PCA and Kriging for the efficient exploration of consistency regions in Uncertainty Quantification

Gianmarco Aversano, John Camilo Parra-Alvarez, Benjamin J. Isaac, Sean T. Smith, Axel Coussement, Olivier Gicquel, Alessandro Parente

Available online at www.sciencedirect.com

ScienceDirect

Proceedings of the Combustion Institute 37 (2019) 4461–4469

Proceedings of the Combustion Institute

www.elsevier.com/locate/proci

PCA and Kriging for the efficient exploration of consistency regions in Uncertainty Quantification

Gianmarco Aversano [a,b,c,*], John Camilo Parra-Alvarez [d], Benjamin J. Isaac [d], Sean T. Smith [d], Axel Coussement [a,b], Olivier Gicquel [c], Alessandro Parente [a,b,*]

[a] *Université Libre de Bruxelles, École polytechnique de Bruxelles, Aero-Thermo-Mechanics Laboratory, Bruxelles, Belgium*
[b] *Université Libre de Bruxelles and Vrije Universiteit Brussel, Combustion and Robust Optimization Group (BURN), Brussels, Belgium*
[c] *Laboratoire EM2C, CNRS, Centrale-Supélec, Université ParisSaclay, 8–10 rue Joliot-Curie, Gif-sur-Yvette 91190, France*
[d] *Department of Chemical Engineering, University of Utah, Salt Lake City, UT, USA*

Received 30 November 2017; accepted 6 July 2018
Available online 22 August 2018

Abstract

For stationary power sources such as utility boilers, it is useful to dispose of parametric models able to describe their behavior in a wide range of operating conditions, to predict some Quantities of Interest (QOIs) that need to be consistent with experimental observations. The development of predictive simulation tools for large scale systems cannot rely on full-order models, as the latter would lead to prohibitive costs when coupled to sampling techniques in the model parameter space. An alternative approach consists of using a Surrogate Model (SM). As the number of QOIs is often high and many SMs need to be trained, Principal Component Analysis (PCA) can be used to encode the set of QOIs in a much smaller set of scalars, called PCA scores. A SM is then built for each PCA score rather than for each QOI. The advantage of reducing the number of variables is twofold: computational costs are reduced (less SMs need to be trained) and information is preserved (correlation among the original variables).
The strategy is applied to a CFD model simulating the Alstom 15 MW_{th} oxy-pilot Boiler Simulation Facility (BSF). In practice, experiments cannot provide full coverage of the pulverized-coal utility boiler due to both practicality and costs. Values of the model's parameters which guarantee consistency with the experimental

* Corresponding authors at: Université Libre de Bruxelles, École polytechnique de Bruxelles, Aero-Thermo-Mechanics Laboratory, Bruxelles, Belgium.
E-mail addresses: Gianmarco.Aversano@ulb.ac.be (G. Aversano), jcparraa@gmail.com (J.C. Parra-Alvarez), benjamin.j.isaac@utah.edu (B.J. Isaac), sean.t.smith@utah.edu (S.T. Smith), axel.coussement@ulb.ac.be (A. Coussement), olivier.gicquel@centralesupelec.fr (O. Gicquel), Alessandro.Parente@ulb.ac.be (A. Parente).

https://doi.org/10.1016/j.proci.2018.07.040

data of this test facility for 121 QOIs are found, by training a SM based on the combination of Kriging and PCA, using only 5 latent variables.

Keywords: PCA; Bound-to-Bound Data Collaboration; Uncertainty Quantification; Surrogate models

1. Introduction

In engineering applications, the ability to make reliable predictions about certain physical systems is granted by the existence of predictive mathematical models, based on the deep understanding of the underlying processes. These mathematical models, even if deterministic, often include uncertainties which limit their predictive abilities (e.g., unknown design or operating parameters). One way to assess a model's capability to predict the correct values for a set of Quantities of Interest (QOIs) is to compare the model's predictions with reference data, i.e., measurements coming from experiments. A mathematical model is usually said to be *consistent* with the data when the error interval between its predictions and the experimental data is in the same range as the uncertainty intervals of the experimental values [1].

On occasions, a model is defined by some parameters whose value is uncertain. Values of these model parameters for which the model's predictions are consistent with experimental data exist, but they are not known [2,3]. A way to find this set of values is to heavily sample the parameter space (e.g., Latin Hypercube Sampling, Monte Carlo random sampling) and evaluate the model's prediction for the QOIs at every location. This strategy can work when the model's output is fast to compute. In the case of computationally costly models, this strategy is prohibitive. Computationally expensive models are dominant in the world of Computational Fluid-Dynamics (CFD). CFD simulations are usually run on many CPUs and, in spite of that, they still need hundreds or thousands of hours of computational time to converge. Heavily sampling the input space of CFD models is not yet feasible. Having a Surrogate Model (SM) that can approximate the model predictions, at a lower computational cost is preferable [4,5]. SMs are mathematical models based on available data that approximate the underlying *hidden* relationship between input and output. SMs are useful when this relationship is either not known or comes in the form of a computationally expensive computer code. SMs are also popular in Uncertainty Quantification (UQ) studies [6–9]. Examples are Polynomial Chaos Expansion (PCE) and Gaussian Process Regression (GPR), which are often employed for the computation of Sobol's indices. SMs are constructed starting from a relatively small set of *training* observations of the predictive model's output, which correspond to a set of *training* locations in the model parameter space. Once a SM is trained, consistency with the experimental data is performed by analyzing the SM's output instead of the actual model's predictions. The uncertain model parameters are then assigned the value for which the difference between the experimental values and the SM predictions is the lowest. Usually, SMs are built for one scalar target. In the presence of many QOIs, as many SMs are needed as the number of QOIs. This is true, for example, if PCE or GPR are chosen as SMs, without any compression/reduction technique. Besides, when dealing with outputs of a deterministic computer code, interpolation might be preferred over regression. A reduction is possible if the original set of QOIs can be represented by a new set of fewer scalars. One reason why one would want to reduce the number of SMs is that the training itself can still be costly. Indeed, SMs are usually defined by a set of hyper-parameters, whose value affect the SM's predictive abilities. Very often, a good estimation for the value of these hyper-parameters comes via the solution of constrained optimization problems that involve local optima. Another reason for reducing the number of SMs is that very often the QOIs are correlated, but their correlation might be lost in the process of building individual SMs for each of them. Taking these factors into account, the advantage of reducing the number of QOIs, and consequently the number of SMs to train, becomes clear.

A very popular method for data compression is Principal Component Analysis (PCA) [10]. PCA is a statistical technique used to find a set of orthogonal low-dimensional basis functions to represent an ensemble of high-dimensional data. In the context of consistency analysis, PCA can be used to find a new, smaller set of uncorrelated variables, often referred to as *PCA scores*, that is representative of the original QOIs. Once these scores are found, a SM can be built for each one of them. Then, the model parameter space can be explored and a consistency region in the model input space can be more easily found.

In this work, we apply this strategy to find the optimal input parameter values of a CFD model [11] for the Alstom 15 MW_{th}, pulverized coal, tangentially fired, oxy-pilot Boiler Simulation Facility (BSF) [12]. Experimental data for this test facility are available, such as temperature and heat-flux mapping measurements. The available CFD model has limited predictive capabilities as it is defined by 3 parameters whose correct values are not known (the value for which consistency with the experimental data is guaranteed). 121 QOIs are to be correctly predicted by the model, namely temperature and heat-fluxes at specific location inside the boiler corresponding to the experimental measurements. 22 simulations are carried out in order to explore the 3-dimensional model parameter space and used as training set. After performing PCA, it is shown that 5 PCA scores can explain over 99% of the original data variance. The set of 121 original variables is thus encoded in a new set of 5 scalars. The consistency region in the 3-dimensional parameter space is found by training a SM based on Kriging for the PCA scores. The CFD model's predictive capabilities for the BSF are improved by choosing values for these 3 parameters that belong to the consistency region, and can be used for the design of larger-scale facilities.

2. Theory

2.1. Bound-to-Bound Data Collaboration

Bound-to-Bound Data Collaboration (B2B-DC) is a mathematical framework that tests consistency between a data-set and a model [13–15].

The basis of B2B-DC is composed of an underlying physical process and associated model, a collection of experimental observations with respective uncertainties, and SMs representing parametric dependence of the physical-model predictions of the QOIs on the uncertain parameters.

Each QOI $y_i \; \forall i = 1, \ldots, N$ is both experimentally measured and predicted by a model. N is the number of QOIs. The set of inequalities

$$|M(\mathbf{x}) - \mathbf{y}_{exp}| \leq \boldsymbol{\sigma}, \tag{1}$$

combines the experimental and modeling information into a single set of constraints. $\mathbf{x}$ is the vector of P uncertain parameters. $M(\mathbf{x})$ is the model's prediction of the QOIs ($\mathbf{y}$) based on the input parameters ($\mathbf{x}$). Thus, $\mathbf{y}(\mathbf{x}) = M(\mathbf{x})$ is the vector of predicted QOIs, by the model, when the input parameters' values are the ones contained in $\mathbf{x}$. $\mathbf{y}^{exp}$ are the measured values. The size of the vectors $\mathbf{y}(\mathbf{x})$ and $\mathbf{y}^{exp}$ is N. The discrepancy between the measurement of one QOI and its model prediction is bounded by σ_i, which is usually the experimental uncertainty. A point $\mathbf{x}$ in the model parameter space is *consistent* with the experimental data if the corresponding model prediction $M(\mathbf{x})$ satisfies the set of constraints (1). The constraints (1) represent a hyperbox in the $\mathbf{y}$-space and limit the allowed discrepancy between experimental measurement and model prediction for each individual QOI y_i. If $\boldsymbol{\Sigma}$ is a diagonal matrix, such that $\boldsymbol{\Sigma} = diag(\boldsymbol{\sigma})$, we can express (1) in matrix form:

$$\boldsymbol{\Sigma}^{-1}|M(\mathbf{x}) - \mathbf{y}_{exp}| \leq \mathbf{1}. \tag{2}$$

The set of N pairs of orthogonal linear constraints (1) or (2) represents a hyper-rectangle in the $\mathbf{y}$-space and it states that the model's predictions for each QOI y_i must lie within this hyper-rectangle in order for the model to be consistent with the reference data.

Rather than a hyper-rectangle, the feasible set can also be bounded using an ellipsoid, which is defined by a single quadratic constraint:

$$\left[M(\mathbf{x}) - \mathbf{y}_{exp}\right]^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}^{-1}\left[M(\mathbf{x}) - \mathbf{y}_{exp}\right] \leq \alpha, \tag{3}$$

where α is a quantity to be determined. Clearly, it is preferable to have the smallest α such that the ellipsoid contains the feasible set. In the case $\alpha = N$, the ellipsoid in the $\mathbf{y}$-space determined by (3) contains the hyper-rectangle defined by (2). If $\alpha = 1$, the opposite is true.

By either using (2) or (3), a region $\mathcal{F}$ of consistency in the $\mathbf{x}$-space can be found. This region is called *consistency region* and represents the region of all possible values of $\mathbf{x}$ for which the model predictions, $M(\mathbf{x})$, respect either (1) or (3). It is worth noting that, in general, a solution to (1) might not exist and in such a case the consistency region would be a null-set.

2.2. Principal Component Analysis (PCA)

PCA is a statistical technique that finds a set of orthogonal low-dimensional basis functions to represent an ensemble of high-dimensional data describing an undesirably complex system [16–18].

For a data-set $\mathbf{Y}(N \times M)$, containing M observations of N original variables, PCA provides an approximation of the original data-set using only $q < N$ linear correlations between the N variables. The quantity q is referred to as *approximation order*.

Data are usually centered and scaled before applying PCA. Centering represents all observations as fluctuations, leaving only the relevant variation for analysis [16]. The centered-scaled data read:

$$\mathbf{Y}_0 = \mathbf{D}^{-1}(\mathbf{Y} - \bar{\mathbf{Y}}), \tag{4}$$

where $\mathbf{D}$ indicates a diagonal matrix of chosen scaling factors, usually standard deviations, and $\bar{\mathbf{Y}}$ a matrix of mean values. The dimension of $\mathbf{Y}_0$ is also $(N \times M)$.

A set of $q < N$ PCA modes or directions can be found, $\mathbf{V_q} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_q\}$, thus the data can be encoded in a set of q scalars called PCA scores or Principal Components (PCs) as follows:

$$a_i(\mathbf{x}_j) = \mathbf{v}_i^T \mathbf{D}^{-1}\left(\mathbf{y}(\mathbf{x}_j) - \bar{\mathbf{y}}\right) \qquad \forall i = 1, \ldots, q. \tag{5}$$

The centered-scaled data can be approximated by $\mathbf{Y}_0 \approx \mathbf{V_q}\mathbf{V_q}^T\mathbf{D}^{-1}\left(\mathbf{Y}-\bar{\mathbf{Y}}\right) = \mathbf{V_q}\mathbf{A_q}$. The data-set $\mathbf{Y}$ can be approximated as:

$$\mathbf{Y} = \bar{\mathbf{Y}} + \mathbf{D}\mathbf{Y}_0 \approx \bar{\mathbf{Y}} + \mathbf{D}\mathbf{V_q}\mathbf{A_q} = \tilde{\mathbf{Y}}_\mathbf{q}, \tag{6}$$

where $\mathbf{A_q} = \{\mathbf{a}(\mathbf{x}_1), \mathbf{a}(\mathbf{x}_2), \ldots, \mathbf{a}(\mathbf{x}_M)\}$ is the matrix storing all the PCA scores for different values of the input parameters; $\mathbf{a}(\mathbf{x}_j) = \{a_1(\mathbf{x}_j), \ldots, a_q(\mathbf{x}_j)\}^T$ is the vector containing observations of all the q PCA scores for $\mathbf{x}_j$ and $\tilde{\mathbf{Y}}_\mathbf{q}$ is the approximation of $\mathbf{Y}$ obtained with PCA, using only q PCs. Equivalently, one observation $\mathbf{y}(\mathbf{x}_j) \in \mathrm{R}^\mathrm{N}$ contained in the data-set matrix $\mathbf{Y}$ can be approximated as: $\mathbf{y}(\mathbf{x}_j) \approx \bar{\mathbf{y}} + \mathbf{D}\mathbf{V_q}\mathbf{a}(\mathbf{x}_j)$.

2.3. Kriging

For a general scalar target y, every realization $y(\mathbf{x})$ is expressed in the Kriging method as a combination of a trend function and a residual [19]:

$$y(\mathbf{x}) = \sum_{i=0}^{p} \beta_i f_i(\mathbf{x}) + z(\mathbf{x}) = \boldsymbol{\beta}^T\mathbf{f}(\mathbf{x}) + z(\mathbf{x}). \tag{7}$$

The trend function is expressed as a weighted linear combination of $p+1$ polynomials,$\mathbf{f}(\mathbf{x}) = [f_0(\mathbf{x}), \ldots, f_p(\mathbf{x})]^T$ with the weights $\boldsymbol{\beta} = \left[\beta_0, \ldots, \beta_p\right]^T$ determined by generalized least squares (GLS). The subscript p also indicates the degree of the polynomial. The residuals $z(\mathbf{x})$ are modeled by a Gaussian process with a kernel or correlation function that depends on a set of hyper-parameters $\boldsymbol{\theta}$ to be evaluated by Maximum Likelihood Estimation (MLE) [19–21]. The natural log of the marginal likelihood is given by:

$$\ln(\mathcal{L}_M) = \frac{M}{2}\ln(2\pi) + \frac{M}{2}\ln(\sigma^2) + \frac{M}{2}\ln(|\mathbf{R}|) + \frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{F}\boldsymbol{\beta})^T\mathbf{R}^{-1}(\mathbf{y}-\mathbf{F}\boldsymbol{\beta}), \tag{8}$$

where $\mathbf{F}$ is the matrix of polynomials evaluated at the training locations, $\mathbf{R}$ is the kernel matrix of the training data, M is the number of training points.

The final form of the Kriging predictor for any realization $y(\mathbf{x})$ is

$$y(\mathbf{x}) = \mathbf{f}(\mathbf{x})^T\tilde{\boldsymbol{\beta}} + \mathbf{r}(\mathbf{x})^T\mathbf{R}^{-1}(\mathbf{y}-\mathbf{F}\tilde{\boldsymbol{\beta}}). \tag{9}$$

In (9), $\mathbf{r}$ is the vector of correlations between the training points and the prediction point $\mathbf{x}$. To make a prediction, only the terms $\mathbf{f}(\mathbf{x})$ and $\mathbf{r}(\mathbf{x})$ need to be updated: $y(\mathbf{x}^*) = \mathbf{f}(\mathbf{x}^*)^T\tilde{\boldsymbol{\beta}} + \mathbf{r}(\mathbf{x}^*)^T\mathbf{R}^{-1}(\mathbf{y}-\mathbf{F}\tilde{\boldsymbol{\beta}})$, where $\mathbf{x}^*$ is the point in the input parameter space for which we wish to make a prediction.

2.4. Reduced-Order Bound-to-Bound Data Collaboration

The constraints (2) and (3) can be reformulated using PCA. The model's outputs and the experimental data can be represented by the $N \times M$ matrix $\mathbf{Y}$. Each column of $\mathbf{Y}$ is encoded in its corresponding set of PCA scores as follows:

$$\mathbf{a}(\mathbf{x}_m) = \mathbf{V_q}^T\mathbf{D}^{-1}(\mathbf{y}(\mathbf{x}_m) - \bar{\mathbf{y}}) \tag{10}$$

$$\mathbf{a}_{exp} = \mathbf{V_q}^T\mathbf{D}^{-1}\left(\mathbf{y}_{exp} - \bar{\mathbf{y}}\right). \tag{11}$$

The subscript $m = 1, \ldots, M$ indicates one of the model's outputs.

For each PCA score a_i, a SM is built using the method introduced in Section 2.3. A very high number of predictions for the PCA scores is generated. The predicted QOIs are recovered from the predicted PCA scores using (6), and consistency is achieved if (2) or (3) is satisfied. Computational savings are achieved because less SMs are trained ($q \ll N$).

Using PCA, the ellipsoid (3) can be approximated by:

$$\Delta\mathbf{a}_m^T\mathbf{V_q}^T\mathbf{D}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{D}\mathbf{V_q}\Delta\mathbf{a}_m \leq \alpha, \tag{12}$$

where $\Delta\mathbf{a}_m = \mathbf{a}(\mathbf{x}_m) - \mathbf{a}_{exp}$. If the matrix $\boldsymbol{\Sigma}^{-1}\mathbf{D}\mathbf{V_q}$ is indicated by $\mathbf{H_q}$, the constraint (12) can be re-expressed as:

$$\Delta\mathbf{a}_m^T\mathbf{H_q}^T\mathbf{H_q}\Delta\mathbf{a}_m \leq \alpha. \tag{13}$$

One can notice that using (13), the quantity $\mathbf{H_q}$, or even $\mathbf{H_q}^T\mathbf{H_q}$, can be pre-computed. This indicates that evaluating (13) involves less operations than evaluating (3).

Similarly, using PCA, the hyper-rectangle (2) can be re-expressed as:

$$-\mathbf{1} \leq \mathbf{H_q}\Delta\mathbf{a}_m \leq \mathbf{1}. \tag{14}$$

A schematic representation of the Reduced-Order B2B DC procedure is reported in Fig. 1: dimension reduction is carried out on a set of observations of the model's output and later combined with an interpolation technique. This leads to the construction of a SM that can be used for parameter exploration and consistency analysis.

3. Application and results

The Alstom Boiler Simulation Facility (BSF) is a 15 $\mathrm{MW_{th}}$ capacity, tangential-fired pilot facility, located at Alstom Windsor, CT. The BSF is an atmospheric pressure, balanced draft combustion test facility designed to replicate the time-temperature stoichiometry history of typical utility boilers [12]. Details about the CFD model used to simulate the BSF can be found in [11,22]. The associated computational cost is roughly 740.000 CPU hours per simulation. Figure 2 shows wall Heat Flux profiles from CFD simulations of the BSF.

An UQ analysis was carried out to identify the parameters which have the highest impact on the predictions of the QOIs. The impact was defined as the product *uncertainty* × *sensitivity*. This study included mesh resolution, the CFL number, spatial

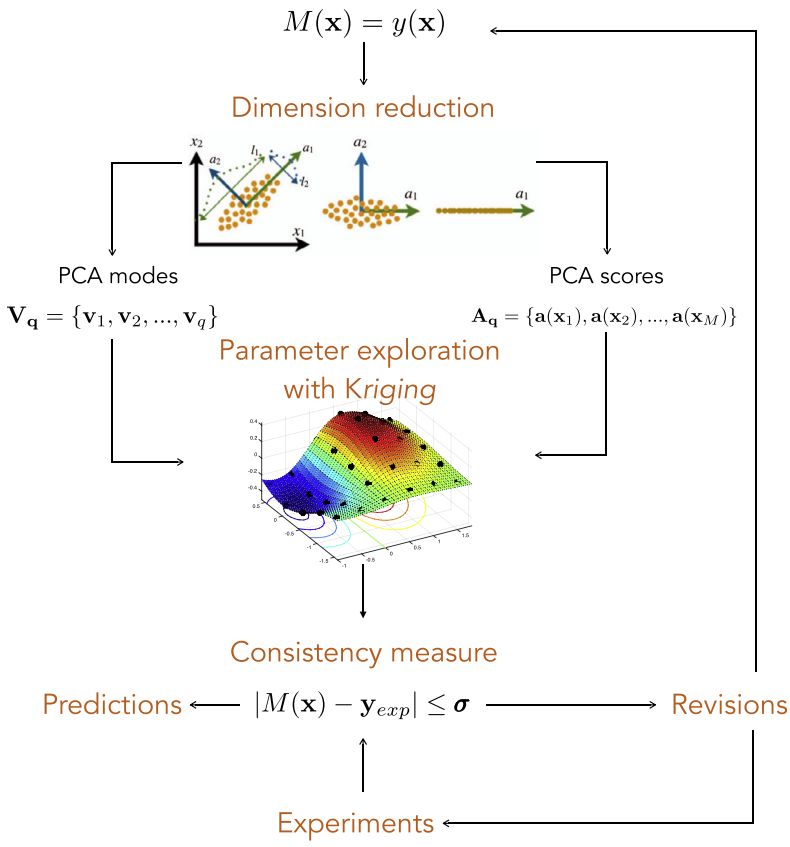

Fig. 1. Flowchart for the Reduced-Order Bound-to-Bound Data Collaboration procedure. Dimension reduction is employed on a set of available observations of the model's output and combined with an interpolation technique. A surrogate model is built for parameter exploration and to find consistency with reference data.

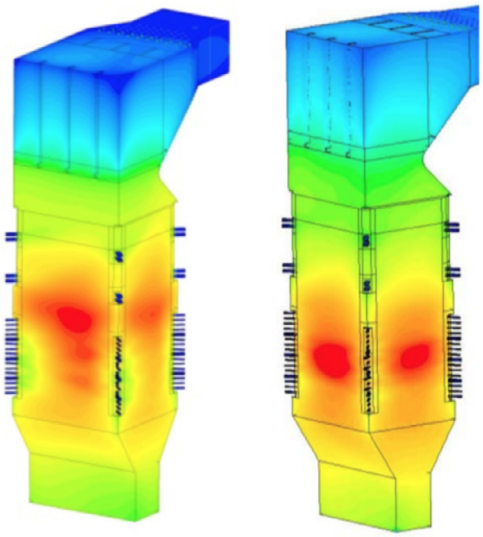

Fig. 2. CFD simulations for Alstom's BSF Heat Flux profiles. Figure from [12].

and temporal schemes, devolatilization parameters such as the swelling factor, char oxidation parameters such as activation energies and pre-exponential factors for O_2, H_2O and CO_2, wall thermal conductivity, scenario parameters such as particle size distribution and particle density. Three parameters are identified for the consistency analysis, namely T_{slag}, k and τ. The ranges associated to these parameters are [1350, 1600] K, [2.5, 4.5] $W/(m \cdot K)$ and [1, 2.5], respectively. In particular, T_{slag} represents the temperature at which the deposits on the wall starts changing phase, from solid to plastic (liquid). The parameter k represents the effective thermal conductivity on the wall [23]. The model's parameter τ represents a constant that scales the activation energies of CO_2, O_2, and H_2O simultaneously from their base values. A value for these parameters needs to be found so that the CFD model's predictions are consistent with the experimental values available for the BSF.

Measurements probes are present in the BSF at specific locations. A number of 22 simulations are run for different values of the described pa-

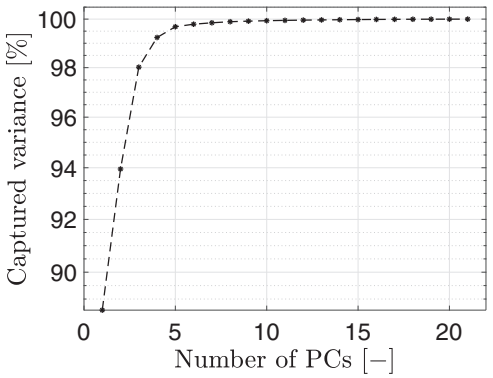

Fig. 3. The cumulative original data variance recovered when using an incresing number of principal components provides a criterion for the selection of the number of latent variables to be retained.

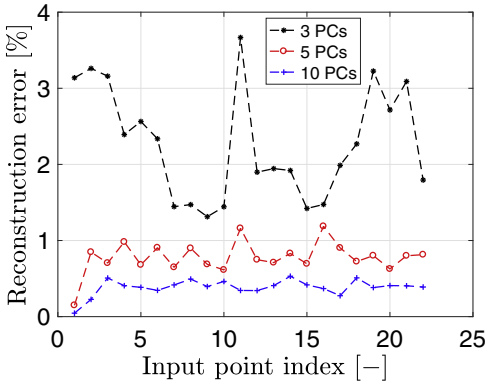

Fig. 4. Error for the reconstruction of the original data by PCA when using 3 PCs (*star*), 5 PCs (*circle*) and 10 PCs (*cross*). This error is zero if all the PCs are kept.

rameters, identified by means of Latin Hypercube Sampling. The parameter region or hypercube $\mathcal{H} = [1350, 1600]\ K \times [2.5, 4.5]\ Wl(m \cdot K) \times [1, 2.5]$ is explored by using the 22 observations as training samples for a SM, and a consistency region is sought. Because there are 121 QOIs that may be correlated, namely 95 Temperature and 26 Heat Flux measurements inside the BSF, PCA is performed in order to identify a set of PCA scores.

Figure 3 shows the cumulative variance of the original data that is recovered for each number of retained PCs. Figure 4 reports the mean relative reconstruction errors of the original data, for each training observations. These errors are reported for a number of 3, 5 and 10 PCs. They are below a value of 1% when 5 or more PCs are used for the compression. These results suggest that the original 121 variable are indeed correlated. A number of 5 PCs is enough to recover over 99% of the total variance. The original data-set $\mathbf{Y}$ of size (121×22) can be compressed into the matrix of PCA scores $\mathbf{A}$ of size (5×22), if 5 PCs are kept. A Kriging model is trained for each of the PCs on the 22 available observations. Once the Kriging models are trained, a consistency analysis is carried out as discussed in Sections 2.1 and 2.4, with the bounds σ_i being the experimental measurement uncertainties. Figure 5 reports the two consistency regions found by the two methodologies using the condition (2), with 5 PCs. A consistency analysis using the constraint (3) on 5 scalars, namely the PCA scores, is able to find the same consistency region of a full consistency analysis carried out on 121 variables. The PCA+Kriging model suggests consistency also for lower values of τ and T_{slag}. Using Eq. (3), the two methodologies provide consistency (Fig. 6) in two regions that differ by 24% in volume. These volumes are computed by means of alpha-shapes [24]. It is worth noting that the input parameters are standardized when building SMs. If this space is standardized, the volume difference between the aforementioned regions is within 5%. The difference between the two consistency regions can be explained by the fact that the manifold found by PCA on the 22 observations recovers 99% of

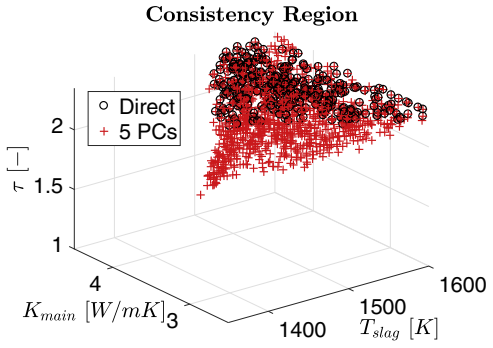

Fig. 5. Consistency region found by direct Kriging (*circles*) and Kriging on the first 5 PCs (*crosses*). Consistency region found by using constraint (2).

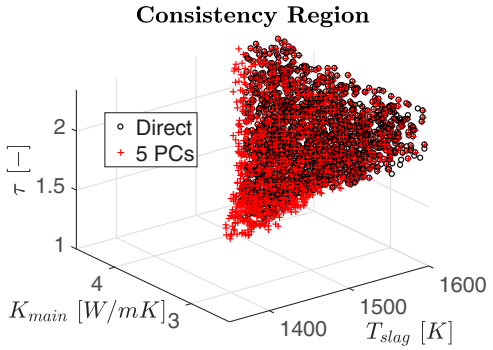

Fig. 6. Consistency regions as in Fig. 5 but using the constraint (3) with $\alpha = 15$.

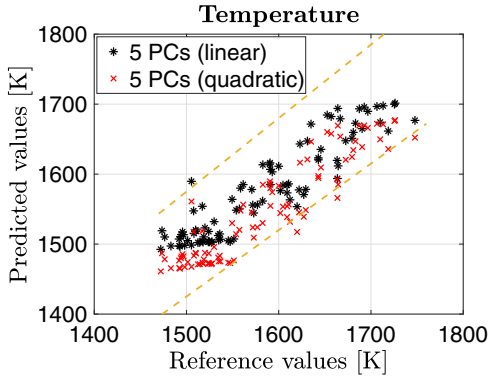

Fig. 7. Parity plot for Temperature when using 5 PCs. Comparison between the reference experimental data and one consistent model's prediction according to the constraint (2) (*linear*) and (3) (*quadratic*). The dashed lines represent the 5% error.

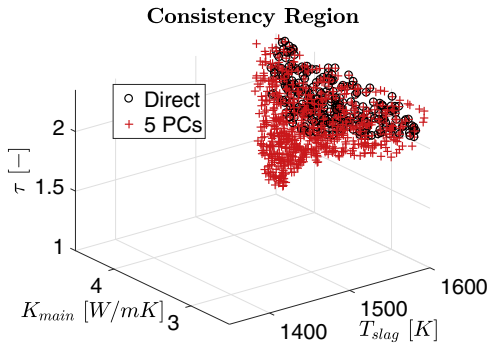

Fig. 8. Consistency regions as in Fig. 5 but using the constraint (3) with $\alpha = 11$.

the data variance, but if more observations were present, that same manifold might not recover as much. The predictions from the SMs trained on the 5 PCA scores are forced to stay on the PCA manifold, while the predictions from SMs trained on the original variables can lie outside of

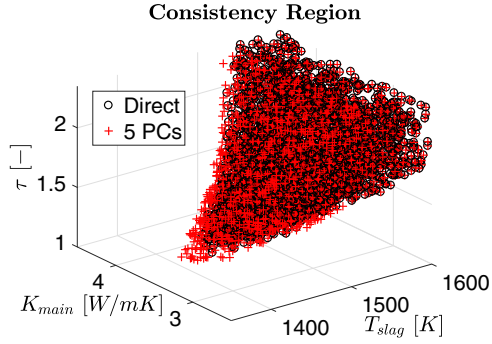

Fig. 9. Consistency regions as in Fig. 5 but using the constraint (3) with $\alpha = 20$.

it. Figure 7 shows a parity plot of the reference experimental data and the predictions from one SM, using the values of the input parameters belonging to the consistency regions shown in Figs 5 and 6. The difference between the predicted and reference temperature data are generally within 5%, confirming that the calibration of the input parameters can improve the model's predictive capabilities. Figures 8 and 9 show how the consistency region found using the constraint (3) changes when changing the value of α. For $\alpha = 11$, a consistency region can still be found. The choice for the value of α depends on how strict a consistency between the model's predictions and the experimental data is sought. From the perspective of the proposed Reduced-Order B2B-DC methodology, in comparison with a classic B2B-DC, the focus is on the fact that the two consistency regions (direct and using a reduced number of PCs) both shrink or grow larger together. In conclusion, there is no need to train 121 SMs, because 5 PCs are enough to perform an accurate consistency analysis. This ensures computational savings and preservation of correlations among variables. In the case of larger data-sets (comparable number of output variables, more than 8 input parameters and more than 10^4 observations), where the training process might cost tens or hundreds of CPU hours, computational savings would be even more relevant. Table 1 reports the

Table 1
Comparison between the computational performances and reconstruction errors of 4 different Kriging models.

	Kriging	10 PCs	5 PCs	3 PCs
Training time	34.5 s	2.78 s	1.76 s	0.93 s
Speed-up	1	12.4	19.6	37.0
Reconstruction Error	–	0.5%	1%	3%
# Hyper-parameters				
Linear trend	484	40	20	12
Gaussian kernel	363	30	15	9
# Coefficients	–	1452	847	605

computational performances of the two methodologies, in the form of reconstruction errors, training times, number of hyper-parameters to train and number of coefficients to store in memory. While no reconstruction error is present for direct Kriging, very small (below 3%) errors are introduced by the PCA compression. On the other hand, it is clear that direct Kriging has more hyper-parameters to train. Although Kriging on 5 PCs has more coefficients to store in memory, namely 2 vectors (121 mean values and 121 scaling factors) and 5 PCA modes of 121 coefficients each, the computation of these coefficients is straightforward compared to the solution of the optimization problems that lead to the estimation of the hyper-parameters.

4. Conclusions

In this work the B2B-DC framework is combined with PCA. Experimental data are available for Temperature and Heat Flux measurements for the Alstom BSF test facility. A CFD model of the BSF is also available [11,22] but not fully defined, as the values of 3 input parameters are uncertain. These parameters are indicated as T_{slag}, k and τ. The model's output is consistent with the experimental data only if suitable values for these parameters are chosen. The latter can be found using the B2B-DC approach, carrying out consistency analysis between the experimental data and the model's output. The available data consist of experimental values for 121 QOIs, namely 95 Temperature and 26 Heat Flux values. A set of 22 full-order CFD simulations was carried out, each time with a different triplet of values for the 3 model parameters. In the classic B2B-DC approach, a consistency analysis is performed with a set of SMs built from these simulations for each of these QOIs, for a total of 121.

In the present work, a consistency analysis is carried out using only 5 trained SMs. This is possible if a reduction technique such as PCA is used to compress the original data. The set of 121 original QOIs is encoded into a set of 5 scalars, namely the PCA scores, and thus only 5 SMs are needed for the consistency analysis. This approach is referred to as Reduced-Order B2B-DC. This is the first time, to the authors knowledge, that a B2B-DC is developed in terms of latent variables rather than original physical variables. Results obtained from the Reduced-Order B2B-DC approach are compared with the standard B2B-DC approach. The results show that the consistency region identified using the Reduced-Order B2B-DC approach is very similar to the one identified using the B2B-DC approach, the difference between the consistency volumes being below 5%, if the input space is standardized. The advantages of the approach include computational savings since less SMs need to be trained: less hyper-parameters need to be found for the construction of the SMs, which is very often not a simple task, especially when the number of input parameters is high. Finally, the CFD model's predictive capabilities for the BSF are improved by defining suitable ranges for the 3 most influential parameters affecting the predictions.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement no 643134 and was also sponsored by the European Research Council, Starting Grant no 714605. This material is based upon work supported by the Department of Energy, National Nuclear Security Administration, under Award no DE-NA0002375.

References

[1] D.R. Yeates, W. Li, P.R. Westmoreland, et al., *Proc. Combust. Inst.* 35 (1) (2015) 597–605, doi:10.1016/j.proci.2014.05.090.
[2] G. Lin, *Spat. Stat.* (2017), doi:10.1016/j.spasta.2017.08.002.
[3] B.A. Khuwaileh, P.J. Turinsky, *Nucl. Eng. Technol.* 49 (6) (2017) 1219–1225, doi:10.1016/j.net.2017.08.007.
[4] J. Jatnieks, M. De Lucia, D. Dransch, M. Sips, *Energy Procedia* 97 (2016) 447–453, doi:10.1016/j.egypro.2016.10.047.
[5] R.E. Edwards, J. New, L.E. Parker, B. Cui, J. Dong, *Appl. Energy* 202 (2017) 685–699, doi:10.1016/j.apenergy.2017.05.155.
[6] N.E. Owen, P. Challenor, P.P. Menon, S. Bennani, *Journal on Uncertainty Quantification* 5 (1) (2017) 403–435, doi:10.1137/15M1046812.
[7] M. De Lozzo, A. Marrel, *Stoch. Environ. Res. Risk Assess.* 31 (6) (2017) 1437–1453, doi:10.1007/s00477-016-1245-3.
[8] S. Dubreuil, M. Berveiller, F. Petitjean, M. Salaün, *Reliab. Eng. Syst. Saf.* 121 (2014) 263–275, doi:10.1016/j.ress.2013.09.011.
[9] A. Marrel, B. Iooss, B. Laurent, O. Roustant, *Reliab. Eng. Syst. Saf.* 94 (3) (2009) 742–751, doi:10.1016/j.ress.2008.07.008.
[10] I.T. Jolliffe, *Principal Component Analysis*, second ed., Springer-Verlag New York, Inc., 2002.
[11] J. Pedel, J.N. Thornock, S.T. Smith, P.J. Smith, *Int. J. Multiph. Flow* 63 (2014) 23–38, doi:10.1016/j.ijmultiphaseflow.2014.03.002.
[12] C. Edberg, A. Levasseur, H. Andrus, J. Kenney, D. Turek, S. Kang, *Pilot Scale Facility Contributions to Alstom's Technology Development Efforts for Oxy–Combustion for Steam Power Plants*, AFRC Industrial Combustion Symposium Kauai, Hawaii, 2013 September 22–25.
[13] R. Feeley, P. Seiler, A. Packard, M. Frenklach, *J. Phys. Chem. A* 108 (44) (2004) 9573–9583, doi:10.1021/jp047524w.

[14] M. Frenklach, A. Packard, P. Seiler, Prediction uncertainty from models and data, American Control Conference 5 (2002) 4135–4140. doi:10.1109/ACC.2002.1024578.
[15] M. Frenklach, A. Packard, P. Seiler, R. Feeley, *Int. J. Chem. Kinet.* 36 (1) (2004) 57–66, doi:10.1002/kin.10172.
[16] A. Parente, J.C. Sutherland, *Combust. Flame* 160 (2) (2013) 340–350, doi:10.1016/j.combustflame.2012.09.016.
[17] A. Coussement, B.J. Isaac, O. Gicquel, A. Parente, *Combust. Flame* 168 (2016) 83–97, doi:10.1016/j.combustflame.2016.03.021.
[18] K. Bizon, G. Continillo, E. Mancaruso, S.S. Merola, B.M. Vaglieco, *Combust. Flame* 157 (4) (2010) 632–640, doi:10.1016/j.combustflame.2009.12.013.
[19] P.G. Constantine, E. Dow, Q. Wang, *SIAM J. Sci. Comput.* 36 (4) (2014) 1500–1524.
[20] S.N. Lophaven, J. Søndergaard, H.B. Nielsen, Technical University of Denmark, Technical Report IMM-TR-2002-12, Informatics and Mathematical Modelling, DTU (2002) 1–28.
[21] M. Seeger, *Gaussian Processes for Machine Learning*, in: C.E. Rasmussen, C.K.I. Williams (Eds.), 14, the MIT Press, 2006 ISBN 026218253X. c 2006 Massachusetts Institute of Technology.
[22] J. Pedel, J.N. Thornock, P.J. Smith, *Combust. Flame* 160 (6) (2013) 1112–1128, doi:10.1016/j.combustflame.2013.01.022.
[23] G.R. Hadley, *Int. J. Heat Mass Transf.* 29 (6) (1986) 909–920, doi:10.1016/0017-9310(86)90186-9.
[24] N. Akkiraju, H. Edelsbrunner, M. Facello, P. Fu, E.P. Mucke, C. Varela, in: International Computer Geometry Software Workshop, 1995, pp. 1–8.