Linear regression-based multifidelity surrogate for disturbance amplification in multiphase explosion

M. Giselle Fernandez-Godino, Sylvain Dubreuil, Nathalie Bartoli, Christian Gogu, Sivaramakrishnan Balachandar, Raphael T. Haftka

Linear Regression Based Multi-fidelity Surrogate for Disturbance Amplification in Multi-phase Explosion

Preprint · April 2019

CITATIONS
0

READS
318

6 authors, including:

Giselle Fernandez
Los Alamos National Laboratory
9 PUBLICATIONS **42** CITATIONS
SEE PROFILE

Sylvain Dubreuil
The French Aerospace Lab ONERA
16 PUBLICATIONS **84** CITATIONS
SEE PROFILE

Nathalie Bartoli
The French Aerospace Lab ONERA
53 PUBLICATIONS **287** CITATIONS
SEE PROFILE

Sivaramakrishnan Balachandar
University of Florida
419 PUBLICATIONS **9,752** CITATIONS
SEE PROFILE

Structral and Multidisciplinary Optimization manuscript No.
(will be inserted by the editor)

Linear Regression-Based Multi-fidelity Surrogate for Disturbance Amplification in Multiphase Explosion

M. Giselle Fernández-Godino · Sylvain Dubreuil · Nathalie Bartoli · Christian Gogu · S. Balachandar · Raphael T. Haftka

Received: date / Accepted: date

Abstract When simulations are very expensive and many are required, as for optimization or uncertainty quantification, a way to reduce cost is using surrogates. With multiple simulations to predict the quantity of interest, some being very expensive and accurate (high-fidelity simulations) and others cheaper but less accurate (low-fidelity simulations), it may be worthwhile to use multi-fidelity surrogates (MFS). Moreover, if we can afford just a few high-fidelity simulations or experiments, MFS become necessary. Co-Kriging, which is probably the most popular MFS, replaces both low-fidelity and high-fidelity simulations by a single MFS. A recently proposed linear-regression-based MFS (LR-MFS) offers the option to correct the LF simulations instead of correcting the LF surrogate in the MFS. When the low-fidelity simulation is cheap enough for use in an application, such as optimization, this may be an attractive option. In this paper, we explore the performance of LR-MFS using exact and surrogate-replaced low-fidelity simulations. The problem studied is a cylindrical dispersal of $100\mu m$ diameter solid particles after detonation and the quantity of interest is a measure of the amplification of the departure from axisymmetry. We find very substantial accuracy improvements for this problem using the LR-MFS with exact low-fidelity simulations. Inspired by these results we also compare the performance of co-Kriging to the use of Kriging to correct exact low-fidelity simulations and find a similar accuracy improvement when simulations are directly used. For this problem, further improvements in accuracy are achievable by taking advantage of inherent parametric symmetries. These results may alert users of MFS to the possible advantages of using exact low-fidelity simulations when this is affordable.

Keywords Multi-fidelity · Surrogates · Symmetries · Linear Regression · Kriging · Co-Kriging

Los Alamos National Laboratory
New Mexico, 87545, United States
Tel.: +1505-667-1627
E-mail: gisellefernandez@lanl.gov

University of Florida
Florida, 32611, United States

ONERA/DTIS, Université de Toulouse
Toulouse F-31055, France

Université de Toulouse, CNRS, UPS, INSA, ISAE, Mines Albi, Institut Clément Ader (ICA)
3 rue Caroline Aigle, Toulouse F-31400, France

Nomenclature

$\delta(\mathbf{x})$ = discrepancy function
$\hat{\delta}(\mathbf{x})$ =discrepancy function surrogate, also known as additive correction
ρ = constant scaling factor
$y_{HF}(\mathbf{x})$ = high-fidelity simulation
$\hat{y}_{HF}(\mathbf{x})$ = high-fidelity surrogate
$y_{LF}(\mathbf{x})$ = low-fidelity simulation
$\hat{y}_{LF}(\mathbf{x})$ = low-fidelity surrogate
$\hat{y}_{a\hat{d}d}(\mathbf{x})$ = multi-fidelity surrogate that uses additive correction and where the prediction is performed using a low-fidelity surrogate
$\hat{y}_{co\hat{m}p}(\mathbf{x})$ = multi-fidelity surrogate that uses comprehensive correction and where the prediction is performed using a low-fidelity surrogate
$\hat{y}_{add}(\mathbf{x})$ = multi-fidelity surrogate that uses additive correction and where the prediction is performed using low-fidelity simulations
$\hat{y}_{comp}(\mathbf{x})$ = multi-fidelity surrogate that uses comprehensive correction and where the prediction is performed using low-fidelity simulations

1 Introduction

Kriging, a Gaussian processes based surrogate, is an increasingly popular surrogate in engineering [17]. Co-Kriging [6, 13–15] is commonly known as the extension of Kriging to include multiple levels of fidelities in the surrogate construction. It replaces both low-fidelity (LF) and high-fidelity (HF) simulations with a single multi-fidelity surrogate (MFS). Co-Kriging approach corrects the LF model using a multiplicative constant plus a discrepancy function between LF and HF models. Co-Kriging is probably one of the most used MFS due to it versatility and good performance for a wider range of applications. Unfortunately, due to its complexity, co-Kriging is generally used as a black box. A recently proposed linear regression-based multi-fidelity surrogate (LR-MFS) [22] offers the option of using exact LF simulations in the MFS. This may offer an advantage if the LF and HF simulations have local fluctuations that are highly correlated between the two levels of fidelity. The objective of this paper is to report on an application which illustrates the usefulness of LR-MFS. Specifically, that niche is when

1. The low-fidelity (LF) function is cheap enough to evaluate without a surrogate for the intended application (such as optimization or UQ) so that replacing it with a surrogate is an unnecessary extra complication.
2. The discrepancy function correcting the LF predictions can be modeled accurately enough by global functions, such as polynomials that are typically used in linear regression.

One advantage of LR-MFS is that it is easy to tailor it to the application. In our case, taking advantage of symmetry proved to be very easy. LR-MFS, as co-Kriging, corrects the LF model using a multiplicative constant plus a discrepancy function between LF and HF models. A second comparison was made by using additive Kriging [11], which is using Kriging surrogate to model the discrepancy between the LF and HF models using also exact LF simulations. In this approach, unlike in co-Kriging and LR-MFS, the LF model is not corrected using a multiplicative factor.

The physical problem studied is a cylindrical two dimensional multiphase explosion. We study multiphase explosion simulations where the distribution of particles is initially nominally axisymmetric. A dense layer of solid particles surrounding a high-energy explosive develops instabilities after detonation. Conjectures as to the cause of these instabilities include imperfections in the casing, inhomogeneities in the initial distribution of particles, characteristics of the particles, and others. It was assumed that the instabilities are due to initial imperfections in the distribution of particles. Therefore, in our simulations, the particle volume fraction is perturbed with azimuthal sinusoidal waves and the distribution of the particles at a later time is studied quantifying the amplification of the departure from axisymmetry. Two single fidelity surrogates and four regression-based MFSs (including LR-MFS) to approximate the amplification of the particle departure from axisymmetry in a multiphase explosion problem are investigated. This comparison was done for two cases, (i) using second order regression basis functions and (ii) using fourth order polynomial regression basis functions. LR-MFS, additive Kriging and co-Kriging performances are also compared. Due to the presence of symmetries in our problem, we explore options for reducing the number of simulations used to construct surrogates while maintaining the desired accuracy by taking advantage of parametric symmetries. These symmetries allow us to obtain free information and, therefore, the possibility of cheaper or more accurate predictions. The inherent parametric symmetries using symmetric basis functions (SBF) and adding permutation points (APP) [9] while building the MFS are imposed and compared with the performance of the MFS without imposing symmetries.

The details of the physical problem studied can be found in Section 2. To quantify the particle departure from axisymmetry, an L^2 metric based on energy has been constructed. An L^2 norm is proportional to the root mean of the vector components squared. The description of the metric can be found in Section 3. The variables considered are the amplitudes and wavelengths of a trimodal sinusoidal perturbation which is used as a perturbation in the particle bed of the problem studied. A variance-based sensitivity analysis has been carried out to select the most relevant variables for surrogate construction (Appendix A). Details of the variable selection and the multi-fidelity (MF) design of experiments are included in Section 4. In Section 5 is described how the two single-fidelity and the four MFSs used are constructed. It is also described here the method to impose symmetries in linear regression-based surrogates. Appendix B shows a study of why substantial accuracy is obtained from using LF simulations instead of LF surrogates for the case where second-order polynomial basis functions are used. In Section 6 the performance of the linear regression-based surrogates is shown and their performance to additive Kriging and co-Kriging is compared and, finally we show how imposing symmetries leads to a computational cost reduction.

2 Physical Problem Description

The physical problem simulated is a two dimensional multiphase[1] detonation. The computational domain is a two dimensional slice of a cylinder of diameter $1.20m$ taken perpendicular to the cylinder axis. This was chosen to contain the blast wave during the entire simulation time of $500\mu s$. This domain is comprised of a $0.0038m$ radius inner circle containing hot, high-pressure gas which is surrounded by an

[1] In this case multiphase refers to the phases, gas and particles.

annular particle bed of outer radius $0.05m$. The remainder of the computational domain, of outer radius $0.6m$, contains ambient air. Except for the inner circle, which contains the high-pressure gas simulating the products of detonation of an initially high-energy charge, the rest of the domain is initially under standard conditions of pressure and temperature. Figure 1(a) shows a schematic of the computational domain.

The initial conditions for the gas phase are as follows. The high-energy explosive in the charge (i.e. $r \leq 0.038m$) is taken to be representative of pentaerythritol tetranitrate (PETN). Using data taken from [7], the gas inside of the charge is set to an initial density of $1770\frac{kg}{m^3}$ and an energy content of $10.089\frac{GJ}{m^3}$ at zero velocity. Outside of the charge, the gas is initialized at standard atmospheric conditions of $1.203\frac{kg}{m^3}$ and $101325Pa$ at zero velocity. For this work, the entire gas phase is governed by the ideal gas equations. For detailed information about the numerical methods of the problem, refer to [10].

The properties for the particle phase at the initial time are as follows. The particles are taken to be made entirely of glass with a density of $2500\frac{kg}{m^3}$ and a diameter of $100\ \mu m$. The heat capacity of the particles is $450\frac{J}{kg \cdot K}$. The particles are initially taken to occupy a volume fraction of 5% in the annular region $0.0038m \leq r \leq 0.05m$. The number of computational particles in all of the simulations is $31,250$, which are randomly distributed within the annular region with uniform probability. The number of physical particles contained in each computational particle is given by the superparticle loading factor. Within each finite volume cell, the superparticle loading factor of the particles inside the cell is adjusted such that the cell averaged particle volume fraction (PVF) within the cell equals the desired PVF. The superparticle loading factor is maintained constant throughout the life of the particle. The ratio of the mass of the particle bed to the initial mass of the charge is 17.9.

The computational domain is divided into two regions for both LF and HF simulations. For the LF simulation, a 64×64 cell Cartesian mesh is forced into the inner circle of the domain which initially contains the high-pressure gas and an outer polar mesh with 125 and 256 cells in the radial and azimuthal directions, respectively. This gives a total of $36,096$ computational cells in both regions. The HF data points are obtained from simulations with a 16 times finer grid, i.e. 256×256 cell Cartesian mesh for the inner grid, and for the outer grid a 500 and $1,024$ cells in the radial and azimuthal directions, respectively. This gives a total of $577,536$ computational cells in both regions. In the LF simulations are used $31,250$ computational particles. In the HF simulations, the number of computational particles has been increased accordingly to maintain the same number of computational particles per cell, in this case, $500,000$. The computational cost of the LF simulations is 96 core hours while the computational cost of the HF simulations is 3,072 core hours in Quartz' Lawrence Livermore National Laboratory high performance computer. Therefore the LF to HF cost ratio is around 3%. For further information about the grids refer to [10].

The base PVF was set to 5%, which is relatively low, to avoid the effects of over compacting particles. The outer annulus contains the blast wave during the entire simulation time, 500 μs. The perturbations imposed to the PVF are inspired by [1]. The base PVF is perturbed using a superposition of up to three sinusoidal waves. Equation (1) provides the mathematical expression of the perturbation while Eq. (2) provides the associated energy constraint of 0.02 ($\sim$ 14%) based on [16]. Note that the PVF perturbation is constant in the radial direction and restricted to the circumferential direction (θ).

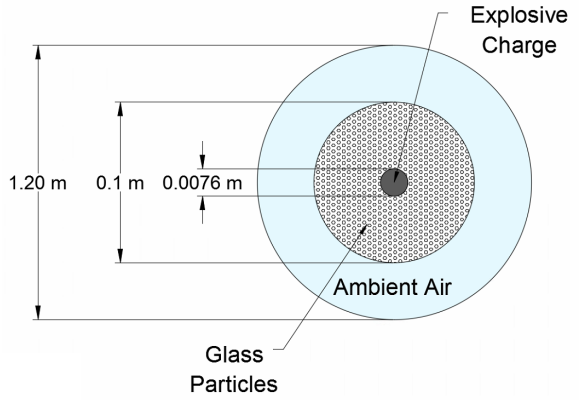
$$\phi^p(\theta) = \phi_0^p[1 + A_1\cos(k_1\theta + \Phi_1) + A_2\cos(k_2\theta + \Phi_2) + A_3\cos(k_3\theta + \Phi_3)], \quad (1)$$

subject to

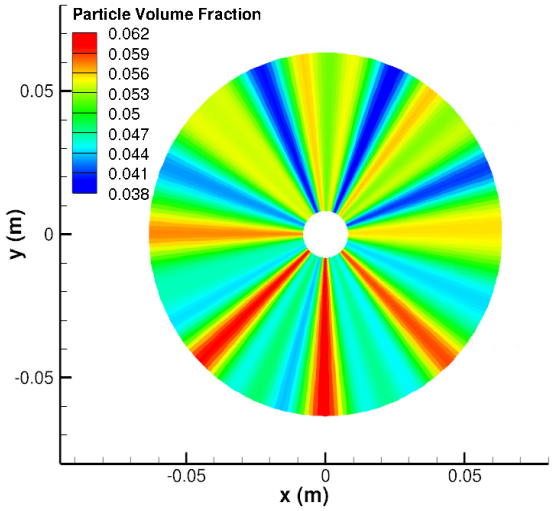
$$\sqrt{A_1^2 + A_2^2 + A_3^2} = 0.02, \quad (2)$$

where ϕ^p is the PVF at a given θ ($0 \leq \theta \leq 2\pi$), and ϕ_0^p is the constant base PVF (here 5%). Mode amplitudes (A_1, A_2 and A_3 where $0 < A_1, A_2, A_3 < \sqrt{0.02}$), phases ($\Phi_1$, Φ_2 and Φ_3 where $0 \leq \Phi_1, \Phi_2, \Phi_3 \leq 2\pi$) and integer wavenumbers (k_1, k_2, and k_3 where $1 \leq k_1, k_2, k_3 \leq 25$) are the perturbation parameters. Furthermore, without loss of generality, we can take $\Phi_1 = 0$ and only the phase of the other two modes with respect to the first one matters. Also, Eq. (2) allows us to write one amplitude in terms of the other two. Therefore even if we start with nine variables, they can be reduced to a seven variable problem, A_1, A_2, k_1, k_2, k_3, Φ_2, and Φ_3. Figure 1(b) shows PVF contours for a case where a trimodal perturbation is imposed.

Figure 2(a) shows the convergence of the shock location for the three grids described in Table 1 and Table 2, LF, intermediate fidelity (IF), and HF. It can be observed that the shock trajectory results are nearly independent of the grid resolution. In Figure 2(b) we show the upstream particle front location as a function of time for the three different resolutions. The convergence is non-monotonic and time-dependent. Note that due to the chaotic motion of the particles, there is statistical variation in the location of upstream and downstream most particle between different runs of the same resolution. Based on these results, and the need for a large number of simulations, the LF grid is deemed accurate enough to be used as the LF simulation. The relative RMSE in the shock location and in the particle upstream front of the LF grid with respect to the HF grid are 2% and 3% respectively. This error was computed using the data points



(a) Schematic of the computational domain (not to scale).



(b) PVF contours at initial time for the perturbed case where A_1=0.13 A_2=0.04 A_3=0.05, k_1=8, k_2=17, k_3=15 and Φ_1=0 Φ_2=2.05 Φ_3=4.5.

Fig. 1 PVF contours at initial time.

of Figure 2 considering values of $t > 0$. The differences between LF and HF simulations in the shock and particle upstream front are at least one order of magnitude higher than the fluctuations in these quantities due to a change of initial particle location within a cell. In this paper, we use two levels of fidelity to construct MFSs. The LF simulation corresponds to the LF grid while the HF simulation corresponds to the HF grid. IF grid results were included in Figure 2 and in Tables 1 and 2 for completeness, however results of IF grid are not used in this work for the surrogate construction.

3 The Normalized Fourier Effective Perturbation

We are interested in measuring the amplification of the departure from axisymmetry that was introduced as an initial perturbation. There are several ways to define this perturbation. Here the variation in the particle volume as a function of θ is chosen. We divide the computational domain into radial sectors of identical volume. In our problem, the volume of the sectors remains a constant and equal to $h\pi R^2/N$, where N is the number of azimuthal divisions, $R = 0.6$m is chosen to be the outer radius of the computational domain, and h=0.02m is the thickness of the computational domain in the axial direction. Note that even though the gas properties are two dimensional and do not vary along the axial direction, particles are distributed within the 3D domain over this axial thickness. For any simulation time t, the total volume of all the particles within each of the N radial sectors defines the variable $\mathrm{PV}(\theta,t)$. Because of the cylindrical nature of the physical problem, PV is a periodic function in $0 \leq \theta \leq 2\pi$. $\mathrm{PV}(\theta,t)$ can be Fourier transformed as follows

$$\mathrm{PV}(\theta,t) = \sum_{k=-N/2}^{N/2} a_k \exp\left(ik\frac{2\pi}{N}\theta\right), \tag{3}$$

where a_k is the Fourier coefficient corresponding to the kth Fourier mode. The Fourier coefficients are complex and are given by

$$a_k = \frac{1}{N}\sum_{k=0}^{N-1} \mathrm{PV}(\theta,t)\exp\left(-ik\frac{2\pi}{N}\theta\right). \tag{4}$$

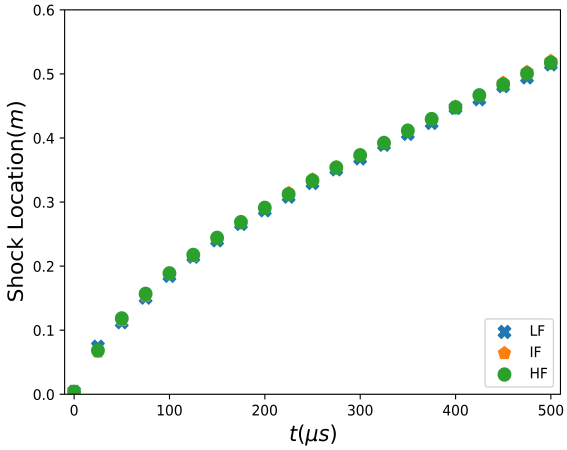
A plot of $|a_k|^2$ as a function of the wavenumber k gives us the energy spectrum of departure from axisymmetry in the periodic signal $\mathrm{PV}(\theta,t)$. Figures 3(a) and 3(b) show the spectrum of the most amplified trimodal case at the initial time and at $t = 500\mu s$, respectively. The spectra are normalized by the squared average value $|a_0|^2$ and the results are symmetric in the wavenumber. In Figure 3(a), besides the constant mean (i.e. $k = 0$), only the initial modes $[k_1,k_2,k_3] =$

Table 1 Number of azimuthal and radial divisions for the low-fidelity, intermediate-fidelity, and high-fidelity grids.

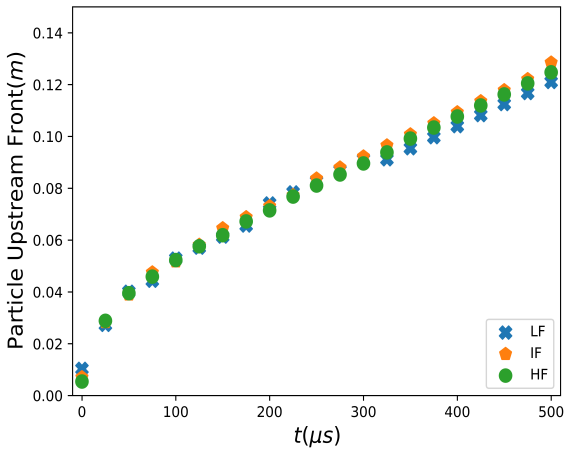
Grid Name	Radial Divisions	Azimuthal Divisions
Low-fidelity (LF)	125	256
Intermediate-fidelity (IF)	250	512
High-fidelity (HF)	500	1024

Table 2 Number of cells, number of computational particles, and relative simulation cost for the low-fidelity, intermediate-fidelity, and high-fidelity grids.

Grid Name	Number of Cells	Comp. Particles	Rel. Sim. Cost
Low-fidelity (LF)	36,096	31,250	0.03
Intermediate-fidelity (IF)	144,384	125,000	0.25
High-fidelity (HF)	577,536	500,000	1

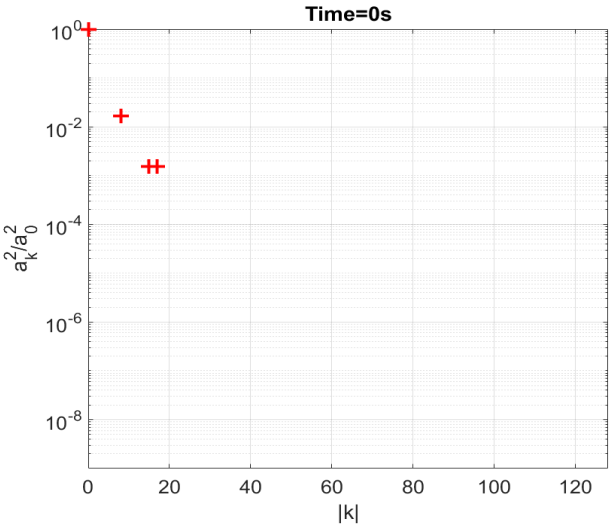


(a) Shock location as a function of time.

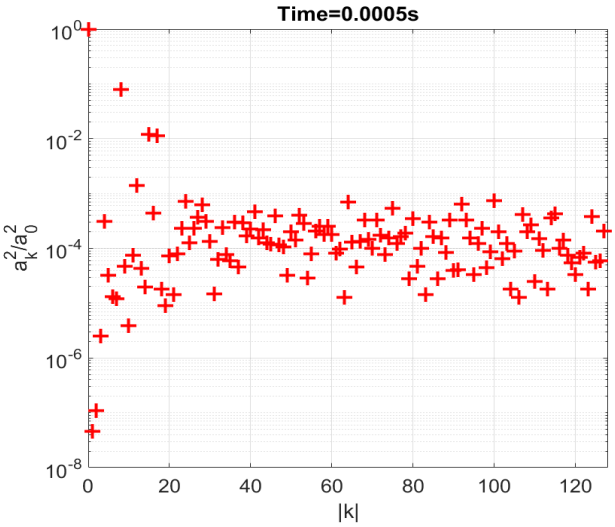


(b) Upstream front of particles as a function of time.

Fig. 2 Simulation metrics as a function of time for the three different grids (LF, IF and HF) described in Table 1 and Table 2 based on a single simulation. The three grids show agreement suggesting that the LF grid can be used to approximate the results of higher fidelity grids.



(a) PV amplitude spectrum squared at the initial time $t = 0$.



(b) PV amplitude spectrum squared at the final time $t = 500\mu s$.

Fig. 3 PV amplitude spectrum squared for $[A_1, A_2, A_3] = [0.130, 0.039, 0.039]$, $[k_1, k_2, k_3] = [8, 17, 15]$, and $[\Phi_1, \Phi_2, \Phi_3] = [0, 2.052, 4.851]$. The spectrum is normalized by $|A_0|^2$.

$[8, 17, 15]$, that are imposed by the initial perturbation, have a non-zero amplitude, while the rest of the modes are zero. At the final simulated time, $t = 500\mu s$, we observe the initial three modes to still remain dominant and their squared amplitudes have substantially grown over time (note the logarithmic y-axis). However, all the other Fourier modes are energized as well. This is partly due to nonlinear interaction between the initial Fourier modes, but also due to circumferential perturbation introduced by the random location and motion of the particles.

We define the normalized effective PV variation as

$$F(t = 500\mu s) = \frac{\sum_{k=1}^{N-1} a_k^2(t = 500\mu s)}{0.02}. \quad (5)$$

Note that in the numerator the sum excludes $k = 0$ and thus only the non-axisymmetric modes contribute to this measure. Also, the denominator is the sum of squared amplitudes of the three modes of the initial perturbation, which in the present simulations has been chosen to be equal to 0.02 (Eq. (2)). The denominator is nothing but the initial value of the numerator for the present set of simulations and therefore we have normalized the above measure of PV variation to yield an initial value of $F(t = 0) = 1$. For example, for the case presented in Figure 3(b) at $500\mu s$ the value of $F(t = 500\mu s)$ is 6.25. When we refer to the metric F we will be actually refereeing to $F(t = 500\mu s)$ unless we explicitly state another evaluation time.

Part of the variation in F can be explained in terms of parametric dependence, but the complex behavior associated with the metric cannot be fully explained in terms of

the controlling parameters alone. One of the contributing factors to the statistical noise is the initial random distribution of computational particles within the annular region. As the parameters of the problem are varied, the importance of these parameters become relevant only if their effect on F is substantially larger than the statistical noise. The relative noise is estimated as two times the standard deviation in F values of 20 realizations normalized by the mean value of these realizations that were run for a few cases. We determined that the noise throughout the metric range is constant and around 5% [10].

4 The Design of Experiments

After exploring and estimating the noise level in the LF simulations we have performed a sensitivity analysis using the HF simulations, motivated by the need to construct a DoE in an as small as possible design space due to the curse of dimensionality. The purpose is to estimate in a more quantitative manner how the seven variables (A_1, A_2, k_1, k_2, k_3, ϕ_2 and ϕ_3) affect the metric F and also how they interact between themselves. For this purpose, a variance based sensitivity analysis using polynomial chaos expansion was performed and the reader can see the details in the Appendix A. After performing the sensitivity analysis we concluded that the amplitude variables A_1 and A_2 do not have a main effect (low first order index) however the variance caused by their interactions is very high (high total index). The wavenumbers k_1, k_2 and k_3 are the most important variables, they have a high first order effect and also high importance when interacting with other variables. Lastly, the phase variables ϕ_2 and ϕ_3 turned out to have a negligible first order index and a negligible total index (less than four orders of magnitude smaller than the other variables). Therefore phase variables are ignored from now on considering in this study only the variables A_1, A_2, k_1, k_2, and k_3.

Two levels of fidelity to construct MFSs are used. The level of fidelity is dictated by the grid resolution. The LF and HF simulations correspond to the LF and HF grid described in Table 1 and Table 2, respectively. As mentioned before, each HF and LF simulation cost 96 and 3,072 core hours in Quartz Lawrence Livermore National Laboratory high-performance computer, respectively.

Our quantity of interest is the amplification of the particle departure from axisymmetry. For this purpose, the metric F, described in Section 3 by Eq. (5), is computed. The surrogates were constructed with the variables A_1, A_2, k_1, k_2, and k_3 using up to 1,415 LF simulations and up to 711 nested HF simulations. The validation data points are 92 additional HF data points due to the fact that increasing the number of validation points has proven not to change substantially the value of the calculated errors (less than 2%).

Knowing the correlation between the LF and HF simulations gives us an idea of what can be expected from the surrogate performance. A high correlation between the HF data points and the LF data points should lead to a good MFS performance. Otherwise, using MFSs may be not only poor but harmful. That is, with low correlation, the MFS may be less accurate than a surrogate built using only HF training data points. The correlation coefficient between the 711 HF data points and the corresponding LF data points for the metric F is 0.92, which indicates a high correlation. The estimated relative RMSE between the LF and HF simulations is 2% for the shock location and 3% for the particle upstream front. The noise in these quantities associated with each simulation is negligible compared with the differences between them, i.e. the shock and the particle upstream front location do not change substantially for different initial location of particles. Moreover, they do not change substantially if the perturbations applied are different.

The algorithm used to obtain the training data points is a nested design of experiments in seven variables based on Latin hypercube sampling (LHS) technique [21]. Nested data points are those computed using the same variables but using different fidelities. The original DoE had 800 HF data points and 1,600 LF data points, however, we eliminate repetitions in the value of the wavenumbers for the same data point. This was done to avoid the presence of bimodal or single modal perturbations in our samples, 185 of the 1600 LF data points and 89 of the 800 HF data points were unimodal or bimodal, and they were eliminated. Finally, we computed 711 HF simulations and 1,415 LF simulations without repetition. The bounds in the variables are $0 < A_i < \sqrt{0.02}$ for $i = 1,2$ and $1 \leq k_j \leq 25$ for $j = 1,2,3$. The wavenumbers, k_j, are by definition integers, therefore, the LHS data were rounded to the nearest integer for these variables. After normalizing the data points between 0 and 1, the Euclidean distance between points was computed as a measure of how well distributed are the data points in the design of experiments. The minimum Euclidean distance between two points is 0.05 (the minimum possible is 0), while the maximum distance is 1.90 (maximum possible is $\sqrt{5}$). The 92 HF validation data points were obtained using LHS.

5 The Linear Regression Surrogates Used

In this section, we show how the data points are used to build single-fidelity surrogates and MFSs. We also describe how the parametric symmetries inherent to our problem can be used to reduce the error if we can afford only few HF data points.

5.1 The single-fidelity surrogates

The LF and HF single-fidelity surrogates were constructed using the classical linear regression approach [18] for two cases, (i) using as basis functions monomials up to a second order polynomial, and (ii) using as basis functions monomials up to a fourth order polynomial. Regression was chosen due to its good performance filtering the noise. This reflects the objective of the paper to explore the niche of LR-MFS for this application.

5.2 The multi-fidelity surrogates

The MFSs were built also using the regression approach but combining more than one fidelity. Four different MF approaches were studied, and they are discussed below. The MFSs were constructed using second-degree polynomials and also fourth degree polynomials as basis functions. The surrogates that use second order polynomials as basis functions were built using MATLAB regress function, while the surrogates that use fouth order polynomials basis functions were built using MATLAB polyfitn function. Given a low-fidelity simulation, $y_{LF}(\mathbf{x})$, and a high-fidelity simulation, $y_{HF}(\mathbf{x})$, their surrogates are denoted as $\hat{y}_{LF}(\mathbf{x})$ and $\hat{y}_{HF}(\mathbf{x})$, respectively. In general, we use $\hat{\cdot}$ to denote a surrogate. The additive correction approach, $\hat{y}_{\hat{add}}$, assumes that the relationship between $y_{LF}(\mathbf{x})$ and $y_{HF}(\mathbf{x})$ is

$$\hat{y}_{\hat{add}}(\mathbf{x}) = \hat{y}_{LF}(\mathbf{x}) + \hat{\delta}(\mathbf{x}), \tag{6}$$

where $\hat{y}_{LF}$ is the regression-based single-fidelity surrogate built using the 1,415 LF data points, and $\hat{\delta}$ is the surrogate constructed using as training data points the difference between the $y_{HF}(\mathbf{x})$ and $y_{LF}(\mathbf{x})$ functions at the nested training data points (in this case up to 711 training data points). In other words $\hat{\delta}$ is the surrogate of δ, the discrepancy function between y_{HF} and y_{LF}[2]. The comprehensive approach can be written as

$$\hat{y}_{\hat{comp}} = \rho \hat{y}_{LF}(\mathbf{x}) + \hat{\delta}(\mathbf{x}), \tag{7}$$

where ρ is a constant. In this paper, the comprehensive surrogate was constructed using LR-MFS [22] which basically consists of adding to the HF surrogate, $\hat{y}_{HF}$, constructed using classical linear regression an extra basis function that depends on the LF simulation evaluated at the HF training data points. The approach is explained below. Lets us consider the surrogate $\hat{y}$,

$$\hat{y} = \rho \hat{y}_{LF}(\mathbf{x}) + \sum_{i=1}^{n} \xi_i(\mathbf{x}) b_i, \tag{8}$$

[2] The multiplicative correction approach is an MFS option that is not included in this paper, however the reader can refer to [11] if interested. This MFS is constructed using as training points the quotient between $y_{HF}(\mathbf{x})$ and $y_{LF}(\mathbf{x})$ functions at the nested training data points.

where $\xi_i(\mathbf{x})$ denotes the ith monomial basis function, b_i is the ith linear regression coefficient to be determined, and n is the number of linear regression coefficients to be determined. The error between the HF simulation, y_{HF}, and the surrogate prediction, $\hat{y}$, at the point $\mathbf{x}_j$ can be written as

$$\begin{aligned} e_j &= y_{HF}(\mathbf{x}_j) - \hat{y}(\mathbf{x}_j) \\ &= y_{HF}(\mathbf{x}_j) - \rho \hat{y}_{LF}(\mathbf{x}) - \sum_{i=1}^{n} \xi_i(\mathbf{x}_j) b_i, \end{aligned} \tag{9}$$

which in vector form we write as

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\mathbf{b}, \tag{10}$$

where

$$\mathbf{e} = \begin{bmatrix} e_1 \\ \vdots \\ e_m \end{bmatrix}, \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix},$$
$$\mathbf{X} = \begin{bmatrix} \hat{y}_{LF}(\mathbf{x}_1) & \xi_1(\mathbf{x}_1) & \dots & \xi_n(\mathbf{x}_1) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{y}_{LF}(\mathbf{x}_m) & \xi_1(\mathbf{x}_m) & \dots & \xi_n(\mathbf{x}_m) \end{bmatrix}, \mathbf{b} = \begin{bmatrix} \rho \\ b_1 \\ \vdots \\ b_n \end{bmatrix}, \tag{11}$$

where m is the number of HF data points used to train the surrogate. The unknown coefficients are then determined, as in the classical linear regression approach, by minimizing the square sum of errors as

$$\underset{\mathbf{b}}{\text{minimize}} \quad \mathbf{e}^T \mathbf{e}. \tag{12}$$

An advantage of the LR-MFS approach is that the $\hat{y}_{LF}(\mathbf{x})$ can be easily replaced by $y_{LF}(\mathbf{x})$ if the LF simulation is cheap enough. However, this option is not usually available for more complex surrogates like co-Kriging due to the fact that a surrogate is constructed internally based on the LF training data points. Therefore, the two approaches presented above (Eqs. (6) and (7)) were also constructed using $y_{LF}(\mathbf{x})$ instead of $\hat{y}_{LF}(\mathbf{x})$. That is, the additive correction, Eq. (6) becomes

$$\hat{y}_{add}(\mathbf{x}) = y_{LF}(\mathbf{x}) + \hat{\delta}(\mathbf{x}). \tag{13}$$

Similarly, for the comprehensive approach (Eq. (7)) we have

$$\hat{y}_{comp} = \rho y_{LF}(\mathbf{x}) + \hat{\delta}(\mathbf{x}), \tag{14}$$

where ρ is a constant. Note that $\hat{\delta}(\mathbf{x})$ coefficients in Eqs. (6) and (13) are the same. Also, $\hat{\delta}(\mathbf{x})$ coefficients, and ρ are identical in Eqs. (7) and (14). These are calculated based on the HF simulation data and LF simulation data at the common points and the difference relies on the surrogate evaluation.

5.3 Imposing Symmetries

The metric F depends on the amplitudes A_1, A_2, A_3 and the wave numbers k_1, k_2, k_3. Note that in our problem the order of the perturbation mode does not matter, therefore, the simulation output of the data points $(A_1, A_2, A_3, k_1, k_2, k_3)$, $(A_2, A_1, A_3, k_2, k_1, k_3)$, $(A_3, A_2, A_1, k_3, k_2, k_1)$, etc. are the identical. Consequently, each simulated data point gives five extra symmetry points to be used to train the surrogate. In this paper we used two techniques for imposing symmetries (i) *Adding permutation points* (APP) is to simply add the permutation points as training data points and (ii) *Symmetric basis functions* (SBF) is to modify the surrogate basis functions making them symmetric. These two approaches are proposed in [9]. Although APP approach is straight forward, SBF needs a more detailed explanation. Therefore, in order to illustrate the SBF technique we include the following example. Let us consider the linear regression surrogate prediction $\hat{y}$ of the function y which can be written as

$$\hat{y}(\mathbf{x}) = \sum_i b_i \xi_i(\mathbf{x}), \tag{15}$$

where ξ_i are the linear regression basis functions, $\mathbf{x}$ is the vector of the function variables, and b_i are the coefficients to be determined. Now, let us assume that linear regression basis functions are the monomials of a pth-degree polynomial. For simplicity, let us consider then a three variable function $y(x_1, x_2, x_3)$ to be of interest where $y(x_1, x_2, x_3) = y(x_1, x_3, x_2) = y(x_2, x_1, x_3)$ and so on. If we approximate y using a third degree polynomial, Eq. (15) can be written as

$$\begin{aligned}\hat{y}(\mathbf{x}) &= b_1 1 + b_2 x_1 + b_3 x_2 + b_4 x_3 + b_5 x_1 x_2 + b_6 x_1 x_3 \\ &+ b_7 x_2 x_3 + b_8 x_1^2 + b_9 x_2^2 + b_{10} x_3^2 + b_{11} x_1 x_2 x_3 \\ &+ b_{12} x_1^2 x_2 + b_{13} x_1 x_2^2 + b_{14} x_1^2 x_3 + b_{15} x_1 x_3^2 \\ &+ b_{16} x_2^2 x_3 + b_{17} x_2 x_3^2 + b_{18} x_1^3 + b_{19} x_2^3 + b_{20} x_3^3. \end{aligned} \tag{16}$$

In order to include the parametric symmetries in surrogates modifying the linear regression basis functions we can rewrite Eq. (16) as

$$\begin{aligned}\hat{y}(\mathbf{x}) &= \tilde{b}_1 1 + \tilde{b}_2 (x_1 + x_2 + x_3) + \tilde{b}_3 (x_1 x_2 + x_1 x_3 + x_2 x_3) \\ &+ \tilde{b}_4 (x_1^2 + x_2^2 + x_3^2) + \tilde{b}_5 x_1 x_2 x_3 + \tilde{b}_6 (x_1^2 x_2 + x_1 x_2^2 + x_1^2 x_3 \\ &+ x_1 x_3^2 + x_2^2 x_3 + x_2 x_3^2) + \tilde{b}_7 (x_1^3 + x_2^3 + x_3^3), \end{aligned} \tag{17}$$

where

$$\begin{aligned} \tilde{b}_1 &= b_1 \\ \tilde{b}_2 &= b_2, b_3, b_4 \\ \tilde{b}_3 &= b_5, b_6, b_7 \\ \tilde{b}_4 &= b_8, b_9, b_{10} \\ \tilde{b}_5 &= b_{11} \\ \tilde{b}_6 &= b_{12}, b_{13}, b_{14}, b_{15}, b_{16}, b_{17} \\ \tilde{b}_7 &= b_{18}, b_{19}, b_{20} \end{aligned} \tag{18}$$

Note that while Eq. (16) has 20 coefficients to be determined, Eq. (17) has only seven. This same analysis can be easily extended to higher order polynomials and a larger number of variables.

As noted in [9], the SBF and APP approaches give the same results for single-fidelity surrogates. For the MF case, this will still apply if we replace the LF simulation with an LF surrogate. However, if the LF simulations are used directly, as in Eq. (14), there is no mechanism that enforces symmetry on the LF simulations if we do not add the symmetry points.

6 Results

In this section, regression-based single-fidelity surrogates and MFSs (additive correction, LR-MFS, additive Kriging and co-Kriging) performances are compared. In addition, imposing symmetries in LR-MFS proved to reduce cost and/or accuracy.

6.1 The Linear Regression-based Surrogate Performance

The performance of the two single-fidelity surrogates (Section 5.1) and the four MF approaches (Section 5.2) used are summarized in Figure 4, where the relative RMSE as a function of the number of HF data points used to train the surrogates is shown. Figure 4 shows results of regression surrogates using up to second order polynomial basis functions. The relative RMSE is calculated based on 92 validation points. Note that the minimum amount of HF data points required to train the second order surrogates is 22, which is the number of coefficients of a quadratic polynomial in five variables plus the extra coefficient for the LF basis function needed to construct the comprehensive MFSs, y_{comp} and $\hat{y}_{comp}$. The LF surrogate is not trained with the HF data points, therefore, its performance is constant and it is included in the plot as a constant relative RMSE line only for graphical comparison. We have also included the relative RMSE in the LF simulation at the HF validation points. The HF surrogate is only trained with HF data points and, as Figure 4 shows, as the number of training points increases the performance improves.

For the comprehensive MFS, $\hat{y}_{comp}$ (Eq. (14)), by construction, the number of LF data points is the same as HF data points, therefore, as the number of HF data points increases, the number of LF data points also increases.

In Figure 4, the MFSs that use the LF surrogate, $\hat{y}_{\hat{add}}$ and $\hat{y}_{\hat{comp}}$, have a performance similar to the HF surrogate, $\hat{y}_{HF}$. The additive and comprehensive MFSs, $\hat{y}_{add}$ and $\hat{y}_{comp}$, can attain errors that are three times lower. Their performance levels off at roughly a relative RMSE of 6%. The best performance is achieved by the comprehensive MFS, $\hat{y}_{comp}$,

built using the LF simulation, y_{LF}. Note also that for approximations that use surrogates ($\hat{y}_{HF}$, $\hat{y}_{a\hat{d}d}$ and $\hat{y}_{co\hat{m}p}$), after adding 100 HF data points, the performance does not change substantially. This is because for the regression, basis functions up to a quadratic polynomial were chosen for both, the correction and the LF surrogate. If the chosen function would have been more complex or the order of the polynomial would have been higher as will be shown in Figure 5, the number of points required until reaching a plateau would have been higher. Note that the computational cost also would have been higher. In Table 3 the relative RMSE

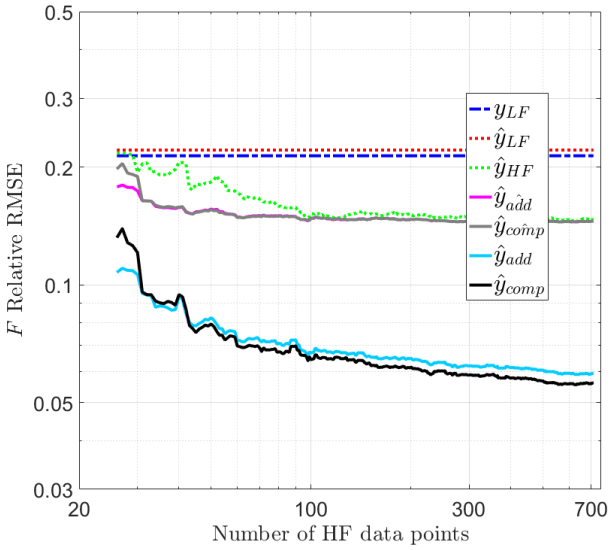


Fig. 4 RMSE for the metric F as a function of the number of HF data points used to train the surrogates. The surrogates are built using up to second-order polynomial basis functions. The relative RMSE was calculated using 92 validation data points.

for the metric F using the maximum number of HF data points available (711) is shown. The coefficient of determination, R^2, was added to the table to show the proportion of the variance of the dependent variable that is predicted by the independent variables. Note that the coefficient was only computed for the case that uses the maximum number of HF data points. Summarizing, the factors that make $\hat{y}_{add}$ and $\hat{y}_{comp}$ to work better than $\hat{y}_{a\hat{d}d}$ and $\hat{y}_{co\hat{m}p}$ are: (i) the noise due to the initial random location of particles is smaller than the fluctuations due to changes in the variables. Therefore, the used second order polynomials are not only filtering the noise, but also filtering meaningful information of the physical model. (ii) One of the variables is discrete (the wavenumbers can take only integer values in simulations) therefore from one data point to the other the fluctuations are substantial, generating spikes that the second-order

[3] The relative RMSE of the LF function at the validation points is 0.214

Table 3 Relative RMSE and coefficient of determination (R^2) for the metric F using 1,415 LF training points and 711 HF training points for each of the surrogates considered using up to second order polynomial basis functions.

Surrogate	Rel. RMSE (F)[3]	$R^2(F)$
$\hat{y}_{LF}$	0.221	0.78
$\hat{y}_{HF}$	0.147	0.62
$\hat{y}_{a\hat{d}d}$	0.143	0.66
$\hat{y}_{co\hat{m}p}$	0.146	0.92
$\hat{y}_{add}$	0.072	0.66
$\hat{y}_{comp}$	0.056	0.92

polynomial filters as noise. (iii) We use a second order polynomial as basis functions of linear regression. When more complex basis functions (as higher order polynomials) are used the performance between using calculations or surrogates becomes closer. This is explored further next by using quartic polynomials and in the next section that examines using Kriging for the discrepancy function. (iv) The correlation between HF and LF is high. (v) The percentage of LF influence in LR-MFS is more than 90% in almost the entire range (see Appendix B).

Next we explore the use of quartic polynomials. These have 126 coefficients and required substantially more data points for good conditioning of the design matrix. Figure 5 shows the same results shown in Figure 4 but including polynomial basis functions up to fourth order. What we observe from the figure is: (i) with quartic polynomials and more than 200 HF points, there is no benefit in using MFS, since no MFS significantly improves on the single fidelity surrogate. (ii) The benefit in using exact LF simulations is reversed, and using the LF surrogate is better, reflecting the noise filtering benefit of the surrogate. (iii) The comprehensive approaches, $\hat{y}_{co\hat{m}p}$ and $\hat{y}_{comp}$, perform poorly for less than 400 HF points, which we traced to ill conditioning and poor selection of ρ.

Combined the lessons of these two figures are that if we can afford a small number of HF simulations, the second order polynomials and exact LF simulations for the MFS are a good way to go. When more than 200 HF simulations are available, single fidelity HF surrogate is the best option.

6.2 Comparison with HF Kriging, Additive Kriging and Co-Kriging

The good accuracy obtained with second order LR-MFS when exact LF simulations are used, raises the question of whether similar gains are available with Kriging. In this section we study the performance of three variants of Kriging. HF Kriging (single fidelity Kriging surrogate using only HF data points), additive Kriging (MF version of Kriging where the LF simulations, and not the LF surrogate, are corrected with a Kriging surrogate of the discrepancy) and co-Kriging (MF

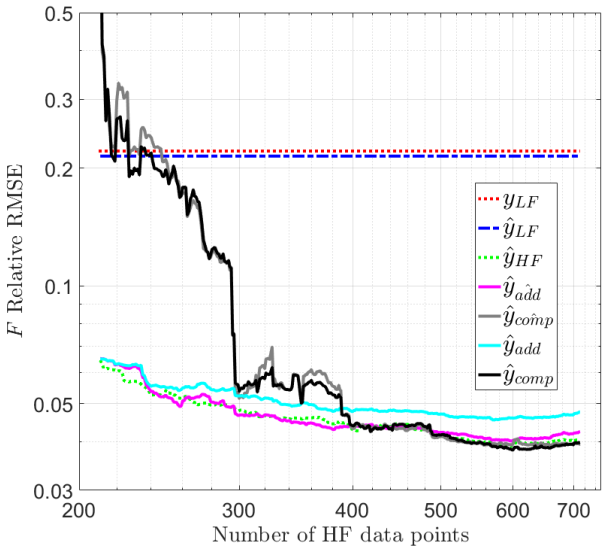


Fig. 5 RMSE for the metric F as a function of the number of HF data points used to train the surrogates. The surrogates are built using up to fourth-order polynomial basis functions. The relative RMSE was calculated using 92 validation data points.

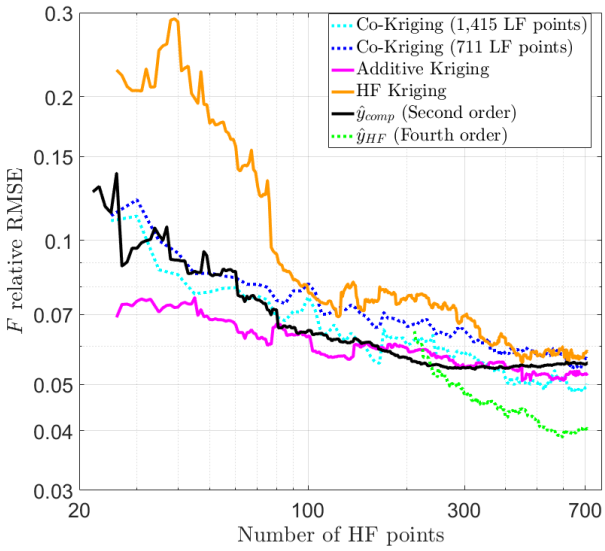


Fig. 6 Relative RMSE error of the metric F as a function of the HF data points used to train HF Kriging, additive Kriging, co-Kriging, $\hat{y}_{comp}$ (second order basis functions) and $\hat{y}_{HF}$ (fourth order basis functions). The plot is presented in log-log scale.

version of Kriging, where an LF surrogate multiplicative correction factor and a discrepancy function are trained internally). Co-Kriging is an MFS commonly used as a black box and therefore cannot be used to correct the exact LF simulations. However, we can use additive Kriging which is basically described by Eq. (13) where the additive correction, $\hat{\delta}$, is the Kriging surrogate constructed using as training data points the difference between $y_{HF}(\mathbf{x})$ and $y_{LF}(\mathbf{x})$ built using the Python library Surrogate Modeling Toolbox (SMT) [4]. Here the LF simulations are corrected using the discrepancy function $\hat{\delta}$. Co-Kriging [14] was aslo built using the SMT Python library.

Figure 6 shows the relative RMSE of the metric F for two single-fidelity surrogates and four MFS. The single-fidelity surrogates included are HF Kriging and $\hat{y}_{HF}$ (using fourth order basis functions). The MFS included are two surrogates built using co-Kriging, one using additive Kriging, and one using $\hat{y}_{comp}$ (using second order basis functions). The difference between the two co-Kriging surrogates is the amount of LF points used for training.

For a low number of HF points (≤ 100), additive Kriging clearly outperforms the others, while co-Kriging and $\hat{y}_{comp}$ (quadratic polynomial) are comparable and $\hat{y}_{HF}$ (quartic polynomial) could not be fit. For $150 <$ HF points < 250, $\hat{y}_{comp}$ and additive Kriging perform the best, while for more than 250 HF points, $\hat{y}_{HF}$ outperforms the others. The excellent performance of additive Kriging for very small number of HF points indicates the potential of using exact LF simulations when the number of HF simulations is severely restricted. Indeed, with 30 LF points and 30 HF points (training cost of 31 HF simulations), additive Kriging achieves similar performance to co-Kriging with 100 HF points and 711 LF points (training cost of 121 HF simulations). This difference will offset up to 3,000 evaluations of the additive Kriging for an application such as optimization. An important conclusion that can be extracted from Figure 6 is that on our application, MFSs are useful for low number of HF points, in this case less than 250.

6.3 Using Symmetries to Further Reduce the Number of HF Points

LR-MFS works better if only few HF points are needed. In this section, we go further and the amount of HF points needed is reduced by using the parametric symmetries associated with our problem [9]. After identifying that the quadratic MF comprehensive surrogate using y_{LF} for prediction (i.e. $\hat{y}_{comp}$) works better than the approaches $\hat{y}_{LF}, \hat{y}_{HF}, \hat{y}_{add}, \hat{y}_{\hat{add}}$, and $\hat{y}_{co\hat{m}p}$, for the chosen metric, symmetries to improve its performance are imposed. As mentioned in Section 5.3 we have implemented two approaches, APP and SBF. Figure 7 presents the relative RMSE error resulting of adding permutation points ($\hat{y}^{APP}_{comp}$), and also the results of modifying regression basis functions to impose symmetries ($\hat{y}^{SBF}_{comp}$). The relative RMSE error resulting from $\hat{y}_{comp}$ is also included in the figure for direct comparison.

Figure 7 shows that for a low number of HF points ($\leq$ 200 HF data points) it is clear that imposing symmetries offers a higher error reduction. An outstanding performance is achieved by $\hat{y}^{APP}_{comp}$ which reduces the relative RMSE to less

than 6% using less than 30 HF points. This is 50% of the error of $\hat{y}_{comp}$ for the same number of HF points. The rest of the surrogates in the figure, $\hat{y}_{comp}$ and $\hat{y}^{APP}_{comp}$, achieve an error less than 6% only after using more than 150 HF points, this is a five-fold cost reduction. Although $\hat{y}^{SBF}_{comp}$ error reduction for low number of HF data points is not as impressive as $\hat{y}^{APP}_{comp}$, it allows an error of 7% with less than 30 HF points. $\hat{y}_{comp}$ needs at least 80 HF points for achieving this level of error, which represents a three-fold cost reduction. Note also, that if the LR-MFS with symmetries saves us 100 HF calculations ($3,072 \times 100$ core hours), it allows additional 3,200 LF simulations ($3,072 \times 100/96$) for carrying UQ or optimization.

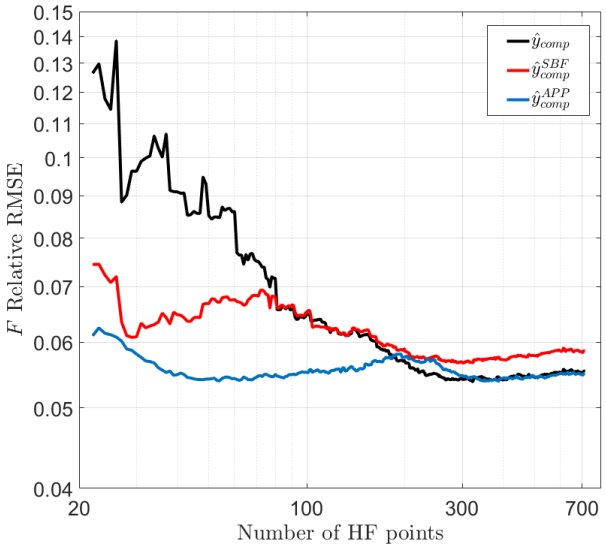


Fig. 7 RMSE for the metric F as a function of the number of HF data points used to train the quadratic surrogates using $\hat{y}_{comp}$, $\hat{y}^{APP}_{comp}$, and $\hat{y}^{SBF}_{comp}$. The plot is presented in log-log scale.

7 Conclusion

In this paper, various surrogates were constructed to predict the metric F that measures the amplification of the departure from axisymmetry of the particle cloud after a multiphase explosion. Low-fidelity and high-fidelity simulations were used to construct low-fidelity, high-fidelity, and multi-fidelity surrogates. The performance of the surrogates has been studied and compared. First, linear regression surrogates with monomial basis functions up to a quadratic polynomial were used. It was found that, for the studied problem with the chosen surrogate and basis functions, (i) multi-fidelity surrogates that use low-fidelity surrogates for prediction perform similarly to the high-fidelity surrogate, (ii) multi-fidelity surrogates that use low-fidelity simulations for prediction performed substantially better. This is due to the complex behavior of both low-fidelity and high-fidelity functions, and the high degree of correlation between them which allows correcting the low-fidelity function with a low-order polynomial. The multi-fidelity surrogates whose predictions were performed using low-fidelity simulations (instead of low-fidelity surrogates) achieved a performance in terms of relative RMSE close to 5%. This is not true when basis functions up to fourth order polynomial are used, but instead we observed that with quartic polynomials and more than 200 HF points, there is no benefit in using MFS, since no MFS significantly improves on the single fidelity surrogate. The benefit in using exact LF simulations is reversed, and using the LF surrogate is better, reflecting the noise filtering benefit of the surrogate. The regression-based comprehensive approaches, i.e. linear regression surrogates that include both, additive and multiplicative corrections, performs poorly for less than 400 HF points, which we traced to ill conditioning and poor selection of ρ.

The superior performance of using exact LF simulations, motivated us to try the same approach for Kriging. Co-Kriging, which replaces the LF simulations with a surrogate, was compared to additive Kriging, which corrects the LF simulations with a Kriging discrepancy function. Similar large gains were observed.

Taking advantage of the parametric symmetries available in our problem, we imposed them using two approaches. LR-MFS built with and without taking advantage of symmetries were compared. We found that if symmetries are used, an accuracy of $\approx 5\%$ (noise level) is achieved using less than 30 high-fidelity data points, instead of 150 high-fidelity data points needed if symmetries are not used, reducing the simulation cost roughly five times. This way, the very small number of required high-fidelity simulations would allow the use of the low-fidelity simulations instead of the low-fidelity surrogates for applications requiring several thousands of evaluations of the multi-fidelity surrogate.

A Appendix: Identification of the significant variables

A.0.1 Methodology

The objective of this appendix is to present the sensitivity analysis of the quantity of interest studied in the paper (the normalized effective Fourier effective perturbation described in Section 3) with respect to the variables used to parametrized the PVF perturbation and described in Section 2 by Eq. (1). We recall that this perturbation is modeled by,

$$\phi^p(\theta) = \phi^p_0[1 + A_1\cos(k_1\theta + \Phi_1) + A_2\cos(k_2\theta + \Phi_2) + A_3\cos(k_3\theta + \Phi_3)],$$

with the constraint,

$$\sqrt{A_1^2 + A_2^2 + A_3^2} = 0.02.$$

Moreover as the variables Φ_1, Φ_2, Φ_3 are used to model the phase shift, one can arbitrary set $\Phi_1 = 0$ and consider the phase shift with respect to the first mode. As a consequence the problem counts seven independent variables concatenated into the vector $X = \{A_1, A_2, k_1, k_2, k_3, \Phi_2, \Phi_3\}$. The quantity of interest can thus be expressed as,

$$F = \mathscr{M}(X)$$

in which the mapping $\mathscr{M}$ involves the numerical resolution of the multi-phase explosion and the post processing of the solution as explained in Section 2 and Section 3.

The methodology used to study the sensitivity of the quantity of interest F to the input parameters X is to perform a global sensitivity analysis by computing variance based sensitivity indices. Consequently it is assumed that X is a vector of seven independent random variables (the probability distribution will be discussed in the following) and F is a random variable of unknow probability distribution. Then the approach proposed in [20] and improved in [2] for the computation of sensitivity indices (Sobol' indices [19]) by sparse polynomial chaos expansion (PCE) [3] is applied. Assuming that F is a second order random variable, it can be shown [5] that,

$$F = \sum_{i=0}^{\infty} C_i \phi_i(X)$$

where $\{\phi_i\}_{i\in\mathbb{N}}$ is a polynomial basis orthogonal with respect to the probability distribution of X and C_i are unknown coefficients.
Sparse PCE consists in the construction of a sparse polynomial basis $\{\phi_i\}_{\alpha\in\mathscr{A}}$, where $\alpha = (\alpha_1, \cdots, \alpha_n)$ is a multi index used to identify the polynomial acting with the power α_i on the variable X_i and $\mathscr{A}$ is a set of indices α. In practice $\mathscr{A}$ is a subset of the set $\mathscr{B}$ which contains all the indices α up to a dimension d, i.e. $card(\mathscr{B}) = \frac{(d+n)!}{d!n!}$. Objective of sparse approach is to find an accurate polynomial basis $\{\phi_i\}_{\alpha\in\mathscr{A}}$ such that $card(\mathscr{A}) << card(\mathscr{B})$. In the present case this is achieved by Least Angle Regression, i.e. unknown coefficients C_i are computed by iteratively solving a mean square problem and selecting, at each iteration, the polynomial which is the most correlated with the residual (see [3] for details).
Finally one gets the following approximation,

$$F \approx \hat{F} = \sum_{\alpha\in\mathscr{A}} C_\alpha \phi_\alpha(X)$$

from which the sensitivity index can be derived. Indeed the orthogonality of the polynomial basis $\{\phi_i\}_{\alpha\in\mathscr{A}}$ allows to write the expectation and the variance in the following form,

$$E[\hat{F}] = C_0$$
$$Var[\hat{F}] = \sum_{\alpha\in\mathscr{A}} C_\alpha^2 E[\phi_\alpha^2(X)]$$

In addition, the idea pointed out in [20] is to identify the PCE with the ANOVA decomposition, from which one can show that, the first order sensitivity index of the variable X_i reads,

$$\hat{S}_i = \frac{\sum_{\alpha\in L_i} C_\alpha^2 E[\phi_\alpha^2(X)]}{Var[\hat{Y}]}$$

where $L_i = \{\alpha \in \mathscr{A},\ \forall\ j \neq i\ \alpha_j = 0\}$, i.e. only the polynomials acting exclusively on the variable X_i are considered.
The total sensitivity index is also available by,

$$\hat{S}_{T_i} = \frac{\sum_{\alpha\in L_i^+} C_\alpha^2 E[\phi_\alpha^2(X)]}{Var[\hat{Y}]}$$

where $L_i^+ = \{\alpha \in \mathscr{A},\ \alpha_i \neq 0\}$, i.e. all the polynomials acting on the variable X_i are considered (allows to consider interactions between X_i and the other variables).

One can note that the approximation of the sensitivity index obtained by sparse PCE relies on an *accurate* approximation of the surrogate response by the sparse PCE, however, the link between the accuracy of the PCE approximation and the accuracy of the approximated sensitivity index is not straightforward. In order to access the quality of the sensitivity index computed by PCE, a bootstrap approach proposed in [8] is set up and detailed in the next section.

A.0.2 Application to the high fidelity computation of normalized Fourier effective perturbation

The probabilistic surrogate of the seven independent input parameters is detailed in Table 4. Uniform distributions are assumed for each component of the random vector X and variability ranges are defined, which basically define the domain of interest for the sensitivity analysis.

Table 4 Probabilistic surrogate of seven independent input parameters

Variable	Distribution	Lower Boundary	Upper Boundary
A_1	Continuous uniform	0	$\sqrt{0.02}$
A_2	Continuous uniform	0	$\sqrt{0.02}$
k_1	Discrete uniform	1	25
k_2	Discrete uniform	1	25
Φ_2	Continuous uniform	0	2π
Φ_3	Continuous uniform	0	2π

Then, a design of experiments of 711 points is drawn by LHS in order to estimate the sensitivity index by sparse PCE. Maximum order of the polynomials is set to $d = 4$. In order to assess the accuracy of the obtained sensitivity index, the following bootstrap procedure is proposed. Among the 711 points, 611 are used as a training set to compute the PCE approximation (least angle regression approach) and 100 are used as a validation set. The training and validation set are randomly chosen among the 711 points. The bootstrap approach consists in repeating B times this procedure changing each time the training and the validation set. This leads to B different PCE approximations and thus to a sample of B sensitivity indices. This sample is further used to estimate the coefficient of variation of the sensitivity index estimators obtained by sparse PCE. Moreover, for each bootstrap sample, the relative L_2 norm of the relative error (ε^2) is computed on the validation set as well as the coefficient of determination (R^2) computed on the training set.

$$\varepsilon^2 = \frac{||\hat{F}_{validation} - F_{validation}||^2}{||F_{validation}||^2}$$
$$R^2 = \frac{||\hat{F}_{train} - E[F_{train}]||^2}{||F_{train} - E[F_{train}]||^2}$$

Over the B bootstrap repetitions the estimated mean values are $E[\varepsilon^2] \approx 1.54 \times 10^{-3}$ and $E[R^2] \approx 9.74 \times 10^{-1}$ and the coefficients of variation are, $cv_{\varepsilon^2} \approx 1.57 \times 10^{-1}$ and $cv_{R^2} \approx 1.86 \times 10^{-3}$. These first results allowed to be confident in the accuracy of the PCE approximation. Note that the relatively large coefficient of variation cv_{ε^2} should be considered with respect to its very low mean value. Concerning the sensitivity index, Table 5 presents the first order and total index with the mean values and coefficient of variation estimated by bootstrap.

First of all the results presented in Table 5 show that when a sensitivity index has a significant value (values highlighted in bold font) its estimation is quite accurate as the coefficient of variation is relatively low (less than 3%). One can also note that for low values of sensitivity indices the coefficients of variation are quite large, however, as these

Table 5 Estimation of the first order sensitivity index and total sensitivity index estimated by sparse PCE, mean values and coefficient of variation estimated by bootstrap with $B = 500$ repetitions.

variable	$E[\hat{S}_i]$	$cv_{\hat{S}_i}$	$E[\hat{S}_{T_i}]$	$cv_{\hat{S}_{T_i}}$
A_1	3.05×10^{-4}	5.14×10^{-1}	$\mathbf{2.58 \times 10^{-1}}$	1.50×10^{-2}
A_2	1.45×10^{-4}	1.11	$\mathbf{2.74 \times 10^{-1}}$	1.70×10^{-2}
k_1	$\mathbf{1.38 \times 10^{-1}}$	2.1×10^{-2}	$\mathbf{2.91 \times 10^{-1}}$	1.88×10^{-2}
k_2	$\mathbf{1.40 \times 10^{-1}}$	2.1×10^{-2}	$\mathbf{2.92 \times 10^{-1}}$	1.89×10^{-2}
k_3	$\mathbf{1.69 \times 10^{-1}}$	2.2×10^{-2}	$\mathbf{4.39 \times 10^{-1}}$	1.16×10^{-2}
Φ_2	1.34×10^{-4}	1.10	1.37×10^{-3}	4.60×10^{-1}
Φ_3	6.20×10^{-5}	1.31	1.01×10^{-3}	4.55×10^{-1}

sensitivity indices are, at least, two orders of magnitude lower than the significant ones their poor estimation is not detrimental for the purpose of sensitivity analysis. With respect to the results of the sensitivity analysis, one can conclude that the variance of the quantity of interest is mainly driven by the first five variables namely A_1, A_2, k_1, k_2, k_3. It is also interesting to note the strong interaction between the amplitude variables (that have almost no first-order effect) with the wavenumbers which is consistent with the shape of the perturbation (in Eq. (1) the interactions between A_1, k_1 and A_2, k_2 clearly appear) and the constraint on the amplitude (Eq. (2)) which explains the interaction between A_1, A_2 and k_3 and justify that $\hat{S}_{T_{k_3}}$ has the highest value.

Based on this result it has been decided to consider only the five variables A_1, A_2, k_1, k_2, k_3 for the construction of the surrogate of the quantity of interest F.

B Appendix: Percentage contribution of low-fidelity and high-fidelity data in LR-MFS using up to second order polynomial basis functions

The contribution of the LF data points to the MFSs is studied in order to understand why the performance of the surrogates that use y_{LF} instead of the LF surrogate, $\hat{y}_{LF}$, for prediction, worked overwhelmingly better for the metric F. First, notice that in our case $n = 21$ in Eq. (8), which is the number of coefficients of a quadratic polynomial in five variables. Also, note that for prediction we can choose between using y_{LF} or $\hat{y}_{LF}$, however for training purposes y_{LF} is used. Therefore ρ and $\hat{\delta}(x)$ coefficients in Eq. (7) and in Eq. (14) are identical. Equation (8) shows explicitly the contribution of each surrogate, HF and LF, to the comprehensive MF approximation using the $\hat{y}_{LF}$, however, we can choose to use y_{LF}, i.e. the LF simulations directly, instead. The first term represents the contribution of the LF simulation, $\rho\hat{y}_{LF}(x)$ (using LF surrogate) or $\rho y_{LF}(x)$ (using LF simulations), to the MFS. The second term, $\sum_{1=1}^{p} X_i(x)b_i$, represents the contribution of the HF simulation to the MFS.

Figure B1 shows the mean contribution in percentage of the LF and HF information to the comprehensive MFS predicted using y_{LF}, i.e. $\hat{y}_{comp}$. Using y_{LF} or $\hat{y}_{LF}$ for predicting the MF comprehensive correction gives the same LF and HF mean contributions, therefore, only one plot was included. The mean contribution was calculated averaging the contribution of each of the 92 validation points. The contribution of the LF model is defined as $\rho\hat{y}_{LF}$, while the contribution of the HF model is defined as $\sum_{1=1}^{p} X_i(x)b_i$. A negative percentage indicates, in this case, that the contribution has negative sign. Values higher than 100% indicates that the mean of the contribution is higher than the metric value in average, i.e. calculating the term $\rho\hat{y}_{LF}$ for each of the validation points and then taking the average. Therefore, ρ does not need to be higher than one, instead, the term $\rho\hat{y}_{LF}$ needs to be higher than one on average. Naturally, the sum of the LF and HF contributions adds to 100%. For more than 50 HF data points, the contribution is

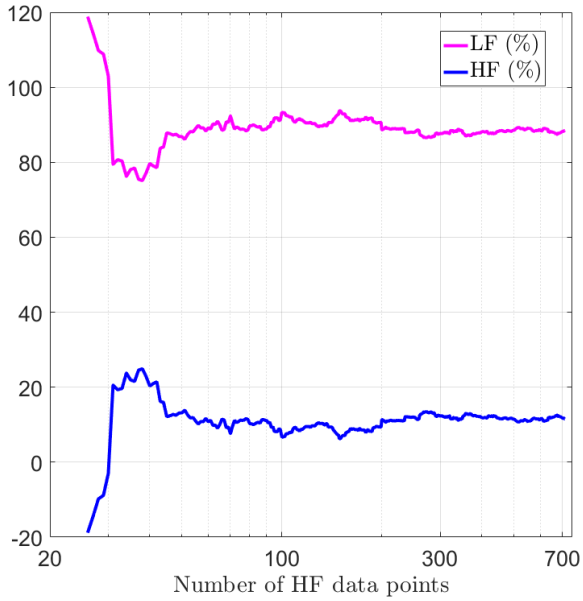


Fig. B1 Mean contribution of the LF ($\rho\hat{y}_{LF}(x)$) and of the HF ($\sum_{1=1}^{p} X_i(x)b_i$ surrogates from Eq. (8)) to the comprehensive MFS ($\hat{y}_{comp}$) prediction (using LF simulations) in percentage as a function of the HF data points used. The plots are presented in a linear-log scale. A negative percentage indicates that the contribution has negative sign. Values higher than 100% indicates that the mean of the contribution is higher than the metric value in average.

dominated by the LF information ($\approx 85\%$). This helps to explain why when the LF simulation output is used directly (without constructing a surrogate) the improvement is substantial in Figure 4. That is, for metric F the results of the MF comprehensive prediction using y_{LF}, $\hat{y}_{comp}$, performs substantially better than the ones that use the LF surrogate, $\hat{y}_{comp}$. When the source of LF is changed for prediction instead of a surrogate, the results change drastically, which is expected due to the high correlation between HF and LF simulations (0.92).

Acknowledgment

The authors would like to thank to the reviewers of the paper for their wonderful suggestions that made the manuscript a better contribution.

Funding Sources

This work was partially supported by the Center for Compressible Multiphase Turbulence, the U.S. Department of Energy, National Nuclear Security Administration, Advanced Simulation and Computing Program, as a Cooperative Agreement under the Predictive Science Academic Alliance Program, under Contract No. DE-NA0002378.

This work was partially supported by the French National Research Agency (ANR) through the ReBReD project under grant ANR-16-CE10-0002 and by a ONERA internal project MUFIN dedicated about multi-fidelity.

This work was partially performed under U.S. Government contract 89233218CNA000001 for Los Alamos National Laboratory (LANL), which is operated by Triad National Security, LLC for the U.S. Department of Energy/National Nuclear Security Administration. Approved for public release LA-UR-19-22491.

Conflict of Interest

The authors declare that they have no conflict of interest.

Replication of Results

The HF data points, the LF data points, and the validation data points are included in the supplementary material.

References

1. Annamalai, S., Rollin, B., Ouellet, F., Neal, C., Jackson, T.L., Balachandar, S.: Effects of initial perturbations in the early moments of an explosive dispersal of particles. Journal of Fluids Engineering **138**(7), 070903 (2016). DOI 10.1115/1.4030954
2. Blatman, G., Sudret, B.: Efficient computation of global sensitivity indices using sparse polynomial chaos expansions. Reliability Engineering & System Safety **95**(11), 1216 – 1229 (2010). DOI https://doi.org/10.1016/j.ress.2010.06.015. URL http://www.sciencedirect.com/science/article/pii/S0951832010001493
3. Blatman, G., Sudret, B.: Adaptive sparse polynomial chaos expansion based on least angle regression. Journal of Computational Physics **230**(6), 2345 – 2367 (2011). DOI http://dx.doi.org/10.1016/j.jcp.2010.12.021. URL http://www.sciencedirect.com/science/article/pii/S0021999110006856
4. Bouhlel, M.A., Hwang, J.T., Bartoli, N., Lafage, R., Morlier, J., Martins, J.R.R.A.: A Python surrogate modeling framework with derivatives. Advances in Engineering Software (2019). (In press)
5. Cameron, R.H., Martin, W.T.: The Orthogonal Development of Non-Linear Functionals in Series of Fourier-Hermite Functionals. Annals of Mathematics **48**(2), 385–392 (1947). URL http://www.jstor.org/stable/1969178
6. Cressie, N.: Statistics for Spatial Data: Wiley Series in Probability and Statistics. Wiley: New York, NY, USA (1993). DOI 10.1002/9781119115151
7. Dobrat, B., Crawford, P.: Handbook, LLNL explosives. Lawrence Livermore National Laboratory (1981). OSTI Identifier 6530310
8. Dubreuil, S., Berveiller, M., Petitjean, F., Salaün, M.: Construction of bootstrap confidence intervals on sensitivity indices computed by polynomial chaos expansion. Reliability Engineering & System Safety **121**(Supplement C), 263 – 275 (2014). DOI https://doi.org/10.1016/j.ress.2013.09.011. URL http://www.sciencedirect.com/science/article/pii/S0951832013002688
9. Fernández-Godino, M.G., Balachandar, S., Haftka, R.: On the use of symmetries in building surrogate models. Journal of Mechanical Design (2018). DOI 10.1115/1.4042047
10. Fernandez-Godino, M.G., Ouellet, F., Haftka, R., Balachandar, S.: Early time evolution of circumferential perturbation of initial particle volume fraction in explosive cylindrical multiphase dispersion. Journal of Fluids Engineering **141**, 0913021–09130220 (2019). DOI 10.1115/1.4043055
11. Fernández-Godino, M.G., Park, C., Kim, N.H., Haftka, R.T.: Issues in deciding whether to use multifidelity surrogates. AIAA Journal pp. 1–16 (2019)
12. Gray, J.S., Hwang, J.T., Martins, J.R., Moore, K.T., Naylor, B.A.: Openmdao: An open-source framework for multidisciplinary design, analysis, and optimization. Structural and Multidisciplinary Optimization pp. 1–30 (2019)
13. Kennedy, M.C., O'Hagan, A.: Predicting the output from a complex computer code when fast approximations are available. Biometrika **87**(1), 1–13 (2000). DOI 10.1093/biomet/87.1.1
14. Le Gratiet, L., Garnier, J.: Recursive co-kriging model for design of computer experiments with multiple levels of fidelity. International Journal for Uncertainty Quantification pp. 365–386 (2014)
15. Myers, D.E.: Matrix formulation of co-kriging. Mathematical Geology **14**(3), 249–257 (1982). DOI 10.1007/BF01032887
16. Ouellet, F., Annamalai, S., Rollin, B.: Effect of a bimodal initial particle volume fraction perturbation in an explosive dispersal of particles. In: AIP Conference Proceedings, vol. 1793, p. 150011. AIP Publishing (2017). DOI 10.1063/1.4971740
17. Queipo, N.V., Haftka, R.T., Shyy, W., Goel, T., Vaidyanathan, R., Tucker, P.K.: Surrogate-based analysis and optimization. Progress in aerospace sciences **41**(1), 1–28 (2005). DOI 10.1016/j.paerosci.2005.02.001
18. Seber, G.A., Lee, A.J.: Linear regression analysis, vol. 329. John Wiley & Sons (2012). ISBN 978-0-471-41540-4
19. Sobol, I.M.: Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. Mathematics and computers in simulation **55**(1), 271–280 (2001). URL http://www.sciencedirect.com/science/article/pii/S0378475400002706
20. Sudret, B.: Global sensitivity analysis using polynomial chaos expansions. Reliability Engineering & System Safety **93**(7), 964 – 979 (2008). DOI http://dx.doi.org/10.1016/j.ress.2007.04.002. URL http://www.sciencedirect.com/science/article/pii/S0951832007001329. Bayesian Networks in Dependability
21. Vauclin, R.: Développement de modèles réduits multifidélité en vue de l'optimisation de structures aéronautiques. In: Rapport Institut Supérieur de l'Aéronautique et de l'Espace – École Nationale Supérieur des Mines de Saint-Étienne (2014)
22. Zhang, Y., Kim, N.H., Park, C., Haftka, R.T.: Multifidelity surrogate based on single linear regression. AIAA Journal **56**(12), 4944–4952 (2018). DOI 10.2514/1.J057299