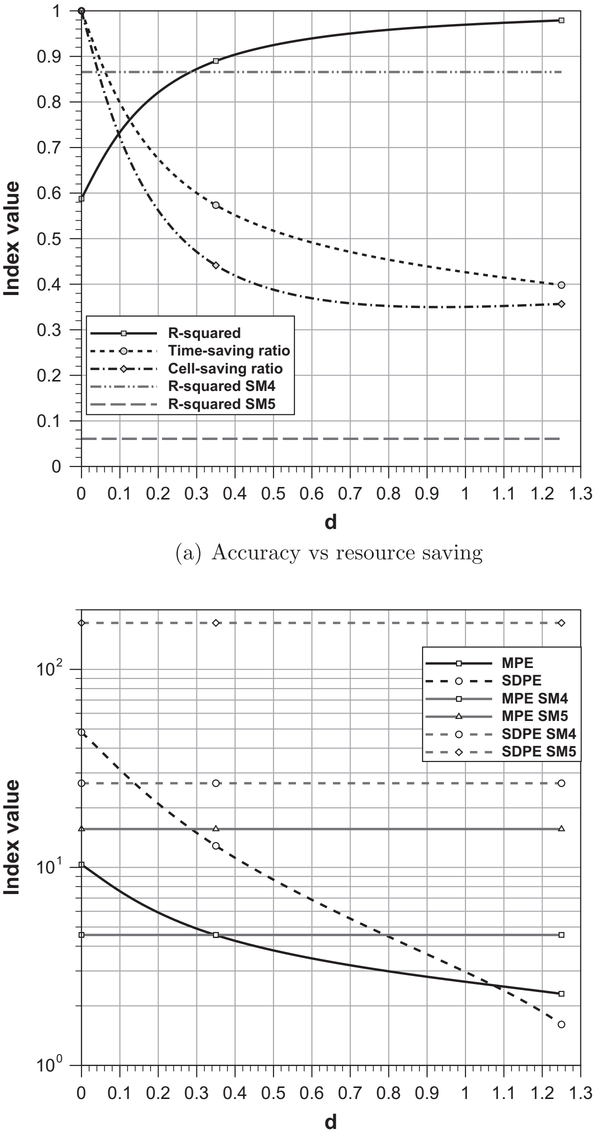
$$TS = \frac{T_{FULL} - T_{SM}}{T_{FULL}} \tag{19}$$

$$CS = \frac{N_{SM}}{N_{FULL}} \tag{20}$$

where $T$ and $N$ are respectively the computational time for 1000 CFD iterations and the number of solved computational cells. The subscripts *FULL* and *SM* refer to the full grid CFD computation and the CFD computation on the smaller FOM domain. In Fig. 9a the three indices (*R*-squared, *TS* and *CS*) are plotted against the distance *d* of the FOM/ROM interface from the airfoil leading edge. It shows that a clear trade-off exists between accuracy and time/cells saving and provides useful guidelines to tailor the choice of the best POD model to the basic requirements of the target application. For instance, if the target is to do a pre-screening of the objective space, one could use a faster and less accurate POD model which however guarantees the preservation of the physics. Fig. 9b proposes a comparison between surrogate models in terms of the mean percentage error (MPE) and its standard deviation (SDPE). The plot shows a graphical picture of the results in Table 3 concerning the POD models accuracy with moving FOM/ROM interface and the comparison with more classical meta-models.

## 5. Strategies for improving the POD model quality: adaptive sampling

The main conclusion of the previous section was that, provided an initial ensemble of flow solutions, feeding the POD model with the flow fields over the whole computational domain leads to inaccurate prediction in the validation phase. In other words,



(a) Accuracy vs resource saving

(b) Percentage errors

**Fig. 9.** Surrogate models performance as a function of the FOM/ROM interface positioning.

SM1 model showed poor results in the evaluation of design sites not belonging to the initial ensemble. On the other hand, such a POD-based model is the most efficient computationally as it does not require any CFD evaluation in the prediction phase. Indeed, one of the key point of the research is to recover the accuracy issues by "optimally" selecting the training candidates. In the proposed example, we selected 180 sites to *a priori* sample the 16-dimensional design space, but in principle we do not have any information about the appropriate size and locations of the sample points. Intuitively, we would like to have a sampling strategy which would fill the space in an efficient manner and would allocate more points in regions of the design space where the simulation response is strongly non-linear or it is likely to find an optimum. In the industrial practice, this would mean, given a computational budget, to improve the quality of the POD surrogate by "intelligently" choosing the training samples. Conversely, given a POD model with a certain quality level, the rationale would be to reach the same accuracy with less high-fidelity computations.

Adaptive sampling strategies can be properly designed to account for these requirements by means of the so-called "in-fill" criteria. While *a priori* sampling techniques do not use any information about the model prediction, adaptive techniques incrementally select new sampling points by exploiting the input/output relation observed at the previous stages. Hence, some adaptive DoE strategies for POD-based reduced order models are proposed. The main reference are the works by Goblet and Lepot [47] and Sainvitu et al. [48]. Generally speaking, with reference to the nomenclature used in previous sections and to Eq. (11), the quality of the POD/ROM models basically depends on: (1), the quality of the modal basis $\{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_M\}$ and (2), the quality of the RBF models built on the modal coefficient $\{\alpha_1(\mathbf{w}), \alpha_2(\mathbf{w}), \ldots, \alpha_M(\mathbf{w})\}$. Indeed, they both depend on the choice of the snapshot dataset. In the following, two methods are proposed to properly balance the improvement of the POD model and the space-filling properties. Both methods are based on the leave-one-out cross-validation technique.

### 5.1. Method 1: improvement of the modal basis

The first method is aimed at improving the modal basis, which represents the core of the POD modelling. Indeed, given a POD model built on a snapshot ensemble $\{\mathbf{s}(w_1), \mathbf{s}(w_2), \ldots, \mathbf{s}(w_M)\}$, the aim is to find a new point $\mathbf{w}_{\text{new}}$ in the design space so that the new POD model basis, built on the new set $\{\mathbf{s}(w_1), \mathbf{s}(w_2), \ldots, \mathbf{s}(w_M), \mathbf{s}(w_{\text{new}})\}$, will provide for improved predictions and better exploration of the design space at the same time. For the sake of clarity, the superscript $^{-j}$ indicates that the modelled element (basis vector, coefficient model, SVD matrices) has been obtained by means of the leave-one-out process, i.e. by removing the $j$th sample from the training set and re-computing the model. The first step is to estimate the relative influence of the $j$th snapshot on the modal basis as:

$$I_b^r(\mathbf{w}_j) = \frac{I_b(\mathbf{w}_j)}{\sum_{k=1}^{M} I_b(\mathbf{w}_k)} \tag{21}$$

where

$$I_b(\mathbf{w}_j) = \sum_{i=1}^{M} \sigma_i \left( \frac{1}{|(\boldsymbol{\phi}_i, \boldsymbol{\phi}_i^{-j})|} - 1 \right) \tag{22}$$

is the influence of the $j$th snapshot on the modal basis, $\boldsymbol{\phi}_i^{-j}$ is the $i$th column vector of $\mathbf{U}^{-j}$ and $\mathbf{U}^{-j}\boldsymbol{\Sigma}^{-j}\mathbf{V}^{-j^T}$ is a thin SVD of the reduced deviation matrix

$$\mathbf{P}^{-j} = (\, \mathbf{s}_1 - \bar{\mathbf{s}} \quad \mathbf{s}_2 - \bar{\mathbf{s}} \quad \cdots \quad \mathbf{s}_{j-1} - \bar{\mathbf{s}} \quad \mathbf{0} \quad \mathbf{s}_{j+1} - \bar{\mathbf{s}} \quad \cdots \quad \mathbf{s}_M - \bar{\mathbf{s}}\, )$$

The scalar product $(\boldsymbol{\phi}_i, \boldsymbol{\phi}_i^{-j})$ gives the projection of one of the two vectors on the other, hence, if they are almost orthogonal, the quantity will be nearly zero and the influence term will go to infinity, while if they are almost parallel, the influence contribution will be almost zero. The relative influence is normalized with the sum of the influence of the whole set of snapshots and weighted with the singular values $\sigma_i$ as they reflect the importance of each mode with respect to the whole basis. Computing these quantities would require the solution of $M$ thin SVD of $N \times M$ matrices where $N = s \times q$ (number of flow variables × number of mesh points) can be huge as it is related to the dimension of the data set. As detailed in Goblet and Lepot [47], it is possible to get a cheaper evaluation by computing $M$ thin SVD on $M \times M$ matrices.

Once known the relative influence of each snapshot on each modal basis vector, we need to decide where and how to choose the new sampling point. From a theoretical point of view, we would need to sample "near" design sites where the relative influence is higher. But "how near" is still not known. Moreover, choosing a new sample near a known design site would prevent from exploring undiscovered regions. To address the first issue, the design space is heavily sampled with a new LHS technique, e.g. a hundred times the dimension $t$ of the design space. Then, the Euclidean distance of each new sampled point $\mathbf{y}_i$, $i = 1, \ldots, l = 100t$ from each of the snapshot sites $\mathbf{w}_k$, $k = 1, \ldots, M$ is computed and, for each $\mathbf{y}_i$, the distance from the nearest snapshot $\mathbf{w}_{\bar{k}}$ is stored as $d(\mathbf{w}_{\bar{k}}, \mathbf{y}_i)$. The second question is more subtle as it concerns the trade-off between local accuracy and global design space exploration. Knowing the relative distances between new sampling points and snapshot sites, a different need would be to sample far away from the known points in order to potentially enrich the global prediction of the POD model. The idea is to combine the information about the snapshot's relative influence and the "nearest" distance in order to satisfy both requirements. Hence, for each new candidate $\mathbf{y}_i$ a potential of enrichment $V_\phi$ can be identified by weighting the distance $d(\mathbf{w}_{\bar{k}}, \mathbf{y}_i)$ with the relative influence of the $\bar{k}^{th}$ snapshot on the modal basis:

$$V_\phi(\mathbf{y}_i) = d(\mathbf{w}_{\bar{k}}, \mathbf{y}_i) \mathrm{I}_\mathrm{b}^\mathrm{r}(\mathbf{w}_{\bar{k}}) \tag{23}$$

Finally, a new sample point is selected at $\mathbf{w}_{\text{new}} = \arg\max_{\mathbf{y}_i} V_\phi(\mathbf{y}_i)$.

### 5.2. Method 2: improvement of the modal coefficients

As already described in Section 3.2.3, the POD modal coefficients are provided with a global approximation through Radial Basis Functions. The second adaptive method is then conceived to improve the quality of these RBF models. Given a POD model built on a snapshot ensemble $\{\mathbf{s}(\mathbf{w_1}), \mathbf{s}(\mathbf{w_2}), \ldots, \mathbf{s}(\mathbf{w_M})\}$, the aim is to find a new point $\mathbf{w}_{new}$ in the design space so that the new modal coefficient models, built on the new set $\{\mathbf{s}(\mathbf{w_1}), \mathbf{s}(\mathbf{w_2}), \ldots, \mathbf{s}(\mathbf{w_M}), \mathbf{s}(\mathbf{w}_{\text{new}})\}$, will provide for improved predictions and better exploration of the design space at the same time. Two sub-strategies are proposed: the first aims at improving the worst modal coefficient, the second is designed to improve all coefficients simultaneously.

#### 5.2.1. First sub-strategy

This strategy is applied when one of the coefficient model $\{\alpha_1(\mathbf{w}), \ldots, \alpha_M(\mathbf{w})\}$ exhibits low quality with respect to the others. For the sake of clarity, the symbol $\alpha_i^{-j}(\mathbf{w})$ indicates the RBF model built on the $i$th modal coefficients by leaving the $j$th snapshot out, i.e. based on the following correspondence:

$$\{\mathbf{w}_1, \ldots, \mathbf{w}_{j-1}, \mathbf{w}_{j+1}, \ldots, \mathbf{w}_M\} \rightarrow \{\alpha_i(\mathbf{w}_1), \ldots, \alpha_i(\mathbf{w}_{j-1}), \alpha_i(\mathbf{w}_{j+1}), \ldots, \alpha_i(\mathbf{w}_M)\} \tag{24}$$

First of all, we need to evaluate the quality of the coefficient model, possibly taking into account the relative importance of the mode itself. This is done by computing the Pearson correlation coefficient of the model $\alpha_i(\mathbf{w})$ as

$$\mathrm{R}(\alpha_i) = \frac{\mu\left(\alpha_i \alpha_i^{-j}\right) - \mu(\alpha_i)\mu\left(\alpha_i^{-j}\right)}{\sqrt{\mu(\alpha_i\alpha_i) - [\mu(\alpha_i)]^2}\sqrt{\mu\left(\alpha_i^{-j}\alpha_i^{-j}\right) - \left[\mu\left(\alpha_i^{-j}\right)\right]^2}} \tag{25}$$

where

$$\mu(\alpha_i) = \frac{1}{M}\sum_{j=1}^{M} \alpha_i(\mathbf{w}_j) \tag{26}$$

$$\mu\left(\alpha_i^{-j}\right) = \frac{1}{M}\sum_{j=1}^{M} \alpha_i^{-j}(\mathbf{w}_j) \tag{27}$$

$$\mu(\alpha_i\alpha_i) = \frac{1}{M}\sum_{j=1}^{M} \alpha_i(\mathbf{w}_j)\alpha_i(\mathbf{w}_j) \tag{28}$$

$$\mu\left(\alpha_i\alpha_i^{-j}\right) = \frac{1}{M}\sum_{j=1}^{M} \alpha_i(\mathbf{w}_j)\alpha_i^{-j}(\mathbf{w}_j) \tag{29}$$

$$\mu\left(\alpha_i^{-j}\alpha_i^{-j}\right) = \frac{1}{M}\sum_{j=1}^{M} \alpha_i^{-j}(\mathbf{w}_j)\alpha_i^{-j}(\mathbf{w}_j) \tag{30}$$

The Pearson coefficient (25) provides a statistical measure of the accuracy of the model $\alpha_i$. But, each POD mode has its own energy contribution which the quality parameter has to be related to. Therefore, a “weighted” quality of the model $\alpha_i(\mathbf{w})$ is defined as

$$\mathrm{R}_w(\alpha_i) = \frac{\sigma_i}{\sum_{j=1}^{M}\sigma_j}\mathrm{R}(\alpha_i)$$

Then, the coefficient model with the lowest weighted quality is selected and denoted with the symbol $\alpha_{\bar{i}}$. The relative influence of the $j$th snapshot on the $\bar{\imath}^{th}$ modal coefficient is defined as the absolute error of the model when leaving the $j$th snapshot out:

$$\mathrm{I}_{\alpha_{\bar{i}}}(\mathbf{w}_j) = |\alpha_{\bar{i}}(\mathbf{w}_j) - \alpha_{\bar{i}}^{-j}(\mathbf{w}_j)|$$

From here on, the procedure is the same as for the modal basis improvement: the design space is heavily sampled with a LHS technique, e.g. 100 times the dimension $t$ of the design space. Then, the Euclidean distance of each new sampled point $\mathbf{y}_i$, $i = 1, \ldots, l = 100t$ from each of the snapshot sites $\mathbf{w}_k$, $k = 1, \ldots, M$ is computed and, for each $\mathbf{y}_i$, the distance from the nearest snapshot $\mathbf{w}_{\bar{k}}$ is stored as $d(\mathbf{w}_{\bar{k}}, \mathbf{y}_i)$. Hence, for each new candidate $\mathbf{y}_i$ a potential of enrichment $V_{\alpha_{\bar{i}}}$ (with respect to the worst coefficient model) can be identified by weighting the distance $d(\mathbf{w}_{\bar{k}}, \mathbf{y}_i)$ with the relative influence of the $\bar{k}$th snapshot on the $\bar{\imath}$th modal coefficient:

$$V_{\alpha_{\bar{i}}}(\mathbf{y}_i) = d(\mathbf{w}_{\bar{k}}, \mathbf{y}_i)\mathrm{I}_{\alpha_{\bar{i}}}(\mathbf{w}_{\bar{k}}) \tag{31}$$

Finally, a new sample point is selected at $\mathbf{w}_{\text{new}} = \arg\max_{\mathbf{y}_i} V_{\alpha_{\bar{i}}}(\mathbf{y}_i)$.

The leave-one-out procedure allows to determine the quantities $\alpha_i^{-j}(\mathbf{w})$, but its cost is high as it would apparently require the building of $M \times M$ new approximations. However, when using RBF network interpolators for POD coefficient models, the leave-one-out procedure can be performed at low-cost by using the efficient formula provided by Rippa [39]. Indeed, the author showed that computing the term $|\alpha_i(\mathbf{w}_j) - \alpha_i^{-j}(\mathbf{w}_j)|$ can be done with no extra cost as the required model parameters have been already computed during the construction of $\alpha_i(\mathbf{w})$.

#### 5.2.2. Second sub-strategy

This sub-strategy is used when the quality of all coefficient models are comparable and it is very similar to the improvement of the modal basis. The relative influence of the $j$th snapshot on the whole set of coefficient models is computed as

$$I_c^r(\mathbf{w}_j) = \frac{I_c(\mathbf{w}_j)}{\sum_{k=1}^{M}\mathrm{I}_c(\mathbf{w}_k)} \tag{32}$$

where $I_c(\mathbf{w}_j)$ is the influence of the $j$th snapshot on the coefficient models weighted with the corresponding singular values:

$$I_c(\mathbf{w}_j) = \sum_{i=1}^{M}\sigma_i I_{\alpha_i}(\mathbf{w}_j) = \sum_{i=1}^{M}\sigma_i|\alpha_i(\mathbf{w}_j) - \alpha_i^{-j}(\mathbf{w}_j)| \tag{33}$$

It must be noted that now the worst $\alpha_i$ model does not have to be computed as the aim is to improve the coefficients all together. Therefore, the design space is heavily sampled with a LHS technique, e.g. 100 times the dimension $t$ of the design space. Then, the Euclidean distance of each new sampled point $\mathbf{y}_i$, $i = 1, \ldots, l = 100t$ from each of the snapshot sites $\mathbf{w}_k$, $k = 1, \ldots, M$ is computed and, for each $\mathbf{y}_i$, the distance from the nearest snapshot $\mathbf{w}_{\bar{k}}$ is stored as $d(\mathbf{w}_{\bar{k}}, \mathbf{y}_i)$. Hence, for each new candidate $\mathbf{y}_i$ a potential of enrichment $V_\alpha$ (with respect to all POD coefficients) can be identified by weighting the distance $d(\mathbf{w}_{\bar{k}}, \mathbf{y}_i)$ with the relative influence $I_c^r(\mathbf{w}_j)$

$$V_\alpha(\mathbf{y}_i) = d(\mathbf{w}_{\bar{k}}, \mathbf{y}_i)\mathrm{I}_\mathrm{c}^\mathrm{r}(\mathbf{w}_{\bar{k}}) \tag{34}$$

Finally, a new sample point is selected at $\mathbf{w}_{\text{new}} = \arg\max_{\mathbf{y}_i} V_\alpha(\mathbf{y}_i)$.

## 6. Surrogate-based evolutionary optimization

The POD surrogates as well as the adaptive sampling techniques described in the previous sections have been combined within an evolutionary optimization loop and used to validate the presented methodology in transonic flow. Several approaches are proposed, differing for the key ingredients of the methodology: the construction of the POD model (full/zonal approach), the strategy chosen to compute the training sample (*a priori*, auto-adaptive) and the strategy to exploit the optimization results (single optimization, real-time updating). The optimization approaches share the same target, i.e. to improve the performance of the scaled RAE2822 airfoil.

### 6.1. The surrogate-based shape optimization framework

The workflow of the surrogate-based shape optimization (SBSO) is depicted in Fig. 10. Basically, it consists of an *a priori* design of experiment module (a Latin Hypercube sampler), the CST parameterization module, an in-house developed automatic mesh generator, the ZEN CFD flow solver, the POD/ROM module, which also encloses the adaptive sampling techniques, and the in-house ADGLIB genetic optimization library [49,50]. The integrating platform is called GAPOD, which controls the various modules through internal or external calls. For example, the geometric modeller, the mesh generator and the flow solver can be easily replaced as they appear in the text input files and they are invoked through system calls. On the other hand, the POD module and the optimizer are software library modules directly linked with the launcher. Moreover, the POD module can be activated with two modes: a “standalone” mode and a “coupled/zonal” mode [51]. The first one is switched on when the surrogate prediction directly provides the objective/constraint functions, e.g. without any further call of the high-fidelity code. This is, for example, the case when the POD model is trained on the whole flow field. Objective/constraint functions related to the aerodynamic performance can be retrieved by using Eqs. (13) and (14). The “coupled/zonal” mode is designed to apply the zonal CFD/POD approach. Once computed: (1) a POD model and (2) an airfoil geometry from a selected design site, the POD model is evaluated to provide the ROM (outer) solution in the form of a predicted snapshot. Then, the full volume mesh around the selected geometry is calculated, the volume mesh in the FOM (inner) domain is extracted and written in the ZEN code file format. A specifically designed procedure detects the FOM/ROM interface in the snapshot definition and writes a boundary condition file suitable for the ZEN solver. Finally, the flow solver is launched to compute the flow field in the FOM inner domain with the surrogate-derived boundary condition. Detailed information about the evaluation of the ZEN flow solver on the RAE 2822 case can be found in Catalano and Amato [52].

The workflow in Fig. 10 shows two internal cycles: the adaptive sampling and the real-time optimization. The first one (adaptive updating), based on the techniques described in Section 5, is aimed at improving the surrogate model before the optimization phase by providing a new design candidate $\mathbf{w}_{\text{new}}$ to be added to the ensemble database. The condition to exit from this internal loop is based either on pre-defined levels of improvement for POD modal basis/coefficients or on computational budget considerations. The second cycle (real-time updating) is related to the inclusion of the surrogate-based optimal design sites $\mathbf{w}_{\text{opt}}$ into the POD ensemble database to accurately explore design space regions which may potentially enclose the “true” minima. The loop terminates either when the residual of the objective function of the predicted optima falls below a pre-defined threshold or when the required number of iterations is performed.

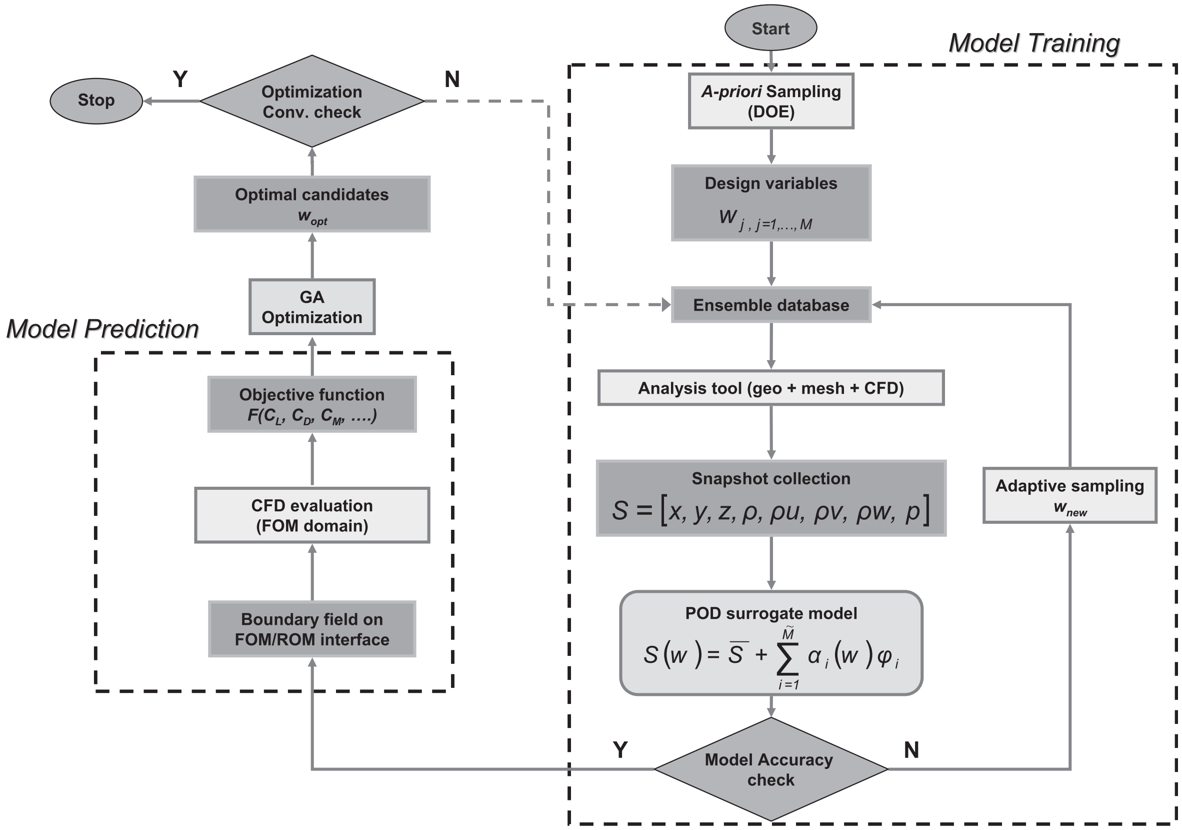


**Fig. 10.** Workflow of CFD/POD-based genetic optimization.

### 6.2. Problem definition

The definition of the geometry parameters, parameterization technique and design variable ranges are those described in the previous sections. The design point is summarized in Table 2. Here, we define the airfoil shape optimization problem in terms of objective/constraint functions specification as

$$\begin{aligned} &\underset{\mathbf{w}\in DW\subset\mathbb{R}^{16}}{\text{minimize}} \; -\frac{C_l}{C_d} \\ &\text{subject to } \left(\frac{t}{c}\right)_{max} = 0.14 \\ &\qquad C_l \geqslant 0.5 \\ &\qquad C_m \geqslant -0.05 \\ &\qquad C_m \leqslant 0.05 \end{aligned}$$

In other words, the goal is to maximize the aerodynamic efficiency $\frac{C_l}{C_d}$ while keeping a minimum level of lift generation ($C_l \geqslant 0.5$) and of pitching moment controllability ($|C_m| \leqslant 0.05$). Moreover, a geometric constraint is added in order to set the airfoil maximum thickness-to-chord ratio $\frac{t}{c}$ at 14%: this constraint is implicitly treated within the parameterization. The constraint functions are actually treated as quadratic penalties, hence the constrained optimization is transformed into the following unconstrained problem:

$$\begin{aligned} &\text{minimize}_{\mathbf{w}\in DW\subset\mathbb{R}^{16}} -\frac{C_l}{C_d} + K[\min(C_l - 0.5, 0)]^2 \\ &\quad + K[\min(C_m + 0.05, 0)]^2 + K[\min(-C_m + 0.05, 0)]^2 \qquad (35) \end{aligned}$$

where $K$ is a constant weight (equal to $10^4$) which amplifies the relative importance of possible constraint violations. For instance, a unit penalty will be applied to the objective function in the case of an airfoil having a pitching moment of ±0.06.

### 6.3. Optimization strategies and set up

Several optimization approaches have been set up and tested in order to possibly cover all the issues concerning surrogate/ROM training and prediction. Table 5 summarizes the characteristics of each optimization in terms of: fitness evaluator, optimization algorithm, POD energy threshold (when using POD as surrogate), high-fidelity computational budget, i.e. the total number of computations with the ZEN RANS solver during the optimization

**Table 5**
Optimization approaches.

| Opt Tag | Fitness evaluator | Optimizer | POD energy | Budget hi-fi | $M_{\text{apr}}$ | $M_{\text{adp}}$ | $M_{\text{opt}}$ |
|---|---|---|---|---|---|---|---|
| DGA | ZEN | ADGLIB | – | 9600 | 0 | 0 | 0 |
| FPGA1 | standalone POD | ADGLIB | 85% | 180 | 180 | 0 | 0 |
| FPGA2 | standalone POD | ADGLIB | 95% | 180 | 180 | 0 | 0 |
| FPGA3 | standalone POD | ADGLIB | 99% | 180 | 180 | 0 | 0 |
| MPGA1 | zonal POD | ADGLIB | 95% | 180 | 180 | 0 | 0 |
| MPGA2 | zonal POD | ADGLIB | 99% | 180 | 180 | 0 | 0 |
| KGA | Kriging | Dakota SOGA | – | 190 | 180 | 0 | 10 |
| EGO | Kriging | Dakota EGO | – | 553 | 153 | 400 | 0 |
| AFPGA1 | standalone POD | ADGLIB | 99% | 96 | 32 | 16 | 48 |
| AFPGA2 | standalone POD | ADGLIB | 99% | 96 | 16 | 32 | 48 |
| AFPGA3 | standalone POD | ADGLIB | 99% | 96 | 4 | 44 | 48 |
| AMPGA1 | zonal POD | ADGLIB | 99% | 112 | 8 | 56 | 48 |
| AMPGA2 | zonal POD | ADGLIB | 99% | 96 | 8 | 40 | 48 |

process, number of *a priori* LHS samples $M_{\text{apr}}$, number of adaptively added samples $M_{\text{adp}}$ and number of surrogate-based optima $M_{\text{opt}}$ which are iteratively added to the ensemble database. It must be noted that not all the optimization strategies use POD as surrogate: in particular, optimizations tagged as KGA and EGO have been performed by respectively using a Kriging method as fitness evaluator and the EGO (Efficient Global Optimization) algorithm [1], based on Kriging and Expected Improvement evaluation, to compute new optimal samples. The EGO algorithm represents one of the modern state-of-art methods in surrogate-based global optimization. In the following with the term "true" we will indicate the results obtained with the ZEN CFD solver as it is adopted as the reference high-fidelity simulation tool. Each optimization method is here described in detail:

DGA – a plain, brute-force genetic optimization with the full high-fidelity solver ZEN called as fitness evaluator;

FPGA1 – a surrogate-based optimization where the aerodynamic analysis is carried out through a POD model built on the complete flow field of a set of 180 initial samples. This case corresponds to the POD-driven "standalone" mode and the surrogate POD evaluator is the one presented as SM1. No zonal approach is used. The POD energy content is 85%. The snapshot size *N* is 201,488;

FPGA2, FPGA3 – same as FPGA1, but the POD models are defined by increasing the energy content (95% and 99%, respectively);

MPGA1 – a surrogate-based optimization where the zonal CFD/POD model is trained on the initial design space sampling (180 snapshots) and adopted as objective function evaluator throughout the optimization cycle. The FOM domain is defined at a distance $d = 1.25$ chord length from the airfoil leading edge. The POD model used here has been already validated as SM3 in previous sections. The POD energy threshold is set at 95%. The snapshot size is 75,232;

MPGA2 – same as MPGA1, but the POD energy content is increased up to 99%;

KGA – a surrogate-based optimization where a Kriging metamodel, built on the objective function, is coupled to the genetic optimization. The JEGA library [53] from the DAKOTA package was used for optimization purposes. In particular, the Single-Objective Genetic Algorithm (SOGA) was used to perform optimization on a single objective function with general constraints. The Kriging is initially trained on the 180 samples dataset. Then, a surrogate-based iterative optimization scheme is performed consisting in the following steps:

1. adding points to the sample set used to create the surrogate;
2. rebuilding the surrogate;
3. performing a global optimization on the new surrogate;
4. finding of minimizers of the surrogate model;
5. passing a selected optimal subset (in the present case, just the optimum candidate) to the next iteration;
6. re-evaluation of the surrogate points with the "true" (CFD) model;
7. adding to the set of points upon which the next surrogate is constructed and return back to 1.

This procedure offers a more accurate surrogate to the minimizer at each subsequent iteration, presumably driving to optimality quickly. In the present optimization, 10 SBO iterations are performed.

EGO – the particular response surface used is a Gaussian process (GP). The GP allows one to calculate the prediction at a new input location as well as the uncertainty associated with that prediction. The key idea in EGO is to maximize the Expected Improvement Function (EIF), as previously described in Section 2. The general procedure used here is:

- build an initial Gaussian process model of the objective function on a initial dataset. For the sixteen variable case, 153 design sites are automatically generated by the algorithm, hence we do not use the usual dataset made of 180 samples.
- find the point that maximizes the EIF. If the EIF value at this point is sufficiently small, stop.
- evaluate the objective function at the point where the EIF is maximized. Update the Gaussian process model using this new point. Return to the previous step.

The EGO optimization represents the most interesting algorithm to compare with the newly developed POD-based approaches, as it embeds the concept of adaptivity and trade-off between design space exploration and surrogate model exploitation.

AFPGA1, AFPGA2, AFPGA3 – the surrogate model employed is the same as FPGA3, but the training method is different and an adaptive sampling strategy is added. In particular, we decided to follow a different approach: we want to check if, with a limited computational budget, we get better results by adaptively training the POD model. Hence, we split the surrogate training phase in three contributions: an *a priori* contribution, sampling the design space with the LHS technique and producing $M_{\text{apr}}$ samples; an iterative, adaptive sampling (method 1 from Section 5) aimed at improving the modal basis and enriching the ensemble dataset with $M_{\text{adp}}$ samples; a series of $M_{\text{opt}}$ genetic optimizations, each producing an optimal candidate to update the ensemble and recompute the surrogate. The last phase will be also called real-time updating. The three strategies differ for the relative amount of these three contributions as highlighted in Table 5, keeping fixed the total computational budget. The POD energy content is 99%. The snapshot size *N* is 201,488;

AMPGA1 – the surrogate model employed is the SM2. The FOM/ROM interface is defined at $d = 0.35$ chord length from the airfoil leading edge. However, the training method is different as it embeds *a priori*, auto-adaptive (method 1 from Section 5) and optimal samples as described earlier. The POD energy content is 99%. The snapshot size *N* is 91,792;

AMPGA2 – the surrogate model employed is again the SM3, but it differs from MPGA2 because the training method embeds *a priori*, auto-adaptive (method 1 from Section 5) and optimal samples as described before. The POD energy content is 99%. The snapshot size *N* is 75,232.

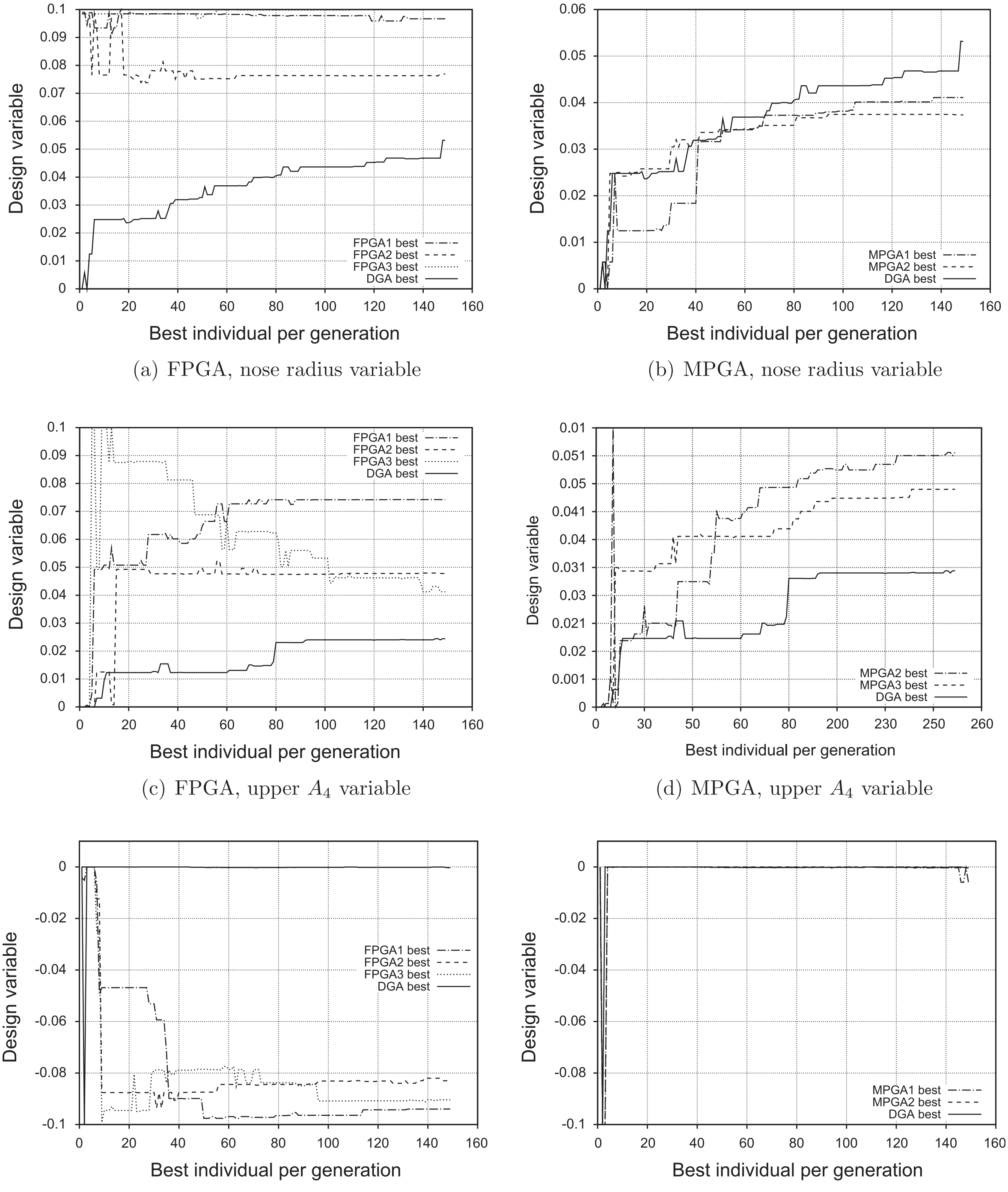
The optimization set up is the same for all the approaches, except for AMPGA1 and AMPGA2. A population of 64 individuals is let evolve for 150 generations with 80% bit crossover rate and 2% bit mutation rate. The genetic evolution is repeated every time a new optimal sample has to be added to the ensemble. Hence, a total number of 9600 evaluations are required for each optimization process. The set up of AMPGA1 and AMPGA2 slightly differ because the adopted surrogate models are more expensive (Ref. Fig. 9a). In order to increase the frequency of model updating stages, a population of 48 individuals is let evolve for just 10 generations and the process is repeated 48 times to iteratively provide new optimal samples. The new feature is that each optimization step is a restart of the previous one with re-evaluation of the population candidates as the surrogate model has been updated meanwhile. In other words, the idea is to update the surrogate model more frequently (after just 10 GA generations instead of 150) even if with smaller amounts of improvement (10 generations are not enough to converge the GA).

The key to understanding the surrogate-based optimizations in Table 5 is not the number of high-fidelity evaluations, but the quality of the solution that a method is able to obtain at its best. Obviously it is very difficult to make a very precise comparison employing this criterion, but the goal is simply to get guidelines, even if approximate, to understand what can be expected from the use of one method over another in the applications.

By looking at the details of the SBO approaches described so far, it seems quite natural to divide them in two main classes: the non-adaptive (FPGAx, MPGAx), i.e. those without any adaptation/real-time updating, and the adaptive optimizations (KGA, EGO, AFPGAx, AMPGAx). Consequently, the presentation of the obtained results will follow this logical sequence.

### 6.4. Non-adaptive optimization results

Fig. 11(a, b), (c, d), and (e, f) shows the convergence history of some design variables (leading edge radius, upper $A_4$ and lower $A_1$) as a function of the progressing generations. It is clearly evident how MPGAs models behave quite well as they approach the design values found by the DGA optimization. The plots in Fig. 11e and f propose an interesting feature: the FPGA optimizations push the lower $A_1$ variable towards the lower bound (−0.1), while the “true” and MPGA models converge towards the variable upper boundary at 0.



(a) FPGA, nose radius variable

(b) MPGA, nose radius variable

(c) FPGA, upper $A_4$ variable

(d) MPGA, upper $A_4$ variable

(e) FPGA, lower $A_1$ variable

(f) MPGA, lower $A_1$ variable

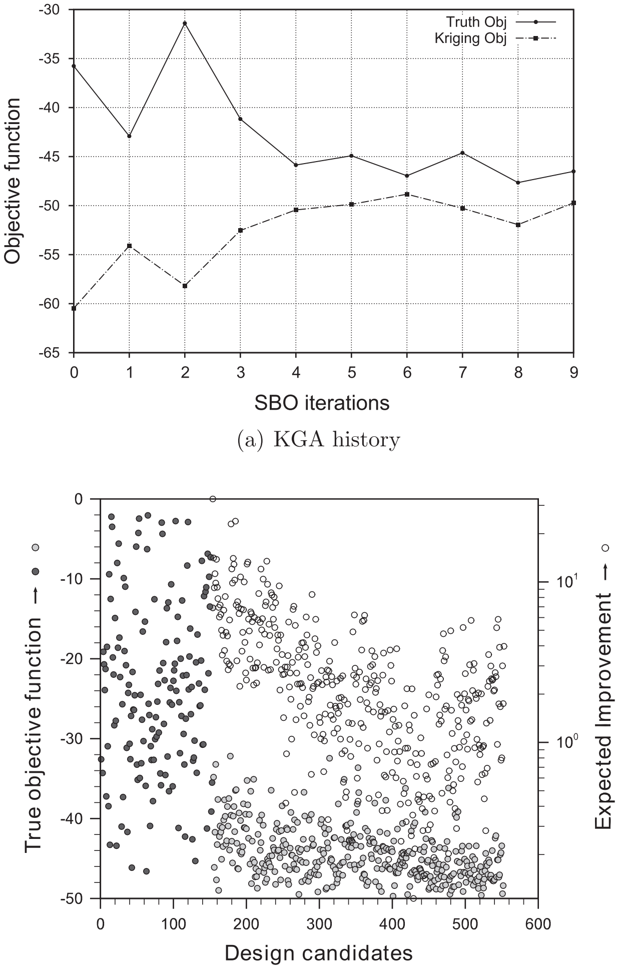
**Fig. 11.** Non-adaptive POD-driven optimization history.

### 6.5. Adaptive optimization results

Fig. 12a shows the convergence history of the iterative SBSO KGA run. As already mentioned, it is made of ten sequential surrogate optimizations: at the end of each of them, the optimal candidate is re-evaluated with the CFD solver and injected in the training set so that an updated surrogate is available. The figure compares the surrogate and “true” prediction of the optimal candidate at each iteration. After about 6–7 SBO iterations, the Kriging model has been improved enough to predict very closely to the CFD solver. Among the SBO minimizers, the ninth iteration shows the lowest “true” objective function value, so that it will be considered as the actual KGA optimum.

Fig. 12b reports the convergence history of the EGO optimization. Grey circles depict the initial DoE sampling (153 candidates), while red circles denote the subsequent 400 candidates found by maximizing the expected improvement. The graph also reports the expected improvement values in blue white-filled circles and logarithmic scale (right axis). It is clearly evident how the progressive decrease of the EIF produces a better quality of the Kriging model which in turn results in a minimization of the “true” objective function.

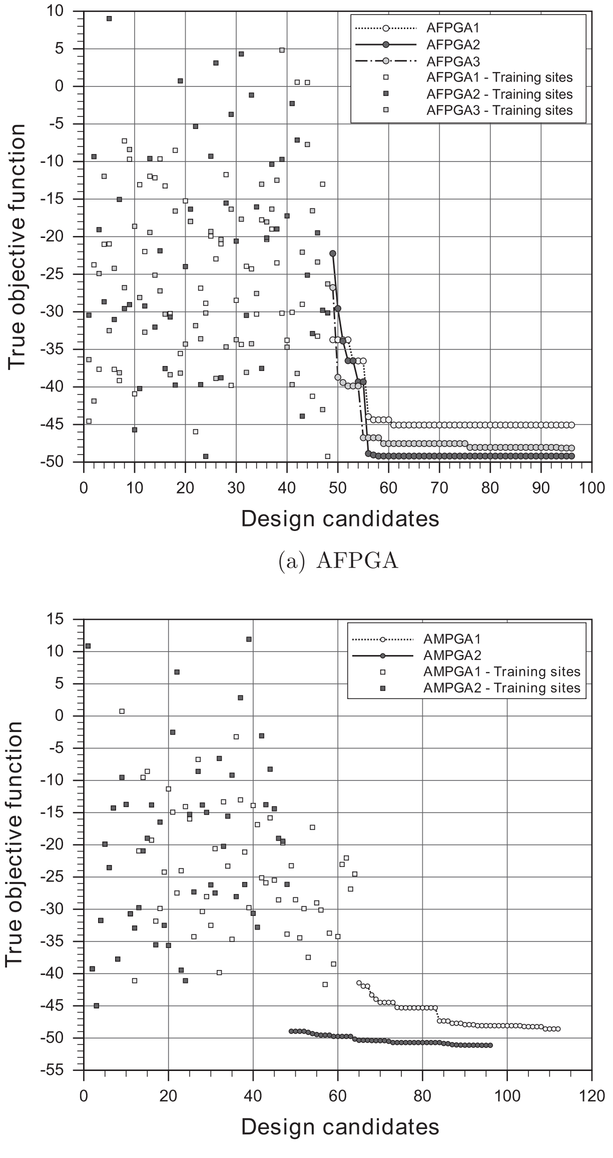
The convergence histories of the AFPGA1, AFPGA2 and AFPGA3 optimizations are reported in Fig. 13a together with the objective function values computed on the training points. In the proposed plot, each point represents a single high-fidelity evaluation, the squares depict the *a priori* and adaptive training sites while the circles connected with solid lines represent the sequence of optima from $M_{\rm opt}$ GA optimizations. It is fairly evident that the adaptive sampling is often helpful as it allows to find sub-optimal solutions even before optimization (see AFPGA1 and AFPGA2). On the other hand, this somewhat disappointing behaviour in the optimization step is due to the fact that the surrogate underestimates the objective function, thus pushing the surrogate-based optimizer to explore uninteresting design space regions. In particular, results show that the more adapted is the initial sampling, the smaller is the underestimation. Hence, the ratio $\frac{M_{\rm apr}}{M_{\rm adp}}$ should be kept low, but another important feature is related to the AFPGA3 method: it shows that, by lowering the ratio $\frac{M_{\rm apr}}{M_{\rm adp}}$ too much (up to 0.09), the performance of the method deteriorates as the final AFPGA3 optimum is worse than the previous ones. Indeed, leaving too much room for adaptive criteria seems to produce a sampling with very poor exploratory capabilities.



(a) KGA history

(b) EGO history

**Fig. 12.** Kriging-based optimizations.



(a) AFPGA

(b) AMPGA

**Fig. 13.** POD-based optimizations convergence history.

These considerations give an helpful hint about the right combination of *a priori* and adaptive sampling: the ratio $\frac{M_{\rm apr}}{M_{\rm adp}}$ should be kept between 0.1 and 0.5. This notable result supports the

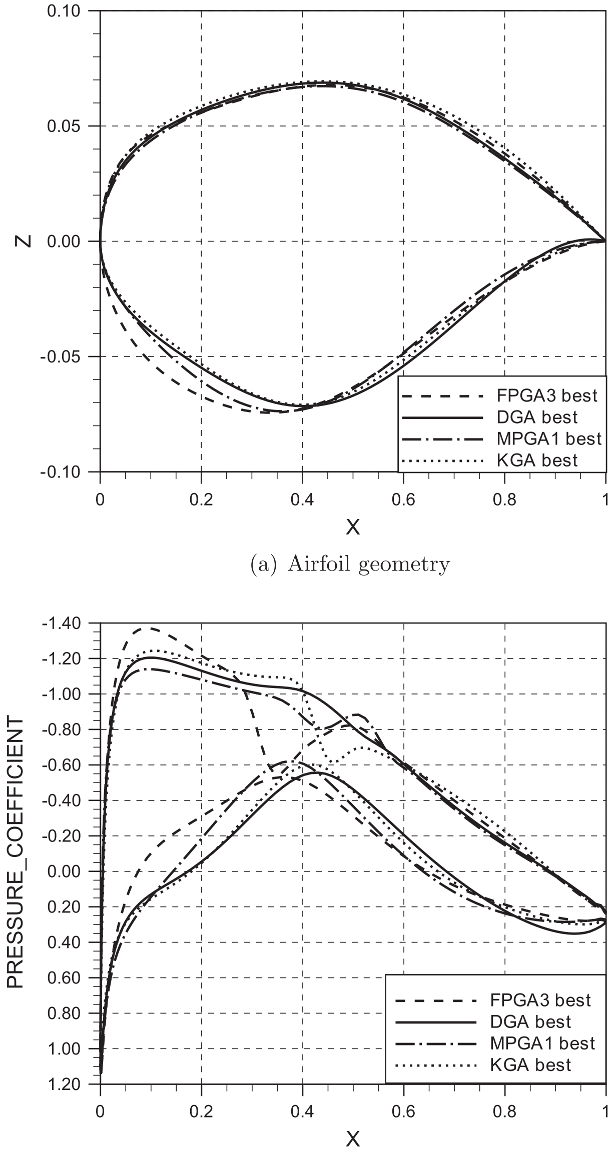
"1/3–2/3" rule proposed by Sóbester et al. [9] and assumes a deeper significance by considering that it was obtained using two different methodologies. This information is exploited in tuning the parameters for AMPGA1 and AMPGA2 optimizations. Fig. 13b shows the "true" objective functions of the training samples and of the sequence of optima candidates. Even if the AMPGA1 performs quite well, it exhibits similar characteristics of the AFPGAx optimization. On the other hand, the AMPGA2 optimum outperforms the optima seen so far and, as it will be clear in the next section, it is the only candidate to get very close to the "true" optimum, i.e. the DGA optimum.

### 6.6. Optima analysis

This section gives details about the optima computed with each of the presented methodologies. In the following, ten optimal candidates will be considered to assess the optimization results, namely the optima from run DGA, FPGA3, MPGA1, KGA, EGO, AFPGA1, AFPGA2, AFPGA3, AMPGA1 and AMPGA2. FPGA3 and MPGA1 have been selected among FPGAx and MPGAx optima because they are the closest to the high-fidelity DGA optimum. The objective function breakdown for each optimal candidate is summarized in Table 6. The table reports both the "true" data, obtained by re-computing each design with the CFD solver, and the predicted objective function as calculated by the surrogate model. Each optimum does not satisfy the pitching moment constraint because the quadratic penalty function and its weight, chosen in the problem definition, purposely do not enforce this constraint strictly to have a less stiff optimization problem. Indeed, getting precisely into the constraint boundaries would have probably penalized too much the aerodynamic efficiency, i.e. the actual objective function, while applying small penalties near a constraint boundary gives more flexibility to the search of the optimal design.

#### 6.6.1. Non-adaptive optima

Among the non-adaptive methods (KGA is here considered as non-adaptive to set a comparison), optimal design coming from MPGA1 and KGA are closer to the plain one in terms of global performance. MPGA1 optimum catches almost perfectly the DGA constraint violation, while KGA design performs even better on pitching moment but at the cost of a slightly lower aerodynamic efficiency. FPGA3 design, although using 75 POD modes, does not belong to an optimal subset but exhibits a small penalty. A more interesting comparison is proposed in Fig. 14 where the optimal airfoils shape (left-hand) and pressure coefficient (right-hand) are depicted. It should be noted that the high-fidelity DGA optimization (black curves) is able to find a shock-less configuration by properly designing the upper airfoil side. The best surrogate solution is the MPGA1, where a weak shock appears on the suction side but at a lower lift level. The plots provide to give an insight explanation of why the FPGA3 candidate shows poor aerodynamic performances: the optimal leading edge radius, as also depicted in Fig. 11a and b, is almost twice the DGA value and this feature causes an over-expansion on the suction side which in turn makes the shock wave occur more upstream and stronger. Another significant design feature is observable on the Kriging-based best candidate, with reduced rear loading to limit nose-down pitching moment and trim drag associated with the rear location of the centre of pressure. This beneficial feature is counterbalanced by the lift production increase in the fore airfoil part and, consequently, by a more pronounced pressure jump across the shock wave. In Fig. 15 the velocity components contour are compared for MPGA1 optimum as obtained from full CFD (solid grey lines) and zonal CFD/POD (dashed black lines) computations. A general agreement can be noticed, even if some slight discrepancies on the CFD/POD boundary interface still exist which may deteriorate the high-fidelity prediction.



(a) Airfoil geometry

(b) Pressure coefficient

**Fig. 14.** Non-adaptive optimal candidates comparison.

**Table 6**
Optimal candidates, obj. function breakdown.

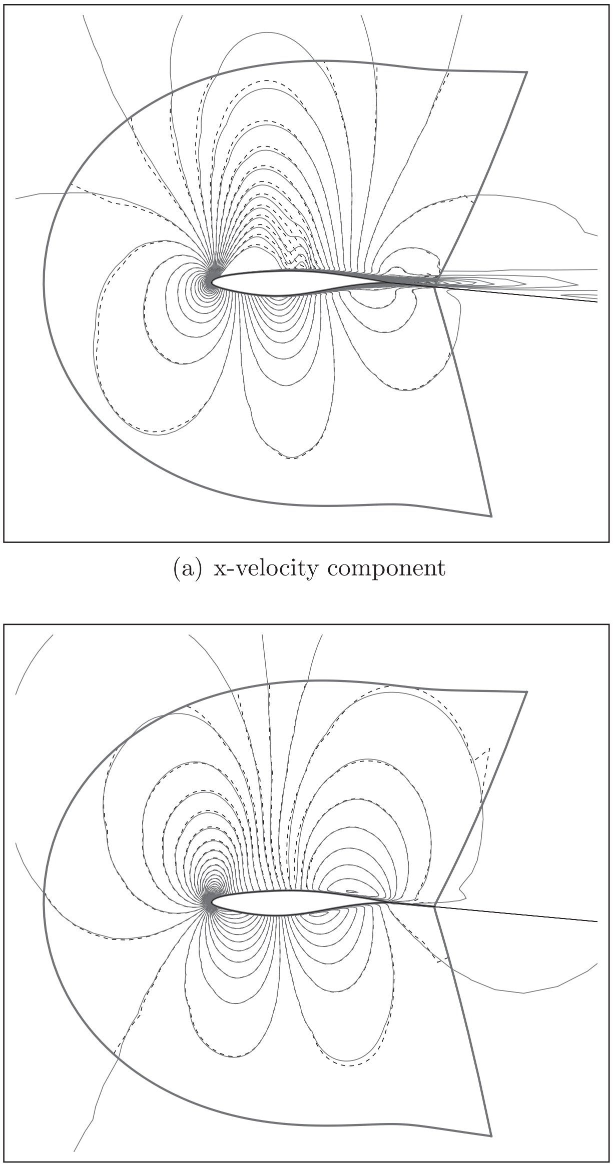
| Opt. run ID | Truth Obj. | Predicted Obj. | Penalty | $C_l$ | $C_d$ | $C_m$ |
|---|---|---|---|---|---|---|
| DGA | −51.18 | −51.18 | 1.025 | 0.619 | 0.0118 | −0.0602 |
| MPGA1 | −48.70 | −50.86 | 1.13 | 0.578 | 0.0116 | −0.0606 |
| FPGA3 | −38.33 | −73.45 | 0.608 | 0.553 | 0.0142 | −0.0578 |
| KGA | −47.65 | −51.94 | 0.585 | 0.612 | 0.0127 | −0.0576 |
| EGO | −49.71 | −49.71 | 0.530 | 0.618 | 0.0123 | −0.0573 |
| AFPGA1 | −49.24 | −47.14 | 1.12 | 0.635 | 0.0126 | −0.0606 |
| AFPGA2 | −49.20 | −52.61 | 0.551 | 0.631 | 0.0127 | −0.0574 |
| AFPGA3 | −48.13 | −47.88 | 1.29 | 0.583 | 0.0118 | −0.0614 |
| AMPGA1 | −48.58 | −44.61 | 0.567 | 0.576 | 0.0117 | −0.0575 |
| AMPGA2 | −51.13 | −50.31 | 0.947 | 0.612 | 0.0117 | −0.0597 |

#### 6.6.2. Adaptive optima

In order to highlight the undertaken improvement path, we report in Fig. 16 a correlation plot. The whole set of optima is reported in the surrogate objective – true objective plane. Two different zooming levels are set, as they reflect the non-adaptive and adaptive process: the FPGA1, FPGA2 and FPGA3 optima show very large discrepancies between true value and surrogate prediction, hence they are located very far from the line of perfect fit. However, a trend is observable as, increasing the POD energy content (i.e., passing from FPGA1 to FPGA3), the best candidate gets closer to the true optimum. By looking at the top part of the figure, a clustering of the remaining optima is observable, so that a closer look is offered in the bottom figure for better understanding. Among the adaptive optima, the AMPGA2 and EGO method produce the best results and demonstrates the benefits of opportunely coupling the zonal approach and an "intelligent" design space sampling. Indeed, these optimal candidates are the closest to the target point in the sense of the Euclidean distance in the objective plane.

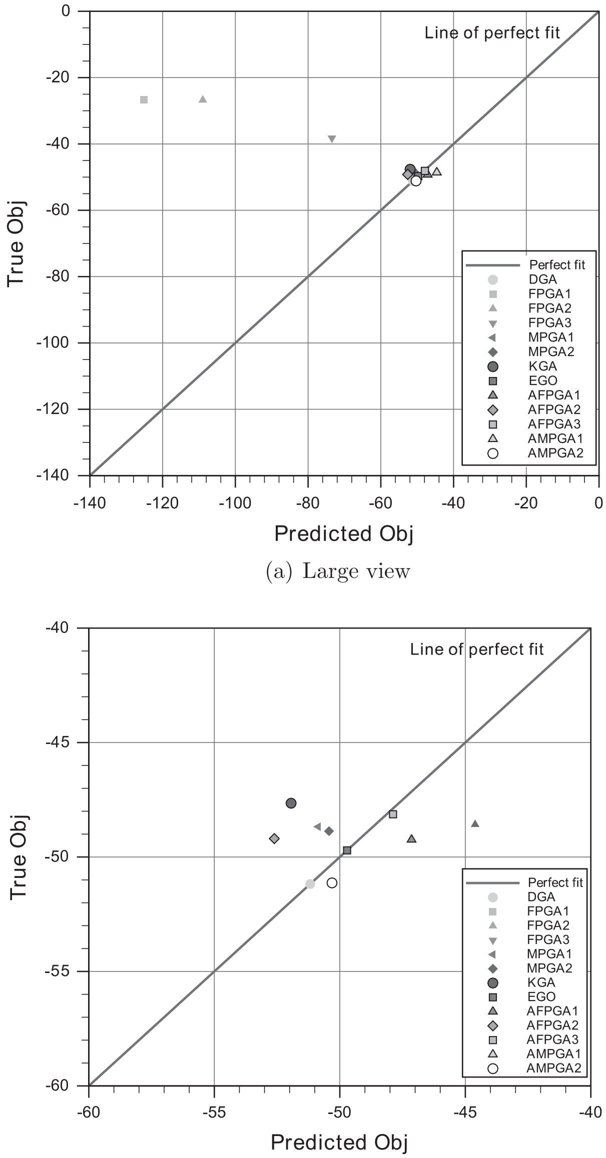
The following Figs. 17 and 18 show some aerodynamic details of each candidate. Surface pressure coefficients as computed by the surrogates and the CFD model are reported for aerodynamic comparison. In particular, it is clearly evident how the AMPGA2 individual presents a shock-less behaviour just as the "true" optimum. Moreover, the pressure distribution is featured with a more flat profile and a more gentle re-compression on the suction side. This is a quite unique feature among the observed optima solutions.

Summing up, from a comparison of surrogate-based optimization results, it emerges that there is a class of methods, namely the adaptive ones that use POD, which should be the first choice when the computational budget is very low, while it is still the algorithm that uses only high-fidelity evaluations that should be preferred if there are no constraints on the computational power available. EGO, instead, is located in the middle, since, while requiring more computational resources than the adaptive methods with POD, it certainly offers a greater ease of setup and allows to reach comparable performance levels.


(a) x-velocity component

(b) z-velocity component

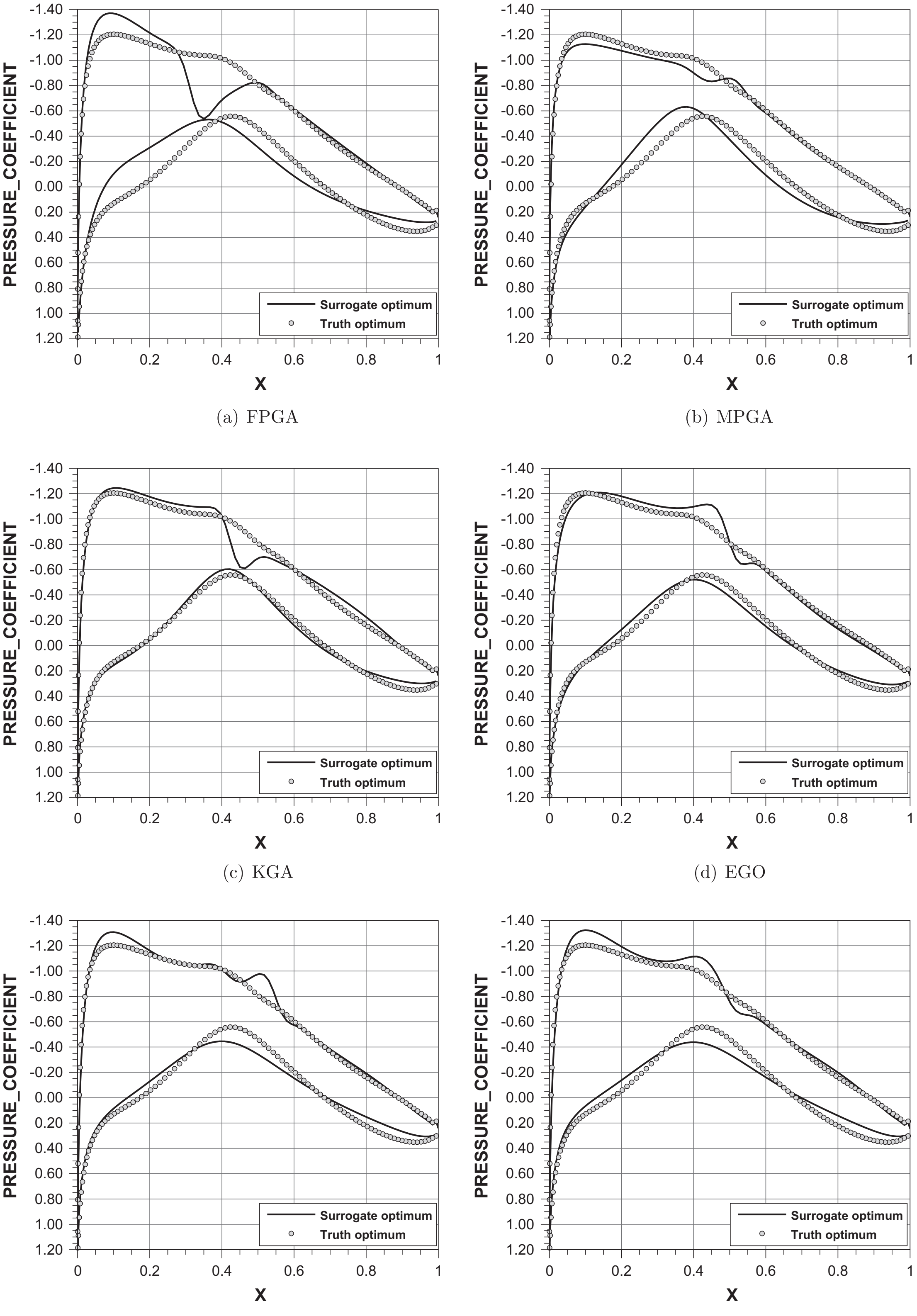
Fig. 15. MPGA1 optimum, velocity contour comparison (POD – dashed lines; CFD – solid lines.



(a) Large view

(b) Zoom near the truth optimum

Fig. 16. Computed optima in the surrogate vs truth objective plane.



(a) FPGA

(b) MPGA

(c) KGA

(d) EGO

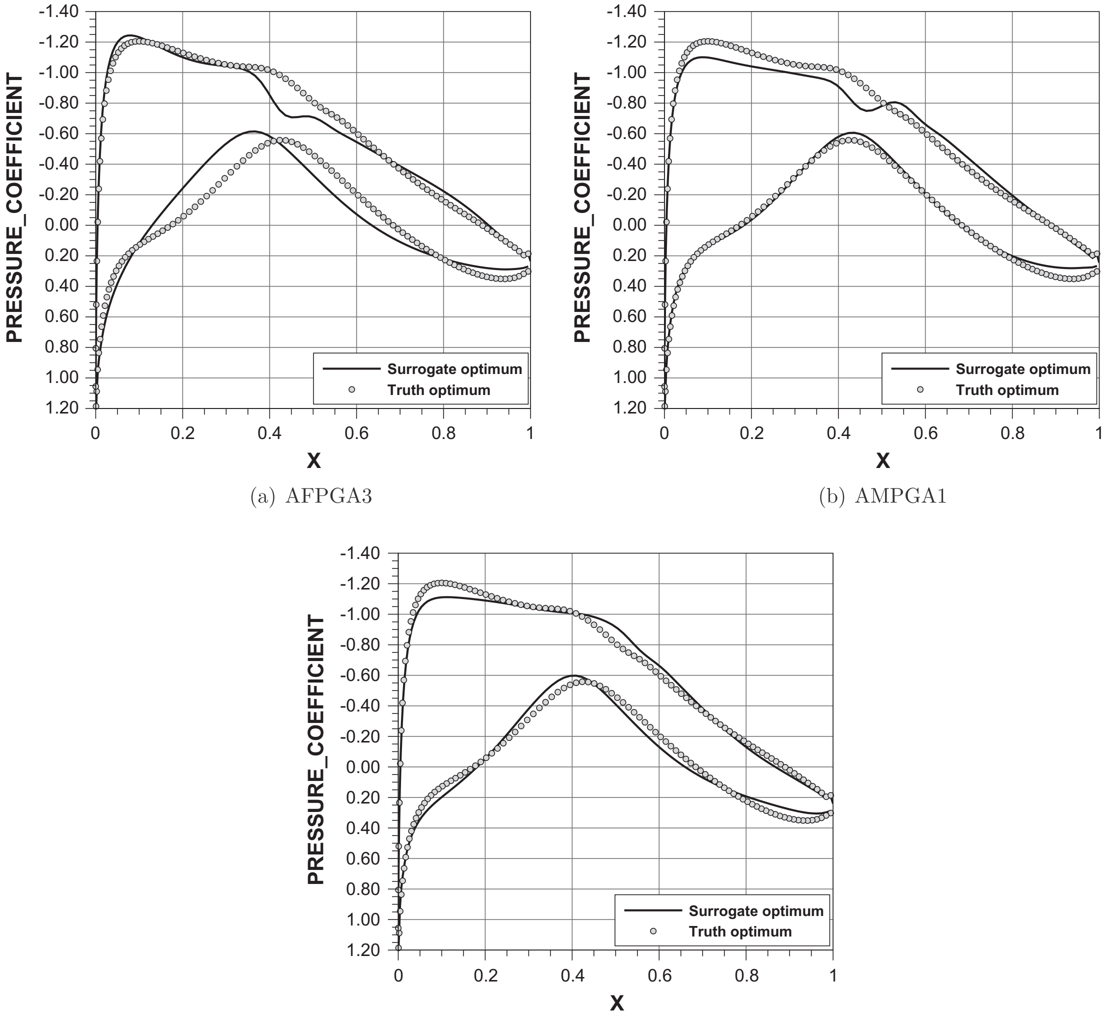
(e) AFPGA1

(f) AFPGA2

**Fig. 17.** Optimal candidates, pressure coefficient distributions.

## 7. Conclusions

The aim of the present research work was to investigate and study *ad hoc* computational techniques to ease the solution of complex aerodynamic shape optimization problems such those commonly encountered in aerospace design at industrial level. Among the various approaches that are subject of research investigation, we chose to focus on *ad hoc* surrogate methods. In



(a) AFPGA3

(b) AMPGA1

(c) AMPGA2

**Fig. 18.** Optimal candidates, pressure coefficient distributions.

particular, we demonstrated that the well-known Proper Orthogonal Decomposition approach is not adequate to provide reliable predictions in peculiar aerodynamic conditions like transonic flow and when the boundary of the computational domain changes like in shape optimization. We proposed a zonal approach to de-couple the strong non-linearities occurring near the body-wall from the POD approximation. This zonal approach proved to give reliable results at a reduced computational cost compared to the full CFD simulation. Furthermore, we showed that the zonal approach can give an accurate approximation of the true optimum when trained with specifically designed adaptive sampling techniques. The latter have been purposely conceived to improve the POD model machinery, namely the basis vectors and coefficients. By using such an "intelligent" design of experiment method, the high-fidelity computational budget can be further reduced and the overall performance of the design loop is increased. The beneficial effects of this approach has been illustrated by comparing several surrogate-based optimization processes on the shape design of a two-dimensional airfoil. The extension of the methodology to complex three-dimensional problems is straightforward and under way. Indeed, one of the main advantages of the proposed methodology is its relative insensitivity to the curse of dimensionality of the design parameter space. On the other hand, the larger snapshot size required by three-dimensional CFD flow fields, where millions of unknowns may be handled, does not represent a big issue with current linear algebra numerical solver technology. Another significant advantage of the zonal approach with respect to other surrogates lies in its favourable scaling property when the third dimension is introduced because the ratio between CFD-solved and POD-predicted points decreases.

Furthermore, zonal POD allows to solve the high-fidelity flow field locally, i.e. only where it is required by geometry-driven considerations. This represents a tremendous benefit when the complexity of the design case grows. As an example, if the goal is to optimally fit a nacelle body in a already optimal wing, the high-fidelity computation zone can be restricted to catch only the wing-nacelle interaction phenomena, leaving the POD model to predict the flow field outside. To further bridge the gap with real-world applications and needs, future works will be focused on validating the proposed methodology in a large-scale, multi-point aerodynamic problems involving huge design parameter spaces and aiming at predicting the aerodynamic characteristics in deep transonic flow. Finally, improvements towards a more efficient exploitation of approximate models in the search algorithm are under investigation. In particular, a bi-objective approach seems promising where the first objective is the true function evaluation and the second is the surrogate one. An asymmetric algorithm, i.e. an algorithm that invokes the surrogate much more times than the high-fidelity one, was already proposed in a multi-fidelity environment by Quagliarella [54] and it is going to be extended to adaptive POD-based surrogate models.

## References

[1] Jones DR, Schonlau M, Welch WJ. Efficient global optimization of expensive black-box functions. J Global Optim 1998;13:455–92.

[2] Schonlau M, Welch WJ, Jones DR. Global versus local search in constrained optimization of computer models. Lect Notes-Monogr Ser 1998;34.

[3] Jones DR. A taxonomy of global optimization methods based on response surfaces. J Global Optim 2001;21:345–83.

[4] Alexandrov NM, Lewis RM, Gumbert CR, Green LL, Newman PA. Optimization with variable-fidelity models applied to wing design. Tech. Rep.; Institute for Computer Applications in Science and Engineering (ICASE), NASA Langley; 1999.

[5] Booker AJ, Dennis JE, Frank PD, Serafini DB, Torczon V, Trosset MW. A rigorous framework for optimization of expensive functions by surrogates. Struct Multidiscip Optim 1999;17:1–13. http://dx.doi.org/10.1007/BF01197708.

[6] Queipo N, Haftka R, Shyy W, Goel T, Vaidyanathan R, Kevintucker P. Surrogate-based analysis and optimization. Prog Aerosp Sci 2005;41(1):1–28.

[7] Simpson TW, Toropov VV, Balabanov V, Viana FAC. Design and analysis of computer experiments in multidisciplinary design optimization: a review of how far we have come – or not. In: Proceedings of the 12th AIAA/ISSMO multidisciplinary analysis and optimization conference. AIAA 2008-5802. American Institue of Aeronautics and Astronautics; 2008. p. 1–22.

[8] Forrester AIJ, Keane AJ. Recent advances in surrogate-based optimization. Prog Aerosp Sci 2009;45(1–3):50–79.

[9] Sóbester A, Leary S, Keane A. A parallel updating scheme for approximating and optimizing high fidelity computer simulations. Struct Multidiscip Optim 2004;27:371–83. http://dx.doi.org/10.1007/s00158-004-0397-9.

[10] Gutmann HM. A radial basis function method for global optimization. J Global Optim 2001;19:201–27.

[11] Goel T, Haftka RT, Shyy W, Queipo NV. Ensemble of surrogates. Struct Multidiscip Optim 2007;33(3):199–216.

[12] Mifsud M. Reduced-order modelling for high-speed aerial weapon aerodynamics. Ph.D. thesis. Cranfield University – College of Aeronautics; 2008.

[13] Loeve M. Probability theory. 4th ed. New York: Springer-Verlag; 1977. ISBN:0387902104 0387902104..

[14] Lumley JL. The structure of inhomogeneous turbulent flows. In: Yaglom AM, Tatarski VI, editors. Atmospheric turbulence and radio propagation. Moscow: Nauka; 1967. p. 166–78.

[15] LeGresley P, Alonso J. Investigation of non-linear projection for pod based reduced order models for aerodynamics. In: Proceedings of the 39th AIAA aerospace sciences meeting and exhibit. AIAA-01-0926, Reno, Nevada; 2001.

[16] Epureanu B. A parametric analysis of reduced order models of viscous flows in turbomachinery. J Fluids Struct 2003;17(7):971–82.

[17] Bui-Thanh T, Damodaran M, Willcox K. Proper orthogonal decomposition extensions for parametric applications in compressible aerodynamics. In: Proceedings of the 21st AIAA applied aerodynamics conference. AIAA 2003-4213, Orlando, Florida; 2003.

[18] Bui-Thanh T, Damodaran M, Willcox K. Aerodynamic data reconstruction and inverse design using proper orthogonal decomposition. AIAA J 2004;42(8):1505–16.

[19] Sirovich L. Turbulence and the dynamics of coherent structures. I – Coherent structures. II – Symmetries and transformations. III – Dynamics and scaling. Quart Appl Math 1987;45:561–71.

[20] Holmes P, Lumley J, Berkooz G. Turbulence, coherent structures, dynamical systems and symmetry. Cambridge: Cambridge University Press; 1996.

[21] Dowell E, Hall K, Thomas J, Florea R, Epureanu B, Heeg J. Reduced order models in unsteady aerodynamics. In: Proceedings of the 40th AIAA/ASME/ASCE/AHS/ASC structures, structural dynamics and materials (SDM) conference. St. Louis, MO; 1999.

[22] Hall KC, Thomas JP, Dowell EH. Reduced-order modeling of unsteady small-disturbance flows using a frequency-domain proper orthogonal decomposition technique. AIAA J 2000;38(10):1853–62.

[23] Everson R, Sirovich L. Karhunen–Loève procedure for gappy data. J Opt Soc Am A 1995;12(8):1657–64.

[24] Tang L, Chen P, Liu D, Gao X, Shyy W, Utturkar Y, et al. Proper orthogonal decomposition and response surface method for TPS/RLV structural design and optimization: X-34 case study. In: Proceedings of the 43rd AIAA aerospace sciences meeting and exhibit, Reno, Nevada; 2005.

[25] LeGresley P, Alonso J. Dynamic domain decomposition and error correction for reduced order models. In: Proceedings of the 41st AIAA aerospace sciences meeting and exhibit. 03-0250, Reno, Nevada: 41st AIAA aerospace sciences meeting and exhibit; 2003.

[26] Lucia DJ, King PI, Beran PS. Reduced order modeling of a two-dimensional flow with moving shocks. Comput Fluids 2003;32(7):917–38.

[27] Lucia DJ. Reduced order modeling for high speed flows with moving shocks. Ph.D. thesis, Air Force Institute of Technology, Wright-Patterson Air Force Base, Ohio; 2001.

[28] Buffoni M, Telib H, Iollo A. Iterative methods for model reduction by domain decomposition. Comput Fluids 2009;38(6):1160–7.

[29] Toal DJJ, Bressloff NW, Keane AJ, Holden CME. Geometric filtration using pod for aerodynamic design optimization. AIAA J 2010;48(5):916–28.

[30] Braconnier T, Ferrier M, Jouhaud JC, Montagnac M, Sagaut P. Towards an adaptive POD/SVD surrogate model for aeronautic design. Comput Fluids 2011;40(1):195–209.

[31] Lieu T, Farhat C. Adaptation of pod-based aeroelastic roms for varying mach number and angle of attack: application to a complete f-16 configuration. In: AIAA paper 2005-7666, U.S. Air Force T&E Days; 2005.

[32] Lieu T, Farhat C, Lesoinne M. Pod-based aeroelastic analysis of a complete f-16 configuration: rom adaptation and demonstration. In: Proceedings of the 46th structures, structural dynamics & materials conference. AIAA paper 2005-2295; 2005.

[33] Reed M, Simon B. Methods of modern mathematical physics I: functional analysis. Academic Press; 1980.

[34] Sirovich L. Turbulence and the dynamics of coherent structures. Part 1: coherent structures. Quart Appl Math 1987;45(3):561–71.

[35] Pettit C, Beran P. Reduced-order modelling for flutter prediction. In: Proceedings of the AIAA/ASME/ASCE/AHS/ASC structures, structural dynamics, and materials conference. No. 00-1446 in AIAA paper, AIAA/ASME/ASCE/AHS/ASC structures, structural dynamics, and materials conference; 2000.

[36] LeGresley P. Application of proper orthogonal decomposition (pod) to design decomposition methods. Ph.D. thesis, Stanford University; 2005.

[37] Chandrashekarappa P, Duvigneau R. Radial basis functions and Kriging metamodels for aerodynamic optimization. Rapport de recherche RR-6151; INRIA; 2007.

[38] Benoudjit N, Archambeau C, Lendasse A, Lee JA, Verleysen M. Width optimization of the gaussian kernels in radial basis function networks. In: ESANN; 2002. p. 425–432.

[39] Rippa S. An algorithm for selecting a good value for the parameter $c$ in radial basis function interpolation. Adv Comput Math 1999;11:193–210. http://dx.doi.org/10.1023/A:1018975909870.

[40] Cook P, Firmin M, McDonald M. Aerofoil RAE 2822: pressure distributions, and boundary layer and wake measurements. Technical memorandum, Royal Aircraft Establishment; 1977.

[41] Kulfan BM. Universal parametric geometry representation method. J Aircraft 2008;45(1):142–58.

[42] Montgomery DC. Design and analysis of experiments. John Wiley & Sons; 2006. ISBN: 0470088109.

[43] McKay M, Conover W, Beckman R. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 1979;21:239–45.

[44] Iuliano E. Towards a pod-based surrogate model for CFD optimization. In: Proceedings of the Eccomas CFD & optimization conference, Antalya, Turkey; 2011.

[45] Amato M, Catalano P. Non linear $\kappa\varepsilon$ turbulence modeling for industrial applications. In: ICAS 2000 congress. Harrogate, UK: IOS Press; 2000.

[46] Eldred M, Bichon B, Adams B, Mahadevan S. Structural design optimization considering uncertainties. In: Structures and infrastructures series. Overview of reliability analysis and design capabilities in DAKOTA with application to shape optimization of MEMS, vol. 1. Taylor & Francis Group; 2008. p. 401–32 [chap.].

[47] Goblet J, Lepot I. Two adaptive doe strategies for pod-based surrogate models. Tech. Rep.; CENAERO, ROM&O project report, CleanSky JTI-GRA, call for proposal JTI-CS-2009-1-GRA-05-004, Proposal number 255779; 2010.

[48] Sainvitu C, Guent M, Lepot I, Goblet J. Adaptive sampling strategies for pod-based surrogate models in an optimization framework. In: Proceedings of the EUROGEN 2011 conference, Capua, Italy; 2011.

[49] Quagliarella D, Vicini A. GAs for aerodynamic shape design I: general issues, shape parametrization problems and hybridization techniques. In: Lecture series 2000–07. Genetic algorithms for optimisation in aeronautics and turbomachinery. Belgium: Von Karman Institute; 2000.

[50] Quagliarella D, Vicini A. GAs for aerodynamic shape design II: multiobjective optimization and multi-criteria design. In: Lecture series 2000–07. Genetic algorithms for optimisation in aeronautics and turbomachinery. Belgium: Von Karman Institute; 2000.

[51] Iuliano E, Quagliarella D. Surrogate-based aerodynamic optimization via a zonal pod model. In: Proceedings of the EUROGEN 2011 conference, Capua, Italy; 2011.

[52] Catalano P, Amato M. An evaluation of rans turbulence modelling for aerodynamic applications. Aerosp Sci Technol 2003;7:493–509.

[53] Eddy JE, Lewis K. Effective generation of pareto sets using genetic programming. In: Proceedings of ASME design engineering technical conference, Pittsburgh, PA; 2001.

[54] Quagliarella D. Airfoil design using Navier–Stokes equations and an asymmetric multi-objective genetic algorithm. In: Evolutionary methods for design, optimization and control applications to industrial and societal problems, Barcelona, Spain: CIMNE; 2003. ISBN: 84-95999-33-1.