Available online at www.sciencedirect.com
SCIENCE DIRECT®
Progress in Aerospace Sciences 41 (2005) 1–28
PROGRESS IN AEROSPACE SCIENCES
www.elsevier.com/locate/pacrosci

Surrogate-based analysis and optimization

Nestor V. Queipo[a,1], Raphael T. Haftka[a], Wei Shyy[a,*], Tushar Goel[a], Rajkumar Vaidyanathan[a,2], P. Kevin Tucker[b]

[a]*Department of Mechanical and Aerospace Engineering, University of Florida, Gainesville, FL, USA*
[b]*NASA Marshall Space Flight Center, Huntsville, AL, USA*

Abstract

A major challenge to the successful full-scale development of modern aerospace systems is to address competing objectives such as improved performance, reduced costs, and enhanced safety. Accurate, high-fidelity models are typically time consuming and computationally expensive. Furthermore, informed decisions should be made with an understanding of the impact (global sensitivity) of the design variables on the different objectives. In this context, the so-called surrogate-based approach for analysis and optimization can play a very valuable role. The surrogates are constructed using data drawn from high-fidelity models, and provide fast approximations of the objectives and constraints at new design points, thereby making sensitivity and optimization studies feasible. This paper provides a comprehensive discussion of the fundamental issues that arise in surrogate-based analysis and optimization (SBAO), highlighting concepts, methods, techniques, as well as practical implications. The issues addressed include the selection of the loss function and regularization criteria for constructing the surrogates, design of experiments, surrogate selection and construction, sensitivity analysis, convergence, and optimization. The multi-objective optimal design of a liquid rocket injector is presented to highlight the state of the art and to help guide future efforts.

Contents

*Corresponding author. Current affiliation: Department of Aerospace Engineering, The University of Michigan, Ann Arbor, MI 48109, USA. Tel.: +1 734 936 0102; fax: +1 734 763 0578.
E-mail address: weishyy@umich.edu (W. Shyy).
[1]Current affiliation: Applied Computing Institute, Faculty of Engineering, University of Zulia, Venezuela.
[2]Current affiliation: General Motors, Bangalore, India.

0376-0421/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.paerosci.2005.02.001

Nomenclature

$a_c(\mathbf{x})$	corrected approximation
$\mathbf{c}$	center of the radial basis function
D	total variance
D_i	partial variance of ith variable x_i
$D_{i,z}$	sum of all partial variances associated with combinations of all variables and ith variable x_i
$D^m\hat{f}$	mth derivative of the function $\hat{f}$
$E(\cdot)$	expected value of the quantity
$E_{\mathrm{ADS}}(.)$	expected value of the quantity considering all data sets
$E_{\mathrm{bias}^2}(\mathbf{x})$	expected value of the square of the bias error at design point $\mathbf{x}$
$E_{\mathrm{var}}(\mathbf{x})$	expected value of the variance at design point $\mathbf{x}$
f	true function to be modeled
$f_p(\mathbf{x})$, $f(\mathbf{x})$	true response at design point $\mathbf{x}$
$\mathbf{f}$	vector of responses at sampled design points
f_{hi}, f_{lo}	high and low fidelity function evaluations respectively
f_i	(Model) Response at ith design point
$\hat{f}$	model estimation
$\mathbf{h(x)}$	vector of radial basis functions
$h_i(\mathbf{x})$	ith radial basis function
$\mathbf{H}$	matrix of radial basis functions for sampled design points
ΔHA	change in hydrogen area
$\mathbf{I}$	identity matrix
$\mathbf{J}$	Hadamard matrix
$\mathbf{K}(\mathbf{x}, \mathbf{x}^{(i)})$	symmetric kernel function matrix
$L(\mathbf{x})$	loss function
N_{dv}	number of design variables
N_{itr}	number of iterations in iterative fractional factorial design
N_{PRG}	number of terms in polynomial regression
N_{RBF}	number of radial basis functions
N_s	number of sampled design points
ΔOA	change in oxygen area
OPTT	oxygen post tip thickness
$\mathbf{P}$	projection matrix
q	number of levels for each variable
$\mathbf{r}$	correlation vector between the sampled design points and prediction design point
$\mathbf{R}$	correlation matrix among the sampled design points
R_a^2	adjusted coefficient of multiple determination in polynomial regression
R_{corr}	partial correlation coefficient
S_i	main factor global sensitivity index of ith variable x_i
s_A^m	sign of Ath variable in mth iteration
t	strength of orthogonal array
TF_{max}	maximum injector face temperature
TT_{max}	maximum injector tip temperature
TW_4	wall temperature at 3'' from the injector face
$V(\cdot)$	variance of the quantity
$\mathbf{w}$	vector of coefficients of linear combinations of radial basis functions
$\hat{\mathbf{w}}$	estimate of coefficients vector $\mathbf{w}$
$\mathbf{x}$	vector of design variables
$\mathbf{x}_c$	vector of current design variables
X_{cc}	combustion length
x_k	kth design variable
$\mathbf{x}^{(i)}$	ith design point
$\mathbf{X}$	$N_s \times N_{\mathrm{PRG}}$ Gramian matrix of basis functions
$\mathbf{z}$	vector of basis functions in polynomial regression
z_j	jth basis function in polynomial regression (e.g. x_1, x_1^2, x_2, …)
$z_j^{(i)}$	jth basis function for ith design point
$\mathbf{1}$	unit column vector

Greek letters

$\boldsymbol{\alpha}$	weight vector
α	hydrogen flow angle (Section 8)
α_i	ith component of the weight vector
$\boldsymbol{\beta}$	vector of coefficients in polynomial regression
β_i, β_{ij}	coefficients of the basis functions in polynomial regression
$\hat{\boldsymbol{\beta}}$	estimate of coefficient vector $\boldsymbol{\beta}$ in polynomial regression
γ	index of the orthogonal array
δ	radius of the radial basis function
$\boldsymbol{\varepsilon}$	vector of errors in polynomial regression
$\varepsilon(\mathbf{x})$, ε_i	error in surrogate
θ	probability density function
λ	regularization parameter
μ	mean of the responses at sampled design points
σ^2	variance of the response at sample design points
σ_a	adjusted root mean square error in polynomial regression
ϕ_k	degree of correlation among data points along kth direction

1. Introduction

Major risks in the successful full-scale development of aerospace systems are related to effectively addressing the competing needs of improving performance, reducing costs, and enhancing safety. Typically, the analysis of the components of such systems, such as air-breathing or rocket propulsion devices is expensive, hindering the search for optimal designs. Surrogate-based analysis and optimization (SBAO) has been shown to be an effective approach for the design of computationally expensive models such as those found in aerospace systems, involving aerodynamics, structures, and propulsion, among other disciplines. Successful applications include the multidisciplinary optimal design of aerospike rocket nozzles, injectors and turbines for liquid rocket propulsion, and supersonic business aircrafts.

For example, SBAO has been applied to rotor blade design and optimization [108], high speed civil transport [1], airfoil shape optimization [2–4], diffuser shape optimization [5], supersonic turbine [6–8], and injectors [7–11].

This review provides a comprehensive discussion of the fundamental issues that arise in SBAO, emphasizing the concepts behind the methods and techniques, and practical implications of an integrated approach to SBAO for aerospace components and systems.

The first part of the review is structured following the key steps in surrogate modeling (Fig. 1):

1. *Design of experiments* (*DOE*). The design of experiment is the sampling plan in design variable space. The key question in this step is how we assess the goodness of such designs, considering the number of samples is severely limited by the computational expense of each sample. A discussion of the most

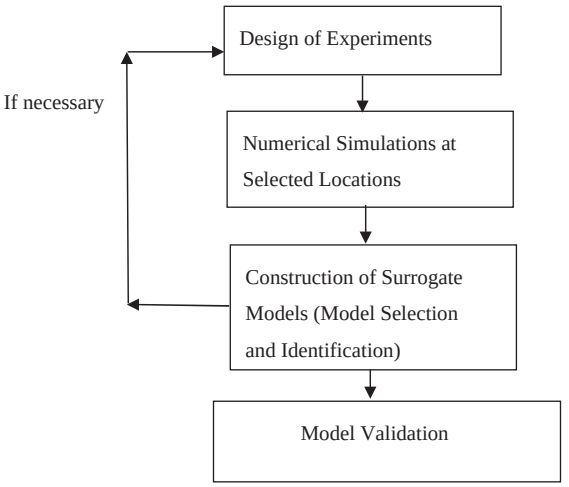

Fig. 1. Key stages of the surrogate-based modeling approach.

prominent approaches related to DOE in SBAO is presented in Section 3.

2. *Numerical simulations at selected locations*. Here, the computationally expensive model is executed for all the values of the input variables in the DOE specified in the previous step.
3. *Construction of surrogate model*. Two questions are of interest in this step: (a) what surrogate model(s) should we use (model selection) and, (b) how do we find the corresponding parameters (model identification)? A formal description of the problem of interest is discussed in Section 2. A framework for the discussion and mathematical formulation of alternative surrogate-based modeling approaches is the subject of Section 4.
4. *Model validation*. This step has the purpose of establishing the predictive capabilities of the surrogate model away from the available data (generalization error). Section 5 discusses schemes for estimating the generalization error for model validation purposes.

Then, SBAO are discussed in Sections 6 and 7, respectively. A case study associated with the multi-objective optimal design of a liquid-rocket injector is used to illustrate the different issues that arise when conducting SBAO and particular methods and techniques selected.

2. Overview of surrogate modeling

With reference to Fig. 2, surrogate modeling can be seen as a non-linear inverse problem for which one aims to determine a continuous function (f) of a set of design variables from a limited amount of available data (**f**). The available data **f** while deterministic in nature can represent exact evaluations of the function f or noisy observations and in general cannot carry sufficient information to uniquely identify f (multiple surrogates may be consistent with the available data as illustrated in Fig. 3). Thus, surrogate modeling deals with the twin problems of: (a) constructing a model $\hat{f}$ from the available data **f** (model estimation), and (b) assessing the errors ε attached to it (model appraisal). A general description of the anatomy of inverse problems can be found in Snieder [12].

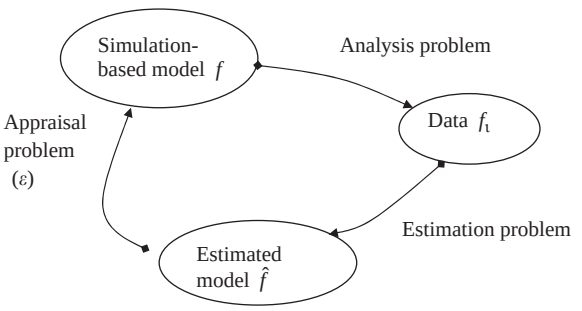

Fig. 2. Anatomy of surrogate modeling: model estimation + - model appraisal. The former provides an estimate of function f while the latter forecasts the associated error.

Hence, using the surrogate modeling approach the prediction of the simulation-based model output is formulated as $f_p(\mathbf{x}) = \hat{f}(\mathbf{x}) + \varepsilon(\mathbf{x})$. The prediction expected value and its variance $V(f_p)$ are illustrated in Fig. 4, with θ being a probability density function. Note that in Fig. 3 it is assumed that the expected value of $\varepsilon(\mathbf{x})$ is zero.

Different model estimation and model appraisal components of the prediction have been shown to be effective in the context of SBAO (see for example

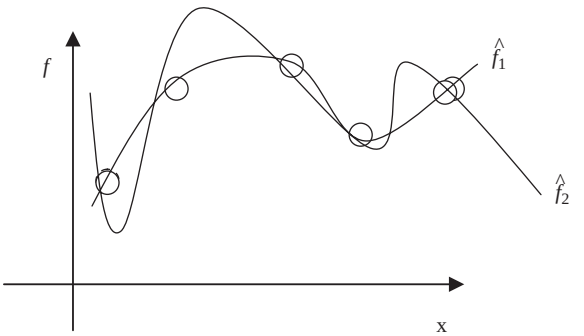

Fig. 3. Multiple surrogates may be consistent with the data. For a given problem we may have preference for one or the other (e.g., additional information about the function f is available), but often which is the best surrogate is not clear a priori.

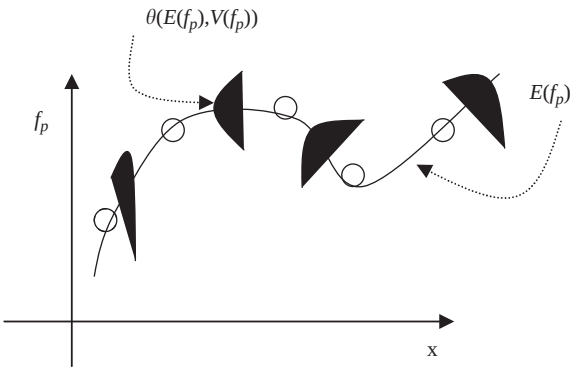

Fig. 4. A surrogate modeling scheme provides the expected value of the prediction $E(f_p)$ (solid line) and the uncertainty associated with that prediction illustrated here using a probability density function θ.

[13–16]), namely polynomial regression (PRG), Gaussian radial basis functions (GRF), and (ordinary) Kriging (KRG) as described by Sacks et al. [17]. Model estimation and appraisal components of these methods are presented in Section 4.

A good paradigm to illustrate how particular solutions ($\hat{f}$) to the model estimation problem can be obtained is provided by regularization theory (see for example [18,19]), which imposes additional constraints on the estimation. More precisely, $\hat{f}$ can be selected as the solution to the following Tikhonov regularization problem:

$$\min_{\widehat{f}\in \mathrm{H}} \mathbb{Z}(\hat{f}) = \frac{1}{N_s}\sum_{i=1}^{N_s} L(f_i - \widehat{f}(\mathbf{x}^{(i)})) + \lambda \int ||D^m \widehat{f}||_{\mathrm{H}}\, \mathrm{d}x, \tag{1}$$

where H is the family of surrogate models under consideration, $L(x)$ is a loss or cost function used to quantify the so called empirical error (e.g., $L(x) = x^2$), λ is a regularization parameter, and $D^m\widehat{f}$ represents the value of the m-derivative of the proposed model at location $\mathbf{x}$. Note that $D^m\widehat{f}$ represents a penalty term; for example, if m is selected equal to 2, it penalizes high local curvature. Hence, the first term enforces closeness to the data (goodness of fit), while the second term addresses the smoothness of the solution with λ (a real positive number) establishing the trade-off between the two. Increasing values of λ provide smoother solutions. The purpose of the regularization parameter λ is hence to help implement Occam's razor principle [20] which favors parsimony or simplicity in model construction. A good discussion on statistical regularization of inverse problems can be found in Tenorio [21].

The quadratic loss function (i.e., L_2 norm) is the most commonly used in part because it typically allows easy estimation of the parameters associated with the surrogate model; however, it is very sensitive to outliers. The linear (also called Laplace) loss function takes the absolute value of its argument (i.e., L_1 norm); on the other hand, the Huber loss function is defined as quadratic for small values of its argument and linear otherwise. The so-called ε loss function has received considerable attention in the context of the support vector regression surrogate [22,23] and assigns an error equal to zero (interpolation) if the true and estimated values are within an ε distance. Fig. 5 illustrates the cited loss functions.

Methods for identifying the regularization parameter λ are typically based on generalization error estimates (e.g., cross-validation); as a result, they will be discussed in the context of Validation (Section 5).

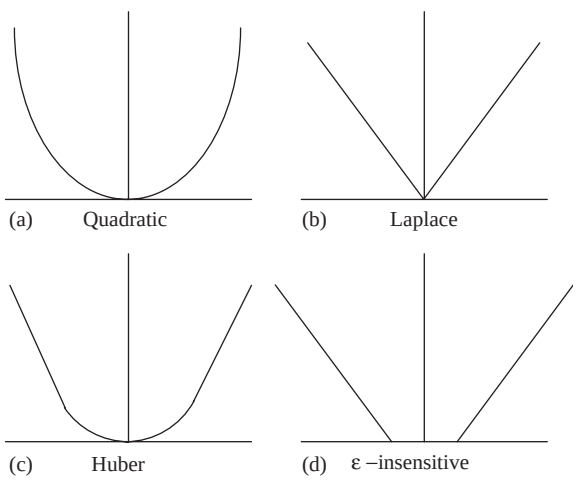

Fig. 5. Alternative loss functions for the construction of surrogate models.

3. Design of experiments (DOE)

As stated in the Introduction, the design of experiment is the sampling plan in design variable space and the key question in this step is how we assess the goodness of such designs. In this context, of particular interest are sampling plans that provide a unique value (in contrast to random values) for the input variables at each point in the input space, and are model-independent; that is, they can be efficiently used for fitting a variety of models.

With reference to (a) in most applications, where the assumed model is in doubt (see Sections 2 and 4), and the data is collected from deterministic computer simulations, the primary interest is minimizing the *bias error* because the numerical noise is small, and a DOE is selected accordingly. Brief descriptions for bias and variance components of the empirical error, and the corresponding expressions for the particular case of average error formulation and quadratic loss function are provided next. In the following discussion it is assumed that the function values (data set) have some noise or random component in them (e.g., due to numerical noise).

Bias: quantifies the extent to which the surrogate model outputs (i.e., $\hat{f}(\mathbf{x})$) differ from the true values (i.e., $f(\mathbf{x})$) calculated as an average over all possible data sets D.

Variance: measures the extent to which the surrogate model $\hat{f}(\mathbf{x})$ is sensitive to particular data set D. Each data set D corresponds to a random sample of the function of interest.

For an average error formulation with a quadratic loss function the expressions for bias and variance are shown in Eqs. (2) and (3), respectively. In both

expressions E_{ADS} denotes the expected value considering all possible data sets.

$$E_{bias^2}(\mathbf{x}) = \{E_{ADS}[\hat{f}(\mathbf{x})] - f(\mathbf{x})\}^2, \quad (2)$$

$$E_{var}(\mathbf{x}) = E_{ADS}[\hat{f}(\mathbf{x}) - E_{ADS}[\hat{f}(\mathbf{x})]]^2. \quad (3)$$

There is a natural trade-off between bias and variance. A surrogate model that closely fits a particular data set (lower bias) will tend to provide a larger variance. We can decrease the variance by smoothing the surrogate model but if the idea is taken too far then the bias error becomes significantly higher. In principle, we can reduce both the bias (can choose more complex models) and the variance (each model more heavily constrained by the data) by increasing the number of points provided the latter increases more rapidly than the model complexity.

In practice, the number of points in the data set is severely limited (e.g., due to computational expense) and often during the construction of the surrogate model a balance between bias and variance errors is sought. This balance can be achieved, for example, by reducing the bias error while imposing penalties on the model complexity (e.g., Tikhonov regularization).

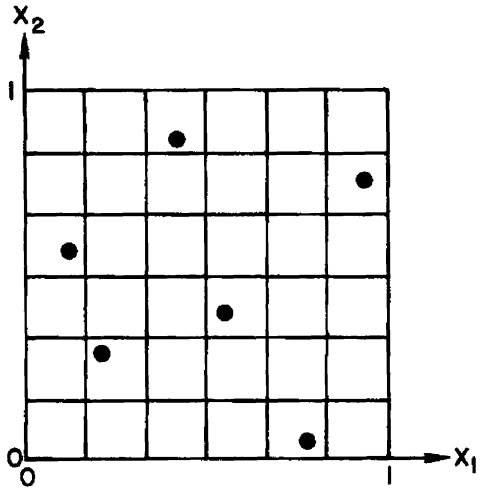

Fig. 6. A Latin Hypercube Design with $N_s = 6$, $N_{dv} = 2$ for X uniformly distributed on the unit square.

When polynomial regression is used, and the true model is assumed to be a polynomial of a given order, there is good theory for obtaining minimum bias designs (e.g., [24], Section 9.2) as well as some implementations in low dimensional spaces (e.g., [25]). For the more general case, the bias error can be reduced through a DOE that distributes the sample points uniformly in the design space [26,27] as referenced in Tang [28]) but computational expense does not allow to conduct dense full factorial designs so, instead, some forms of fractional factorial designs (adequately chosen fraction of the full factorial design) are used. The uniformity property in designs is sought by (issue b), for example, maximizing the minimum distances among design points [29], or by minimizing correlation measures among the sample data [30,31]. Practical implementation of these strategies includes orthogonal arrays (OA, e.g., [32]) and Latin Hypercube sampling (LHS, e.g., [33]); the former produces uniform designs but can generate particular forms of point replications while the latter does not produce replicates but can lack uniformity. As a result, OA-based LHS [28,34] and other optimal LHS schemes [35,36] have been proposed. These practical implementations are discussed next.

3.1. Latin hypercube sampling (LHS)

Stratified sampling ensures that all portions of a given partition are sampled. LHS [33] is a stratified sampling approach with the restriction that each of the input variables (x_k) has all portions of its distribution represented by input values. Following McKay et al.'s notation, a sample of size N_s can be constructed by dividing the range of each input variable into N_s strata of equal marginal probability $1/N_s$ and sampling once from each stratum. Let us denote this sample by $x_k^{(j)}$, $j = 1, \ldots, N_s$; $k = 1, \ldots, K$. The sample is made of components of each of the x_k's matched at random. Fig. 6 illustrates a LHS design.

While LHS represents an improvement over unrestricted stratified sampling [37] it can provide sampling plans with very different performance in terms of uniformity (issue b) measured by, for example, maximum minimum-distance among design points, or by correlation among the sample data. Fig. 7 illustrates this

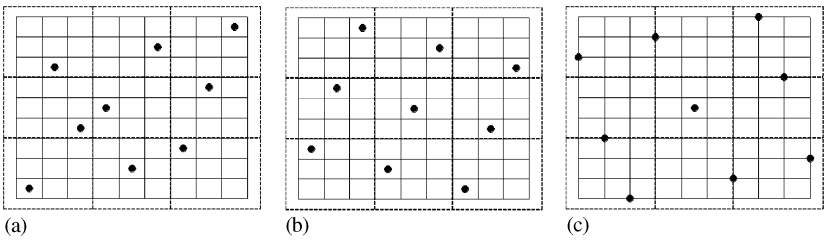

Fig. 7. LHS designs with significant differences in terms of uniformity [36].

shortcoming; the LHS plan in Fig. 7(c) is significantly better than that in Fig. 7(a).

3.2. Orthogonal arrays (OA)

These arrays were introduced by C.R. Rao in the late 40s and can be defined as follows. An OA of strength t is a matrix of N_s rows and N_{dv} columns with elements taken from a set of q symbols, such that in any $N_s \times t$ submatrix each of the q^t possible rows occurs the same number γ (index) of times. The number of rows (N_s) and columns (N_{dv}) in the OA definition represent the number of samples and variables or factors under consideration, respectively. The q symbols are related to the levels defined for the variables of interest, and the strength t is an indication of how many effects can be accounted for (to be discussed later in this section) with values typically between 2 and 4 for real-life applications. Such an array is denoted by $\mathrm{OA}(N_s, N_{dv}, q, t)$. Note that by definition a LHS is an OA of strength one, $\mathrm{OA}(N_s, N_{dv}, N_s, 1)$.

There are two limitations on the use of OA for DOE:

- *Lack of flexibility*: Given desired values for the number of rows, columns, levels, and strength, the OA may not exist. For a list of available orthogonal arrays, theory and applications, see, for example, Owen [38], Hedayat et al. [32] and references therein.
- *Point replicates*: OA designs projected onto the subspace spanned by the effective factors (most influential design variables) can result in replication of points. This is undesirable for deterministic computer experiments where the bias of the proposed model is the main concern.

3.3. Optimal LHS, OA-based LHS, optimal OA-based LHS

Different approaches have been proposed to overcome the potential lack of uniformity of LHS. Among those, most of them adjust the original LHS by optimizing a spreading measure (e.g., minimum distance or correlation) of the sample points. The resulting designs have been shown to be relatively insensitive to the optimal design criteria [39]. Examples of this strategy can be found in the works of Johnson et al. [29], Iman and Conover [30] and Owen [31]. Tang [28] presents the construction of strength t OA-based LHS which stratify each t-dimensional margin and shows that they offer a substantial improvement over standard LHS. Other strategies optimize a spreading measure of the sample points, but restrict the search of LHS designs which are orthogonal arrays, resulting in so called optimal OA-based LHS [36]. Adaptive DOE in which model appraisal information is used to place additional samples have also been proposed [40–42].

If feasible, two sets of DOE are generated, one (so-called training data set) for the construction of the surrogate (Section 4) and one for assessing its quality (Validation as discussed in Section 5).

Given the choice of surrogate, the DOE can be optimized to suit a particular surrogate. This has been done extensively for minimizing variance in polynomial regression (e.g., [24], Chapter 8) and to some extent for minimizing bias (e.g., [25]). However, for non-polynomial models, the cost of the optimization of a surrogate-specific DOE is usually prohibitive, and so is rarely attempted.

4. Construction of the surrogate model

There are both parametric (e.g., polynomial regression, Kriging) and non-parametric (e.g., projection-pursuit regression, radial basis functions) alternatives for constructing surrogate models. The parametric approaches presume the global functional form of the relationship between the response variable and the design variables is known, while the non-parametric ones use different types of simple, local models in different regions of the data to build up an overall model. This section discusses the estimation and appraisal components of the prediction of a sample of both parametric and non-parametric approaches, and the rationale of addressing the issue of model uncertainty (not knowing which surrogate may perform the best) using multiple surrogates [43].

Specifically, the model estimation and appraisal components corresponding to Polynomial Regression (PRG), Kriging (KRG), and Radial Basis Functions (RBF) surrogate models are discussed next, followed by a discussion of a more general non-parametric approach called Kernel-based regression. Throughout this section a square loss function is assumed unless otherwise specified, and given the stochastic nature of the surrogates, the available data is considered a sample of a population.

4.1. Polynomial regression model (PRG)

The regression analysis is a methodology that studies the quantitative association between a function of interest f, and N_{PRG} basis functions z_j, where there are N_s sample values of the function of interest f_i, for a set of basis functions $z_j^{(i)}$ [44]. For each observation i a linear equation is formulated:

$$f_i(\mathbf{z}) = \sum_{j=1}^{N_{PRG}} \beta_j z_j^{(i)} + \varepsilon_i, \quad E(\varepsilon_i) = 0, \quad V(\varepsilon_i) = \sigma^2, \tag{4}$$

where the errors ε_i are considered independents with expected value equal to zero and variance σ^2.

For example, a second-order polynomial ($N_{\mathrm{dv}} = 2$; $N_{\mathrm{PRG}} = 6$) can be expressed as

$$\hat{f}(\mathbf{x}) = \beta_0 + \sum_{i=1}^{N_{\mathrm{dv}}} \beta_i x_i + \sum_{i=1}^{N_{\mathrm{dv}}} \sum_{j \leqslant i}^{N_{\mathrm{dv}}} \beta_{ij} x_i x_j. \tag{5}$$

The estimated parameters $\hat{\boldsymbol{\beta}}$ (by least squares) are unbiased and have minimum variance.

The set of equations specified in Eq. (4) can be expressed in matrix form as

$$\mathbf{f} = \mathbf{X}\boldsymbol{\beta} + \varepsilon, \quad E(\varepsilon) = 0, \quad V(\varepsilon) = \sigma^2 I, \tag{6}$$

where $\mathbf{X}$ is a $N_{\mathrm{s}} \times N_{\mathrm{PRG}}$ matrix of basis functions with the design variable values for sampled points. The vector of the estimated parameters then can then be found as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{f} \tag{7}$$

Considering a new set of basis function vector $\mathbf{z}$ for design point P, the variance of the estimation is

$$V(\hat{f}(\mathbf{z})) = \sigma^2(\mathbf{z}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{z}). \tag{8}$$

4.2. Kriging modeling (KRG)

It is named after the pioneering work of D.G. Krige (a South African mining engineer), and was formally developed by Matheron [45]. More recently, Sacks et al. [46,47], and Jones et al. [40] made it well-known in the context of the modeling, and optimization of deterministic functions, respectively. The Kriging method in its basic formulation estimates the value of a function (response) at some unsampled location as the sum of two components: the linear model (e.g., polynomial trend) and a systematic departure representing low (large scale) and high frequency (small scale) variation components, respectively.

The systematic departure component represents the fluctuations around the trend, with the basic assumption being that these are correlated and the correlation depends only on the distance between the locations under consideration. More precisely, it is represented by a zero mean, second-order, stationary process (mean and variance constant with a correlation depending on a distance) as described by a correlation model.

Hence, these models (Ordinary Kriging) suggest estimating deterministic functions as

$$\begin{aligned} &f_p(\mathbf{x}) = \mu + \varepsilon(\mathbf{x}), \quad E(\varepsilon) = 0, \\ &\mathrm{cov}(\varepsilon(\mathbf{x}^{(i)}),\ \varepsilon(\mathbf{x}^{(j)})) \neq 0, \quad \forall i, j, \end{aligned} \tag{9}$$

where μ is the mean of the response at sampled design points, and ε is the error with zero expected value, and with a correlation structure that is a function of a generalized distance between the sample data points. A possible correlation structure [46] is given by

$$\mathrm{cov}(\varepsilon(\mathbf{x}^{(i)}), \varepsilon(\mathbf{x}^{(j)})) = \sigma^2 \exp\left(-\sum_{k=1}^{N_{\mathrm{dv}}} \phi_k (\mathbf{x}_k^{(i)} - \mathbf{x}_k^{(j)})^2 \right). \tag{10}$$

where N_{dv} denotes the number of dimensions in the set of design variables $\mathbf{x}$; σ, identifies the standard deviation of the response at sampled design points, and, ϕ_k is a parameter which is a measure of the degree of correlation among the data along the kth direction. Specifically, given a set of N_{s} input/output pairs ($\mathbf{x}$, f), the parameters, μ, σ, and ϕ are estimated such that a likelihood function is maximized [46]. Given a probability distribution and the corresponding parameters, the likelihood function is a measure of the probability of the sample data being drawn from it. The model estimates at unsampled points is

$$E(f_p(\mathbf{x})) = \overline{\mu} + \mathbf{r}^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{f} - \mathbf{1}\overline{\mu}), \tag{11}$$

where the bar above the letters denotes estimates, $\mathbf{r}$ identifies the correlation vector between the set of prediction points and the points used to construct the model, $\mathbf{R}$ is the correlation matrix among the N_{s} sample points, and $\mathbf{1}$ denotes an N_{s}-vector of ones.

On the other hand, the estimation variance at unsampled design points is given by

$$V(f_p(\mathbf{x})) = \sigma^2 \left[1 - \mathbf{r}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{r} + \frac{(1 - \mathbf{1}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{r})}{\mathbf{1}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{1}} \right]. \tag{12}$$

Gaussian Processes [48,49], another well-known approach to surrogate modeling, can be shown to provide identical expressions for the prediction and prediction variance to those provided by Kriging, under the stronger assumption that the available data (model responses) is a sample of a multivariate normal distribution [50].

4.3. Radial basis functions (RBF)

Radial basis functions have been developed for the interpolation of scattered multivariate data. The method uses linear combinations of N_{RBF} radially symmetric functions, $h_i(\mathbf{x})$, based on Euclidean distance or other such metric, to approximate response functions as

$$f_p(\mathbf{x}) = \sum_{i=1}^{N_{\mathrm{RBF}}} w_i h_i(\mathbf{x}) + \varepsilon_i, \tag{13}$$

where $\mathbf{w}$ represents the coefficients of the linear combinations, $h_i(\mathbf{x})$ the radial basis functions and ε_i independent errors with variance σ^2.

The flexibility of the model, that is its ability to fit many different functions, derives from the freedom to choose different values for the weights. Radial basis functions are a special class of functions with their main

feature being that their response decreases (or increases) monotonically with distance from a central point. The center, the distance scale, and the precise shape of the radial function are parameters of the model.

A typical radial function is the Gaussian which, in the case of a scalar input, is expressed as

$$h_i(\mathbf{x}) = \exp\left(-\frac{(\mathbf{x}-\mathbf{c})^2}{\delta^2}\right). \tag{14}$$

The parameters are its center $\mathbf{c}$ and its radius δ. Note that the response of the Gaussian RBF decreases monotonically with the distance from the center, giving a significant response only in the center neighborhood.

Given a set of N_s input/output pairs (sample data) a radial basis functions model can be expressed in matrix form as,

$$\mathbf{f} = \mathbf{H}\mathbf{w}, \tag{15}$$

where $\mathbf{H}$ is a matrix given by,

$$\mathbf{H} = \begin{bmatrix} h_1(\mathbf{x}^{(1)}) & h_2(\mathbf{x}^{(1)}) & \cdots & h_{N_{\mathrm{RBF}}}(\mathbf{x}^{(1)}) \\ h_1(\mathbf{x}^{(2)}) & h_2(\mathbf{x}^{(2)}) & \cdots & h_{N_{\mathrm{RBF}}}(\mathbf{x}^{(2)}) \\ \vdots & \vdots & \ddots & \vdots \\ h_1(\mathbf{x}^{(N_s)}) & h_2(\mathbf{x}^{(N_s)}) & \cdots & h_{N_{\mathrm{RBF}}}(\mathbf{x}^{(N_s)}) \end{bmatrix}. \tag{16}$$

Similarly to the polynomial regression method, by solving Eq. (15) the optimal weights (in the least squares sense) can be found to be,

$$\hat{\mathbf{w}} = \mathbf{A}^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{f}, \tag{17}$$

where $\mathbf{A}^{-1}$ is a matrix given by,

$$\mathbf{A}^{-1} = (\mathbf{H}^{\mathrm{T}}\mathbf{H})^{-1}. \tag{18}$$

The error variance estimate can be shown to be given by,

$$\hat{\sigma}^2 = \frac{\mathbf{f}^{\mathrm{T}}\mathbf{P}^2\mathbf{f}}{\mathrm{tr}(\mathbf{P})}, \tag{19}$$

where $\mathbf{P}$ is a projection matrix,

$$\mathbf{P} = \mathbf{I} - \mathbf{H}\mathbf{A}^{-1}\mathbf{H}^{\mathrm{T}}. \tag{20}$$

The RBF model estimate for a new set input values is given by,

$$\hat{f}(\mathbf{x}) = \mathbf{h}^{\mathrm{T}}\hat{\mathbf{w}}, \tag{21}$$

where, $\mathbf{h}$ is a column vector with the radial basis functions evaluations,

$$\mathbf{h} = \begin{Bmatrix} h_1(\mathbf{x}) \\ h_2(\mathbf{x}) \\ \vdots \\ h_{N_{\mathrm{RBF}}}(\mathbf{x}) \end{Bmatrix}. \tag{22}$$

On the other hand, the prediction variance is the variance of the estimated model $\hat{f}(\mathbf{x})$ plus the error variance and is given by

$$V(f_p(\mathbf{x})) = V(\mathbf{h}^{\mathrm{T}}\hat{\mathbf{w}}) + V(\varepsilon) = (\mathbf{h}^{\mathrm{T}}(\mathbf{H}^{\mathrm{T}}\mathbf{H})^{-1}\mathbf{h} + 1)\frac{\mathbf{f}^{\mathrm{T}}\mathbf{P}\mathbf{f}}{N_s - N_{\mathrm{RBF}}}. \tag{23}$$

The basic idea of RBF can be generalized to consider alternative loss functions and basis functions, in a scheme known as kernel-based regression. With reference to Eq. (1), it can be shown that independent of the form of the loss function L, the solution of the variational problem has the form (the Representer Theorem; see [23,51]):

$$\hat{f}(\mathbf{x}) = \sum_{i=1}^{N_s} \alpha_i \mathbf{K}(\mathbf{x}, \mathbf{x}^{(i)}), \tag{24}$$

where $\mathbf{K}$ is a (symmetric) kernel function that determines the smoothness properties of the estimation scheme. Table 1 shows the kernel functions of selected estimation schemes with the kernel parameters being estimated by model selection approaches (see next section for details). If the loss function L is quadratic, the unknown coefficients in Eq. (24) can be found by solving the linear system:

$$(N_s\lambda\mathbf{I} + \mathbf{K})\alpha = \mathbf{f}, \tag{25}$$

where $\mathbf{I}$ is the identity matrix, $\mathbf{K}$ a square matrix with elements $K_{i,j} = \mathbf{K}(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$, and λ is a regularization parameter as discussed in Eq. (1). Note that the linear system is well posed since $(N_s\lambda\mathbf{I} + \mathbf{K})$ is strictly positive and well conditioned for large values of $N_s\lambda$. If loss function L is non-quadratic, the solution of the variational problem still has the form of Eq. (24) but the coefficients α_i are found by solving a quadratic programming problem in what is known as support vector regression [22]. Comparative studies have shown that depending on the problem under consideration, a particular modeling scheme (e.g., polynomial regression, Kriging, radial basis functions) may outperform the others and in general, it is not known a priori which one should be selected. See for example, the works of Friedman and Stuetzle [52], Yakowitz and Szidarovsky [53], Laslett [54], Giunta et al. [55], Simpson et al. [13], and Jinet et al. [16]. Considering plausible alternative surrogate models can reasonably fit the available data, and the cost of constructing surrogates is small compared to the cost of the simulations, using multiple

Table 1
Examples of kernel functions and related estimation schemes

Kernel function	Estimation scheme
$K(x;x_i) = (1 + x.x_i)^d$	Polynomial of degree d (PRD)
$K(x;x_i) = \|\|x - x_i\|\|$	Linear splines (LSP)
$K(x;x_i) = \exp(-\|\|x - x_i\|\|^2/\gamma_i)$	Gaussian RBF (GRF)

surrogates may offer advantages compared to the use of a single surrogate. Recently, multiple SBAO approaches have been suggested by Zerpa et al. [43], based on the model averaging ideas of Perrone and Cooper [56], and Bishop [57]. The multiple surrogate-based analysis approach is based on the use of weighted average models which can be shown to reduce the prediction variance with respect to that of the individual surrogates. The idea of multiple surrogate-based optimization is discussed in a later section of the paper.

5. Model selection and validation

Generalization error estimates assess the quality of the surrogates for prediction and can be used for model selection among alternative models and establish whether they are ready to use in analysis and optimization studies (validation). This section discusses the most prominent approaches in the context of SBAO.

5.1. Split sample (SS)

In this scheme the sample data is divided into *training* and *test* sets. The former is used for constructing the surrogate while the latter, if properly selected, allows computing an unbiased estimate of the generalization error. Its main disadvantages are that the generalization error estimate can exhibit a high variance (it may depend heavily on which points end up in the training and test sets) and that it limits the amount of data available for constructing the surrogates.

5.2. Cross-validation (CV)

It is an improvement on the split sample scheme that allows the use of most if not all of the available data for constructing the surrogates. In general, the data is divided into k subsets (k-fold cross-validation) of approximately equal size. A surrogate model is constructed k times, each time leaving out one of the subsets from training, and using the omitted subset to compute the error measure of interest. The generalization error estimate is computed using the k error measures obtained (e.g., average). If k equals the sample size, this approach is called leave-one-out cross-validation (known also as PRESS in the polynomial regression terminology). Eq. (26) represents a leave-one-out calculation when the generalization error is described by the mean square error (GMSE).

$$\mathrm{GMSE} = \frac{1}{k}\sum_{i=1}^{k}(f_i - \hat{f}_i^{(-i)})^2, \tag{26}$$

where $\hat{f}_i^{(-i)}$ represents the prediction at $\mathbf{x}^{(i)}$ using the surrogate constructed using all sample points except $(\mathbf{x}^{(i)}, f_i)$. Analytical expressions are available for that case for the GMSE without actually performing the repeated construction of the surrogates for both polynomial regression [24], Section 2.7 and Kriging [58].

The advantage of cross-validation is that it provides nearly unbiased estimation of the generalization error and the corresponding variance is reduced (when compared to split-sample) considering that every point gets to be in a test set once, and in a training set $k-1$ times (regardless of how the data is divided); the variance of the estimation though may still be unacceptably high in particular for small data sets. The disadvantage is that it requires the construction of k surrogate models; this is alleviated by the increasing availability of surrogate modeling tools. A modified version of the CV approach called GCV-generalized cross-validation, which is invariant under orthogonal transformations of the data (unlike CV) is also available [59].

If the Tikhonov regularization approach for regression is adopted, the regularization parameter λ can be identified using one or more of the following alternative approaches: CV—cross-validation (leave-one-out estimates), GCV (smoothed version of CV), or the L-curve (explained below). While CV and GCV can be computed very efficiently [60,61], they may lead to very small values of λ even for large samples (e.g., very flat GCV function). The L-curve [62] is claimed to be more robust and have the same good properties of GCV. The L-curve is a plot of the residual norm (first term) versus the norm $\int ||D^m\widehat{f}||_{\mathrm{H}}$ of the solution for different values of the regularization parameter and displays the compromise in the minimization of these two quantities. The best regularization parameter is associated with a characteristic L-shaped "corner" of the graph.

5.3. Bootstrapping

This approach has been shown to work better than cross-validation in many cases [63]. In its simplest form, instead of splitting the data into subsets, subsamples of the data are considered. Each subsample is a random sample with replacement from the full sample, that is, it treats the data set as a population from which samples can be drawn. There are different variants of this approach [64,65] that can be used not only for model identification, but also for identifying confidence intervals for surrogate model outputs. This may come, though at the expense of considering several dozens or even hundreds of subsamples.

For example, in the case of polynomial regression (given a model) regression parameters can be estimated for each of the subsamples and a probability distribution (and then confidence intervals) for the parameters can be identified. Once the parameter distributions are estimated confidence intervals on model outputs of interest (e.g., mean) can also be obtained.

Bootstraping has been shown to be effective in the context of neural network modeling; recently, its performance in the context of model identification in regression analysis is also being explored [66,67].

6. Sensitivity analysis

Sensitivity, in this context, is a measure of the contribution of an independent variable to the total variance of the dependent data. Sensitivity analysis allows addressing settings such as:

- Can we safely fix one or more of the input variables without significantly affecting the output variability (Screening)?
- How can we rank a set of input variables according to their contribution to the output variability (Variables Prioritization)?
- If we could eliminate the uncertainty of one or more of the input variables which ones should be chosen (Variable Selection for Maximum Uncertainty Reduction)?
- If and which (group of) parameters interact with each other (Parameter Interactions)?
- What are the main regions of interest in the parameter space if additional samples become available?
- Does the model reproduce well known behavior of the process of interest (Model Validation)?

There are alternative approaches for sensitivity analysis, differing, for example, in scope (local vs. global), nature (qualitative vs. quantitative), and in whether they assume a particular model. Table 2 shows a sample of such methods. In this section, we discuss the Morris method [71], Iterated Fractional Factorial Design [72,73] frequently used for *screening* purposes and the Sobol's method [74] a variance-based non-parametric approach for analysis applications. Note that a *screening* effort is expected to be economical (require a relatively small number of computationally expensive simulations) and potentially reduce the number of variables considered in the construction of the surrogate model.

6.1. Morris method [71]

This approach is model independent and is particularly useful when the number of available simulations for screening purposes is of the order of the number of design variables. Its main purpose is to determine, within reasonable uncertainty, whether the effect of particular design variables are negligible, linear and additive, or non-linear (interaction with other design variables are present).

In its simplest form, the empirical distribution of the sensitivities associated with each of the design variables (x_i) is estimated by computing a number ($2N_{dv}p$) of first-order sensitivities of the model response with respect to each of the N_{dv} design variables at a set of p random locations. Each x_i is then characterized by measures of central tendency (e.g., mean) and spread (e.g., standard deviation). A large (absolute) value of central tendency for x_i indicates a design variable with an important (global) influence on the model response. On the other hand, a large measure of spread is indicative of a design variable whose influence is highly dependent on the value of the other design variables (interactions/non-linear effects). While this formulation of the Morris method requires $N_s = 2N_{dv}p$ simulations, there are variations that result in more economical designs with some of the simulations used in computing more than one sensitivity. A further reduction of the number of simulations can be obtained using cluster sampling (a concept to be discussed next). Details can be found in Morris [71].

6.2. Iterated fractional factorial design (IFFD) method [72,73]

The IFFD method was designed for screening purposes (identifying influential variables) for situations

Table 2
Different techniques in sensitivity analysis, and application examples, can be found in the works of Iman and Helton [68], Kleijnen [69], and Frey and Patil [70], and references therein

Method	Scope	Nature	Model independent[a]
Local sensitivity analysis	Local	Quantitative	Yes
Morris method	Global	Qualitative	Yes
Scatter plots	Global	Qualitative	No
Correlation coefficients	Global	Quantitative	No[b]
Variance-based parametric (e.g., IFFD)	Global	Quantitative	No
Variance-based non-parametric	Global	Quantitative	Yes

[a]Independent from assumptions about the model being linear, etc.
[b]There are exceptions e.g., the Spearman rank-order correlation coefficient.

requiring fewer runs than there are design variables with the model response dominated by a few influential variables. Assuming a particular model, IFFD establishes as influential design variables with significant linear, quadratic, or non-linear/interaction effects.

IFFD belongs to the so called *cluster sampling* designs with the design variables randomly aggregated into a small number of clusters. These clusters are then investigated using orthogonal fractional factorial designs in multiple iterations (composite design), with different groupings of design variables. A fractional factorial design is defined as a factorial experiment in which only an adequately chosen fraction of the combinations required for the complete factorial experiment is selected to be run while the orthogonal property makes reference to balanced designs in which the number of different combination of values for two or three design variables appear with equal frequency. Considering an influential group must contain an influential design variable, the influential variables are expected to lie in the intersection of influential groups in the iterations.

The steps involved in the IFFD method can be summarized as follows:

1. The design variables are sampled at three different levels, L (low), M (medium) and H (high) ensuring that the sampling is balanced. The number of clusters or groups of variables k is set. The value of k is assumed a power of two without a loss of generality; commonly used values of k are 8 and 16.
2. A basic two-level (L&H) resolution IV design matrix is constructed. The resolution of a design makes reference to its ability to discriminate between main factors (the individual effect of design variables), two-factor interactions, three factor interactions, etc. In a resolution IV design, main factors are confounded at worst with three-factor interactions; two factor interactions are confounded with certain other two-factor interactions though. A Hadamard matrix can describe these designs with the value of -1 or $+1$ at each entry representing the L and H values, respectively. Each column in the matrix represents the values of the variables assigned to it, while each row denotes a computational experiment to be conducted. This Hadamard matrix will be denoted by $J_k[i, j]$ and has a size of $2k$ rows and k columns.
3. Randomly assign (with equal probability) a design variable to a column in the two-level resolution IV design matrix; the corresponding variable value will be given by the design matrix multiplied by a sign (i.e., $+1$ or -1 with equal probability). The sign is denoted by s_A^m; the sign of variable A in iteration m.
4. Repeat step $3N_{\text{itr}}$ times assuming the number of available simulations for screening purposes is approximately $2kN_{\text{itr}}$. During a fraction of the iterations (usually 1/4) each of the design variables values are set to zero, hence creating a three-level design, even though each of the iterations has either a single middle level M or two extreme levels (L, H) per individual variable.
5. The model responses corresponding to the rows in the design matrix in each of the iterations are calculated. The model response in iteration m for a particular row is denoted by f_i^m.

The global effect of a variable x_j that takes its value from column j in the design matrix in a given iteration can be calculated as

$$\mathrm{ME}_m(x_j, \mathbf{f}^m) = \frac{1}{k}\sum_{i=1}^{2k} \mathbf{J}_k[i,j] f_i^m. \tag{27}$$

The global effect of a variable denoted with letter A throughout the entire design is given by

$$\begin{aligned}\mathrm{ME}(A, \mathbf{f}) &= \mathrm{avg}_m(s_A^m\ \mathrm{ME}(x_{C_A^m}, \mathbf{f}^m) | s_A^m \neq 0) \\ &= \frac{\sum_{m=1}^{M} s_A^m\ \mathrm{ME}(x_{C_A^m}, \mathbf{f}^m)}{\sum_{m=1}^{M} |s_A^m|},\end{aligned} \tag{28}$$

where C_A^m is the column that has assigned variable A in iteration m. Quadratic effects are estimated by

$$\begin{aligned}\mathrm{QE}(A, \mathbf{f}) &= \mathrm{avg}(\mathbf{f} | s_A = 0) - \mathrm{avg}(\mathbf{f} | s \neq 0), \\ \mathrm{QE}(A, \mathbf{f}) &= \frac{\sum_{m=1}^{N_{\text{itr}}} (1 - |s_A^m|) \sum_{i=1}^{2k} f_i^m}{2k \sum_{m=1}^{N_{\text{itr}}} (1 - |s_A^m|)} \\ &\quad - \frac{\sum_{m=1}^{N_{\text{itr}}} |s_A^m| \sum_{i=1}^{2k} f_i^m}{2k \sum_{m=1}^{N_{\text{itr}}} |s_A^m|}.\end{aligned} \tag{29}$$

Provided the model response is dominated by a few influential variables, and their contribution mostly limited to linear and quadratic effects, the IFFD method has been found to be effective in the screening of hundred and even thousands of variables with an order of magnitude less simulations.

6.3. Sobol's method [74]

To understand the concept, assume a surrogate model of a square integrable objective, $f(\mathbf{x})$, as a function of a vector of design variables, $\mathbf{x}$, whose values have been scaled between zero and one (this assumes that the design domain is box-like). This surrogate model represented in Eq. (30) can be decomposed as the sum of functions of increasing dimensionality as

$$\begin{aligned}f(x) = f_0 &+ \sum_i f_i(x_i) + \sum_{i<j} f_{ij}(x_i, x_j) \\ &+ \cdots + f_{12\ldots K}(x_1, x_2, \ldots, x N_{\text{dv}}).\end{aligned} \tag{30}$$

If the following condition:

$$\int_0^1 f_{i_1 \ldots i_s} \, \mathrm{d}x_k = 0 \tag{31}$$

is imposed for $k = i_1, \ldots, i_s$, where $1 \leqslant i_1 < \cdots < i_s \leqslant N_{\mathrm{dv}}$, the decomposition described in Eq. (30) is unique and each term in the sum can be obtained by computing the following integrals:

$$\int f(\mathbf{x}) \prod \mathrm{d}x_k = f_0, \tag{32}$$

$$\int f(\mathbf{x}) \prod_{k \neq i} \mathrm{d}x_k = f_0 + f_i(x_i), \tag{33}$$

from which $f_i(x_i)$ can be found, and

$$\int f(\mathbf{x}) \prod_{k \neq i,j} \mathrm{d}x_k = f_0 + f_i(x_i) + f_j(x_j) + f_{i,j}(x_i, x_j), \tag{34}$$

from which $f_{i,j}(x_i, x_j)$ can be obtained. The higher dimensional summands are similarly found except for the last one that is calculated using Eq. (30).

Furthermore, the summands are orthogonal (ensured by the condition expressed in Eq. (31)), and, square integrable following the same condition for $f(\mathbf{x})$. Therefore the partial variances, that is, the contribution of each of the summands to the total variance observed in the response, can be shown to be:

$$D_{i_1 \ldots i_s} = \int f^2_{i_1 \ldots i_s} \, \mathrm{d}x_{i_1} \ldots \mathrm{d}x_{i_s} \tag{35}$$

with the total variance being equal to:

$$D = \int f^2(\mathbf{x}) \, \mathrm{d}\mathbf{x} - f_0^2, \tag{36}$$

which can also be expressed as,

$$D = \sum_{s=1}^{N_{\mathrm{dv}}} \sum_{i_1 < \cdots < i_s}^{N_{\mathrm{dv}}} D_{i_1 \ldots i_s}. \tag{37}$$

Each partial variance gives a measure of the contribution of each independent variable or set of variables to the total variance, and provides an indication of their relative importance. Note that all the required integrations are conducted on the surrogate (fast) model and can in principle be calculated accurately provided an integration numerical procedure is available (e.g., Gaussian quadrature).

The relative importance of a design variable is quantified by a set of indices, namely, main factors, and total sensitivity indices. The former refer to the fraction of the total variance contributed by a particular variable in isolation, while the latter represents the contribution (expressed as a fraction) of all the partial variances in which the variable of interest is involved. The influence of a design variable, say x_i, to an objective variability without accounting for any of its interactions with other variables is denoted as a main factor index and given as

$$S_i = D_i / D. \tag{38}$$

To calculate the total sensitivity of any design variable, say x_i, the design variable vector x is divided into two complementary subsets, x_i and $\mathbf{Z}$ where $\mathbf{Z}$ is a vector containing $x_1, x_2, x_3, \ldots, x_n (n \neq \mathrm{i})$. The purpose of using these subsets is to isolate the influence of x_i on the $f(\mathbf{x})$ variability from the influence of the remaining design variables included in $\mathbf{Z}$. The total sensitivity index for x_i is then defined as

$$S_i^{\mathrm{total}} = D_i^{\mathrm{total}} / D, \tag{39}$$

where

$$D_i^{\mathrm{total}} = D_i + D_{i,\mathbf{Z}}, \tag{40}$$

D_i is the partial variance associated with x_i, and $D_{i,\mathbf{Z}}$ is the sum of all partial variances associated with any combination of the remaining variables representing the interactions between x_i and $\mathbf{Z}$. Similarly the sum of all the partial variances associated with the variables denoted by $\mathbf{Z}$ can be defined as $D_{\mathbf{Z}}$, hence the total variance can be written as

$$D = D_i + D_{\mathbf{Z}} + D_{i,\mathbf{Z}}. \tag{41}$$

Formulations of the Sobol's method that account for non-rectangular domains and correlated inputs are available; see, for example, Jacques et al. [75], and Mack et al. [76] for a recent application. A detailed discussion of global sensitivity methods and applications can be found in Sobol [74], Homma and Saltelli [77], Saltelli et al. [78] and the references therein.

7. Surrogate-based optimization

Surrogate model based optimization refers to the idea of speeding optimization processes by using *surrogates* for the objectives and constraints functions. The surrogates also allow for the optimization of problems with non-smooth or noisy responses, and can provide insight into the nature of the design space [10,11,76,79]. SBAO has shown to be effective in both multidisciplinary and multi-objective optimization. This section discusses the basic unconstrained SBAO algorithm, a newly proposed multiple surrogate-based optimization approach, the use of surrogate management frameworks to obtain provably convergent methods, and ways of addressing general non-linear constraints.

7.1. Basic unconstrained SBAO

It can be summarized as follows:

1. Construct a surrogate model from a set of known data points.

2. Estimate the function minimizer using the surrogate function.
3. Evaluate the true function value at the estimated minimum (*checking phase*).
4. Check for convergence; if achieved, *stop*.
5. Update surrogate using new data points.
6. Iterate until convergence.

where possible convergence criteria include the achievement of a target for the objective function, or finding that the best point found so far remains unchanged (within a given tolerance) from one iteration to the next. There are trust-region approaches, to be discussed below where the region where the surrogate is constructed and the optimization is conducted, is shrunk or expanded depending on the results of the checking phase in order to guarantee convergence to a local optimum (e.g., [80]). In the so called *one-shot solution* approach though, a single iteration is used.

In general, the algorithm above should be focused on identifying trends in the objective with the accuracy of the surrogates being important only when in the vicinity of a minimizer. Estimating the function minimizer is conducted using standard optimization methods and updating the surrogate using new data points can be accomplished by, for example, using model appraisal information and merit functions (e.g., [40,81,41,81–84]). Jones et al. [40], for example, use a generalized expected improvement function whereby points with low objective function value or high uncertainty are given precedence. A discussion of other sampling criteria in the context of SBAO is presented by Sasena et al. [41] as proposed by Watson and Barnes [85].

7.2. Multiple surrogates SBAO

For most challenging problems, given the sparseness of the data used to construct the surrogates, alternative surrogates can provide reasonable approximations while giving different uncertainty estimates throughout the design space. On the other hand, the surrogate construction requires small computational resources compared to the cost of simulations and, as previously discussed, a variety of parametric and non-parametric alternative loss functions (e.g., quadratic, and ε-sensitive) can be implemented.

The use of multiple surrogates will bring in several advantages at the optimization and decision-making levels of the proposed approach, namely:

- The predictions made by each surrogate are checked in the checking phase and when simulations are performed in additional cycles (iterations). Thus, we should be able to rank surrogates based on a small number of cycles and select those that appear to fit best the problem at hand.
- The additional surrogate optima analyzed in the checking phase may increase the global search capability of the optimization in that they may correspond to different local optima.
- A weighted averaged model may be constructed that is expected to provide a prediction with lower variance than any of the individual surrogates.
- Large variability in the estimated values and variances among surrogates at a point in design space may indicate that the uncertainty at that point is higher than predicted.

Some preliminary results obtained on the use of multiple surrogates for optimization has been reported by Zerpa et al. [43].

With reference to the basic SBAO algorithm, when multiple surrogates are available changes need to be made in the proposed approach. At the *optimization level*, in general there will be multiple suggested optima, and the checking phase will include not one but as many simulations as surrogates. This new data set is available to all surrogates and their expressions for the predictions and variances can be updated. At the decision making level, the designer can make a qualitative assessment of the uncertainty in the results, and has more information to decide whether or not to undertake another cycle and on how to proceed (the number and location of additional sample points). Note that significant differences in the predictions and variances at particular locations can be used to identify regions where the uncertainty may be higher than expected. Hence, it may be required to sample those regions and subsequently update the predictions and variances.

7.3. Convergence and SBAO

There are two approaches that are provably globally convergent to local optima or solutions of the original (computationally expensive) optimization problem of interest that involve surrogates. These are called the Approximation Model Management Framework (AMMF) [80,86,87] and Surrogate Management Framework (SMF) [88], and inherit/exploit the convergence properties of trust region and pattern search methods, respectively. They both use variable-fidelity models (*surrogate* and *computationally expensive*) and differ in whether they can use derivative information of the high-fidelity function (if available). A description of each of these frameworks is next.

7.4. Approximation model management framework (AMMF)

The approach is typically associated with gradient-based optimization algorithms and relies under the

general assumption that the surrogate model is accurate enough for the purpose of finding a good direction of improvement for the higher fidelity (computationally expensive) model. The AMMF replaces the local, Taylor series approximation, typical of conventional optimization, by an arbitrary model required to satisfy consistency conditions. The consistency conditions can be of different orders (zero, first, or second); for a first-order model the consistency conditions are:

$$a(\mathbf{x}_c) = f(\mathbf{x}_c),$$
$$\nabla a(\mathbf{x}_c) = \nabla f(\mathbf{x}_c), \tag{42}$$

where a and f denote the corrected approximation (its expression is shown below) and high-fidelity models, respectively. The consistency conditions can be imposed using a correction technique:

$$\beta(\mathbf{x}) = f_{hi}(\mathbf{x})/f_{lo}(\mathbf{x}), \tag{43}$$

where $f_{hi}(\mathbf{x})$ and $f_{lo}(\mathbf{x})$ corresponds to computationally expensive and surrogate evaluations, respectively. Furthermore, a first-order model β_c of β about a current design variable vector $\mathbf{x}_c$ can be written as

$$\beta_c(\mathbf{x}) = \beta(\mathbf{x}_c) + \nabla\beta(\mathbf{x}_c^t)(\mathbf{x} - \mathbf{x}_c). \tag{44}$$

The local model of β is then used to correct $f_{lo}(\mathbf{x})$ to obtain a better approximation $a(\mathbf{x})$ of $f_{hi}(\mathbf{x})$:

$$f_{hi}(\mathbf{x}) = \beta(\mathbf{x})f_{lo}(\mathbf{x}),$$
$$a(\mathbf{x}) = \beta_c(\mathbf{x})f_{lo}(\mathbf{x}). \tag{45}$$

It can be shown that the approximation model $a(\mathbf{x})$ satisfies the first-order consistency conditions. The basic AMMF algorithm can then be summarized as follows:

Until convergence do:

- Choose the corrected approximation a_k so that it satisfies the consistency conditions.
- Find an approximate solution $\mathbf{s}_k$ to the subproblem:

$$\min a_k(\mathbf{x}_k + \mathbf{s}),$$
$$\text{subject} ||\mathbf{s}|| \leqslant \Delta_k. \tag{46}$$

- Compare the actual and predicted decrease in f_{hi}.
- Update $\mathbf{x}_k$ and Δ_k accordingly.

As in the classical trust region methods [89] the length of the steps is regulated based on how well the corrected model predicts the decrease in f_{hi}. A significant mismatch in the actual and predicted decrease in f_{hi} may lead to a reduction in the trust radius and/or to update the surrogate model approximation. The AMMF approach inherits convergence properties of the classical trust regions algorithms for non-linear optimization. More precisely, first order AMMF methods can be shown to converge to local optima of the high-fidelity problem under appropriate standard conditions of continuity and boundedness of the functions and derivatives, given that the consistency conditions are imposed at each major iteration. The AMMF can be adapted to handle non-linear constraints and successful applications related to variable fidelity models have been reported [90,91]. While the guaranteed convergence of the algorithm is valuable, it comes at the cost of requiring derivative calculation for the expensive simulation at one point in the design space (normally the optimum of the previous cycle).

7.5. Surrogate management framework (SMF)

It is a framework based on pattern search methods (see, for example [92]) which incorporates surrogates to make the optimization cost effective. Being based on pattern search methods, the SMF framework: (i) can be used for scenarios where the objective functions are non-differentiable, or for which sensitivities are difficult or expensive to attain, (ii) can be easily adapted to either a parallel or distributed computing environment, (iii) is less likely to be trapped by non-global minimizers than are traditional non-linear optimization algorithms, (iv) can be extended to handle general non-linear constraints, and (v) convergence theory for the SMF can be found in the convergence of pattern search methods. The following SMF description follows that provided by Booker et al. [88] and Marsden et al. [93].

The SMF algorithm (illustrated in Fig. 8) is mesh-based (all points evaluated are restricted to lie on a mesh) and consists of two steps, SEARCH and POLL. The exploratory SEARCH step uses the surrogate to aid in the selection of points likely to minimize the objective function. The SEARCH step provides means for local and/or global exploration of the parameter space, but it is not strictly required for convergence. Hence, this step can be adapted to the particular engineering problem under consideration.

The convergence of the SMF is guaranteed by the POLL step, in which points neighboring the current best points on the mesh are evaluated in a positive spanning set of directions (positive basis) to check whether the current best point is a mesh local optimizer. A positive basis in $R^{N_{dv}}$ is a set of vectors whose non-negative linear combinations span $R^{N_{dv}}$, but for which no proper subset has that property. The relevance of a positive basis in the context of the SMF is that it ensures that if the gradient of the high-fidelity function f at the best current solution x is not zero, then at least one vector in the positive basis defines a descent direction from f at $\mathbf{x}$. This can be guaranteed without any knowledge of the gradient. Any positive basis has at least N_{dv}+1 (minimal) and at most $2N_{dv}$ (maximal) vectors; for example, in three dimensions, such a basis can be given by (1,0,0), (0,1,0), (0,0,1), and (−1,−1,−1). For unconstrained problems, a minimal positive basis is sufficient to

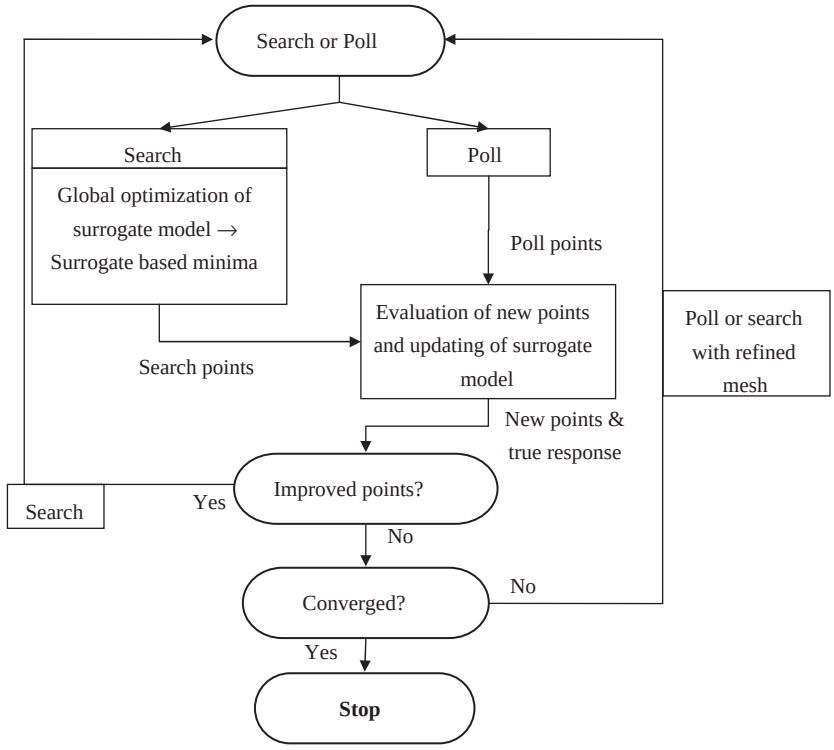

Fig. 8. The basic SMF algorithm.

guarantee convergence. The POLL set is made by the set of mesh points adjacent to the current best point in the above-referenced directions.

In the SEARCH step, the high-fidelity model is evaluated at one or more minimizers as predicted by the surrogate. If a lower objective function value is found among the recent evaluations, the search is considered successful, the surrogate is updated, and another search step is performed. If the SEARCH step fails to find an improved point, then it is considered unsuccessful and a POLL step is performed. In this step, the set of POLL points are evaluated. As soon as an improved point is found, a SEARCH step can be conducted on the current mesh. If no improved points are found, then the current best point is considered a so-called *mesh local optimizer*. The local term acknowledges the fact that the search was limited to the POLL set. For greater accuracy the mesh can be refined, at which point the algorithm continues with a SEARCH. Convergence is reached when a local minimizer on the mesh is found, and the mesh has been refined to the desired accuracy. Each time new data points are found in a SEARCH or POLL step, the surrogate model is updated.

Convergence of the SMF framework for constrained problems can be guaranteed (provided the function is continuously differentiable in the feasible region) if a generalized pattern search (GPS) strategy is adopted. The GPS strategy step halves the mesh if the POLL step is unsuccessful. Convergence properties of the SMF based on the generalized pattern search strategy can be found in Booker et al. [88]. Note, however, that the number of high-fidelity simulations performed (the POLL stage) is relatively small compared to those used for constructing surrogates, so that more iterations are required, and parallel processing opportunities are more limited.

7.6. Constrained SBAO

Once the constraints are modelled using surrogates, constrained SBAO is typically solved using well-known non-linear programming algorithms [94] for constrained optimization based on the concept of *penalty function*—penalty method, classical and modified barrier methods, augmented Lagrangian methods—which *approximate* the constrained problem to an unconstrained one. The penalty function is a linear combination of the objective and some measure of the constraint violation. The approximation is accomplished in the case of penalty methods by adding to the objective function a term that prescribe a high cost for constraint violations and in the case of barrier methods by adding a term that favors points interior to the feasible region over those near the boundary. A related idea is that corresponding to an

augmented Lagrangian function in which a penalty term is added to a Lagrangian function. While successful applications of these methods in the context of SBAO have been reported, in general, there are some difficulties associated with the use of penalty functions and the selection of the penalty parameter. Namely, there is usually a threshold value for the penalty parameter below which the unconstrained problem associated with the penalty function does not have a local minimum at the solution of the constrained problem, and the threshold value is not known a priori. If the penalty parameter is too low the solution can be unfeasible, and if it is too high it can damped out the effect of the objective which can result in very slow convergence.

Methods to avoid the above-referenced difficulties have been proposed. In particular, Fletcher and Leyffer [95] recently presented a so-called *filter* approach that does not require estimating a potentially troublesome penalty or barrier parameter. In this method, instead of combining the *objectives* and a measure of *constraint satisfaction* into a single minimization problem (as with penalty functions), these are treated as two separate objectives (multi-objective optimization). The set of points that are non-dominated (none is better than any other in both objective function value and constraint violation) is called a *filter*, and an iteration of an algorithm, selects a filter point, generates a new one (e.g., using gradient information) and accept the point (added to the filter) if the new point is non-dominated with respect to those in the filter, or reject it otherwise. Note the contrast with the penalty function methods where a point in an iteration is accepted only if there is a decrease in the penalty function. While the filter approach was developed in the context of gradient-based optimization (e.g., SLP, SQP) the method has been extended to derivative-free methods [96] and holds promise to be useful in industrial applications (see, for example [97,98]).

8. Case study: multi-objective liquid-rocket injector design

In this section, we will use an injector design case study, motivated by rocket propulsion, to illustrate the issues and usefulness of SBAO for multi-objective optimization. The materials are extracted largely from the recent conference papers by Vaidyanathan et al. [9,10] and Goel et al. [11].

Two types of injectors are commonly used in space propulsion. These are (i) coaxial injectors and (ii) impinging type injectors. In coaxial injectors, the propellant streams flow in parallel. There are different types of coaxial injectors namely, shear coaxial, and swirl coaxial injectors. In the shear coaxial injector element mixing is accomplished through a shear-mixing process [99]. In impinging type injectors, the mixing occurs by direct impingement of the propellant streams at an acute angle. Calhoon et al. [99] conducted a large number of cold-flow and hot-fire tests over a range of propellant mixture ratios, propellant velocity ratios and chamber pressure for shear coaxial, swirl coaxial, impinging, and premixed elements. The data were correlated directly with injector/chamber design parameters, which were recognized from both theoretical and empirical standpoints as the controlling variables. A schematic diagram of a single element injector proposed by Boeing is shown in Fig. 9.

Fig. 9. Schematic of Hybrid Boeing Element (US Patent 6253539).

8.1. Problem description

The injector design has two primary objectives: improvement of performance and life. As discussed by Vaidyanathan et al. [9] the performance of the injector is indicated by the axial length of the thrust chamber, while the life of the injector is associated with the thermal field inside the thrust chamber. A visual representation of the objectives is shown in Fig. 10.

In summary, the objectives are listed as the following:

Combustion length (X_{cc}): This is defined as the distance from the inlet where 99% of the combustion is complete. It is desirable to keep the combustion length as small as possible as this directly affects the size of the combustor and hence the weight of the spacecraft.

Wall temperature (TW_4): This is defined as the wall temperature at $3''$ from the injector face. Higher values of the wall temperature have negative impact on the life of the injector, so this objective has to be minimized.

Face temperature (TF_{max}): This is defined as the maximum temperature of the injector face. It is desirable to reduce temperature to increase the life of the injector.

Tip temperature (TT_{max}): This is defined as the maximum temperature on the post tip of the injector. It is desirable to keep this temperature as low as possible.

The independent variable ranges considered in this study are shown in Table 3.

It can be seen that the dual goal of maximizing the performance and the life is now cast as a four-objective design problem. Improved performance of the injector causes development of higher temperatures in the thrust chamber, so as discussed by Goel et al. [11], some of these objectives are conflicting in nature, hence there cannot be a single optimal solution for this problem.

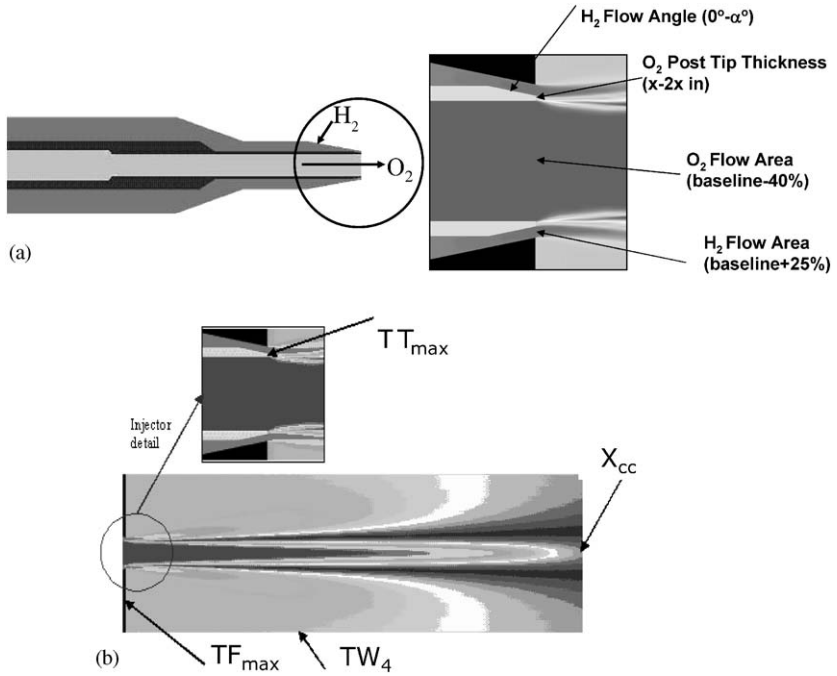

Fig. 10. Design variables and objectives of the single element rocket injector: (a) design variables; (b) objective functions.

Table 3
Range of design variables (α is an acute angle in degrees and X is the thickness of OPTT in inches)

Variable	Minimum		Maximum	
	Actual	Scaled	Actual	Scaled
α	0°	0	α^0	1
ΔHA	Baseline	0	Baseline + 25%	1
ΔOA	Baseline − 40%	0	Baseline	1
OPTT	X in	0	$2X$ in	1

There are four primary design variables for the injector design problem shown in Fig. 10. These variables are given as follows:

Flow Angle (α): This variable represents the hydrogen flow angle. The maximum angle varies between 0^0 to α^0.

Hydrogen area (ΔHA): This variable represents the increment with respect to the baseline cross-section area of the tube carrying hydrogen. The increment varies from 0% to 25% of the baseline hydrogen area.

Oxygen area (ΔOA): This variable represents the decrement with respect to the baseline cross-section area of the tube carrying oxygen. The area varies between 0% and (−40)% of the baseline area.

Oxidizer post tip thickness (*OPTT*): Oxidizer post tip thickness varies between X'' to $2X''$, where X is a dimension, which cannot be disclosed because of confidentiality restrictions.

All the variables are linearly normalized between 0 and 1. More information about the design variables can be found in Vaidyanathan et al. [9].

The objective functions values corresponding to a particular design can be obtained through computationally expensive CFD simulations. Each CFD simulation can be obtained from a pressure-based Navier–Stokes solver, FDNS500-CVS [100–102]. In addition to the Favre-averaged Navier–Stokes equations, the two-equation turbulence model, and kinetic equations are solved. Steady-state solutions are reached by an implicit Euler time marching scheme. The chemical species transport equations represented the H_2–O_2 chemistry with the aid of a 7-species and 9-reaction set [100–102]. The simulation domain and the boundary conditions used in all the CFD cases are shown in Fig. 11. Because of the very large aspect ratio, both the injector and chamber have been shortened (at the cross hatched areas) for clarity. This is a computationally expensive simulation-based problem so for optimization purposes it is advisable to develop the surrogate models for the objective functions. It was shown by Vaidyanathan et al. [9,103], Shyy et al. [8], and Papila et al. [6,104] that accurate response surfaces for complex problems, like single element injectors, can be developed.

Comparing two of the evaluated designs (using CFD), Vaidyanathan et al. [9] have shown the motivation for using surrogate models and an efficient optimization technique in the design process. The independent design variables, normalized between 0 and 1, are shown for the

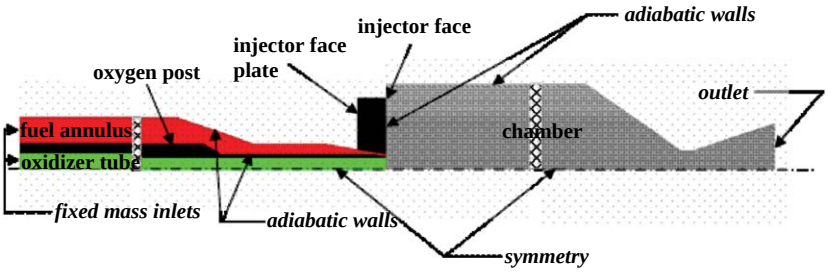

Fig. 11. Simulation domain and boundary conditions of the injector flow model.

Table 4
Independent variables and objectives for Cases X and Y from CFD computations (normalized values shown)

Case	α	ΔHA	ΔOA	OPTT	TF_{max}	TW_4	TT_{max}	X_{cc}
X	1.0	0.0	0.0	0.0	0.998	0.927	0.128	−0.004
Y	0.0	0.5	0.5	1.0	0.285	0.395	0.923	0.567

two cases in Table 4. In terms of the design space evaluated, these two designs (X and Y) are seen to be quite different.

The chamber wall and injector face temperatures (TW_4 and TF_{max}, respectively) for Case Y are low or moderately low, while for Case X, they are high. It was seen that a large recirculation zone located between the injector and the chamber wall strips hot gases from the flame and causes them to flow back along the chamber wall and injector face [9]. This phenomenon regulates the chamber wall temperature TW_4 and the injector face temperature TF_{max}. The other life-indicating variable, the maximum oxidizer post tip temperature, TT_{max} has essentially the opposite trend as compared to the other two temperatures. The performance indicator, combustion length X_{cc} is seen (Table 5) to be at a minimum level for Case X (shorter combustion lengths indicate better mixing elements) and at a moderate level for Case Y. It is clear that the performance and environment indicators exhibit competing trends such that no design is the "best" for all the indicators of performance and environment.

These comparisons confirm the earlier statement that changes in the injector design details have major effects on injector performance and injector-generated environments. These, and other relevant geometric details, can be effectively addressed by modern computation-based design tool.

8.2. Design of experiments (DOE)

The selected DOE scheme was orthogonal arrays (Section 3). Specifically, an OA (54, 4, 3, 2) was used to generate fifty-four (54) designs, considering four (4) factors, three (3) levels, with strength 2. For model selection and validation the split-sample approach (Section 5) was selected so from the fifty-four (54) designs, fourteen (14) were set aside for testing purposes. The fifty-four (54) designs specified by the orthogonal array were computed on an axisymmetric geometry with 336×81 nodes. Only 33 out of the 40 training cases gave valid results. Results of the remaining seven cases contained unsteady features, which did not represent solutions of the steady-state model employed. Based on the quality of the approximation obtained using the selected surrogate model (polynomial regression) for TT_{max} and X_{cc}, it was noticed that the grid distribution in the combustion zone was insufficient. The grid was then refined to a 430×81 grid (Fig. 12). The refined grid was found to be appropriate and was selected for the purposes of the optimization study. Only 2 out of the forty (40) designs to be used for fitting the surrogate model gave unacceptable results (i.e., unsteady features were present). Note that in addition to facilitating design optimization, surrogate models can also help check the adequacy of CFD simulations via identification of outliers.

Hence, the final data set included thirty-eight (38) designs for fitting the surrogate model and fourteen (14) to test their predictive capabilities. The fitting and testing design points can be found in Vaidyanathan et al. [9].

8.3. Construction of the surrogate model

The parametric polynomial regression approach (Section 4) was selected as surrogate model. Before constructing the surrogates, all the design variables were scaled between zero and one based on their upper and lower bounds. For example, when the ΔOA is 1 or 0, the O_2 flow area is reduced by 0% or 40%, respectively, as compared to the baseline area. Similarly, the objective values were scaled between zero and one based on the upper and lower bounds observed in the sample data. Once the polynomial approximations were available, the scaled objective values were normalized using the surrogate-based minimum and maximum values for the objectives.

Table 5
Minimizing individual objectives. Value in parenthesis (1) indicates which objective function is minimized (normalized values shown)

Opt-case	α	ΔHA	ΔOA	OPTT	TF_{max}	TW_4	TT_{max}	X_{cc}
1	0	1	0.592	1	0.000 (1)	0.0725 (0)	0.914 (0)	0.769 (0)
CFD					−0.00207	0.0656	0.936	0.758
Error (%)					0.110	0.080	0.700	0.250
2	0	1	0	1	0.0309 (0)	0.000 (1)	1.0 (0)	0.440 (0)
CFD					0.0910	0.0461	0.911	0.568
Error (%)					0.900	0.570	2.84	2.93
3	1	0	1	0	0.944 (0)	0.976 (0)	0.0 (1)	0.153 (0)
CFD					0.943	0.969	0.103	0.158
Error (%)					0.0100	0.0800	4.46	0.120
4	0.917	0	0	0	0.987 (0)	0.926 (0)	0.182 (0)	0.000 (1)
CFD					0.981	0.919	0.119	−0.004
Error (%)					0.0800	0.0800	2.69	0.120

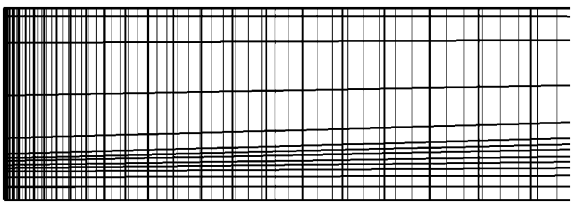

Fig. 12. Comparing the unrefined (336 × 81) (thicker lines) and refined (430 × 81) (thinner lines) grids.

The polynomial-based approximations were created using standard least-squares regression [24], and terms with insignificant influence (based on t-statistics) on the prediction of the response/objective were discarded, thereby improving the parsimony of the surrogate model. The statistical analysis software JMP [106] was used for the generation of the polynomials. The quality of alternative polynomial approximations was evaluated by comparing their adjusted root mean square (rms)-error σ_a, and adjusted coefficient of multiple determinations, R_a^2 [24] corresponding to the training set, and those with the best performance were selected (model selection). The model evaluation process was conducted using the rms-error σ for the test set. Details of the model selection and validation (Section 5) process can be found in Vaidyanathan et al. [9].

The polynomial approximations corresponding to each of the objectives are presented below. Note that the best approximations for objectives TF_{max}, TW_4 and X_{cc} are quadratic polynomials, while for objective TT_{max} a reduced cubic polynomial represented the best choice.

$$\begin{aligned} TF_{max} = {} & 0.692 + 0.477(\alpha) - 0.687(\Delta HA) \\ & - 0.080(\Delta OA) - 0.0650(OPTT) \\ & - 0.167(\alpha)^2 - 0.0129(\Delta HA)(\alpha) \\ & + 0.0796(\Delta HA)^2 - 0.0634(\Delta OA)(\alpha) \\ & - 0.0257(\Delta OA)(\Delta HA) + 0.0877(\Delta OA)^2 \\ & - 0.0521(OPTT)(\alpha) + 0.00156(OPTT)(\Delta HA) \\ & + 0.00198(OPTT)(\Delta OA) \\ & + 0.0184(OPTT)^2, \end{aligned} \tag{47}$$

$$\begin{aligned} TW_4 = {} & 0.758 + 0.358(\alpha) - 0.807(\Delta HA) \\ & + 0.0925(\Delta OA) - 0.0468(OPTT) \\ & - 0.172(\alpha)^2 + 0.0106(\Delta HA)(\alpha) \\ & + 0.0697(\Delta HA)^2 - 0.146(\Delta OA)(\alpha) \\ & - 0.0416(\Delta OA)(\Delta HA) + 0.102(\Delta OA)^2 \\ & - 0.0694(OPTT)(\alpha) - 0.00503(OPTT)(\Delta HA) \\ & + 0.0151(OPTT)(\Delta OA) + 0.0173(OPTT)^2, \end{aligned} \tag{48}$$

$$\begin{aligned} TT_{max} = {} & 0.370 - 0.205(\alpha) + 0.0307(\Delta HA) \\ & + 0.108(\Delta OA) + 1.019(OPTT) \\ & - 0.135(\alpha)^2 + 0.0141(\Delta HA)(\alpha) \\ & + 0.0998(\Delta HA)^2 + 0.208(\Delta OA)(\alpha) \\ & - 0.0301(\Delta OA)(\Delta HA) - 0.226(\Delta OA)^2 \\ & + 0.353(OPTT)(\alpha) - 0.0497(OPTT)(\Delta OA) \\ & - 0.423(OPTT)^2 + 0.202(\Delta HA)(\alpha)^2 \\ & - 0.281(\Delta OA)(\alpha)^2 - 0.342(\Delta HA)^2(\alpha) \\ & - 0.245(\Delta HA)^2(\Delta OA) + 0.281(\Delta OA)^2(\Delta HA) \\ & - 0.184(OPTT)^2(\alpha) \\ & + 0.281(\Delta HA)(\alpha)(\Delta OA), \end{aligned} \tag{49}$$

$$X_{cc} = 0.153 - 0.322(\alpha) + 0.396(\Delta HA) + 0.424(\Delta OA) + 0.0226(OPTT) + 0.175(\alpha)^2 + 0.0185(\Delta HA)(\alpha) - 0.0701(\Delta HA)^2 - 0.251(\Delta OA)(\alpha) + 0.179(\Delta OA)(\Delta HA) + 0.0150(\Delta OA)^2 + 0.0134(OPTT)(\alpha) + 0.0296(OPTT)(\Delta HA) + 0.0752(OPTT)(\Delta OA) + 0.0192(OPTT)^2. \quad (50)$$

8.4. Global sensitivity analysis

While it is well established that variations in injector geometry can have a significant impact on performance and environmental objectives such as combustion chamber wall and injector face temperatures and heat fluxes [107], these sensitivities are seldom quantified. As shown in Vaidyanathan et al. [10], once surrogates become available, these sensitivities can be computed using Sobol's method (Section 6).

In the context of the case study, a design variable is considered essential if it is responsible for at least 5% of the objective variability. Fig. 13 shows the percentage of main factor (S_i) contribution of different design variables to individual objectives. The variability of TF_{max} is largely influenced by ΔHA and moderately by α (Fig. 13a). The effect of the other design variables is marginal, suggesting that they are non-essential and in principle could be fixed. The variability of TW_4 is considerably influenced only by ΔHA (Fig. 13b). TT_{max} is influenced considerably by OPTT and marginally by α (Fig. 13c). For X_{cc}, ΔHA, ΔOA and α have considerable influence (Fig. 13d). From the comparison of total sensitivity indices (S_i^{total}) with the main factors (S_i) it is concluded that the contributions of the cross-interactions among the design variables to the objectives variability are negligible.

The sensitivity information cited above can be used for screening purposes and therefore, ease the search for optimum designs. In fact, surrogate models that included only essential variables (with the non essential

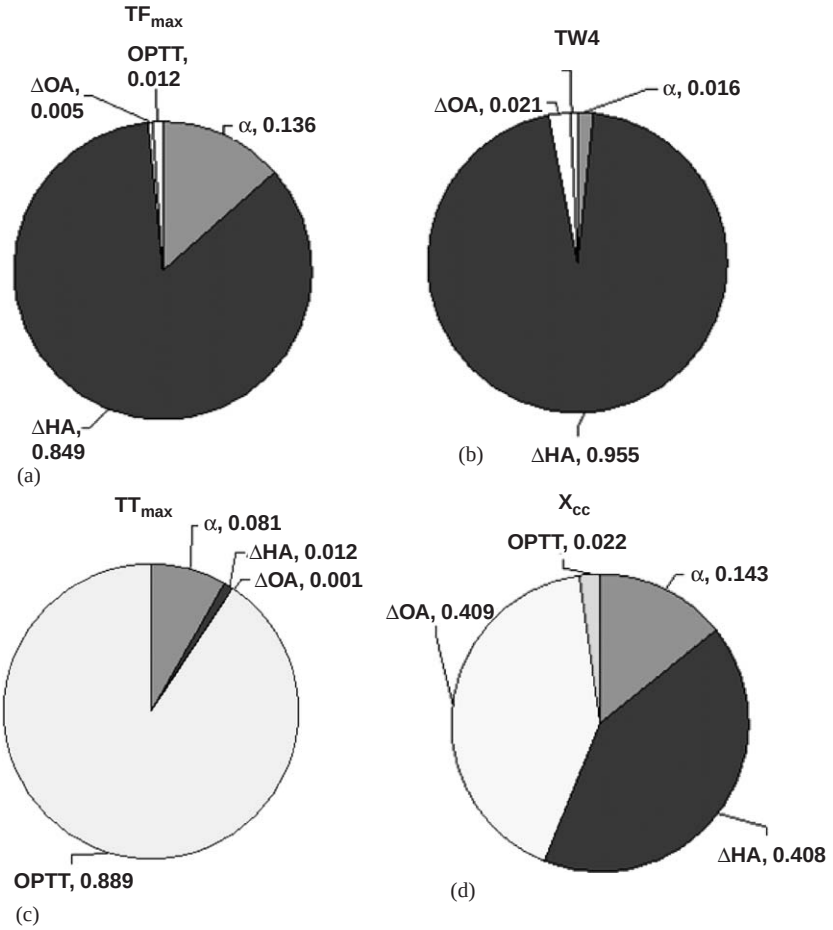

Fig. 13. Main factor (S_i) influence on objective variability: (a) TF_{max}, (b) TW_4, (c) TT_{max} and (d) X_{cc}.

fixed at their mean values) were constructed and differences in the predictions were found insignificant when compared with those provided by the original (all design variables included) polynomial approximations. However, since the number of design variables in the case study is small, the original polynomial approximations were adopted. Note that the sensitivity analysis provides an insight into the physics of a design problem by highlighting the design features that govern the individual objectives.

8.5. Surrogate-based optimization

Considering the problem of interest is one of multi-objective optimization, the optimal solutions are known as Pareto-optimal and their objective function space representation is called the Pareto-optimal front (POF). A solution is called Pareto-optimal (or efficient) solution, if there is no other solution for which at least one objective has a better value while values of the remaining objectives are the same or better. Since surrogates for each of the objectives are available, any standard multi-objective optimization methods could be adopted. In the context of the case study, though, the results obtained by Goel et al. [11] using an evolutionary multi-objective optimization procedure (multi-objective genetic algorithm with ε-constraint strategy) are presented.

Extreme values of the Pareto-optimal front, that is, the results of optimizing (here, minimizing) each objective separately are shown in Table 5. A number one in parentheses indicates the objective selected for optimization. As expected based on physical grounds, the results show there is a strong correlation between TF_{max} and TW_4; a correlation analysis shows a statistically significant correlation coefficient close to 1. As a result, the multi-objective problem can be addressed using only three of the original objectives (TF_{max}, TT_{max}, X_{cc}). In addition, note that the geometrical characteristics of the optimal designs are aligned with the objectives, and as a result, the design that minimizes TF_{max} corresponds to a shear-coaxial injector, that is, $\alpha = 0$, and hydrogen flow area, ΔHA, and oxidizer post tip thickness, OPTT, at their maximum values; other objectives exhibit relative high values as expected for this type of injector. The results also show a correlation between TT_{max} and X_{cc}, although not as strong as the one for TF_{max} and TW_4. When TT_{max} is minimized, X_{cc} exhibits a low value, and vice versa. The optimal designs associated with the minimization of TT_{max} and X_{cc} correspond to impinging-like injectors with the hydrogen flow angle, α, at or near its maximum value and ΔHA and OPTT at their minimum values, yielding significantly shorter combustion lengths than those corresponding to shear coaxial designs. Both minimizations (TT_{max} and X_{cc}) result in very high values for TF_{max}. Considering that the optimization of each objective in isolation resulted in alternative designs, and that objectives are in clear conflict (TF_{max}, TT_{max}; TF_{max}, X_{cc}) a full examination of the Pareto-optimal set is required. Fig. 14 displays the two hundred and forty nine (249) Pareto-optimal solutions found. In order to analyze the POF a representative set of solutions (9) was identified using a hierarchical clustering algorithm. Values of the design variables and objectives for these designs are shown in Table 6. Graphically, these representative solutions are highlighted on the POF in Fig. 14; note that the solutions are uniformly distributed along the POF. Box plots for the design variables and objectives in selected clusters (1, 3, 6, 9) are shown in Figs. 15 and 16, respectively. These plots highlight the variability of the design variables and objectives in each cluster.

For cluster 1 it is seen that the value of α is fixed at 0 (shear coaxial injector) (Fig. 15a) and ΔHA is fixed at 1 (Fig. 15b). This suggests that in this cluster, the designs are sensitive to α and ΔHA, both of which reach their extreme values. The remaining two design variables

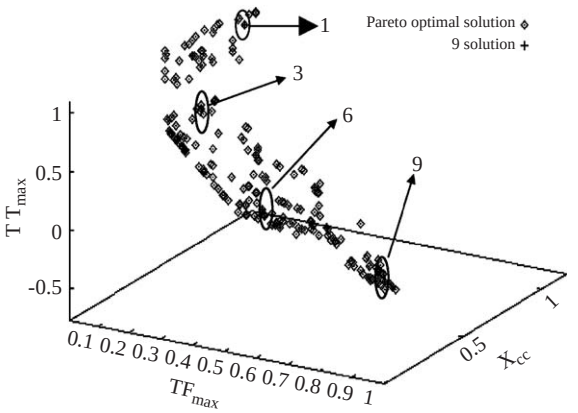

Fig. 14. Pareto optimal solution set and nine (9) representative solutions from the same set. The circles identify the representative of a specific cluster.

Table 6
Objective function and design variables of nine (9) representative designs from the Pareto optimal solution

Cluster	α	ΔHA	ΔOA	OPTT	TF_{max}	X_{cc}	TT_{max}
1	0.000	1.000	0.842	0.712	0.0231	1.090	0.880
2	0.000	1.000	0.356	0.587	0.0276	0.749	0.890
3	0.000	1.000	0.442	0.0144	0.0541	0.750	0.452
4	0.0939	1.000	0.000	0.0146	0.126	0.453	0.466
5	0.668	1.000	0.732	0.000	0.259	0.681	0.229
6	0.600	0.670	0.000	0.000	0.489	0.264	0.226
7	0.295	0.108	0.000	0.354	0.719	0.129	0.641
8	0.314	0.0656	0.000	0.0554	0.776	0.0969	0.357
9	1.000	0.0140	0.680	0.000	0.935	0.138	−0.0432

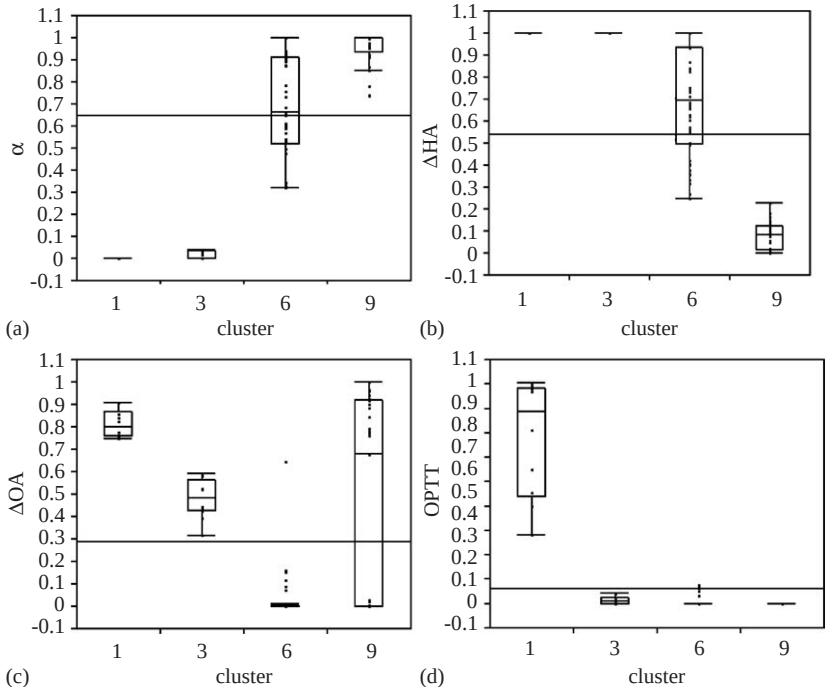

Fig. 15. Box plots for the design variables in cluster 1, 3, 6 and 9, (a) α (b) ΔHA (c) ΔOA and (d) OPTT. The horizontal line shows the mean value for all the variables on POF.

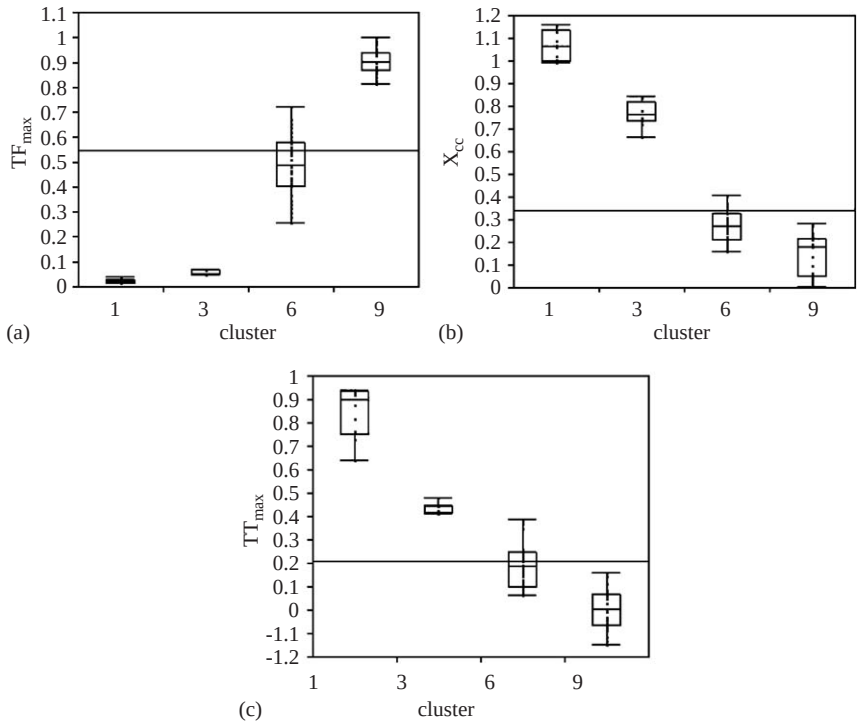

Fig. 16. Box plots for the objectives in cluster 1, 3, 6 and 9, (a) TF_{max} (b) X_{cc} and (c) TT_{max}. The horizontal line shows the mean value of the objectives on POF.

ΔOA (Fig. 15c) and OPTT (Fig. 15d), vary over a range. It is observed that TF_{max} is minimized (Fig. 16a) where as X_{cc} and TT_{max} lie near their maximum and have little variability. Hence the designs in cluster 1 tend to minimize TF_{max} and represent shear coaxial injector designs. The design variables ΔOA and OPTT do not influence TF_{max} but affect the remaining objectives, X_{cc} and TT_{max}. Partial correlation coefficients are estimated to obtain the relationship between these design variables and objectives (Table 7). It is noticed that as ΔOA increases, X_{cc} increases ($R_{corr} = 1.00$) and TT_{max} decreases ($R_{corr} = -0.638$). As OPTT decreases both X_{cc} and TT_{max} decrease ($R_{corr} = 1.00$ for both).

Similar observations can be made for clusters 3, 6 and 9. Figs. 16a–c show that in different clusters, X_{cc} and TT_{max} decrease with increments in TF_{max} (based on the median of the box plots). This highlights the trade-off between the objectives. Cluster 9 provides information about the opposing trend as to what was observed in cluster 1. As objectives X_{cc} and TT_{max} are minimized (Figs. 16b and c) an impinging injector design is obtained ($\alpha \sim 1$, Fig. 15a) with TF_{max} exhibiting high values (Fig. 16a). The ΔHA is near minimum (Fig. 15b) contrary to the design in cluster 1 where high ΔHA minimized TF_{max}. The ΔOA has considerable variation which suggests that the variability of objectives, X_{cc} and TT_{max} are not largely affected by this design variable. Fig. 15d shows that X_{cc} and TT_{max} are minimized for the minimum value of OPTT. Table 7 gives the partial correlation coefficients for the set of design variables and the objectives in each cluster which shows considerable variation. The partial correlation coefficients for the combinations left out are effectively zero.

Finally, the Pareto fronts of the objectives known to be in conflict (TF_{max}, X_{cc}) and (TF_{max}, TT_{max}) are shown. Fig. 17a shows the relation between TF_{max} and X_{cc}. The O_2 post-tip temperature, TT_{max} is ignored. It can be seen that the POF in this case is linear over a large region. A small increase in the value of TF_{max} ($\approx 10\%$) reduces the combustion length, X_{cc}, by nearly 50%. Fig. 17b shows the relation between TF_{max} and TT_{max}. The combustion length, X_{cc} is ignored. It is obvious that the POF is non-convex. It is also seen that a small drop in the value of the face temperature ($\approx 10\%$) can reduce the tip temperature TT_{max} by nearly 60%. Hence at a small cost of TF_{max} both X_{cc} and TT_{max} can improve considerably.

Table 7
Partial correlation coefficients (R_{corr}) of design variables vs. objectives for different clusters along the POF

Cluster	Design variable	TF_{max}	X_{cc}	TT_{max}
1	ΔOA	—	1.000	−0.638
	OPTT	—	1.000	0.991
3	ΔOA	—	1.000	—
6	α	0.982	−0.735	−0.983
	ΔHA	−0.999	0.994	−0.729
9	α	0.877	−0.203	−0.769
	ΔHA	−0.992	0.983	0.816
	ΔOA	−0.977	0.997	−0.940

9. Summary and conclusions

The fundamental issues that arise in the SBAO of computationally expensive models such as those found in aerospace systems were discussed. The issues included the selection of the loss function and regularization criterion to identify the surrogates, design of experiments, surrogate selection and identification, sensitivity analysis, and how to incorporate surrogates in optimization efforts when constraints are present. Some of these issues were demonstrated through the multi-objective optimal design of a liquid rocket injector.

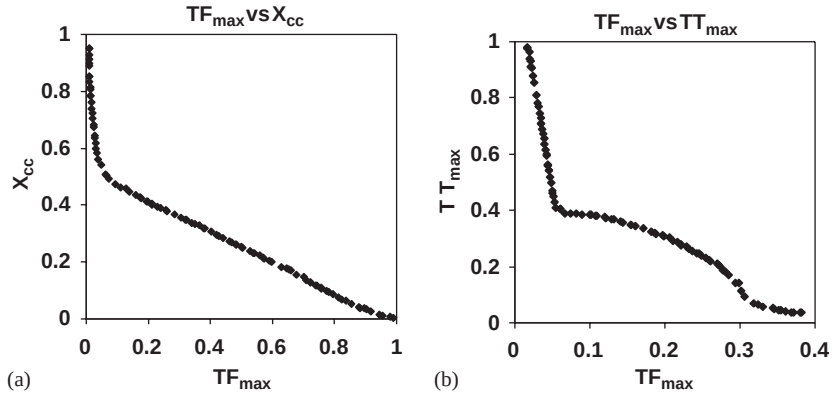

Fig. 17. Pareto optimal front: (a) TF_{max} vs. X_{cc} (b) TF_{max} vs. TT_{max}.

In the context of the case study, the SBAO approach demonstrated its effectiveness by solving a model for a critical problem in space propulsion: the multi-objective optimization of liquid injector rocket injector designs.

New, potentially more effective methods and techniques are increasingly available to address some of the above-referenced issues, including: improvements on the commonly used orthogonal arrays and Latin hypercubes for design of experiments, merit functions for sampling that account for model appraisal information, alternative loss functions for modeling (e.g., e-sensitive) which has led to the development of more robust approximation schemes (e.g., support vector regression), the use of multiple surrogates to address the issue of model uncertainty, more efficient strategies for model selection and evaluation (e.g., bootstrapping), efficient global sensitivity methods for screening and general sensitivity evaluations, and multiple-surrogate based optimization strategies.

SBAO is an active area of research, and has made substantial progress in addressing the analysis and optimization of a variety of aerospace systems. It has the potential of playing a vital role in the successful full-scale development of modern aerospace systems while effectively considering the competing needs of improving performance, reducing costs, and enhancing safety.

Acknowledgments

The present efforts have been supported by the Institute for Future Space Transport, under the NASA Constellation University Institute Program (CUIP), Ms. Claudia Meyer program monitor.

References

[1] Knill DL, Giunta AA, Baker CA, Grossman B, Mason WH, Haftka RT, Watson LT. Response surface models combining linear and euler aerodynamics for supersonic transport design. J Aircr 1999;36(1):75–86.

[2] Rai MM, Madavan NK. Aerodynamic design using neural networks. Proceedings of the seventh AIAA/USAF/NASA/ISSMO symposium on multidisciplinary analysis and optimization, St. Louis, September 1998: Paper no. 98–4928, AIAA.

[3] Rai MM, Madavan NK. Improving the unsteady aerodynamic performance of transonic turbines using neural networks. Proceedings of the 38th AIAA aerospace sciences meeting and exhibit, Reno, NV, January 2000. Paper no. 2000–0169, AIAA.

[4] Madavan NK, Rai MM, Huber FW. Neural net-based redesign of transonic turbines for improved unsteady aerodynamic performance. AIAA/SAE/ASME/ASEE 35th joint propulsion conference, June 1999. Paper no. 99–2522, AIAA.

[5] Madsen JI, Shyy W, Haftka RT. Response surface techniques for diffuser shape optimization. AIAA J 2000;38:1512–8.

[6] Papila N, Shyy W, Griffin L, Dorney DJ. Shape optimization of supersonic turbines using global approximation methods. J Propulsion Power 2002;18:509–18.

[7] Shyy W, Papila N, Tucker PK, Vaidyanathan R, Griffin L. Global optimization for fluid machinery applications; keynote paper. Proceedings of the second international symposium on fluid machinery and fluid engineering (ISFMFE), Beijing, China, October 2000: p. 1–10.

[8] Shyy W, Papila N, Vaidyanathan R, Tucker PK. Global design optimization for aerodynamics and rocket propulsion components. Prog Aerospace Sci 2001;37:59–118.

[9] Vaidyanathan R, Tucker PK, Papila N, Shyy W. CFD-based optimization for a single element rocket injector. J Propulsion Power 2004;20(4):705–17.

[10] Vaidyanathan R, Goel T, Shyy W, Haftka RT, Queipo NV, Tucker PK. Global sensitivity and trade-off analyses for multi-objective liquid rocket injector design. Proceedings of the 40th AIAA/ASME/SAE/ASEE joint propulsion conference and exhibit, Fort Lauderdale, FL, July 2004. Paper no. 2004–4007, AIAA.

[11] Goel T, Vaidyanathan R, Haftka RT, Queipo NV, Shyy W, Tucker PK. Response surface approximation of pareto optimal front in multi-objective optimization. In: Proceedings of the 10th AIAA/ISSMO multidisciplinary analysis and optimization conference, Albany, NY, August–September 2004. Paper no. 2004–4501, AIAA.

[12] Snieder R. The role of non-linearity in inverse problems. Inverse Problems 1998;14:387–404.

[13] Simpson TW, Mauery TM, Korte JJ, Mistree F. Kriging models for global approximation in simulation-based multidisciplinary design optimization. AIAA J 2001; 39(12):2233–41.

[14] McDonald D, Grantham W, Tabor W, Murphy M. Response Surface model development for global/local optimization using radial basis functions. Proceedings of the eighth AIAA/USAF/NASA/ISSMO symposium on multidisciplinary analysis and optimization, Long Beach, CA, September 2000. Paper no. 2000–4776, AIAA.

[15] Chung HS, Alonso JJ. Comparison of approximation models with merit functions for design optimization. Proceedings of the eigthth AIAA/USAF/NASA/ISSMO symposium on multidisciplinary analysis and optimization, Long Beach, CA, September 2000. Paper no. 2000–4754, AIAA.

[16] Jin R, Chen W, Simpson TW. Comparative studies of metamodeling techniques under multiple modeling criteria. Struct Multidisciplinary Optim 2001;23:1–13.

[17] Sacks J, Welch WJ, Mitchell TJ, Wynn HP. Design and analysis of computer experiments. Stat Sci 1989;4:409–35.

[18] Tikhonov AN, Arsenin VY. Solutions to ill-posed problems. New York: Wiley; 1977.

[19] Morozov VA. Methods for solving incorrectly posed problems. Berlin: Springer; 1984.

[20] Ariew R. Ockham's Razor: a historical and philosophical analysis of Ockham's Razor Principle of Parsimony. Philosophy, University of Illinois, Champane-Urbana; 1976.

[21] Tenorio L. Statistical regularization of inverse problems. SIAM Rev 2001;43:347–66.
[22] Vapnik VN. Statistical learning theory. New York: Wiley; 1998.
[23] Girosi F. An equivalence between sparse approximation and support vector machines. Neural Comput 1998;10(6):1455–80.
[24] Myers RH, Montgomery DC. Response surface methodology—process and product optimization using designed experiment. New York: Wiley; 1995.
[25] Qu X, Venter G, Haftka RT. New formulation of minimum-bias central composite experimental design and Gauss quadrature. Struct Multidisciplinary Optim 2004;28(4):231–42.
[26] Box G, Draper N. A basis for the selection of a response surface design. J Am Stat Assoc 1959;54:622–54.
[27] Sacks J, Ylvisaker D. Some model robust designs in regression. Ann Stat 1984;12:1324–48.
[28] Tang B. Orthogonal array-based latin hypercubes. J Am Stat Assoc 1993;88:1392–7.
[29] Johnson M, Moore L, Ylvisaker D. Minimax and maximin distance designs. J Stat Plann Inference 1990;26:131–48.
[30] Iman R, Conover W. A distribution-free approach to inducing rank correlation among input variables. Commun Stat, Part B—Simulat Comput 1982;11:311–34.
[31] Owen AB. Controlling correlations in latin hypercube samples. J Stat Assoc 1994;89:1517–22.
[32] Hedayat A, Sloane N, Stufken J. Orthogonal arrays: theory and applications. Springer, Series in Statistics, Berlin: Springer, 1999.
[33] 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.
[34] Ye K. Orthogonal column latin hypercubes and their application in computer experiments. J Am Stat Assoc 1998;93:1430–9.
[35] Palmer K, Tsui K. A minimum bias latin hypercube design. IIE Trans 2001;33:793–808.
[36] Leary S, Bhaskar A, Keane A. Optimal orthogonal array-based latin hypercubes. J Appl Statist 2003;30:585–98.
[37] Stein M. Large sample properties of simulations using latin hypercube sampling. Technometrics 1987;29:143–51.
[38] Owen A. Orthogonal arrays for computer experiments, integration and visualization. Stat Sin 1992;2:439–52.
[39] Ye K, Li W, Sudjianto A. Algorithmic construction of optimal symmetric latin hypercube designs. J Stat Plann Inference 2000;90:145–59.
[40] Jones D, Schonlau M, Welch W. Expensive global optimization of expensive black-box functions. J Global Optim 1998;13:455–92.
[41] Sasena M, Papalambros P, Goovaerts P. Metamodeling sampling criteria in a global optimization framework. Proceedings of the eighth AIAA/NASA/USAF/ISSMO symposium on multidisciplinary optimization, Long Beach, CA, September 2000. Paper no. 2000–4921, AIAA.
[42] Williams B, Santner T, Notz W. Sequential design of computer experiments to minimize integrated response functions. Stat Sin 2000;10:1133–52.
[43] Zerpa L, Queipo N, Pintos S, Salager J. An optimization methodology for alkaline–surfactant–polymer flooding processes using field scale numerical simulations and multiple surrogates. Proceedings of the 14th SPE/DOE symposium on improved oil recovery, Tulsa, OK, 2004, SPE-89387.
[44] Draper NR, Smith H. Applied regression analysis, 3rd Edition. New York: Wiley; 1998.
[45] Matheron G. Principles of geostatistics. Econ Geol 1963;58:1246–66.
[46] Sacks J, Schiller S, Welch W. Designs for computer experiments. Technometrics 1989;31:41–7.
[47] Sacks J, Welch W, Mitchell T, Wynn H. Design and analysis of computer experiments. Stat Sci 1993;4: 409–23.
[48] Williams CKI, Rasmussen CE. Gaussian processes for regression. In: Touretzky DS, Mozer MC, Hasselmo ME, editors. Advances in neural information processing systems 8. Cambridge: MIT Press; 1996. p. 514–20.
[49] Gibbs M. Bayesian Gaussian processes for regression and classification. Ph.D. thesis, Cambridge University, 1997.
[50] Rao CR. Linear statistical inference and its applications, 2nd ed. New York: Wiley; 2002.
[51] Poggio T, Smale S. The mathematics of learning: dealing with data. Not Am Math Soc 2003;50:537–44.
[52] Friedman J, Stuetzle W. Projection pursuit regression. JASA: Theory Methods 1981;76:817–23.
[53] Yakowitz S, Szidarovsky F. A comparison of kriging with non-parametric regression. J Multivariate Anal 1985;16: 21–53.
[54] Laslett G. Kriging and Splines: an empirical comparison of their predictive performance in some applications. JASA: Appl Case Stud 1994;89:391–400.
[55] Giunta A, Watson L, Koehler J. A comparison of approximation modeling techniques: polynomial versus interpolating models. Proceedings of the seventh AIAA/USAF/NASA/ISSMO symposium on multidisciplinary analysis and optimization, St. Louis, MO, 1998. Paper no. 98–4758, AIAA.
[56] Perrone M, Cooper L. When networks disagree: ensemble methods for hybrid neural networks. In: Mammone RJ, editor. Artificial neural networks for speech and vision. London: Chapman & Hall; 1993. p. 126–42.
[57] Bishop C. Neural networks for pattern recognition. Oxford: Oxford University Press; 1995. pp. 364–9.
[58] Martin JD, Simpson TW. On the use of kriging models to approximate deterministic computer models. AIAA J 2005 in press.
[59] Golub G, Heath M, Wahba G. Generalized cross-validation as a method for choosing a goodridge parameter. Technometrics 1979;21:215–23.
[60] Wahba G. Bayesian confidence interval for the cross-validated smoothing spline. J R Stat Soc B 1983;45: 133–50.
[61] Hutchinson M, de Hoog F. Smoothing noisy data with spline functions. Numer Math 1985;47:99–106.
[62] Hansen PC. Analysis of discrete ill-posed problems by means of the L-curve. SIAM Rev 1992;34:561–80.
[63] Efron B. Estimating the error rate of a prediction rule: improvement on cross-validation. J Am Stat Assoc 1983;78:316–31.

[64] Hall P. On the bootstrap and confidence intervals. Ann Stat 1986;14:1431–52.
[65] Efron B, Tibshirani R. Introduction to the bootstrap. New York: Chapman & Hall; 1993.
[66] Othani K. Bootstrapping R^2 and adjusted R^2 in regression analysis. Econ Modell 2000;17:473–83.
[67] Kleijnen J, Deflandre D. Validation of regression metamodels in simulation: bootstrap approach. Eur J Oper Res 2004 in press.
[68] Iman R, Helton J. An investigation of uncertainty and sensitivity analysis techniques for computer models. Risk Anal 1988;8:71–90.
[69] Kleijnen J. Sensitivity analysis and related analyses: a review of some statistical techniques. J Stat Comput Simulat 1997;57:111–42.
[70] Frey H, Patil S. Identification and review of sensitivity analysis methods. Risk Anal 2002;22:553–78.
[71] Morris M. Factorial sampling plans for preliminary computational experiments. Technometrics 1991;33: 161–74.
[72] Andres T, Hajas W. Using iterated fractional factorial design to screen parameters in sensitivity analysis of a probabilistic risk assessment model. Joint international conference on mathematical methods and supercomputing in nuclear applications, vol. 2. Kuster, Stein and Werner, 1993. 328–37.
[73] Saltelli A, Andres T, Homma T. Sensitivity analysis of model output. Performance of the iterated fractional factorial design method. Comput Stat Data Anal 1995;20: 387–407.
[74] Sobol I. Sensitivity estimates for non-linear mathematical models. Math Modeling Comput Exp 1993;4: 407–14.
[75] Jacques J, Lavergne C, Devictor N. Sensitivity analysis in presence of model uncertainty and correlated inputs. Special issue of reliability engineering and system safety. SAMO 2004.
[76] Mack Y, Goel T, Shyy W, Haftka RT, Queipo NV. Multiple surrogates for the shape optimization of bluff body-facilitated mixing. Proceedings of the 43rd AIAA aerospace sciences meeting and exhibit, Reno, NV, 2005. Paper no. 2005–0333, AIAA.
[77] Homma T, Saltelli A. Importance measures in global sensitivity analysis of nonlinear models. Reliab Engng Syst Saf 1996;52:1–17.
[78] Saltelli A, Tarantola S, Chan K. A quantitative model-independent method for global sensitivity analysis of model output. Technometrics 1999;41(1):39–56.
[79] Goel T, Mack Y, Haftka RT, Shyy W, Queipo NV. Interaction between grid and design space refinement for bluff body-facilitated mixing. Proceedings of the 43rd AIAA aerospace sciences meeting and exhibit, Reno, NV, 2005. Paper no. 2005–0125, AIAA.
[80] Alexandrov NM. A trust region framework for managing approximation models in engineering optimization. Proceedings of the sixth AIAA/NASA/USAF multidisciplinary analysis & optimization symposium, Bellevue, WA, 1996. Paper no. 96–4102, AIAA.
[81] Torczon V, Trosset M. Using approximations to accelerate engineering design optimization. Proceedings of the seventh AIAA/USAF/NASA/ISSMO symposium on multidisciplinary analysis and optimization, St. Louis, Missouri, 1998. Paper no. 98–4800, AIAA.
[82] Queipo N, Goicochea J, Pintos S. Surrogate modeling-based optimization of SAGD processes. J Pet Sci Eng 2002;35:83–93.
[83] Queipo N, Verde A, Canelon J, Pintos S. Efficient global optimization for hydraulic fracturing treatment design. J Pet Sci Eng 2002;35:151–66.
[84] Queipo N, Pintos S, Rincon N, Contreras N, Colmenares J. Surrogate modeling-based optimization for the integration of static and dynamic data into a reservoir description. J Pet Sci Eng 2002;35:167–81.
[85] Watson A, Barnes R. Infill sampling criteria to locate extremes. Math Geol 1995;27(5):589–608.
[86] Lewis R. A trust region framework for managing approximation models in engineering optimization. Proceedings of the sixth AIAA/NASA/USAF multidisciplinary analysis and optimization symposium, Bellevue, WA, 1996. Paper no. 96–4101, AIAA.
[87] Rodriguez J, Renaud J, Watson L. Convergence of trust region augmented lagrangian methods using variable fidelity approximation data. Struct Optim 1998;15: 141–56.
[88] Booker A, Dennis J, Frank P, Serafini D, Torczon V, Trosset M. A rigorous framework for optimization of expensive functions by surrogates. Struct optim 1999;17: 1–13.
[89] Dennis J, Schnabel R. Numerical methods for unconstrained optimization and nonlinear equations. Englewood Cliffs, NJ: Prentice-Hall; 1983.
[90] Alexandrov N, Nielsen E, Lewis R, Anderson W. First order model management with variable fidelity physics applied to multi-element airfoil optimization. Proceedings of the eighth AIAA/USAF/NASA/ISSMO symposium on multidisciplinary analysis & optimization, Long Beach, CA, 2000. Paper no. 2000–4886, AIAA.
[91] Alexandrov N, Lewis R, Gumbert C, Green L, Newman P. Approximation and model management in aerodynamic optimization with variable fidelity models. J Aircr 2001;38(6):1093–101.
[92] Torczon V. On the convergence of pattern search algorithms. SIAM J Optim 1997;7:1–25.
[93] Marsden A, Wang M, Dennis J, Moin P. Optimal aeroacoustic shape design using the surrogate management framework. Optim Eng 2004;5:235–62.
[94] Luenberger D. Linear and non-linear programming, 2nd ed. New York: Addison-Wesley; 2004.
[95] Fletcher R, Leyffer S. Nonlinear programming without a penalty function. Math Programming Ser A 2002;91: 239–69.
[96] Audet C, Dennis J. A pattern search filter method for nonlinear programming without derivatives. SIAM J Optim 2004;14(4):980–1010.
[97] Audet C, Dennis J, Moore D, Booker A, Frank P. A surrogate-model-based method for constrained optimization. Proceedings of the eighth AIAA/USAF/NASA/ISSMO symposium on multidisciplinary analysis & optimization, Long Beach, CA, 2000. Paper no. 2000–4891, AIAA.
[98] Marsden A, Wang M, Dennis J. Constrained aeroacoustic shape optimization using the surrogate management

framework. Center for Turbulence Research, Stanford University, Annual Research Briefs 2003. pp. 399–412.

[99] Calhoon DF, Ito JI, Kors DL. Investigation of gaseous propellant combustion and associated injector-chamber design guidelines. Aerojet Liquid Rocket Company, NASA CR-121234, Contract NAS3-13379, 1973.

[100] Chen YS. Compressible and incompressible flow computation with a pressure-based method. Proceedings of the AIAA 28th aerospace sciences meeting, 1989. Paper no. 89–0286, AIAA.

[101] Wang TS, Chen YS. A United Navier-Stokes flowfield and performance analysis of liquid rocket engines. Proceedings of the AIAA 26th joint propulsion conference, 1990. Paper no. 90–2494, AIAA.

[102] Chen YS, Farmer RC. CFD analysis of baffle flame stabilization. Proceedings of the AIAA 27th joint propulsion conference, 1991. Paper no. 91–1967, AIAA.

[103] Vaidyanathan R, Papila N, Shyy W, Tucker PK, Griffin LW, Haftka RT, Fitz-Coy N. Neural network and response surface methodology for rocket engine component optimization. Proceedings of the eighth AIAA/NASA/USAF/ISSMO symposium on multidisciplinary analysis and optimization, Long Beach, CA, September 2000. Paper no. 2000–4880, AIAA.

[104] Papila N, Shyy W, Griffin LW, Huber F, Tran K. Preliminary design optimization for a supersonic turbine for rocket propulsion. Proceedings of the 36th AIAA/ASME/SAE/ASEE joint propulsion conference and exhibit, Huntsville, Alabama, 2000. Paper no. 2000–3242, AIAA.

[106] JMP™. The statistical discovery software™, Version 5, Copyright: 1989–2002. Cary, NC, USA:SAS Institute Inc.

[107] Gill DS, Nurick WH. Liquid rocket engine injectors, 1976. NASA SP–8089, NASA Lewis Research Center.

[108] Booker AJ, Dennis JE, Frank PD, Serafini D, Torczon V. Optimization using surrogate objectives on a helicopter test example. Computational methods in optimal design and control. Boston: Birkhauser, 1998. p. 49–58.