Development of physics-based reduced-order models for reacting flow applications

Gianmarco Aversano

université PARIS-SACLAY

CentraleSupélec

ULB UNIVERSITÉ LIBRE DE BRUXELLES

NNT : 2019SACLC095

Thèse de doctorat

Development of physics-based reduced-order models for reacting flow applications

Thèse de doctorat de l'Université Paris-Saclay et de l'Université Libre de Bruxelles préparée à CentraleSupélec

École doctorale n°579 Sciences mécaniques et énergétiques, matériaux, géosciences (SMEMAG)
Spécialité de doctorat : Combustion

Thèse présentée et soutenue à Bruxelles, le 15/11/2019, par

GIANMARCO AVERSANO

Composition du Jury :

Francesco Contino Prof., Vrije Universiteit Brussel	Président
Ronan Vicquelin Prof., CentraleSupélec (EM2C)	Rapporteur
Alessandro Parente Prof., Université libre de Bruxelles (ATM)	Directeur de thèse
Olivier Gicquel Prof., CentraleSupélec (EM2C)	Co-directeur de thèse
Sean Smith Prof., University of Utah	Invité
Caroline Sainvitu Dr., Cenaero	Invité

université PARIS-SACLAY
ÉCOLE DOCTORALE
Sciences mécaniques et énergétiques, matériaux et géosciences (SMEMAG)

Titre : Développement de modèles d'ordre réduit basés sur la physique pour les applications d'écoulement réactif

Mots clés : Combustion, Unsupervised learning, supervised learning

Résumé : L'objectif final étant de développer des modèles d'ordre réduit pour les applications de combustion, des techniques d'apprentissage automatique non supervisées et supervisées ont été testées et combinées dans les travaux de la présente thèse pour l'extraction de caractéristiques et la construction de modèles d'ordre réduit. Ainsi, l'application de techniques pilotées par les données pour la détection des caractéristiques d'ensembles de données de combustion turbulente (simulation numérique directe) a été étudiée sur deux flammes H2 / CO: une évolution spatiale (DNS1) et une jet à évolution temporelle (DNS2). Des méthodes telles que l'analyse en composantes principales (ACP), l'analyse en composantes principales locales (LPCA), la factorisation matricielle non négative (NMF) et les autoencodeurs ont été explorées à cette fin. Il a été démontré que divers facteurs pouvaient affecter les performances de ces méthodes, tels que les critères utilisés pour le centrage et la mise à l'échelle des données d'origine ou le choix du nombre de dimensions dans les approximations de rang inférieur. Un ensemble de lignes directrices a été présenté qui peut aider le processus d'identification de caractéristiques physiques significatives à partir de données de flux réactifs turbulents. Des méthodes de compression de données telles que l'analyse en composantes principales (ACP) et les variations ont été combinées à des méthodes d'interpolation telles que le krigeage, pour la construction de modèles ordonnées à prix réduits et calculables pour la prédiction de l'état d'un système de combustion dans des conditions de fonctionnement inconnues ou des combinaisons de modèles valeurs de paramètre d'entrée. La méthodologie a d'abord été testée pour la prévision des flammes 1D avec un nombre croissant de paramètres d'entrée (rapport d'équivalence, composition du carburant et température d'entrée), avec des variantes de l'approche PCA classique, à savoir PCA contrainte et PCA locale, appliquée aux cas de combustion la première fois en combinaison avec une technique d'interpolation. Les résultats positifs de l'étude ont conduit à l'application de la méthodologie proposée aux flammes 2D avec deux paramètres d'entrée, à savoir la composition du combustible et la vitesse d'entrée, qui ont donné des résultats satisfaisants. Des alternatives aux méthodes non supervisées et supervisées choisies ont également été testées sur les mêmes données 2D. L'utilisation de la factorisation matricielle non négative (FNM) pour l'approximation de bas rang a été étudiée en raison de la capacité de la méthode à représenter des données à valeur positive, ce qui permet de ne pas enfreindre des lois physiques importantes telles que la positivité des fractions de masse d'espèces chimiques et comparée à la PCA. Comme méthodes supervisées alternatives, la combinaison de l'expansion du chaos polynomial (PCE) et du Kriging et l'utilisation de réseaux de neurones artificiels (RNA) ont été testées. Les résultats des travaux susmentionnés ont ouvert la voie au développement d'un jumeau numérique d'un four à combustion à partir d'un ensemble de simulations 3D. La combinaison de PCA et de Kriging a également été utilisée dans le contexte de la quantification de l'incertitude (UQ), en particulier dans le cadre de collaboration de données lié (B2B-DC), qui a conduit à l'introduction de la procédure B2B-DC à commande réduite. comme pour la première fois, le centre de distribution B2B a été développé en termes de variables latentes et non en termes de variables physiques originales.

Title : Development of physics-based reduced-order models for reacting flow applications

Keywords : Combustion, Unsupervised learning, supervised learning

Abstract : With the final objective being to develop reduced-order models for combustion applications, unsupervised and supervised machine learning techniques were tested and combined in the work of the present Thesis for feature extraction and the construction of reduced-order models. Thus, the application of data-driven techniques for the detection of features from turbulent combustion data sets (direct numerical simulation) was investigated on two H2/CO flames: a spatially-evolving (DNS1) and a temporally-evolving jet (DNS2). Methods such as Principal Component Analysis (PCA), Local Principal Component Analysis (LPCA), Non-negative Matrix Factorization (NMF) and Autoencoders were explored for this purpose. It was shown that various factors could affect the performance of these methods, such as the criteria employed for the centering and the scaling of the original data or the choice of the number of dimensions in the low-rank approximations. A set of guidelines was presented that can aid the process of identifying meaningful physical features from turbulent reactive flows data. Data compression methods such as Principal Component Analysis (PCA) and variations were combined with interpolation methods such as Kriging, for the construction of computationally affordable reduced-order models for the prediction of the state of a combustion system for unseen operating conditions or combinations of model input parameter values. The methodology was first tested for the prediction of 1D flames with an increasing number of input parameters (equivalence ratio, fuel composition and inlet temperature), with variations of the classic PCA approach, namely constrained PCA and local PCA, being applied to combustion cases for the first time in combination with an interpolation technique. The positive outcome of the study led to the application of the proposed methodology to 2D flames with two input parameters, namely fuel composition and inlet velocity, which produced satisfactory results. Alternatives to the chosen unsupervised and supervised methods were also tested on the same 2D data. The use of non-negative matrix factorization (NMF) for low-rank approximation was investigated because of the ability of the method to represent positive-valued data, which helps the non-violation of important physical laws such as positivity of chemical species mass fractions, and compared to PCA. As alternative supervised methods, the combination of polynomial chaos expansion (PCE) and Kriging and the use of artificial neural networks (ANNs) were tested. Results from the mentioned work paved the way for the development of a digital twin of a combustion furnace from a set of 3D simulations. The combination of PCA and Kriging was also employed in the context of uncertainty quantification (UQ), specifically in the bound-to-bound data collaboration framework (B2B-DC), which led to the introduction of the reduced-order B2B-DC procedure as for the first time the B2B-DC was developed in terms of latent variables and not in terms of original physical variables.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France

DECLARATION

This thesis is submitted to the Université Libre de Bruxelles (ULB) for the degree of philosophy doctor. This doctoral work has been performed at the Université Libre de Bruxelles, École polytechnique de Bruxelles, Aero-Thermo-Mechanics Laboratory, Bruxelles, Belgium with Professor Alessandro Parente and at the Laboratoire EM2C, CNRS, Centrale-Supélec, Université ParisSaclay, 8-10 rue Joliot-Curie 91190 Gif-sur-Yvette, France with Professor Olivier Gicquel.

Brussels, September 2019

Gianmarco Aversano

To my loved ones.

ABSTRACT

Modern society will have to meet its energy demands while ensuring low or virtually zero emissions in order to meet future challenges associated to air pollution, climate change and energy storage. Very often, renewable sources cannot be directly employed because of their intermittent nature and because many applications such as transport and other industrial processes require high energy densities. Therefore, novel storage solutions for the energy that renewable sources contribute to produce is necessary and the transformation of this energy into chemical compounds represents the best choice in order to meet the aforementioned demands, which requires novel combustion technologies, such as Moderate and Intense Low-oxygen Dilution (MILD) combustion, to be efficient and fuel-flexible. In order to develop such technologies, several studies are being proposed and terabytes of data collected as more and more experiments and high-fidelity simulations are carried out. However, there are two main challenges to this: the huge amount of data available makes it hard for the researcher to distinguish useful from redundant data, with the risk that useful information might stay hidden; the production process of these data-sets requires substantial resources as combustion process are multi-physics, multi-scale and thus require high-fidelity computationally-intensive simulations and experiments over a wide range for their operating conditions or input parameters. Digital twins and Artificial Intelligence (AI) are shaping the fourth industrial revolution by building data-driven models that make use of machine learning. It makes sense then to extend this approach to combustion applications in order to alleviate the two aforementioned issues: the use of machine learning techniques can help automate the process of data interpretation as well as provide a low-dimensional representation of the high-dimensional data produced by either experiments or simulations; they can speed up the data production process by building reduced-order models that can foresee the outcome of a certain simulation with reduced or negligible computational cost. Besides, such reduced-order models are the foundations for the development of virtual counterparts of real physical systems, which can be employed for system control, non-destructive testing and visualization.

With the final objective being to develop reduced-order models for combustion applications, unsupervised and supervised machine learning techniques were tested and combined in the work of the present Thesis for feature extraction and the construction of reduced-order models. Thus, the application of data-driven techniques for the detection of features from turbulent combustion data sets (direct numerical simulation) was investigated on two H_2/CO flames: a spatially-evolving (DNS1) and a temporally-evolving jet (DNS2). Methods such as Principal Component Analysis (PCA), Local Principal Component Analysis (LPCA), Non-negative Matrix Factorization (NMF) and Autoencoders were explored for this purpose. It was shown that various factors could affect the performance of these methods, such as the criteria employed for the centering and the scaling of the

original data or the choice of the number of dimensions in the low-rank approximations. A set of guidelines was presented that can aid the process of identifying meaningful physical features from turbulent reactive flows data. Data compression methods such as Principal Component Analysis (PCA) and variations were combined with interpolation methods such as Kriging, for the construction of computationally affordable reduced-order models for the prediction of the state of a combustion system for unseen operating conditions or combinations of model input parameter values. The methodology was first tested for the prediction of 1D flames with an increasing number of input parameters (equivalence ratio, fuel composition and inlet temperature), with variations of the classic PCA approach, namely constrained PCA and local PCA, being applied to combustion cases for the first time in combination with an interpolation technique. The positive outcome of the study led to the application of the proposed methodology to 2D flames with two input parameters, namely fuel composition and inlet velocity, which produced satisfactory results. Alternatives to the chosen unsupervised and supervised methods were also tested on the same 2D data. The use of non-negative matrix factorization (NMF) for low-rank approximation was investigated because of the ability of the method to represent positive-valued data, which helps the non-violation of important physical laws such as positivity of chemical species mass fractions, and compared to PCA. As alternative supervised methods, the combination of polynomial chaos expansion (PCE) and Kriging and the use of artificial neural networks (ANNs) were tested. Results from the mentioned work paved the way for the development of a digital twin of a combustion furnace from a set of 3D simulations. The combination of PCA and Kriging was also employed in the context of uncertainty quantification (UQ), specifically in the bound-to-bound data collaboration framework (B2B-DC), which led to the introduction of the reduced-order B2B-DC procedure as for the first time the B2B-DC was developed in terms of latent variables and not in terms of original physical variables.

ABSTRACT (FRENCH)

L'objectif final étant de développer des modèles d'ordre réduit pour les applications de combustion, des techniques d'apprentissage automatique non supervisées et supervisées ont été testées et combinées dans les travaux de la présente thèse pour l'extraction de caractéristiques et la construction de modèles d'ordre réduit. Ainsi, l'application de techniques pilotées par les données pour la détection des caractéristiques d'ensembles de données de combustion turbulente (simulation numérique directe) a été étudiée sur deux flammes H2 / CO : une évolution spatiale (DNS1) et une jet à évolution temporelle (DNS2). Des méthodes telles que l'analyse en composantes principales (ACP), l'analyse en composantes principales locales (LPCA), la factorisation matricielle non négative (NMF) et les autoencodeurs ont été explorées à cette fin. Il a été démontré que divers facteurs pouvaient affecter les performances de ces méthodes, tels que les critères utilisés pour le centrage et la mise à l'échelle des données d'origine ou le choix du nombre de dimensions dans les approximations de rang inférieur. Un ensemble de lignes directrices a été présenté qui peut aider le processus d'identification de caractéristiques physiques significatives à partir de données de flux réactifs turbulents. Des méthodes de compression de données telles que l'analyse en composantes principales (ACP) et les variations ont été combinées à des méthodes d'interpolation telles que le krigeage, pour la construction de modèles ordonnées à prix réduits et calculables pour la prédiction de l'état d'un système de combustion dans des conditions de fonctionnement inconnues ou des combinaisons de modèles valeurs de paramètre d'entrée. La méthodologie a d'abord été testée pour la prévision des flammes 1D avec un nombre croissant de paramètres d'entrée (rapport d'équivalence, composition du carburant et température d'entrée), avec des variantes de l'approche PCA classique, à savoir PCA contrainte et PCA locale, appliquée aux cas de combustion la première fois en combinaison avec une technique d'interpolation. Les résultats positifs de l'étude ont conduit à l'application de la méthodologie proposée aux flammes 2D avec deux paramètres d'entrée, à savoir la composition du combustible et la vitesse d'entrée, qui ont donné des résultats satisfaisants. Des alternatives aux méthodes non supervisées et supervisées choisies ont également été testées sur les mêmes données 2D. L'utilisation de la factorisation matricielle non négative (FNM) pour l'approximation de bas rang a été étudiée en raison de la capacité de la méthode à représenter des données à valeur positive, ce qui permet de ne pas enfreindre des lois physiques importantes telles que la positivité des fractions de masse d'espèces chimiques et comparée à la PCA. Comme méthodes supervisées alternatives, la combinaison de l'expansion du chaos polynomial (PCE) et du Kriging et l'utilisation de réseaux de neurones artificiels (RNA) ont été testées. Les résultats des travaux susmentionnés ont ouvert la voie au développement d'un jumeau numérique d'un four à combustion à partir d'un ensemble de simulations 3D. La combinaison de PCA et de Kriging a également été utilisée dans le contexte de la quan-

tification de l'incertitude (UQ), en particulier dans le cadre de collaboration de données lié (B2B-DC), qui a conduit à l'introduction de la procédure B2B-DC à commande réduite. comme pour la première fois, le centre de distribution B2B a été développé en termes de variables latentes et non en termes de variables physiques originales.

PUBLICATIONS

— Gianmarco Aversano, Aurélie Bellemans, Zhiyi Li, Axel Coussement, Olivier Gicquel and Alessandro Parente, *Application of reduced-order models based on PCA & Kriging for the development of digital twins of reacting flow applications*, Computers and Chemical Engineering 121 (2019), Pages 422-441.

— Gianmarco Aversano, John Camilo Parra-Alvarezd, Benjamin J.Isaa, Sean T.Smith, Axel Coussement, Olivier Gicquel and Alessandro Parente, *PCA and Kriging for the efficient exploration of consistency regions in Uncertainty Quantification*, Proceedings of the Combustion Institute 37 (2019), Pages 4461-4469.

— Simone Giorgetti, Coppitters Diederik, Ward De Paepe, Laurent Bricteux, Francesco Contino, Gianmarco Aversano and Alessandro Parente, *Surrogate-assisted modeling and Robust Optimization of a micro Gas Turbine plant with Carbon Capture*, Conference Proceedings of the ASME Turbo EXPO 2019 (2019).

— Aurélie Bellemans, Gianmarco Aversano, Axel Coussement and Alessandro Parente, *Feature extraction and reduced-order modelling of nitrogen plasma models using principal component analysis*, Computers and Chemical Engineering 115 (2018), Pages 504-514.

— Magnus Fürst, Pino Sabia, Marco Lubrano Lavadera, Gianmarco Aversano, Mara de Joannon, Alessio Frassoldati and Alessandro Parente, *Optimization of Chemical Kinetics for Methane and Biomass Pyrolysis Products in Moderate or Intense Low-Oxygen Dilution Combustion*, Energy and Fuels 32 (2018), Pages 10194-10201.

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my thanks to those who helped me with various aspects of conducting research and the writing of this thesis.

First and foremost, Prof. Alessandro Parente for his guidance, patience and support throughout this research and the writing of this thesis.

Prof. Olivier Giquel and Prof. Ronan Vicquelin for their supervision during my secondment period at CentraleSupelec.

Prof. Sean Smith and his team for their precious teachings and enjoyable talks during my stay at the University of Utah.

My colleagues for their help during the various stages of my research.

The CLEAN-Gas project for giving me the opportunity to carry out this research.

CONTENTS

LIST OF FIGURES

1 INTRODUCTION

The present study is part of Marie Sklodowska-Curie CLEAN-Gas "European Joint Doctorate" program. It is funded by the European Community through the Horizon 2020 Actions. The acronym CLEAN-Gas is the abbreviation for "Combustion for Low Emission Applications of Natural Gas". The study was also sponsored by the European Research Council, Starting Grant No 714605.

1.1 GENERAL BACKGROUND AND MOTIVATION

Development of PCA-based reduced models for natural gas combustion

One of the objective of the modern society is to ensure a healthy environment for future generations, which is considered possible only if more and more affordable and sustainable energy is employed. However, the intermittent nature of renewable sources requires the development of novel storage solutions that can guarantee the availability of the required energy supply when renewable sources are not available. Because many applications such as air and ground transportation require high energy density and as a consequence cannot rely on the direct use of renewable energy, it is energy storage in the form of chemical compounds that will lead to a true integration between renewable sources and existing infrastructure for energy conversion, such as combustion systems. Energy density is the key feature that makes the use of fuels inescapable for energy demanding applications, such as transportation and other industrial processes. This means that the transformation of excess renewable energy into energy carriers is an appealing solution for energy storage. Fuel flexibility also poses technical challenges and indicates the need for advanced combustion technologies. Such technologies have to be fuel flexible, highly efficient and non-polluting, operating in conditions which substantially differ from those characterizing traditional combustion systems.

Moderate and Intense Low-oxygen Dilution (MILD) [25] combustion represents a very attractive solution for its fuel flexibility and capability to deliver very high combustion efficiency with virtually zero pollutant emissions. MILD combustion ensures large fuel flexibility, representing an ideal technology for low-calorific value fuels, high-calorific industrial wastes and hydrogen-based fuels, as well as for liquid and solid fuels. These new technologies result from a change in perspective in the analysis of combustion problems, from purely energetic to chemical aspects. The combustion process becomes a chemical reactor, which must be optimized from the perspective of flexibility, conversion efficiency and emissions. This requires a deep understanding of the kinetic aspects related to combustion reactions, interactions between chemical kinetics and turbulence, and heat exchange phenomena, in particular radiation, meaning that both high-fidelity simulations and experimental techniques have to be used in a unified framework, to optimize the operation

of existing systems and develop new combustion systems. Besides, the experiments and simulations needed for the production of informative data with the objective of acquiring usable knowledge for the development of new combustion technologies come with associated costs, which limit their production. Thus, the need for automated algorithms that can help data interpretation and speed up the process of data production becomes clear. The development of virtual models, also referred to as digital twins, of industrial systems opens up a number of opportunities, such as the use of data to anticipate the response of a system and brainstorm malfunctioning, and the use of simulations to develop new technologies, i.e. virtual prototyping. A definition of digital twin is "An integrated multi-physics, multi-scale, probabilistic simulation of an as-built system, enabled by digital thread, that uses the best available models, sensor information, and input data to mirror and predict activities/performance over the life of its corresponding physical twin" [14]. Digital twins are a disruptive technology that creates a living model of a physical system that can also be used for predictive maintenance. The digital twin will continually adapt to changes in the environment or operation using real-time sensory data and can forecast the future of the corresponding physical system and act as soft sensor [28]. Importantly, digital twins can also be used for non-destructive testing, which can undoubtedly benefit industrial protagonists. For the above reasons, the need for digital twins is becoming imperative.

Digital twins, Internet of Things (IoT), and Artificial Intelligence (AI) are shaping the fourth industrial revolution [22]. A digital twin comprises and combines sensor and measurement technologies, simulation and modeling, and machine learning. Data and information fusion is often the key to propel a digital twin as it facilitates the flow of information from raw sensory data to high-level understanding and insights. Digital twins generally incorporate different kind of fusions: fusion of sensory data coming from different sources can be used to achiever better signal quality; sensor and physics-based model fusion is used to build an adaptive physics-based model, where the real-time sensory feedback is usually incorporated in the physics model via Kalman filters [41]; sensor and data model fusion is employed to build robust data-driven models which make use of machine learning. The challenges to the construction of digital twins of real systems are many. They need to be reliable, i.e. able to predict a certain system's state for a wide range of operating conditions, easy to update in case new data become available, and they have to provide a fast response in order to be beneficial for real-time use. The data needed for their development can come from either experiments or simulations, or both. Clearly, the quality of these data-sets is crucial for the development of data-driven digital twins.

In engineering applications, the ability to make reliable predictions about complex physical systems is granted by the existence of predictive mathematical models that are based on the deep understanding of the underlying processes. These mathematical models, together with experiments, are the main source of data production and usually come in the form of high-fidelity expensive simulations. Combustion systems are characterized by very complex physical interactions, between chemistry, fluid-dynamics and heat transfer processes, for which expensive simulations are needed. The input parameters for these

simulations can either be actual operating conditions or model parameters. The outcome of these simulations will be different each time the values of these parameters are changed, sometimes drastically so due to the non-linearity of combustion-related phenomena. In order to acquire knowledge on how the model responds to different inputs, the expensive model needs to be run several times with different inputs. There are two scenarios where this is wanted. The first scenario is when the model is known to be predictive of a certain complex system, but either a specific operating condition is sought that maximizes some objective function or the prediction of the system's state is needed in real-time, based on a change in some operating parameters whose values are tracked by sensors. The second scenario is when a specific combination of the model parameters is sought so that the model's prediction for a set of Quantities of Interest (QOIs) is consistent with reference data, usually measurements coming from experiments. The problem is that in both scenarios, the expensive model needs to be evaluated several times, which is unfeasible as this requires substantial computational resources. The development of advanced reduced-order models (ROMs) that can accurately represent the behavior of complex reacting systems in a wide range of input parameters, without the need for expensive simulations to be run, is therefore necessary. This is also fundamental for the development of digital twins of real systems, with application in monitoring, diagnostics and prognostics [43, 48]. In order to develop such predictive models, two phases are necessary. The first phase consists of the use of unsupervised techniques that can help data interpretation and offer a compressed representation of the training data that will be used to build a data-driven digital twin. As mentioned above, these training data-sets are usually high-dimensional, thus a compressed representation for them by a few scalars or extracted *features* is a necessity. The second phase consists of supervised techniques that can infer the values attained by these features for not yet explored operating conditions. By doing so, the system's state can be predicted in the explored range of input parameters with reduced computational costs.

1.2 HIGH-DIMENSIONALITY OF AVAILABLE DATA

High-dimensional, digitally-stored data sets are collected in large quantities in many branches of science via either experiments or numerical simulations. With the increasing need to reduce and analyze these data, various data science and machine learning methods have been developed over the last decades and applied in domains such as fluid dynamics, astrophysics or psychology, to name a few.

In the recent years, much research work has showed potential of identifying low-dimensional manifolds in chemical state-space with Principal Component Analysis (PCA) [34, 35, 38]. The authors showed also that the PCA representation of the data indicate invariance to certain parameters such as the Reynolds number. PCA coupled with a rotation method was also applied to a study of NASA database for nitrogen shock flows [2] to aid data interpretation.

Recently, data science techniques were applied to turbulent reacting systems to reduce the complexity of the chemical mechanisms and drive the development of combustion

models. Some of the popular techniques that have been used in conjunction with combustion systems are Principal Component Analysis (PCA) [1, 33, 47], Non-negative Matrix Factorization (NMF) [11, 32] or Dynamic Mode Decomposition (DMD) [15, 40]. These techniques offer low-rank approximations of high-dimensional data, thus they encode these data into a fewer number of components. They also allow for seeking patterns in large data-sets and hence provide information that would otherwise be hidden to the researcher. These techniques rely on the fact that there are often underlying low-dimensional structures (manifolds) in high-dimensional data-sets. They offer a linear representation (approximation) of the data, in the form $\mathbf{X} \approx \mathbf{U}\mathbf{V}^T$, where $\mathbf{X}(m \times n)$ is a data matrix for which a low-rank approximation is sought, $\mathbf{V}(n \times q)$ is a matrix of data-driven basis functions or modes, and $\mathbf{U}(m \times q)$ is the matrix of coordinates of the original data (rows of $\mathbf{X}$) on the used basis functions, also referred to as coefficients, scores or features. A low-rank approximation of $\mathbf{X}$ is found if $q < \min(n, m)$ and, thus, the symbol $\approx$ is used instead of the equality symbol because reconstruction errors for the original matrix are involved when such an approximation is achieved. These errors are zero only if $q = \min(n, m)$.

Because of the reconstruction errors that are involved in low-rank approximations such as the one provided by PCA, important physical laws such as positivity of the chemical species mass fractions might be violated. PCA finds a low-rank approximation of $\mathbf{X} \approx \mathbf{U}\mathbf{V}^T$ where the columns of $\mathbf{V}$ are orthogonal. Besides, the following property holds: $\mathbf{V}^T\mathbf{V} = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix. The matrix $\mathbf{U}$ is then evaluated as projection of the data onto $\mathbf{V}$: $\mathbf{U} = \mathbf{X}\mathbf{V}$. Constrained PCA (CPCA) has been shown to be capable to alleviate or solve this problem [1, 52]. In such an approach, PCA is still used to find the set of basis functions $\mathbf{V}$ to represent the data, but the coefficients $\mathbf{U}$ are estimated via a constrained minimization problem (minimizing the reconstruction errors, subject to a set of physical constraints). For positive data-set, such as combustion ones, the approach of finding reduced rank non-negative factors to approximate a given non-negative data matrix might be a natural choice as well. In fact, NMF offers a low-rank approximation where a matrix $\mathbf{X}$ is factorized into (usually) two matrices $\mathbf{U}$ and $\mathbf{V}$, with the property that all three matrices have no negative elements. This is different from PCA, where the constraint is that the columns of $\mathbf{V}$ (the PCA modes) be orthogonal.

The use of low-rank approximations leads to a reduced-order data representation that can be exploited both for data interpretation and for the subsequent development of surrogate models.

1.3 DATA REGRESSION

In the approach of the present Thesis, a specific computationally-expensive CFD simulation or computer code, referred to as Full-Order Model (FOM) [6, 7], is treated as a black box that generates a certain output $\mathbf{y}$ (e.g. the temperature field) given a set of input parameters $\mathbf{x}$ (e.g. the equivalence ratio) and indicated by $\mathcal{F}(\cdot)$:

$$\mathbf{y} = \mathcal{F}(\mathbf{x}). \tag{1}$$

The evaluation of the function $\mathcal{F}(\cdot)$ usually requires many hours of computational time. After enough observations of the FOM's output are available, $\mathbf{y}(\mathbf{x}_i) \quad \forall i = 1, ..., M$, a regression model can be trained and the output $\mathbf{y}^*$ for a particular set of unexplored inputs $\mathbf{x}^*$ can be predicted without the need to evaluate $\mathcal{F}(\mathbf{x}^*)$ and, thus, no simulation is run. The function $\mathcal{F}(\cdot)$ is therefore approximated by a new function $\mathcal{M}(\cdot)$ whose evaluation is very cheap compared to $\mathcal{F}(\cdot)$:

$$\mathbf{y}^* = \mathcal{F}(\mathbf{x}^*) \approx \mathcal{M}(\mathbf{x}^*). \tag{2}$$

Regression models are predictive mathematical models based on available data that try to approximate the underlying *hidden* relationship between input and output. These predictive models are constructed or *trained* from a relatively small set of *training* observations of the model's output, which correspond to a set of training locations or points in the model input parameter space. Once trained, regression models allow for a fast evaluation of the system's state over a wide range of their input parameters. Therefore, they are very appealing in the context of optimization studies as well as for Uncertainty Quantification (UQ) [26] and global optimization problems [31, 39]. In [13], regression models are used to optimize the performance of chemical kinetics with respect to MILD combustion. Ideally, the trained regression models should preserve the physics of the investigated phenomena, and be developed from a limited number of expensive function evaluations. Examples of regression models are Radial Basis Functions, Kriging and Polynomial Chaos Expansion [10]. Examples of regression models used in combustion applications can be found in [23].

1.4 REDUCED-ORDER MODELS

Regression or predictive models are generally constructed directly on the analyzed system's output, i.e. directly on the variables of interest such as the velocity and temperature fields. For each individual output variable a regression model is trained and a response surface is found, indicating the relationship between the variable and the input parameters. If the number of variables of interest is high, many predictive models need to be trained. Besides, any correlation between these variables of interest might be lost in the process of training individual regression models: the information about the physics of the phenomena involved is lost. Reducing the number of predictive models to train is possible if the original set of variables can be represented by a new set of fewer scalars. This corresponds to the idea that the original variables are actually realization of unknown *latent* variables [4].

Data compression methods such as Principal Component Analysis (PCA) [20] offer the potential of preserving the physics of the system while reducing the size of the problem. Data compression methods are techniques used to find a set of data-driven basis functions to represent an ensemble of high-dimensional data. Data compression methods find a new, smaller set of hopefully uncorrelated variables, often referred to as *scores* or extracted *features*, which is representative of the original variables of interest. Once these scores are found, a regression model can be built for each one of them.

Regression models usually include interpolation or regression techniques which depend on the choice of some particular design functions. These design functions are defined by a set of so-called hyper-parameters (or also length-scales) whose values affect the regression model'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. As shown in [50], ROMs are less sensitive to the particular design functions chosen for their construction, which is desirable. ROMs also have a reduced number of variables for which a regression model needs to be trained. This means that fewer optimization problems are solved in order to estimate feasible values for the hyper-parameters of the design functions, which also means that updating a trained predictive model when in the event of new available data is less computationally demanding. In addition, in [50] it is also shown how ROMs usually scale better than classic regression models for parallel computing. These features are what makes ROMs very attractive for the development of physics-preserving surrogate models.

Combustion problems are well-known for being characterized by a set of strongly inter-dependent variables. In fact, PCA has been employed in [18, 19] to re-parameterize the thermo-chemical state of a reacting system by a small number of progress variables, drastically reducing the number of transport equations to solve, and in the process showing the intrinsic lower-dimensionality of these systems, which will be exploited in the present work. PCA has also been employed in the context of turbulent combustion in [30], for the a-posteriori validation of a turbulent combustion model based on the solution of transport equations for the principal components [12] and for on-line process monitoring and fault diagnostics [54].

1.5 OBJECTIVE AND CHOSEN APPROACH

The objective is to pave the way for the development of digital twins, trained on a reduced number of full simulations, able to predict the full system state at unexplored conditions indicated by physical sensors, without running any expensive simulation, for real-time implementation. To this end, an approach based on the combination of a data-compression method and a regression technique was chosen. The dimensionality reduction step was used to extract the invariant (w.r.t. the input parameters) physics-related information of an investigated combustion system and identify the system's coefficients which instead depend on the operating conditions, referred to as scores. The regression method was then able to find a response surface for these scores. With this strategy it was possible to build a ROM that grants the possibility of parameter exploration with reduced computational cost. Regression methods that provide a distribution for the prediction value, rather than just one value as the prediction, thus also capturing the model's uncertainty, might be preferred. Some methods such as Kriging allow the user to add prior knowledge on the model by selecting different kernel functions. The use of Kriging for Computational Fluid Dynamics (CFD) data has also produced encouraging results. In fact, Kriging was employed for the shape optimization of a car engine intake port in [51]

and for aerodynamical shape optimization problems as shown in [52, 53]. However, the application was limited to non-reacting flows.

In the present Thesis, the approach of combining a data compression technique with a regression method is extended to combustion applications, to develop a ROM that can faithfully reproduce the temperature and chemical species mass fraction fields in a reacting flow simulation. The objective of the present work is to demonstrate the applicability of the proposed methodology for the development of reduced-order models of multi-scale and multi-physics computer models. In this perspective, this work paves the way for the development of digital twins of realistic engineering systems [17].

1.6 METHODOLOGY

The methodology used in the present Thesis is sketched in figure 2. Consider that a certain high-fidelity simulation model or Full Order Model (FOM) $\mathbf{y}(\mathbf{x}) = F(\mathbf{x}) \in \mathbb{R}^N$ is available, such as a CFD-combustion solver. For one value of the input parameter(s) $\mathbf{x}$, the solver returns a vector $\mathbf{y}^{(j)}$ of observations of all the involved physical variables at every grid point:

$$\mathbf{y}^{(j)} = [T(r_1, \mathbf{x}_j), \dots, T(r_L, \mathbf{x}_j), Y_{CH_4}(r_1, \mathbf{x}_j), \dots, Y_{CH_4}(r_L, \mathbf{x}_j), \dots]^T, \tag{3}$$

where L is the total number of grid points, r_i is the i-th spatial location and $\mathbf{x}_j$ is the j-th point in the input parameter space. This FOM is solved for a limited amount $M < N$ of training points in the input parameter space $\mathbf{X} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, ..., \mathbf{x}^{(M)}\} \in \mathcal{D}$, where $\mathcal{D}$ is the region spanned by the training points. Thus, only M simulations are available, one for each of those points: $\mathbf{Y} = \{\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, ..., \mathbf{y}^{(M)}\}$. The full exploration of the region $\mathcal{D}$ is possible only by running the expensive CFD-combustion solver $F(\cdot)$ for every $\mathbf{x} \in \mathcal{D}$. From the data-set $\mathbf{Y}$ of run simulations, a certain data compression method is able to extract a set of basis functions $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_q\}$, with $q < N$ usually, called modes that are invariant with respect to the input parameters $\mathbf{x}$. A set of coefficients $\mathbf{u}(\mathbf{x}) = \{u_1(\mathbf{x}), u_2(\mathbf{x}), ..., u_q(\mathbf{x})\}$, called scores and depending on $\mathbf{x}$, is consequently found. An illustrative example is reported in Figure 1, where a spatial field (e.g. a temperature field) is represented as a set of coefficients that weight a set of basis functions. These coefficients are less in number than the original number of variables as $q < N$ and can be regressed or interpolated in order to acquire knowledge about the system's state for any unexplored point $\mathbf{x}^* \in \mathcal{D}$.

One advantage of this approach is that a much smaller number of variables, namely q scores, are interpolated instead of N original variables. Another advantage is that the N original variables might be correlated. The application of dimensionality reduction for the detection of latent variables preserves this correlation, which might be lost if each original variable is interpolated independently. One additional remark is that considering, for example, $T(r_i, \mathbf{x}_j)$ and $T(r_j, \mathbf{x}_j)$ as two separate variables (rather than using the spatial locations r_i as additional input parameters) also reduces the computational costs.

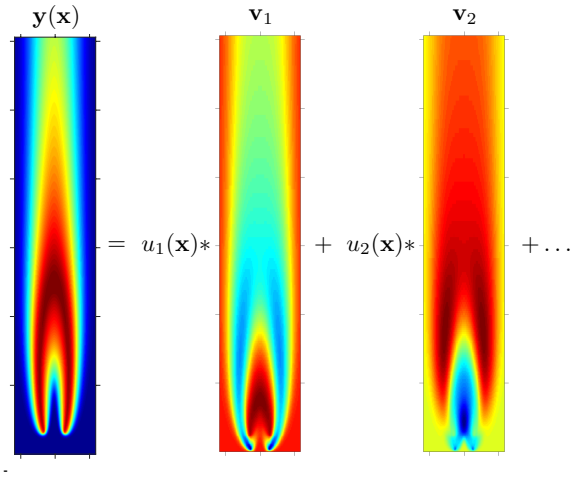


Figure 1 – Illustrative example of modal representation: one particular spatial field is represented by a set of coordinates (the coefficients u_i, called scores) on the basis functions or found by a data compression method.

$$\begin{bmatrix} \mathbf{y}^{(1)} = \mathcal{F}(\mathbf{x}^{(1)}) \\ \mathbf{y}^{(2)} = \mathcal{F}(\mathbf{x}^{(2)}) \\ \dots \\ \mathbf{y}^{(M)} = \mathcal{F}(\mathbf{x}^{(M)}) \end{bmatrix} \rightarrow \text{Data compression} \rightarrow \begin{bmatrix} u_1^{(1)}, \dots, u_q^{(1)} \\ u_1^{(2)}, \dots, u_q^{(2)} \\ \dots \\ u_1^{(M)}, \dots, u_q^{(M)} \\ \mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_q) \end{bmatrix} \longrightarrow \text{Response surface for the scores}$$

Figure 2 – Data compression finds the set of basis functions $\mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_q)$ and encodes each observation $\mathbf{y}^{(i)} \in \mathbb{R}^N$ into a small set of scalars $u_1^{(i)}, ..., u_q^{(i)}$ for $q < N$. A response surface is then found for these scalars.

1.7 PERSPECTIVES

The methodology outlined in the present Thesis shows the advantages of reduced-order modeling in terms of predictive capabilities and computational efficiency. Indeed, these features are necessary for the development of predictive models of engineering systems, which can be employed for visualization, real-time control, optimization and troubleshooting.

As 3D simulations of practical combustion systems usually require significant amount of CPU hours, having a low-order model that can reliably and instantaneously predict the outcome of these simulations is precious. Moreover, the promptness of the ROM's predictions is paramount for the development of digital twins for real systems. A correctly trained ROM also grants the possibility of performing sensitivity analysis of the investigated system w.r.t. its input parameters and can be employed to solve optimization problems in the context of system design, where the evaluation of the objective function

is the computational burden. The training costs of ROMs are also lower in comparison to a predictive model with no compression, which is very useful when new training processes are continuously needed in order to update the developed ROM in the event of new available data. Hopefully, the work carried out in the present Thesis will lead to the development of digital twins for real industrial systems.

2

METHODS

This chapter provides a summary of all the methods that have been used in the activities included in the present doctoral thesis. These methods are divided into three main categories: unsupervised learning techniques used for feature extraction and data compression; supervised learning techniques used to perform interpolation or regression and predict the values of quantities of interest for some unexplored input values; sampling strategies.

2.1 DATA COMPRESSION

Data compression or low-rank approximation is a lower dimensional representation of a higher-dimensional data-set. Thus, a certain vector $\mathbf{y}$ of n entries can be encoded into (represented by) a vector $\mathbf{z}$ of $q < n$ entries. The methods that allow to perform data compression are usually called *unsupervised* as they only need one data-set in order for them to be trained: a data-set $\mathbf{Y}$ of size $(m \times d)$ with m observations of n variables. An unsupervised technique has to find a lower-dimensional representation $\mathbf{Z}$ of size $(m \times k < n)$ of the original data matrix $\mathbf{Y}$. Naturally, the method also learns the mapping from the n-dimensional space to the k-dimensional manifold, and viceversa. Such a mapping can be linear, locally linear or non-linear, as this Chapter will explain. The low-rank approximation methods considered in this doctoral thesis are *lossy*: as said, it is possible to compress $\mathbf{Y}$ into $\mathbf{Z}$ using the mapping that has been learned from the data; however, some of the original information is lost in this process and this the inverse operation of going from the low-dimensional space to the original higher-dimensional one comes with an approximation (or reconstruction) error. Thus, this means that once $\mathbf{Z}$ is found, it is only possible to recover an approximation of the original input, say $\tilde{\mathbf{Y}}$, and not $\mathbf{Y}$. Usually, the higher the difference between the original and the recovered data matrix, the worse the performance of the unsupervised method. An illustrative example of data compression is reported in Figure 3. The possibility of finding a lower representation for $\mathbf{Y}$ exists if the n variables are actually (linearly or non-linearly) dependent. As shown in Figure 3, where an example with 2-dimensional observations (two variables) is reported, the original data points (blue) are not found all over the 2-dimensional plane: given one point and its horizontal coordinate (the value of one variable), an idea can already be had about its vertical coordinate (the value of the second variable). If these observations were scattered all over the plane (perhaps forming a filled square with arbitrarily long sides), then no acceptable one-dimensional representation would be possible as the two original variable would be completely independent of one another.

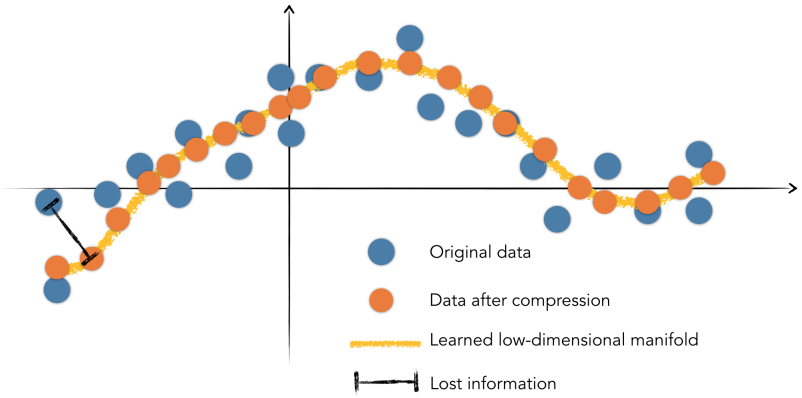


Figure 3 – Illustrative example of a 2-dimensional set of observations or objects being encoded as points on a 1-dimensional manifold, i.e. the lower-dimensional manifold learned by an unsupervised method (a curved, non-linear one in this case). The distance between the original blue points and their corresponding image or projection on the reduced manifold, the orange points, is the information that is lost in the data compression process.

Because of the discrete nature of any data-set, it is useful to organize the data into a set (or matrix) of realizations $y(\mathbf{r}_i, \mathbf{x}_j)$, where $\mathbf{r}_i$ might be considered as coordinates in a 3-dimensional spatial domain and $\mathbf{x}_j$ as the combination of some input parameters for which the realization $y(\mathbf{r}_i, \mathbf{x}_j)$ is observed, with $i = 1, \ldots, n_r$ and $j = 1, \ldots, n_x$. The resulting data matrix $\mathbf{Y}$ of field realizations $\mathbf{y}(\mathbf{r}, \mathbf{x}_j)$ has size $(n_r \times n_x)$. It is not uncommon to organize the data matrix $\mathbf{Y}$ also in the shape of $(n_x \times n_r)$, so to comply to the general practice of having the rows of this matrix corresponding to observations of a spatial field of n_r points. Either organization or shape of $\mathbf{Y}$ is fine and it is just a matter of convenience or personal choice, as long as all the operations to be performed on such a matrix are adjusted accordingly. A linear modal representation or data decomposition of a field realization $\mathbf{y}(\mathbf{r}, \mathbf{x}_j)$ can be written as follows:

$$\mathbf{y}(\mathbf{r}, \mathbf{x}_j) = \sum_{q=1}^{\infty} \sigma_q u_q(\mathbf{x}_j) \mathbf{v}_q(\mathbf{r}), \tag{4}$$

where $\mathbf{v}_q$ is the q-th mode or basis function, u_q is the coefficient for it and σ_q is the energetic contribution of that mode to the global field as the modes are usually normalized. The importance of such a representation is that only the coefficients u_q depend on the parameters $\mathbf{x}$, while the the modes $\mathbf{v}_q$ preserve the spatial dependence and can also be considered as basis spatial profiles. Practically, it is impossible to find an infinite number of modes, thus the modal representation of $\mathbf{Y}$ can be re-written as follows:

$$\mathbf{Y} = \sum_{q=1}^{N} \sigma_q u_q \mathbf{v}_q, \tag{5}$$

with $N = \min(n_x, n_r) < \infty$. This modal representation also implicitly represents a compression for $\mathbf{Y}$ as any field realization $\mathbf{y}(\mathbf{r}, \mathbf{x}_j)$ is expressed by the same modes $\mathbf{v}_q(\mathbf{r})$. Only

the coefficients u_q change when the values of $\mathbf{x}$ change. Because the size of one field realization $\mathbf{y}$ is greater than the number of coefficients u_q, a compressed representation for $\mathbf{y}$ is obtained and $\mathbf{y}$ is said to have been encoded into a reduced set of scalars. An approximation of $\mathbf{Y}$ is achieved if a smaller number of modes $k < N$ is kept and the remainder discarded. In matrix form, such a representation can be formulated as:

$$\mathbf{Y} = \mathbf{U}\mathbf{S}\mathbf{V}^T \tag{6}$$

There exist two main categories for modal representation of field realizations or data-sets more in general. The first category consists of fixing the matrix $\mathbf{V}$ a priori and then finding $\mathbf{U}$ by least-squares regression or by projecting the data onto $\mathbf{V}$. Since this category of modal representations fix one basis regardless of the data-set that is available, they are not data-driven. One example is Discrete Fourier Transform. The second category consists of the use of data-driven approaches for the estimation of $\mathbf{V}$. The matrix $\mathbf{U}$ can then be estimated as already explained or be also an output of the chosen data-driven method. The described modal representation is a linear decomposition of $\mathbf{Y}$ as the mapping from the reduced space of modal coefficients u_q to the original data space is linear. There also exist other data-encoding techniques where this mapping is not linear. In the present Thesis, the focus will be on linear, locally-linear and non-linear data-driven approaches. Figure 4 reports an example of locally-linear and non-linear encoding processes.

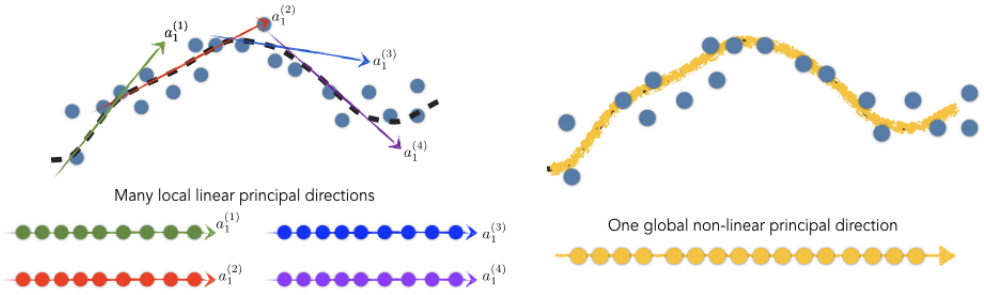


Figure 4 – Illustrative example of 2-dimensional data encoded on a set of locally linear manifolds (*left*) and one globally non-linear one (*right*). In the locally-linear case, the projected data refer to the local coordinate systems $a_1^{(1)}$, $a_1^{(2)}$, $a_1^{(3)}$, $a_1^{(4)}$. In the non-linear case, all data are projected on a common global and curved lower dimensional manifold.

Data are usually centered and scaled before a data compression method is carried out. Centering represents all observations as fluctuations from a chosen center, usually the mean value, leaving only the relevant variation for analysis. Scaling is a crucial operation when dealing with multivariate data-sets. In fact, in the case of combustion-related data-sets, temperature and chemical species mass fractions have different units and vary over different scales. The choice of the appropriate centering and scaling criteria also depends on the subsequent compression method that is employed. Six possible choices for the scaling of the data are here reported. AUTO scaling: each variable is normalized by its standard deviation; RANGE scaling: each variable is normalized by its range; PARETO scaling: each variable is scaled by the square root of its standard deviation; VAST: scaling

each variable is scaled by the standard deviation and coefficient of variation; LEVEL scaling: each variable is normalized by the mean of the data; MAX scaling: each variable is scaled by its maximum value. The choice of scaling usually affects the subsequent data compression process, as shown in [35] where Principal Component Analysis is applied. In the remainder of this Chapter, it is implied that the data matrices have already been appropriately centered and scaled for each method.

2.1.1 *Principal Component Analysis*

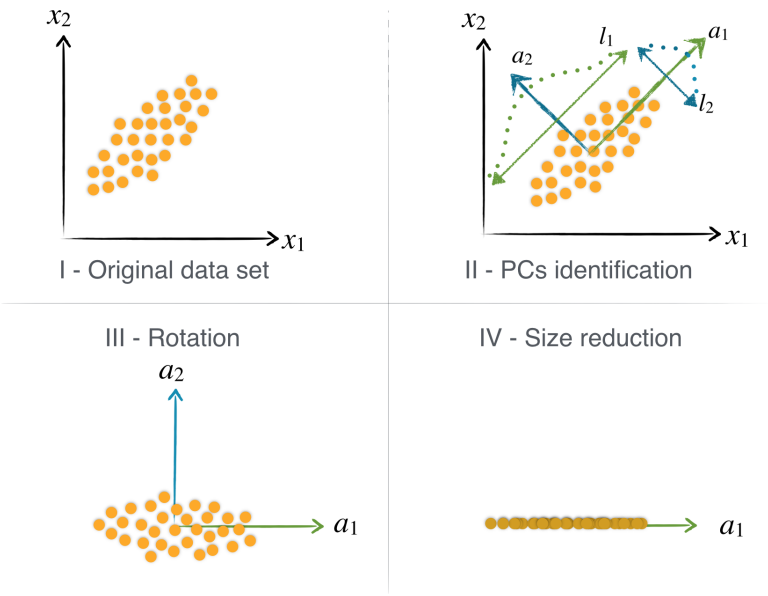


Figure 5 – Illustrative example of Principal Component Analysis (PCA) applied to a 2-dimensional data-set. Axis of maximum variance are identified and the 2-dimensional original points are projected on the direction of maximum variance. Thus, dimensionality reduction is achieved as the original data of 2-dimensional objects are encoded into 1-dimensional objects. In this example, a rotation is also performed.

Principal Component Analysis (PCA) or Proper Orthogonal Decomposition (POD) is a statistical technique able to provide a set of orthogonal low-dimensional basis functions to represent an ensemble of high-dimensional experimental or simulation data describing an undesirably complex system [20], as illustrated in Figure 5. By applying PCA, a compact representation of the data is obtained, that can be used for feature extraction. The key idea of Principal Component Analysis (PCA) is to reduce (compress) a large number of interdependent variables (i.e. independent up to the second-order statistical moments) to a smaller number of uncorrelated variables while retaining as much of the original data variance as possible [5, 9, 33, 35, 47].

The PCA problem can be stated as follows: given a matrix $\mathbf{Y}$ of size $(m \times n)$, find $\mathbf{Z}$ of size $(m \times k)$ and $\mathbf{A}$ of size $(n \times k)$ with $k < \min(m, n)$ such that the functional $f(\mathbf{Z}, \mathbf{A}) = \frac{1}{2}||\mathbf{Y} - \mathbf{Z}\mathbf{A}^T||^2$ is minimized, subject to $\mathbf{A}^T\mathbf{A} = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix. This problem can be solved by computing the singular value decomposition (SVD) of the matrix $\mathbf{Y}$,

which corresponds to finding the eigenvectors and eigenvalues of the matrix $\mathbf{C} = \frac{1}{m-1}\mathbf{Y}^T\mathbf{Y}$. The eigenvectors of $\mathbf{C}$ are the PCA modes and the associated eigenvalues represent their relevance for the low-rank approximation of $\mathbf{Y}$. The PCA (or POD) modes are thus found all at once, and by ordering them in descending order according to their corresponding eigenvalue and retaining only a subset $k < n$ of them, a low-rank approximation of $\mathbf{Y}$ is possible as follows $\mathbf{Y} \approx \mathbf{Z}\mathbf{A}^T = \mathbf{Y}\mathbf{A}\mathbf{A}^T$, where the columns of $\mathbf{A}$ of size $(n \times q)$ are the PCA modes and $\mathbf{Z}$ of size $(m \times k)$ is the matrix of PCA coefficients. Each row of $\mathbf{Z}$ are the k coefficients for the retained k PCA modes so that one particular simulation, or row of $\mathbf{Y}$, can be expressed as $\mathbf{y}(\mathbf{x}_j) = \sum_{i=1}^{k} \mathbf{a}_i z_i(\mathbf{x}_j)$. As the solution of the PCA problem only leads to the evaluation of $\mathbf{A}$, the matrix $\mathbf{Z}$ can be computed only after $\mathbf{A}$ is known, as follows: $\mathbf{Z} = \mathbf{Y}\mathbf{A}$. This holds for new, unseen data as well: $\mathbf{Z}' = \mathbf{Y}'\mathbf{A}$.

2.1.2 *Local PCA*

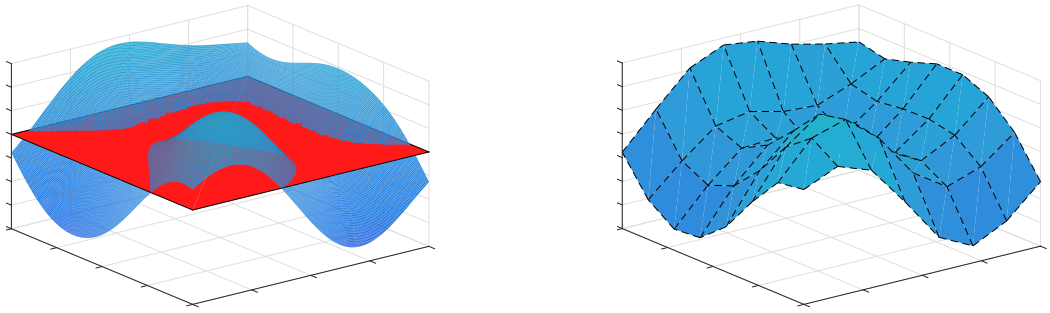


Figure 6 – (*left*) A non-linear hyper-surface is approximated by only one hyper-plane in the data space. (*right*) The same hyper-surface is approximated by a set of local hyper-planes. The application of PCA can lead to better performances if local regions in the data space are detected and PCA is applied locally and independently in each region.

PCA is a linear combination of basis functions. A large number of PCs may be required when applying PCA on highly non-linear systems [33, 47]. Local PCA (LPCA) constructs local models, each pertaining to a different disjoint region of the data space [21]. Within each region, the model complexity is limited, and thus it is possible to construct linear models using PCA [21, 42]. Figure 6 provides a general representation for a set of 3-dimensional observations forming a curved surface. Each axis shows the co-domain for each of the three scalar components that identify the 3-dimensional observations. The figure shows how a local representation of the curved surface can provide a better representation with respect to a single hyper-plane.

The partition in local clusters, where PCA is carried out, is accomplished using a Vector Quantization (VQ) algorithm that minimizes the reconstruction error. The reconstruction error is the squared Euclidean distance from one point or observation in the data-space to

the linear manifold $\mathbf{A}^{(i)}$ that is found by applying PCA in the local region. Mathematically, it can be expressed in a general fashion as:

$$d(\mathbf{y},\ \mathbf{r}^{(i)}) = \left(\mathbf{y}-\mathbf{r}^{(i)}\right)^{T} \mathbf{A}^{(i)\ T}\mathbf{A}^{(i)} \left(\mathbf{y}-\mathbf{r}^{(i)}\right), \tag{7}$$

where $\mathbf{y}$ is the object to be assigned to the cluster $\mathcal{R}^{(i)}$, represented by the reference vector $\mathbf{r}^{(i)}$, defined as the centroid of the i-th region: $\mathbf{r}^{(i)} = E[\mathbf{y} \in \mathcal{R}^{(i)}]$. The cluster i is defined as:

$$\mathcal{R}^{(i)} = \{\mathbf{y} \mid d(\mathbf{y}, \mathbf{r}^{(i)}) \leqslant d(\mathbf{y}, \mathbf{r}^{(j)});\ \forall\, j \neq i\}. \tag{8}$$

2.1.3 *Constrained PCA*

The truncation of the PCA basis may inevitably involve the violation of important physical laws such as the conservation of mass when the original data matrix $\mathbf{Y}$ is reconstructed from the PCA scores: $\mathbf{Y} \approx \mathbf{Z}\mathbf{A}^{T}$. To avoid that, the PCA scores can be evaluated by solving a constrained minimization problem, where the functional to be minimized is the PCA reconstruction error [50]. This approach is usually referred to as Constrained PCA (CPCA). The constraints are the physical laws which are intended not to be violated. This minimization problem can be mathematically expressed as:

$$\begin{aligned} \text{minimize}: \quad & \mathcal{J}(\boldsymbol{\gamma}^{(i)}) = \frac{1}{2}\|\mathbf{y}^{(i)} - \boldsymbol{\gamma}^{(i)}\mathbf{A}^{T}\|^{2} \\ \text{s.t.}: \quad & l_j\left(\boldsymbol{\gamma}^{(i)}\mathbf{A}^{T}\right) = 0 \qquad \forall j = 1, ..., N_c \end{aligned} \tag{9}$$

where $\mathbf{y}^{(i)}$ represents one field realization, $\boldsymbol{\gamma}$ are the CPCA scores for that realization, $l_j(\cdot)$ is the function related to the j-th constraint and N_c is the number of constraints, which can also be inequality constraints. Minimizing the functional $\mathcal{J}$ when no constraints are enforced leads to the estimation of the PCA scores, as explained in Section 2.1.1.

It is preferable that the solution of this system be not too computationally expensive. In [50], the constrained optimization problem has a straightforward solution due to the linearity of the imposed constraints, which allows for a fast evaluation of the CPCA coefficients. If more complex constraints are imposed, the solution of the constrained optimization problem for the evaluation of the CPCA scores might involve the reconstruction of the considered physical fields and the use of more expensive optimization algorithms, making the evaluation of the aforementioned coefficients unfeasible.

2.1.4 *Kernel PCA*

Kernel PCA [4] is a non-linear dimensionality reduction technique that makes use of the kernel methods. In kernel PCA, the original data-set $\mathbf{Y} \in \mathbb{R}^{m\times n}$ where $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_m]^{T}$ is transformed to an arbitrarily high-dimensional feature space in the following way:

$$\mathbf{y}_i \in \mathbb{R}^{n} \longmapsto \boldsymbol{\phi}(\mathbf{y}_i) \in \mathbb{R}^{N} \qquad \forall i = 1, \ldots, m \tag{10}$$

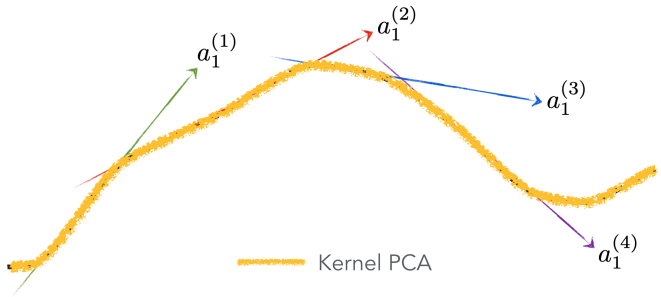


Figure 7 – Illustrative example of Kernel PCA for 2-dimensional data. Kernel PCA attempts to find a non-linear lower-dimensional manifold or hyper-surface (a line in this case) that fits the original data. The locally-linear Local PCA one-dimensional manifolds are also reported for comparison between the two methods.

with $N >> n$, and then PCA is carried out in this new feature space. A non-trivial, arbitrary function ϕ is chosen but is never calculated explicitly, which allows for the possibility to use very high-dimensional feature spaces. In order to avoid working in the feature-space, a m-by-m kernel representing the inner product space (Gramian matrix) can be created:

$$K_{ij} = k(\mathbf{y}_i, \mathbf{y}_j) = \phi(\mathbf{y}_i)^T \phi(\mathbf{y}_j). \tag{11}$$

The non-linearity of the kernel PCA method comes from the use a non-linear kernel function that populates the covariance matrix (known as the *kernel trick*). Thus, the eigenvectors and the eigenvalues of the covariance matrix in the feature space are never actually solved. The choice of the kernel function $k(\mathbf{y}_i, \mathbf{y}_j)$ modeling the covariance between two different points in the feature space is up to the designer. An example is the choice of a Gaussian kernel:

$$k(\mathbf{y}_i, \mathbf{y}_j) = \frac{1}{\sigma} e^{\frac{|\mathbf{y}_i - \mathbf{y}_j|^2}{h}} \tag{12}$$

where h is the kernel width and $\frac{1}{\sigma}$ is the kernel scaling factor. Because no operations are directly carried out in the feature space, kernel PCA does not compute the PCA modes themselves, but the projections of our data onto those components, namely the scores.

PCA in the feature space would require the solution of the following eigenproblem:

$$\lambda \mathbf{a}_k = \mathbf{K}\mathbf{a}_k, \tag{13}$$

where $\mathbf{a}_k$ is the k-th PC in the feature space. All solutions must lie in the span of ϕ-images of the training data, meaning that there exist some coefficients $\boldsymbol{\alpha}^k = (\alpha_1^k, \alpha_2^k, \ldots, \alpha_m^k)$ such that:

$$\mathbf{a}_k = \sum_{i=1}^{m} \alpha_i^k \phi(\mathbf{y}_i). \tag{14}$$

Substituting 14 into 13, and multiplying by $\phi(\mathbf{y})$, leads to:

$$m\lambda_k \boldsymbol{\alpha}^k = \mathbf{K}\boldsymbol{\alpha}^k. \tag{15}$$

Normalizing the solutions $\mathbf{a}$ translates into $\lambda(\alpha^T \cdot \alpha) = 1$. The projection of the ϕ-image of a point $\mathbf{x}_i$ onto the k-th component, the k-th kernel PCA score for point i, is given by:

$$z_i^k = \phi(\mathbf{y}_i)\mathbf{a}_k = \sum_{j=1}^{m} \alpha_i^k k(\mathbf{y}_i, \mathbf{y}_j). \tag{16}$$

One of the main drawbacks of the kernel PCA is that PCA is performed on a covariance matrix that scales with the number of observations m and not the number of variables n as in PCA which increases the computational demands of the method for applications where $m >> n$. In addition, no straightforward operation is available to reconstruct the data from the kernel PCA scores, since the function $\phi(\mathbf{y}_i)$ is never computed explicitly. This is in contrast to linear PCA. In order to reconstruct the data in kernel PCA, the following functional needs to be minimized for $\tilde{\mathbf{y}}_i$:

$$\rho(\tilde{\mathbf{y}}_i) = k(\tilde{\mathbf{y}}, \tilde{\mathbf{y}}) - 2 \sum_{k=1}^{q} z_k \sum_{j=1}^{m} \alpha_i^k k(\tilde{\mathbf{y}}_i, \mathbf{y}_j), \tag{17}$$

where q is the total number of scores. Having to minimize this functional makes the process computationally more expensive in comparison to PCA.

2.1.5 *Non-negative matrix factorization*

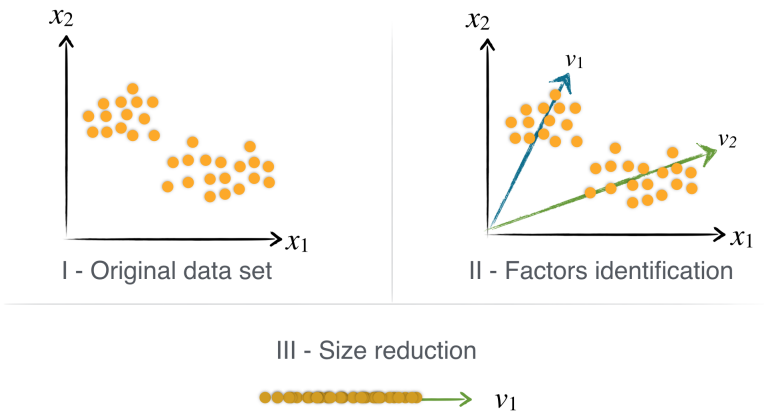


Figure 8 – Illustrative example of Non-negative Matrix Factorization (NMF) applied to a 2-dimensional data-set of positive values. Axis of non-negative components are identified and a part-based representation is learned as original objects are expressed as sum of positive elements. Dimensionality reduction is achieved if the original data of 2-dimensional objects are projected onto one axis found by NMF.

The Non-negative matrix factorization (NMF) problem can be stated as follows: given a non-negative matrix $\mathbf{Y}$ of size $(m \times n)$, find non-negative matrix factors $\mathbf{U}$ of size $(m \times k)$ and $\mathbf{V}$ of size $(n \times k)$ with $k < n$ such that the functional $f(\mathbf{U}, \mathbf{V}) = \frac{1}{2}\|\mathbf{Y} - \mathbf{U}\mathbf{V}^T\|^2$ is minimized, subject to $u_{ij}, v_{ij} > 0 \quad \forall i, j$ [32]. Differently from PCA where the modes are found all at once and only then a subset k of them is chosen, the non-negative factors

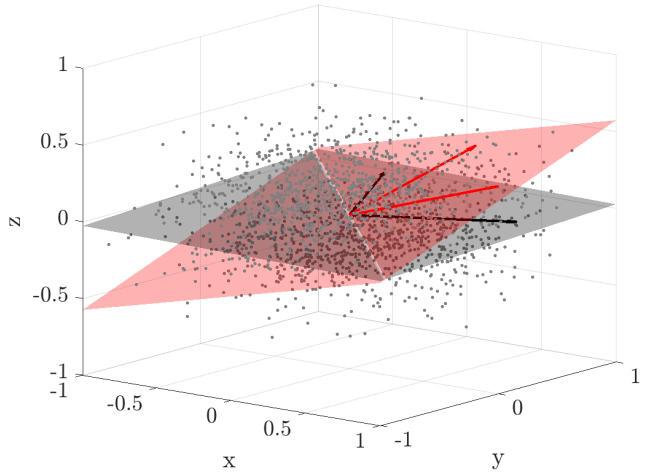


Figure 9 – Illustrative example of PCA and NMF finding 2 different lower-dimensional manifolds (planes) for the compression to 2 dimensions of a 3-dimensional data-set. Red manifold: NMF; grey manifold: PCA. Notice that the PCA directions (arrows) are orthogonal, while the NMF ones are not.

or NMF modes are found for a given approximation order k. For a different value of k, the NMF problem needs to be solved again. As for PCA, once the NMF problem is solved and thus the NMF modes are found, one particular simulation, or row of $\mathbf{Y}$, can be expressed as $\mathbf{y}(\mathbf{x}_j) = \sum_{i=1}^{k} \mathbf{v}_i u_i(\mathbf{x}_j)$. Differently from PCA, both matrices $\mathbf{U}$ and $\mathbf{V}$ are determined by solving the NMF problem. The determination of the NMF scores $\mathbf{U}'$ for new, unseen data $\mathbf{Y}'$ is possible only by solving the least-squares error minimization problem: $\mathbf{U}' = \underset{\mathbf{U}'}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{Y}' - \mathbf{U}'\mathbf{V}^T\|^2$, whose solution is taken as $\mathbf{U}' = \mathbf{Y}\left(\mathbf{V}^T\mathbf{V}\right)^{-1}\mathbf{V}^T$. This solution can lead to negative entries in the matrix $\mathbf{U}'$, which can be set to zero in order to avoid violating the constraint of having positive coefficients.

2.1.6 *Autoencoder*

An Autoencoder is a type of unsupervised artificial neural network (ANN) [27]. The aim of an Autoencoder is to learn a representation (encoding) for a set of data, typically for dimensionality reduction. An Autoencoder learns a lower-dimensional non-linear manifold that fits the original high-dimensional data, similarly to Kernel PCA (Figure 7). Along with the reduction side, a reconstructing side is learnt, where the autoencoder tries to generate a representation from the reduced encoding as close as possible to its original input.

The simplest form of an Autoencoder is a feedforward, non-recurrent neural network having an input layer, an output layer and one or more hidden layers connecting them, where the output layer has the same number of neurons as the input layer, and with the purpose of reconstructing its inputs (minimizing the difference between the input and the output). Therefore, Autoencoders are unsupervised learning models.

Given one hidden layer, the encoder process of an Autoencoder takes to input $\mathbf{x} \in \mathcal{R}^{n_v}$ and maps it to $\mathbf{h} \in \mathcal{R}^q$, with $q < n_v$:

$$\mathbf{h} = f(\mathbf{W}\mathbf{x} + \mathbf{b}). \tag{18}$$

This image $\mathbf{h}$ is usually referred to as *code* or latent representation. Here, $f()$ is an element-wise activation function such as a sigmoid function or a rectified linear unit. $\mathbf{W}$ is a weight matrix and $\mathbf{b}$ is a bias vector. The decoder stage of the Autoencoder maps $\mathbf{h}$ to the reconstruction $\mathbf{x}'$ of the same shape as $\mathbf{x}$:

$$\mathbf{x}' = f'(\mathbf{W}'\mathbf{h} + \mathbf{b}'). \tag{19}$$

where f', $\mathbf{W}'$ and $\mathbf{b}'$ may be unrelated to f, $\mathbf{W}$ and $\mathbf{b}$. An example of ANN is reported in Figure 16, but in the case of an Autoencoder, the net is mirrored and the output layer has as many nodes as the input layer (thus, the output is a reconstructed input and the middle layer corresponds to the lower-dimensional image).

Autoencoders are trained by minimizing the loss function $\mathcal{L}(\mathbf{x}, \mathbf{x}') = ||\mathbf{x} - \mathbf{x}'||^2$. This loss function can be modified to accommodate a regularization and a sparsity term. This is often done to avoid over-fitting, meaning that the trained Autoencoder has only learned to reconstruct seen data but it has failed to find latent structures that can be generalized.

2.2 REGRESSION METHODS

A regression or a predictive model is a function that tries to replicate the relationships among a set of independent variables and a dependent variable. These methods are usually called *supervised* as a regression model usually needs two data-sets to be trained: an input data-set $\mathbf{X}$ of size $(m \times d)$ with m observations of d independent variables, and a data-set of known values of n target output dependent variable(s) $\mathbf{Y}$ of size $(m \times n)$. Usually, a regression model is trained for one target dependent variable at time (one column of $\mathbf{Y}$, thus a scalar output), although Artificial Neural Networks (ANNs) can handle higher dimensional outputs.

2.2.1 *Polynomial regression and over-fitting*

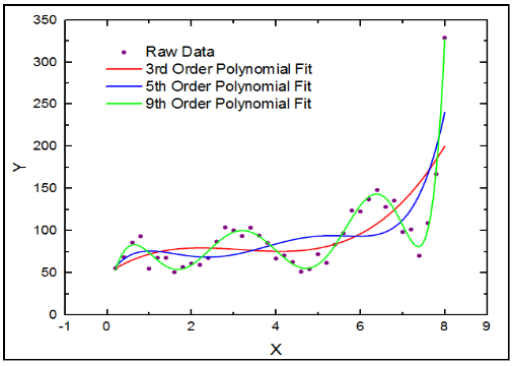


Figure 10 – Illustrative examples of a polynomial regression.

Polynomial regression is one of the most known methods for predictive modeling. It consists in trying to fit available data by correctly choosing the values for the coefficients of a polynomial of any order. This is usually done by ordinary least squares estimation, although other methods also exist [45]. The order of the polynomial is usually lower than the number of available points. When this is true, the polynomial model will not interpolate (pass through) the original training data. Choosing a lower order (a model with lower *capacity*) is done to avoid over-fitting, a condition where a predictive model just remembers the training data but is not able to generalize to new unseen data (the model only knows what he has already seen and has not learned any general pattern). Over-fitting is an issue for all predictive models and needs to be avoided. One way to do so is to use a test data-set, a data-set that is not used for the training of the model but only to test its predictive performance. The prediction errors are tested on both data-sets. The highest model complexity that is chosen (degree of the polynomial regression in this case) is the one for which the error on the test-set starts to grow. Polynomial regression is particularly useful when the relationship to be modeled is not extremely complex or if a very limited amount of data is available. For non-linear data, polynomial regression can be quite challenging to design, as one must have some information about the structure of the data and relationship between feature variables.

2.2.2 *Kriging*

Kriging is an interpolation method in which every realization $y(\mathbf{x})$, where y is a random target, is expressed as a combination of a trend function and a residual [8]:

$$y(\mathbf{x}) = \mu(\mathbf{x}) + s(\mathbf{x}) = \sum_{i=0}^{p} \beta_i f_i(\mathbf{x}) + s(\mathbf{x}) = \mathbf{f}^T(\mathbf{x})\boldsymbol{\beta} + s(\mathbf{x}) \tag{20}$$

The trend function $\mu(\mathbf{x})$ is a low-order polynomial regression and provides a global model in the input space. The term $s(\mathbf{x})$ creates a localized deviation weighting the points in the training set that are closer to the target point $\mathbf{x}$. The trend function $\mu(\mathbf{x})$ is expressed as a weighted linear combination of $p+1$ polynomials $\mathbf{f}(\mathbf{x}) = [f_0(\mathbf{x}),\ ...,\ f_p(\mathbf{x})]^T$ with the weights $\boldsymbol{\beta} = [\beta_0,\ ...,\ \beta_p]^T$ determined by generalized least squares (GLS). The subscript p also indicates the degree of the polynomial. The residuals $s(\mathbf{x})$ 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). The correlation function describes the correlation between two points in the $\mathbf{x}$-space. In other words, it will model the correlation between two different observations of the output variable y based on their corresponding location in the input space. The hyper-parameters $\boldsymbol{\theta}$ (or their inverse) are also called length-scales. Low values for these length-scales will mean that far away observations will not be correlated. Many possible correlation functions are available: linear, quadratic, exponential, Gaussian, Matern 3/2, Matern 5/2, just to name a few [8, 29]. A detailed discussion about these functions can be found in [46]. One of the main differences among these kernels is their smoothness. See Figure 11 and 12 for clarifications.

Name	$\mathcal{R}_j(\theta, d_j)$	
EXP	$\exp(-\theta_j \lvert d_j \rvert)$	
EXPG	$\exp(-\theta_j \lvert d_j \rvert^{\theta_{n+1}})$,	$0 < \theta_{n+1} \leq 2$
GAUSS	$\exp(-\theta_j d_j^2)$	
LIN	$\max\{0,\ 1-\theta_j \lvert d_j \rvert\}$	
SPHERICAL	$1 - 1.5\xi_j + 0.5\xi_j^3$,	$\xi_j = \min\{1,\ \theta_j \lvert d_j \rvert\}$
CUBIC	$1 - 3\xi_j^2 + 2\xi_j^3$,	$\xi_j = \min\{1,\ \theta_j \lvert d_j \rvert\}$
SPLINE	$\varsigma(\xi_j)$, (2.24)	$\xi_j = \theta_j \lvert d_j \rvert$

Figure 11 – Examples of one-dimensional correlation or kernel functions employed in the Kriging formulation [29]. d_j indicates the euclidean distance between two points $|\mathbf{x} - \mathbf{x}'|$. θ_j are the hyper-parameters, also called length-scales, of the kernels.

The Matern correlation function is mainly used as it is a generalization of the exponential and the Gaussian correlation functions. Determining the optimal correlation parameters or hyper-parameters for the chosen kernel function is a complex multi-dimensional minimization problem. Optimization algorithms can be cast into two distinct categories: local and global optimization algorithms. The best optimization algorithm is problem depen-

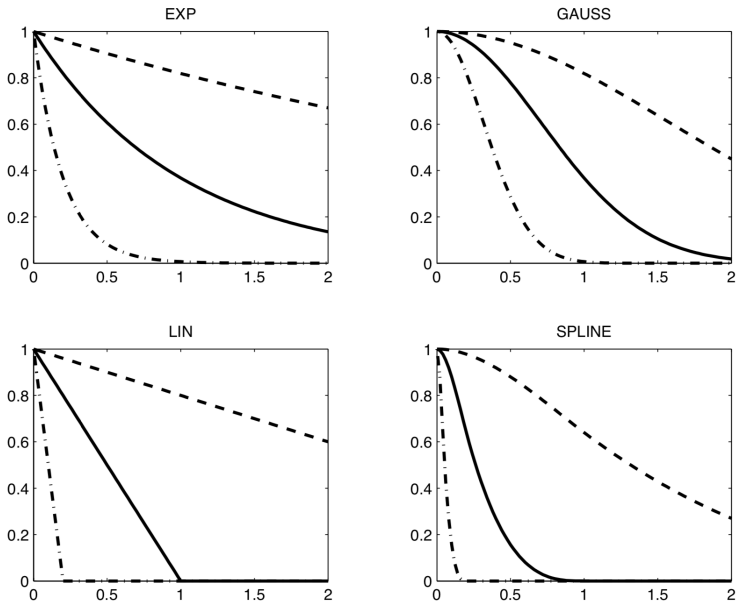


Figure 12 – One-dimensional correlation functions for $\theta = 0.2,\ 1,\ 5$ and $0 < d < 2$ [29].

dent and in many cases not known a-priori. For small dimensional problems, the influence of the function family is low [45].

In the definition of both the trend function and the residual, it is up to the designer to choose the polynomials $\mathbf{f}(\mathbf{x})$ and the correlation model or kernel. In this way, the designer has the possibility to add prior into the problem and subsequently let the data speak for themselves by estimating the hyper-parameters.

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

$$\begin{aligned} y(\mathbf{x}^*) &= \mathbf{f}(\mathbf{x}^*)^T\boldsymbol{\beta} + \mathbf{r}(\mathbf{x}^*)^T\mathbf{R}^{-1}(\mathbf{y} - \mathbf{F}\boldsymbol{\beta}) = \\ &= \mathbf{f}(\mathbf{x}^*)^T\boldsymbol{\beta} + \mathbf{r}(\mathbf{x}^*)^T\mathbf{g}. \end{aligned} \tag{21}$$

In (21), $\mathbf{F}$ is the matrix of polynomials evaluated at the training locations; $\mathbf{R}$ is the kernel matrix or matrix of correlations between the training points; and $\mathbf{r}$ is the vector or correlations between the training points and the point $\mathbf{x}^*$ for which we wish to make a prediction. An important note is that, once the model is trained, the quantities $\boldsymbol{\beta}$ and $\mathbf{g}$ are constant and do not need to be updated anymore. To make a prediction, say $y(\mathbf{x}^*)$, only the terms $\mathbf{f}(\mathbf{x})$ and $\mathbf{r}(\mathbf{x})$ need to be updated:

$$y(\mathbf{x}^*) = \mathbf{f}(\mathbf{x}^*)^T\boldsymbol{\beta} + \mathbf{r}(\mathbf{x}^*)^T\boldsymbol{\gamma}. \tag{22}$$

2.2.3 *Polynomial Chaos Expansion*

The theoretical background for Polynomial Chaos Expansion (PCE) is reported from [44, 45]. Consider a system whose behavior is represented by a computational model $\mathcal{M}$ which maps the d-dimensional input parameter space to the 1-dimensional output space $\mathcal{M} : \mathbf{x} \in \mathbb{R}^d \rightarrow y \in \mathbb{R}$. In Section 1.6, the computational model was indicated by $\mathcal{F}$ and

Distribution	PDF	Orthogonal polynomials	Orthonormal basis
Uniform	$\mathbf{1}_{]-1,1[}(x)/2$	Legendre $P_k(x)$	$P_k(x)/\sqrt{\frac{1}{2k+1}}$
Gaussian	$\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$	Hermite $H_{e_k}(x)$	$H_{e_k}(x)/\sqrt{k!}$
Gamma	$x^a e^{-x}\mathbf{1}_{\mathbb{R}^+}(x)$	Laguerre $L_k^a(x)$	$L_k^a(x)/\sqrt{\frac{\Gamma(k+a+1)}{k!}}$
Beta	$\mathbf{1}_{]-1,1[}(x)\frac{(1-x)^a(1+x)^b}{B(a)B(b)}$	Jacobi $J_k^{a,b}(x)$	$J_k^{a,b}(x)/\mathcal{J}_{a,b,k}$
			$\mathcal{J}^2_{a,b,k} = \frac{2^{a+b+1}}{2k+a+b+1}\frac{\Gamma(k+a+1)\Gamma(k+b+1)}{\Gamma(k+a+b+1)\Gamma(k+1)}$

Figure 13 – Classical orthogonal/orthonormal polynomials. Figure from [44].

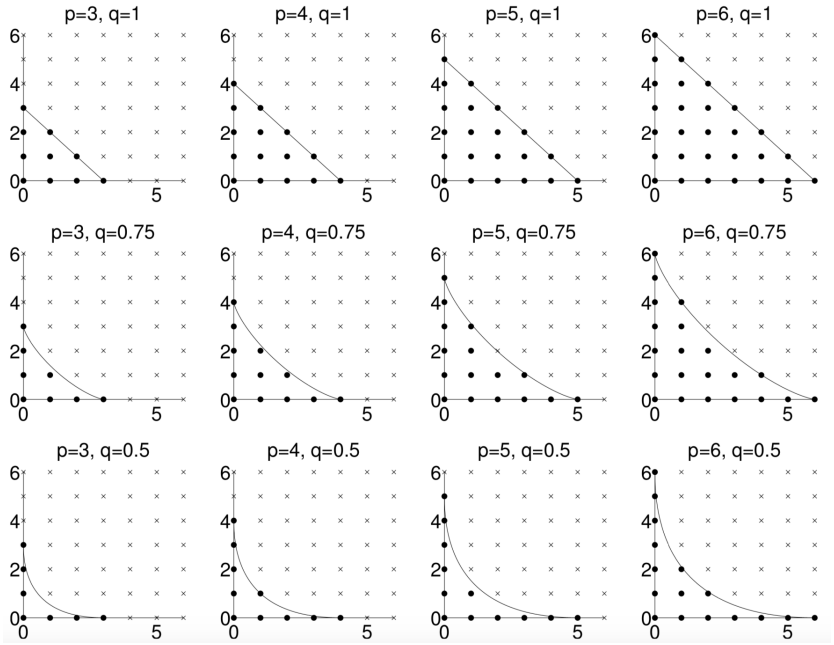


Figure 14 – Representation of a hyperbolic index set $\boldsymbol{\alpha} \in \mathcal{A}_q^{d,N}$ for various N (p in the Figure) and q (d = 2). Figure from [44].

its output was multi-dimensional. For this reason, in this section the symbol $\mathcal{M}$ will be employed. In the present work, the components of the input vector $\mathbf{x} = x_1, \ldots, x_d$ are assumed independent. The case of dependent input variables can easily be addressed as explained in [24]. In the present work, the authors consider that the computational model $\mathcal{M}$ is a deterministic mapping from the input to the output space, i.e. repeated evaluations with the same input values lead to the same output value. As explained in [44], the computational model $\mathcal{M}$ can be approximated by a finite, truncated set of polynomials:

$$y(\mathbf{x}) = \mathcal{M}(\mathbf{x}) \approx \sum_{\boldsymbol{\alpha}\in\mathbb{N}^d} a_{\boldsymbol{\alpha}}\psi_{\boldsymbol{\alpha}}(\mathbf{x}), \tag{23}$$

where $a_{\boldsymbol{\alpha}}$ are the expansion coefficients of the multivariate polynomials $\psi_{\boldsymbol{\alpha}}(\mathbf{x})$ and $\boldsymbol{\alpha}$ is the multi-index. Because of the statistical independence of the input variables, the multivariate polynomials are evaluated as product of uni-variate polynomials $\psi_{\boldsymbol{\alpha}}(\mathbf{x}) = \prod_{i=1}^{d} \psi_{\alpha_i}^{(i)}(x_i)$, where $\psi_{\alpha_i}^{(i)}$ is the polynomial of degree α_i for the i-th variable. Classic

orthogonal univariate polynomials are reported in Figure 13. The total degree of the (multivariate) polynomials is defined by $|\boldsymbol{\alpha}| = \sum_{i=1}^{d} \alpha_i$. The total number of (multivariate) polynomials depends on the adopted truncation scheme. In the present work, the adopted truncation scheme is the *hyperbolic truncation set*: $\boldsymbol{\alpha} \in \mathbb{N} : \|\boldsymbol{\alpha}\|_c \leqslant N$, where N is the total order of the polynomials and the norm $\|\cdot\|_c$ is defined as $\|\boldsymbol{\alpha}\|_c = \left(\sum_{i=1}^{d} \alpha_i^c\right)^{1/c}$, where $0 \leqslant c \leqslant 1$. See Figure 14.

The maximum number of terms in the polynomial basis, attainable for $c = 1$, is given in Eq. (24):

$$p + 1 = \frac{(d+N)!}{d!\,N!}. \tag{24}$$

The expansion coefficients $a_{\boldsymbol{\alpha}}$ can be estimated by a least-square minimization method, least absolute shrinkage operator (LASSO), least angle regression (LAR), and other methods, as explained in [44].

2.2.4 *PC-Kriging*

As explained in [44], Kriging is able to interpolate local variations of the output of the computational model. In contrast, polynomial chaos expansions (PCE) are generally used for approximating global behaviors of computational models. The two techniques can be combined if PCE is used as trend function for the Kriging interpolation method. This approach is referred to as PC-Kriging and its formulation is as follows:

$$y(\mathbf{x}) = \mathcal{M}(\mathbf{x}) \approx \sum_{\boldsymbol{\alpha}\in\mathcal{A}} a_{\boldsymbol{\alpha}} \psi_{\boldsymbol{\alpha}}(\mathbf{x}) + s(\mathbf{x}) = \sum_{\boldsymbol{\alpha}\in\mathcal{A}} a_{\boldsymbol{\alpha}} \psi_{\boldsymbol{\alpha}}(\mathbf{x}) + \mathbf{r}(\mathbf{x}^*)^T \boldsymbol{\gamma}. \tag{25}$$

Building a PC-Kriging meta-model consists of determining the optimal set of polynomials for PCE first and then calibrating the Kriging hyper-parameters.

2.2.5 *Differences between polynomial regression and PCE*

As explained, in both the classic polynomial regression and PCE, the expression for a certain prediction $y(\mathbf{x})$ has the following form:

$$y(\mathbf{x}) = \mathbf{f}(\mathbf{x})^T \boldsymbol{\beta}. \tag{26}$$

In both cases, as explained in [44], there exist many different techniques to evaluate optimal values for the coefficients $\boldsymbol{\beta}$, such as least-square, LASSO, LAR, and other methods. In the case of univariate inputs, tn the classic polynomial expansion the $p+1$ terms of $\mathbf{f}(\mathbf{x})$ would be the terms of a polynomial of degree p: $f_i(x) = x^i \quad \forall i = 0, ..., p$. For PCE, the entries of the vector $\mathbf{f}$ would be: $f_i(\mathbf{x}) = P_i(x) \quad \forall i = 0, ..., p$, where $P_i(x)$ is the Legendre polynomial of degree i. These are reported in Figure 15. Thus, for instance, in the case of polynomial regression f_3 would be x^3 while for PCE it would be $\frac{1}{2}(5x^3 - 3x)$.

n	$P_n(x)$
0	1
1	x
2	$\frac{1}{2}\left(3x^2-1\right)$
3	$\frac{1}{2}\left(5x^3-3x\right)$
4	$\frac{1}{8}\left(35x^4-30x^2+3\right)$
5	$\frac{1}{8}\left(63x^5-70x^3+15x\right)$
6	$\frac{1}{16}\left(231x^6-315x^4+105x^2-5\right)$
7	$\frac{1}{16}\left(429x^7-693x^5+315x^3-35x\right)$
8	$\frac{1}{128}\left(6435x^8-12012x^6+6930x^4-1260x^2+35\right)$
9	$\frac{1}{128}\left(12155x^9-25740x^7+18018x^5-4620x^3+315x\right)$
10	$\frac{1}{256}\left(46189x^{10}-109395x^8+90090x^6-30030x^4+3465x^2-63\right)$

Figure 15 – First 11 Legendre polynomials

2.2.6 *Artificial Neural Networks*

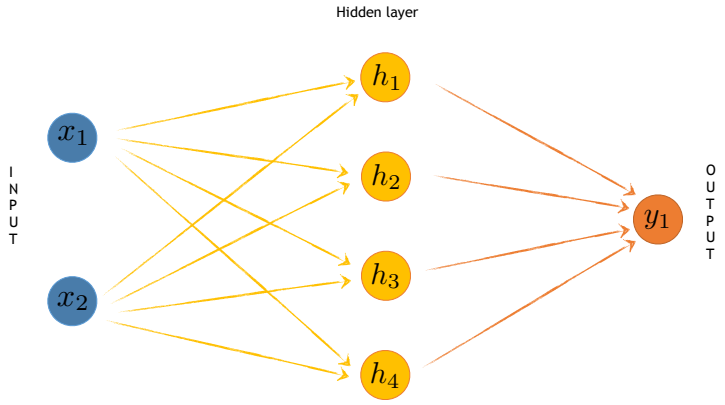


Figure 16 – Illustrative example of an Artificial Neural Network with 2-dimensional input layer, 4-dimensional hidden layer and 1-dimensional output layer. The value of h_i is given by $h_i = \sum_{j=1}^{d} f_l(x_j w_{ji}^{(l)} + b_j^{(l)})$, where d is the size of the input layer, $f_l()$ is the activation function for l-th layer, $w_{ji}^{(l)}$ is the weight from the j-th input to h_i (in layer l) and $b_j^{(l)}$ is the bias for the j-th input (in layer l).

Artificial Neural Networks (ANN) are a supervised method that have certain characteristics in common with biological neural networks and have been developed as generalization of mathematical models [4]. An ANN is characterized by its patterns of connections between neurons (architecture), its method on determining suitable values for the weights which its architecture is composed of (training or learning algorithm), its activation functions. Figure 16 is an illustrative example of an Artificial Neural Network with 2-dimensional input layer, 4-dimensional hidden layer and 1-dimensional output layer. The size of the output layer can even be greater than 1. In such a case, one ANN can be trained for multiple outputs. For instance, only one net can be trained to predict the values of the scores obtained from a data compression method altogether. In the case of Kriging,

PCE or PC-Kriging, this is not possible and the models have to be trained for each score separately.

2.3 SAMPLING STRATEGIES

The choice of the sampling strategy employed to explore a certain region $\mathcal{D}$ in the input parameter space is an important step towards the development of predictive reduced-order models. The construction of high-performing ROMs with a very limited number

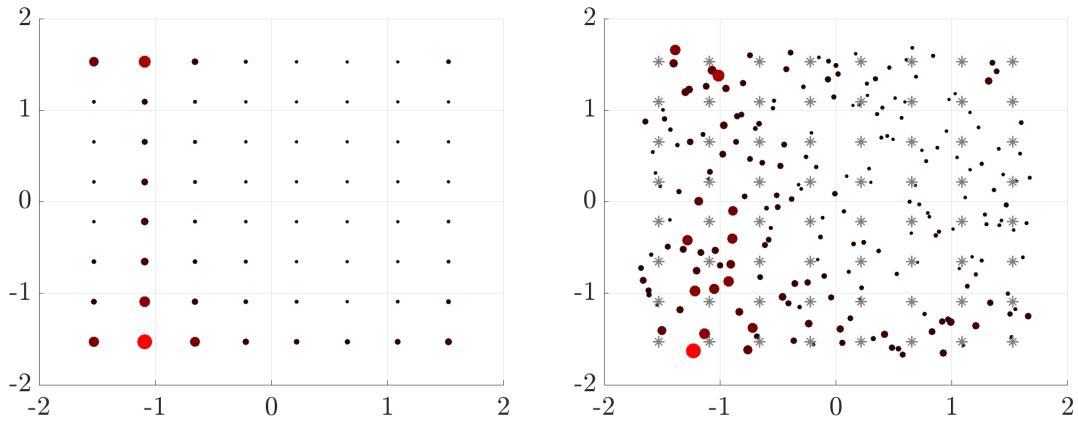


Figure 17 – Illustrative examples of an adaptive sampling strategy. (*left*) A set of 64 (uniformly distributed) points $\mathbf{X}$ in a 2-dimensional input space (samples). By analyzing the corresponding outputs $\mathbf{Y}$ (not shown here), the influence or importance of each sample can be estimated according to some metrics, here represented by the size and color of the circles. (*right*) A set of 200 candidate points in the input space for which to produce the corresponding output (run a simulation). The size and color of the circles represent the *potential of enrichment* of that sample. The original available observations from the left figure are reported here as grey stars. Notice how candidate samples that are close to the most influential original samples have a high potential.

of samples is possible if an effective sampling strategy is chosen or developed. Two main categories for data sampling may be distinguished. The first category regards sampling strategies of the input space which do not take into account the output values. Their objective is only to "fill" the explored region in the input space by following some criteria. The second category consists of sampling strategies which do take into account the values of the output variable(s), and thus their objective is to understand what regions in the input space show strong non-linear behavior in order to try to fill those regions with more samples. They are referred to as adaptive sampling strategies and exploit the knowledge on the values of the output variables and try to associate an importance or influence $\text{Infl}(\mathbf{x}_j)$ with $j = 1, \dots, m$ to each one of the m available samples before choosing the next one. Once the quantities $\text{Infl}(\mathbf{x}_j)$ are estimated, a new sample $\mathbf{x}^*$ (combination of values for the input parameters) is chosen that maximizes a certain *potential of enrichment* $\text{Pot}(\mathbf{x}^*) = d\left(\mathbf{x}^*, \mathbf{x}_j\right) \text{Infl}(\mathbf{x}_j)$, where $j = \texttt{argmin}_k \; d\left(\mathbf{x}^*, \mathbf{x}_k\right)$ and with $d(\cdot, \cdot)$ denoting the Euclidean distance. In other words, a new sample will be chosen as a trade-off between two aspects: the necessity for the new sample not to be too close to the samples already at hand and for it to be as close as possible to the sample $\mathbf{x}_j$ at hand that has the highest importance. An illustrative example of this procedure is reported in Figure 17.

2.3.1 *Random sampling*

Random sampling is perhaps the simplest sampling method. In random sampling, a subset of individuals (a sample) is chosen randomly from a larger set (a population). Each individual is chosen such that each individual has the same probability of being chosen at any stage during the sampling process, and each subset of individuals has the same probability of being chosen for the sample as any other subset of individuals. A simple random sample is an unbiased surveying technique.

2.3.2 *Latin hypercube sampling*

Latin hypercube sampling (LHS) is a statistical method for generating a near-random sample of parameter values from a multidimensional distribution. In the context of statistical sampling, a square grid containing sample positions is a Latin square only if there is only one sample in each row and each column. A Latin hypercube is the generalization of this concept to an arbitrary number of dimensions, whereby each sample is the only one in each axis-aligned hyperplane containing it.

2.3.3 *Adaptive sampling for the improvement of the modal basis*

As a PCA-based model strongly depends on its modal basis, a first step towards the improvement of this kind of model consists in improving the basis [16]. Given a set of (centered-scaled) observations $\mathbf{Y} = \{\mathbf{y}^{(1)},\ ...,\ \mathbf{y}^{(M)}\}$ and its corresponding PCA-based model, we want to choose a new sample point, $\mathbf{x}_{new}$, that will meet the trade-off between the modal basis improvement and the parametric space exploration. Firstly, the influence of each observation on the modal basis is computed. The influence of the j-th observation on the i-th mode is defined by:

$$\mathrm{Infl}_{\nu_i}(\mathbf{x}_j) = \frac{1}{|\mathbf{v}_i^T \mathbf{v}_i^{-j}|} - 1, \tag{27}$$

where $\mathbf{v}_i^{-j}$ is the i-th basis function evaluated from a data set:

$$\mathbf{Y} = \{\mathbf{y}^{(1)},\ ...,\ \mathbf{y}^{(j-1)},\ \mathbf{0},\ \mathbf{y}^{(j+1)},\ ...,\ \mathbf{y}^{(M)}\}. \tag{28}$$

The influence of the observation $\mathbf{y}(\mathbf{x}_j)$ on the modal basis is defined by:

$$\mathrm{Infl}_{\mathrm{Basis}}(\mathbf{x}_j) = \sum_{i=1}^{q} s_i \mathrm{Infl}_{\nu_i}(\mathbf{x}_j), \tag{29}$$

where s_i is the singular value of the i-th mode and q is the number of modes. The relative influence of the j-th observation on the modal basis is given by:

$$\mathrm{Infl}^{\mathrm{Rel}}_{\mathrm{Basis}}(\mathbf{x}_j) = \frac{\mathrm{Infl}_{\mathrm{Basis}}(\mathbf{x}_j)}{\sum_{l=1}^{M} \mathrm{Infl}_{\mathrm{Basis}}(\mathbf{x}_l)}. \tag{30}$$

After the computation of this equation for each $\mathbf{x}_j$, the parametric space is heavily sampled via a LHS technique and the resulting set of samples is denoted by $\mathbf{Q} = \{\boldsymbol{\nu}_1, ..., \boldsymbol{\nu}_b\}$. The size b of this set of samples can be chosen as 100 times the parametric space dimension. Then the potential of enrichment $\text{PotBasis}(\boldsymbol{\nu}_i)$ of each candidate sample is computed with respect to the trade-off between the input space exploration and the improvement of the modal basis as:

$$\text{PotBasis}(\boldsymbol{\nu}_i) = d\left(\boldsymbol{\nu}_i, \mathbf{x}_j\right) \text{Infl}^{\text{Rel}}_{\text{Basis}}(\mathbf{x}_j), \tag{31}$$

where $j = \texttt{argmin}_k \; d\left(\boldsymbol{\nu}_i, \mathbf{x}_k\right)$ with $d(\cdot, \cdot)$ denoting the Euclidean distance. Finally, the new point will be selected to fulfill the following condition:

$$\mathbf{x}_{\text{new}} = \texttt{argmax}_{\boldsymbol{\nu} \in \mathbf{Q}} \text{PotBasis}(\boldsymbol{\nu}). \tag{32}$$

A new sample is chosen as far from the other samples as possible, but at the same time as close to the samples with the highest relative influence as possible.

2.3.4 *Adaptive sampling based on prediction errors*

As sampling strategies are used to estimate the best set of samples for the training of a predictive model, the error made by the trained model when one sample or training observation is left out, also referred to as leave-one-out (LOO) error, can be used to associate an importance to each observation and thus estimate the corresponding influence $\text{Infl}(\mathbf{x}_j)$ of each sample $\mathbf{x}_j$ on the training set. After doing so and estimating the quantities $\text{Infl}(\mathbf{x}_j)$, again a new sample $\mathbf{x}^*$ is chosen so that the certain potential $\text{Pot}(\mathbf{x}^*) = d\left(\mathbf{x}^*, \mathbf{x}_j\right) \text{Infl}(\mathbf{x}_j)$ is maximized. The drawback of this technique is that, usually, training a predictive model several times can be computationally demanding as, given m available observations, m models have to be trained. For some supervised techniques, such as Kriging, training means finding suitable values for some hyper-parameters that are needed to evaluate a kernel functions for the training observations. Thus, the training can be done on all samples and then the LOO error can be estimated by using only $m-1$ training observations for the kernel. By doing so, the estimation of the sample influences is much less demanding computationally. The drawback is that, in the case of a combination between a compression method such as PCA and Kriging, not retraining also means that the PCA modes are also found only once from all samples. The LOO error estimated in this way will therefore not take into account the influence of the samples on the POD basis, which might be needed in the case of reduced-order models that combine unsupervised with supervised learning techniques.

2.4 UNCERTAINTY AND MACHINE LEARNING

Data-driven machine learning techniques build predictive models from available data. In the context of the works of this doctoral thesis, data came from *deterministic* sources such as computer codes. In general, data may also come from experiments or other sources which are not free of random noise, bias or uncertainty as in the case of experimental measurements. A machine learning method has to deal with this type of information, showing a fair degree of robustness with respect to such disturbances. Input data are subject to noise, outliers, and errors. In general, data cleaning techniques are applied in order to get the best out of the available data [49]. One can also remove instances which disproportionately increase model complexity or which have an abnormally large influence on learning. Outlier detection can also deal with this problem [55]. On occasions, input noise can have a positive effect on the generalization capabilities of a machine learning model as the method has to develop a form of invariance and has to abstract from the noise. In fact, one way to deal with random noise is by avoiding over-fitting [3].

Bias is also a problem. The bias-variance trade-off is a well-known problem in the machine learning community. The variance of the machine learning model is in fact an error due to the model's sensitivity to small fluctuations in the training set. A trained method with high variance usually means that the method is trying to model the noise in the data and, as said, avoiding over-fitting deals with this problem. There are several ways to avoid over-fitting: the choice of the training data-set, adding artificial noise to the training data, the use of regularization, the use of test data-sets and validation data-sets, the choice of a machine learning model rather than another (perhaps one with lower capacity: e.g. a linear rather than a cubic regression). However, at the same time these choices are sources of bias. In machine learning, bias can often help to generalize better and make the trained model less sensitive to a single data point. A learning algorithm will not generalize unless it introduces some form of preference or restriction over the space of possible functions [46]. Without any limitation or preference, the learning algorithm can memorize any data set without generalizing. Choosing the right learning model for the problem, choosing a representative training set, the use of different models for the construction of a meta-model (boosting) can reduce model bias.

In this doctoral thesis, the employed data are the outcome of CFD simulations implementing models that attempt to describe and predict reality. However, these models themselves may have bias. Thus, these models are usually validated with respect to experimental data, which usually have uncertainty. Probabilistic or Bayesian frameworks and therefore the possibility to focus on distributions rather than on point estimates can deal with uncertainty. For example, when training a machine learning model, optimal values of the hyper-parameters need to be found by minimizing some loss function (usually, the expected prediction error, with or without regularization terms). The loss function landscape can be strongly non-linear and local optimization methods will perform poorly. Thus, several machine learning models that are optimized from different initial guesses for their hyper-parameters can be combined by means of ensemble methods so to obtain a better

model. In general, once and however the model is trained, it would be useful if a measure could be available of how uncertain the model is when making a prediction. The more your prediction fluctuates with tiny structural changes to the model, the more uncertain that prediction is. Structural changes to the model can be achieved by slightly changing the model's hyper-parameters several times and analyze the corresponding fluctuations on the predictions. These fluctuations can also be analyzed when small changes are applied to the model's inputs, in order to estimate the model's sensitivity to noise, but, as said, this is dealt with by avoiding over-fitting and limiting the model's predictive variance. In the context of Bayesian frameworks, where posterior distributions are available over the optimal hyper-parameters' values or model's predictions, the shape of such distributions will quantify the model's uncertainty, with uniform distribution meaning maximum uncertainty and delta distribution meaning zero uncertainty.

3

RESEARCH CONTRIBUTIONS

This chapter presents a summary of the works that form the basis of this doctoral thesis. Specifically, the first two sections of this chapters summarize works that have already been published in peer-reviewed international journals.

The following sections present work that has been initiated during the present Ph.D. project, but has not yet reached a level that qualifies for publication in a scientific journal. However, more than a valuable foundation was created for future work. Therefore, a summary of these activities is provided here, promoting future work and making existing achievements accessible.

3.1 APPLICATION OF REDUCED-ORDER MODELS BASED ON PCA AND KRIGING FOR THE DEVELOPMENT OF DIGITAL TWINS OF REACTING FLOW APPLICATIONS

This article is based on the paper which was published in Computer and Chemical Engineering: Gianmarco Aversano, Aurélie Bellemans, Zhiyi Li, Axel Coussement, Olivier Gicquel, and Alessandro Parente, Application of reduced-order models based on PCA & Kriging for the development of digital twins of reacting flow applications, Computer and Chemical Engineering 121, 422-411 (2019).

The author of the present doctoral thesis has contributed to all activities presented in this paper. He has contributed to the use of the adaptive sampling strategy, to the implementation of PCA, CPCA and Local PCA and their combination with Kriging on both the 1D and 2D flames, the discussion of the main achievements and the analysis on each method's features. More specifically:

- the author has developed his own implementation of the adaptive sampling strategy and the functions needed to evaluate the quantities necessary for the evaluation of each simulation's influence, such as the ones described in Section 2.3.3;
- the author has developed his own tool for the evaluation of the CPCA scores, for any given set of non-linear constraints;
- the author has parallelized an existing implementation of Local PCA and also developed his own version;
- the author has developed a tool that combines CPCA and Local PCA;
- the author has developed a tool that combines PCA (or its variations) with Kriging, for which the ooDACE toolbox citeLophaven2002 was used and modified;
- the author has developed an object-oriented interface for data post-processing and visualization.

These tools were developed in Matlab, Fortran, and C++.

Detailed numerical simulations of detailed combustion systems require substantial computational resources, which limit their use for optimization and uncertainty quantification studies. Starting from a limited number of CFD simulations, reduced-order models can be derived using a few detailed function evaluations. In this work, the combination of Principal Component Analysis (PCA) with Kriging was chosen to identify accurate low-order models. PCA was used to identify and separate invariants of the system, the PCA modes, from the coefficients that were instead related to the characteristic operating conditions. Kriging was then used to find a response surface for these coefficients. This led to a surrogate model that allowed to perform parameter exploration with reduced computational cost. Variations of the classical PCA approach, namely Local and Constrained PCA, were also presented.

The chosen methodology for the development of accurate and robust ROM generation was demonstrated for a 1D laminar flame produced by OpenSmoke++ with an increasing number of input parameters (equivalence ratio, composition of the fuel, inlet temperature), and for a 2D flame produced by OpenFoam with two input parameters (inlet velocity and inlet fuel composition). In all three cases and for the 2D flame as well, both training and test data were available. The training data were employed to generate the ROM. The test data were used to assess the ROM's predictive capabilities.

The results showed that the combination of PCA with Kriging was a valid solution for the development of physics-based ROMs that can perform accurately with reduced computational cost. Mainly, the following points were remarked:

— The model performed parameter exploration with low prediction errors: $< 10\%$.
— The Local PCA formulation provided an improvement over PCA as it better dealt with the non-linearities of the original system.
— CPCA guaranteed that the imposed physical constraints were not violated when the data were reconstructed. In this work, the imposed constraint was that all variables be positive in value when reconstructed. This was not guaranteed by PCA, although Local PCA alleviated this issue by simply improving the accuracy of the data reconstruction.
— Once the model was correctly trained, instantaneous predictions were possible, making parameter exploration much easier, even for very CPU-intensive systems.

This work represented the first application of the PCA+Kriging methodology to combustion problems. As such, it was intended to be a proof of concept that will pave the way for the application of this methodology to more complex systems. In fact, as 3D simulations of practical combustion systems usually require significant amount of CPU hours, having a low-order model that can reliably and instantaneously predict the outcome of these simulations is precious. Moreover, the promptness of the ROM's predictions is paramount for the development of digital twins for real systems which can be employed for system control and visualization. A correctly trained ROM also grants the possibility of performing sensitivity analysis of the investigated system w.r.t. its input parameters and can be employed to solve optimization problems in the context of system design, where the evaluation of the objective function is the computational burden. The training costs of the

PCA+Kriging ROM were also lower in comparison to a predictive model with no compression. This is very useful when new training processes are continuously needed in order to update the developed ROM in the event of new available data. The predictive capabilities of the developed ROM can also be employed for the initialization of complex simulation, reducing the time needed by the solver to converge. In the application of the proposed methodology to 2D flames, relevant computational savings were present as one 2D simulation needed over 30 CPU hours to run. Despite the simplicity of the test cases, the present work allowed to investigate the advantages and limitations of the method, as well as its potential for applications to more complex combustion systems. A substantial reduction in the system dimensionality was accomplished via PCA (e.g. from $10,780$ to 5 scalars in the one-parameter case), while the use of Kriging allowed to capture the non-linear relation between the reduced set of coefficients and the input parameters, enabling the prediction of non-observed system states.

3.2 PCA AND KRIGING FOR THE EFFICIENT EXPLORATION OF CONSISTENCY REGIONS IN UNCERTAINTY QUANTIFICATION

This article is based on the paper which was published in Proceedings of the Combustion Institute: Gianmarco Aversano, Javier C. Parra-Alvarez, Benjamin Isaac, Sean T. Smith, Axel Coussement, Olivier Gicquel, and Alessandro Parente, PCA and Kriging for the efficient exploration of consistency regions in Uncertainty Quantification, Proceedings of the Combustion Institute 37, 4461-4469 (2018).

The author of the present doctoral thesis has not contributed to running the simulations for the boiler that is object of the paper, but he has contributed to all other activities, especially to the formulation of the Reduced-Order Bound-to-Bound Data Collaboration framework and the main result of the work, which is the finding the of consistency region for the LES model. A tool for the Reduced-Order B2B DC framework has also been developed by the author.

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).

In this work the Bound-to-Bound Data Collaboration (B2B-DC) framework was combined with PCA. Experimental data were available for Temperature and Heat Flux measurements for Alstom's 15 MW_{th} tangentially fired oxy-pilot Boiler Simulation Facility (BSF) situated in CT, USA. The BSF is a test facility for the development of Alstom's combustion systems. A CFD model of the BSF was also available [36, 37] but not fully defined as the value of 3 of its parameters were not known, which are indicated as T_{slag}, k and τ. The model's output could be consistent with the experimental data only if suitable values for these parameters were to be chosen. Suitable values for these parameters can in general be found using the B2B-DC approach and thus carrying out consistency analysis between the experimental data and the model's output. The available data consisted of experimental values for 121 QOIs, namely 95 Temperature and 26 Heat Flux values. A set of 22 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 surrogate models built from these simulations for each of these QOIs, for a total of 121.

In the present work, a consistency analysis was carried out using only 5 trained surrogate models. This was possible because a reduction technique such as PCA was used to compress the original data. The set of 121 original QOIs was encoded into a set of 5 scalars, namely the PCA scores, and thus only 5 surrogate models were needed for the consistency analysis. This approach was named Reduced-Order B2B-DC. This was the first time to the authors' knowledge that a B2B-DC was developed in terms of PCA scores or, more in general, of latent variables and not in terms of the original physical variables involved. Results obtained from the Reduced-Order B2B-DC approach were compared with those of a consistency analysis carried out without the use of PCA. The results showed that the Reduced-Order B2B-DC approach was able to find the consistency region with a difference in volume of about 5% if the input space is standardized and with a smaller set of variables. The advantages of the approach included computational savings as less surrogate models needed to be trained: less hyper-parameters had to be found for the construction of the needed predictive models, 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 were improved by defining suitable ranges for the 3 most influential parameters affecting the predictions.

3.3 COMBINATION OF POLYNOMIAL CHAOS AND KRIGING FOR REDUCED-ORDER MODEL OF REACTING FLOW APPLICATIONS

The work presented in this section has not reached a level that qualifies for publication yet, but its achievements are reported here. The work is accessible in more detail in Chapter 5.

The author of the present doctoral thesis has contributed to all activities carried out in the this work, such as the implementation of the unsupervised and unsupervised methods object of the work and their combination, and the analysis on the strengths and weaknesses of each method. The author of the present doctoral thesis has not contributed to running the simulations of the 2D flames on which the methods are tested. In particular, the author of the present doctoral thesis has updated the tools developed for the work described in Section 3.1 and developed his own tool for the implementation of Polynomial Chaos Expansion and PC-Kriging.

In this work, inspired by [1] and [45], two data compression techniques such as Proper Orthogonal Decomposition (POD) and Non-negative Matix Factorization (NMF) were combined with a predictive model based on the combination of Polynomial Chaos Expansion (PCE) and Kriging for the development of a Reduced-Order Model (ROM) for the prediction of combustion data, with PCE functioning as Kriging trend. In order to compare the performance of the PC-Kriging method, POD and NMF were also combined with PCE only and Kriging only.

As regards the unsupervised techniques used for data compression for the development of the ROM, the results showed that POD could reconstruct the training data with an NRMSE which was ≈ 10 times lower with respect to NMF. On the other hand, the positivity constraint imposed in the NMF formulation guaranteed that positive physical quantities were not reconstructed with negative values. Both methods generalized to unseen data with similar performances in terms of errors for data reconstruction, with POD having reconstruction errors $\approx 10\%$ lower.

As concerns the supervised part of the ROM, results from a leave-k-out analysis showed that the PC-Kriging interpolation performed with lower prediction errors than Kriging for smaller training sizes and low approximation orders. Both PC-Kriging and Kriging outperformed PCE, when these techniques were combined with POD. PC-Kriging performed better (prediction errors lower by $\approx 10\%$) than Kriging also when in combination with NMF. In general, the POD-based ROM performed with lower prediction errors than the NMF-based ROM, by a factor of ≈ 10 on CH_4 and N_2. The PC-Kriging method performed better than Kriging also when the Kriging length-scales were not optimized and set to low values, so that the Kriging would influence predictions only in the regions close to the training data (and interpolate those data), while PCE would contribute in the regions far from the training data. By doing so, the overall prediction errors were higher than the case where the Kriging length-scales were optimized, although the training of such a model was computationally less demanding as it did not involve the optimization procedure for

the evaluation of the length-scales and thus it only involved the evaluation of the PCE coefficients.

The use of ANN as supervised part for a POD-based ROM was also investigated as ANN offers the possibility of training one model for the prediction of all the POD (or NMF) scores, simultaneously. The predictive capabilities of combining POD with ANN were compared to the ones of POD in combination with PC-Kriging. Results showed that, again, POD combined with PC-Kriging had lower prediction errors with respect to the combination of POD and ANN, by a factor of ≈ 10 for nearly all variables involved, as ANN usually requires a high number of observations to perform well, which was not the case in the present work where data from computationally expensive combustion simulations were employed. The training of an ANN also meant that a wide range of design choices for its architecture, such as number of hidden layers or type of activation functions, needed to be explored or cross-validated, which made ANN a more complex choice as supervised method.

3.4 DIGITAL TWIN FOR MILD COMBUSTION FURNACE

The work presented in this section has not reached a level that qualifies for publication yet, but its achievements are reported here. The work is accessible in more detail in Chapter 5.

The author of the present doctoral thesis has contributed to all the activities that were needed for the development of the digital twin discussed in the work, such as the combination of POD with Kriging, the estimation of the importance of each simulation of the MILD combustion furnace on the developed reduced-order model, the use of the leave-k-out to understand the influence of the data-set size and the estimation of the performances of the developed digital twin. The tools developed for the previous works were adapted for this case. A tool for data post-processing and visualization was developed by the author for the specific needs of the case and the tools for the leave-k-out analysis were developed.

In the present work, a reduced-order model (ROM) based on the combination of Proper Orthogonal Decomposition (POD) and Kriging was developed for the prediction of a 3D flame (prediction of the full fields of temperature and main chemical species) in a furnace for different operating conditions, chosen as the fuel composition, which was a mixture of methane and hydrogen, the value of the equivalence ratio and the length of the air injector, with the objective of the present work being the development of a reliable model that can predict the state of the considered combustion system and important scalar quantities such as wall temperature, exhaust gas composition and flame length, within an accuracy of 10%, in real-time and function as digital-twin for it. A set of simulations was produced and used as training data for the development of the ROM. The influence of each simulation on the reduced basis found by POD was estimated, so to identify the most important simulations to retain as training data. Interestingly, this analysis was able to detect a change in the physics of the system that is observed when going from low values of H_2 in the inlet fuel mixture to high ones. The influence of the number of training simulations used for the development of the ROM was also investigated. POD was used for data compression and thus to represent the original data with a reduced number of features, the POD scores. Kriging was used to find a response surface for these scores at unexplored operating conditions. As the mapping from the reduced space to the original space was learned by POD, also the full fields of temperature and main chemical species could be predicted. A leave-k-out analysis was carried out in order to determine how many and which simulations were needed for the training of the ROM and estimate how the developed ROM would generalize to unseen data. Results showed that the developed ROM could predict the fields of temperature and CO_2, O_2, H_2O, CH_4 mass fractions reliably with an overall error of less than 10%. The predictions of scalar quantities such as wall temperature, position of the peak of OH mass fraction, exhaust gas composition and flame length were also obtained with an error of less than 5%. As a consequence, the overall behavior of the considered system was well reproduced by the ROM, suggesting that the proposed methodology for ROM-development was valid and that the developed ROM could be used as a digital-twin of the furnace for real-time predictions of its state when the operating conditions change.

3.5 FEATURE EXTRACTION IN COMBUSTION APPLICATIONS

The work presented in this section has not reached a level that qualifies for publication yet, but its achievements are reported here. The work is accessible in more detail in Chapter 5.

The present article is based on a work which was carried out in collaboration with other researchers. The author of the present thesis has contributed to the application of unsupervised learning techniques such as Non-negative Matrix Factorization (NMF), Kernel PCA, and the Procrustes analysis needed to compare results from Local PCA and Autoencoder.

The present work investigated the application of data-driven techniques for the detection of features from turbulent combustion data sets (direct numerical simulation). Two H_2/CO flames were analyzed: a spatially-evolving (DNS1) and a temporally-evolving jet (DNS2). Methods such as Principal Component Analysis (PCA), Local Principal Component Analysis (LPCA), Non-negative Matrix Factorization (NMF) and Autoencoders were explored for this purpose. It was shown that various factors could affect the performance of these methods, such as the criteria employed for the centering and the scaling of the original data or the choice of the number of dimensions in the low-rank approximations. A set of guidelines was presented in this paper that can aid the process of identifying meaningful physical features from turbulent reactive flows data.

PCA-based techniques were firstly explored, such as Global PCA, Local PCA and Kernel PCA. The effect of the various scaling methods on the feature extraction capability of PCA was assessed. Local PCA was shown to be able to automatically detect the most important features of the data-set. As the mixture fraction was known to be an important feature for DNS1, it was remarkable to see that Local PCA could partition the data into zones of low and high values of mixture fraction even if this variable was not present in the data analyzed by Local PCA. Besides, the locally-linear nature of Local PCA , in contrast with the globally linear nature of PCA, meant that the method could better approximate the original data manifold, confirmed by the lower reconstruction errors obtained for a given number of retained principal components, and thus it could detect features that only had a local importance in the original data manifold, that were not detected by global PCA, such as O_2 branching reactions. The interpretability of the factors extracted with the Non-negative Matrix Factorization (NMF) technique showed to be a more challenging task. Due to the NMF positive constraint, some features, such as mixture fraction, could not be represented by a single mode, while this was possible with PCA. Non-linear techniques such as Kernel PCA and Autoencoders were also tested for feature extraction purposes. Kernel PCA could detect features such as mixture fraction or the reactive layer. The Autoencoder was used to find a lower two-dimensional manifold in the original data space, which proved to be a reliable choice to separate well the reactants on opposing ends of the manifold and the reactive layer in-between, indicating that such manifold could be linked to a reaction progress variable. Besides, the manifold detected by a locally-linear method

such Local PCA was shown to converge to the one found by a non-linear method such as the Autoencoder.

In parallel with detecting features, the potential of the mentioned techniques for data compression purposes, thus the quality of reconstruction of the original data with a fewer number of dimensions, was also assessed, as achieving both tasks well is a decisive factor for selecting a given data-driven method for accurate modeling of combustion systems. It was shown that Local PCA was capable of achieving excellent reconstruction with $q = 2$ principal components when the number of clusters was sufficiently high. Autoencoders performed comparably well when used with the ELU activation function. The reconstruction errors associated to these two methods were lower for Local PCA only when a high (128) number of clusters was used, showing the ability of the method to more easily approximate non-linear manifolds in the original data space as, to do so, the user would just need to change, and more specifically further increase, one free parameter, while decreasing the reconstruction errors of an Autoencoder would require the investigation of a wider range of design choices. However, it is true that using a very high number of clusters had some drawbacks: data interpretability became more complex as the number of features to analyze increased (e.g., 2 features in each one of the 128 clusters); a high number of matrices, one for each cluster, containing the local PCA modes had to be stored.

A new way to cluster data-sets based on the possibility to a priori decide what features to find in it was also presented and named Feature-assisted Clustering.

Future work will include the application of different data-driven methods in the feature detection approach as well as generalization of the findings presented in this paper to a broader range of turbulent combustion data sets.

4

CONCLUDING REMARKS

The present Chapter provides a brief discussion of the main accomplishments and original contributions of the present doctoral thesis.

A methodology has been proposed for the development of physics-based reduced-order models (ROMs) for reacting flow applications based on the combination of unsupervised methods such as Principal Component Analysis (PCA), also knwon as Proper Orthogonal Decomposition (POD), or variations of it, namely Local PCA and Constrained PCA, with supervised techniques such as Kriging, with the objective of developing digital twins for reacting flow applications. The use of other unsupervised (Non-negative Matrix Factorization (NMF)) and supervised (Polynomial Chaos Expansion (PCE)) techniques was also investigated. The work carried out in the present Thesis represented the first application of the PCA+Kriging methodology to combustion problems as well as the first application of Local PCA, Constrained PCA and NMF in combination with Kriging. As such, it is intended to be a proof of concept that will pave the way for the application of this methodology to more complex systems and the development of digital twins for industrial applications.

4.1 REDUCED-ORDER MODELS FOR THE DEVELOPMENT OF DIGITAL TWINS

The methodology was firstly demonstrated for 1D laminar flames with an increasing number of input parameters (equivalence ratio, composition of the fuel, inlet temperature), and for a 2D flame with two input parameters (inlet velocity and inlet fuel composition). The results showed that the combination of PCA with Kriging could perform accurately. In comparison to a surrogate model that does not make use of an unsupervised technique for data compression, the training costs of the developed model were lower, which is very useful when new training processes are continuously needed in order to update the developed ROM in the event of new available data.

The use of alternative unsupervised and supervised methods was also investigated. In particular, NMF was tested as alternative data compression method and Polynomial Chaos Expansion (PCE), PC-Kriging and Artificial Neural Networks (ANNs) as alternative supervised methods. The methodology was demonstrated again for 2D flames with two input parameters (inlet velocity and inlet fuel composition). A small summary is provided here about each method's strengths and weaknesses:

Unsupervised methods

— PCA/POD proved to be the easiest method at providing linear low-rank approximations based on uncorrelated components. Its strengths include small reconstruction errors for the training set, orthogonal modes or *eigenflames*, which may be preferable, and a small number of parameters to control (choice of centering and scaling). However, no physical criterion is taken into account by the method. Although the method showed to be able to detect variables not present in the data-set, such as mixture fraction, and offers an explicit link between the original and the reduced space, the physical interpretability of the PCA basis functions is not always straightforward.
— CPCA/CPOD could force the spatial fields reconstructed by POD to respect given physical laws such as positivity of mass fractions or conservation of mass. However, this is possible only by solving a constrained optimization problem that may require high computational times. The CPCA coefficients lie within a feasible regions determined by the constrained optimization problem that led to their evaluation. When combined with a supervised technique for the prediction of the CPCA coefficients, there is no guarantee that the supervised method will not predict values for these coefficients that lie outside of such a feasible region.
— Local PCA/POD finds local *eigenflames* and has the potential of achieving lower reconstruction errors than POD. The weaknesses of the method in the context of developing physics-based ROM were that the clustering times may be long, and that the local eigenflames showed strong discontinuities at the borders connecting them.
— NMF is a natural choice for non-negative data such as the chemical state space. When combined to a supervised method, when negative values were predicted for the NMF coefficients, these values could easily be *corrected* by manually setting them to zero. The method was computationally more expensive than other linear counterparts such as POD, and results showed that it performed with higher reconstruction error on test data. NMF detected different features from POD, both when applied to 2D flames (snapshot matrix), or to a DNS data-set, where it failed to detect mixture fraction.

Supervised methods

— Kriging is an interpolation method that provides a distribution over possible output values rather than only one value as output. It can incorporate external knowledge in its prior and choice of kernels. However, it can be memory and computationally expensive for large data-sets and its training needs the solution of an non-linear optimization problem. Results showed that the method was sensitive to design choices such as degree of the trend function and choice of kernel function. In the application to 1D flames and in combination with PCA/POD, the Gaussian kernel performed 10 times more poorly than the other kernels.

— PCE is a regression method and is relatively cheaper to train in comparison to Kriging. Results showed that the total degree of the polynomial basis and the degree of polynomial interactions were the main parameters which the model was sensitive to. Results showed that, when combined with POD or NMF, its predictive capabilities were limited w.r.t. Kriging.
— PC-Kriging was combined with a POD and NMF for the prediction of 2D flames, but it usually exhibited higher prediction errors than Kriging. A leave-k-out analysis on the prediction errors by a POD-based ROM showed that PC-Kriging with $N = 5$ could achieve lower prediction errors than Kriging for small sizes of the training set ($k > 15$). When the Kriging hyper-parameters were not optimized and set to low values, the model offered two main advantages: it could predict test data with the same accuracy of PCE while being able to interpolate training data; its training costs were lower than Kriging's.
— ANN was considered as it offers the possibility to train one model or net for the prediction of all targets, namely the POD or NMF coefficients, simultaneously. Results showed that in the context of predicting spatial fields of combustion systems where only a small number of training simulations can be provided, the method performed poorly. Besides, the method requires several design choices to be validated for its architecture.

Development of a digital twin for a MILD combustion surface

A ROM based on the combination of POD and Kriging was developed for the prediction of a 3D flame (prediction of the full fields of temperature and main chemical species) in a furnace for different operating conditions, chosen as the fuel composition, which was a mixture of methane and hydrogen, the value of the equivalence ratio and the length of the air injector, with the objective of the present work being the development of a reliable model that can predict the state of the considered combustion system and important scalar quantities such as wall temperature, exhaust gas composition and flame length, within an accuracy of 10%, in real-time and function as digital-twin for it. A leave-k-out analysis was carried out in order to determine how many and which simulations were needed for the training of the ROM and estimate how the developed ROM would generalize to unseen data. Results showed that the developed ROM could predict the fields of temperature and CO_2, O_2, H_2O, CH_4 mass fractions reliably with an overall error of less than 10%. The predictions of scalar quantities such as wall temperature, position of the peak of OH mass fraction, exhaust gas composition and flame length were also obtained with an error of less than 5%. As a consequence, the overall behavior of the considered system was well reproduced by the ROM, meaning that it could be used a digital-twin of the furnace for real-time predictions of its state when the operating conditions change.

4.2 UNCERTAINTY QUANTIFICATION

In the present Thesis, the consistency analysis from the Bound-to-Bound Data Collaboration (B2B DC) framework was also carried out by making use of a reduction technique such as PCA in order to compress the original data: a set of original QOIs was encoded into a much smaller set of scalars or features extracted by PCA. A reduced number of surrogate models was thus necessary for the consistency analysis. The approach was named Reduced-Order Bound-to-Bound Data Collaboration (Reduced-Order B2B DC). This was the first time to the author's knowledge that a B2B-DC was developed in terms of principal components or, more in general, of extracted features and not in terms of the original physical variables involved. Results obtained from the Reduced-Order B2B-DC approach were compared with those of a consistency analysis carried out without the use of PCA and the results showed that the Reduced-Order B2B-DC approach was able to find the consistency region with a smaller set of variables. Computational savings were one of the advantages of the developed approach as less optimization problems had to be solved for the estimation of the hyper-parameters of the needed surrogate models, which is very often not a simple task, especially when the number of input parameters is high.

4.3 ANALYSIS OF DNS DATA-SETS

The application of data-driven techniques for the detection of features from turbulent combustion data sets (direct numerical simulation) was also investigated during the present Thesis. Two H_2/CO flames were analyzed: a spatially-evolving (DNS1) and a temporally-evolving jet (DNS2). Methods such as Principal Component Analysis (PCA), Local Principal Component Analysis (LPCA), Non-negative Matrix Factorization (NMF) and Autoencoders were explored for this purpose. PCA-based techniques were firstly explored, such as Global PCA, Local PCA and Kernel PCA. Local PCA was shown to be able to automatically detect the most important features of the data-set. As the mixture fraction was known to be an important feature for DNS1, it was remarkable to see that Local PCA could partition the data into zones of low and high values of mixture fraction even if this variable was not present in the data analyzed by Local PCA. Besides, the locally-linear nature of Local PCA, in contrast with the globally linear nature of PCA, meant that the method could better approximate the original data manifold, confirmed by the lower reconstruction errors obtained for a given number of retained principal components, and thus it could detect features that only had a local importance in the original data manifold, that were not detected by global PCA. The interpretability of the factors extracted with the Non-negative Matrix Factorization (NMF) technique showed to be a more challenging task. Due to the NMF positive constraint, some features, such as mixture fraction, could not be represented by a single mode, while this was possible with PCA. Non-linear techniques such as Kernel PCA and Autoencoders were also tested for feature extraction purposes. Kernel PCA could detect features such as mixture fraction or the reactive layer. The Autoencoder was used to find a lower two-dimensional manifold in the original data space,

which proved to be a reliable choice to separate well the reactants on opposing ends of the manifold and the reactive layer in-between, indicating that such manifold could be linked to a reaction progress variable. Procrustes analysis was used to show that LPCA and the Autoencoder detected similar lower dimensional manifolds.

In parallel with detecting features, the potential of the mentioned techniques for data compression purposes, thus the quality of reconstruction of the original data with a fewer number of dimensions, was also assessed, as achieving both tasks well is a decisive factor for selecting a given data-driven method for accurate modeling of combustion systems. It was shown that Local PCA was capable of achieving excellent reconstruction with $q = 2$ principal components when the number of clusters was sufficiently high. Autoencoders performed comparably well when used with the ELU activation function. The reconstruction errors associated to these two methods were lower for Local PCA only when a high (128) number of clusters was used, showing the ability of the method to more easily approximate non-linear manifolds in the original data space as, to do so, the user would just need to change, and more specifically further increase, one free parameter, while decreasing the reconstruction errors of an Autoencoder would require the investigation of a wider range of design choices. A new way to cluster data-sets based on the possibility to a priori decide what features to find in it was also presented and named Feature-assisted Clustering.

5

SELECTED PUBLICATIONS

5.1 APPLICATION OF REDUCED-ORDER MODELS BASED ON PCA AND KRIGING FOR THE DEVELOPMENT OF DIGITAL TWINS OF REACTING FLOW APPLICATIONS

Application of Reduced-Order Models based on PCA & Kriging for the development of digital twins of reacting flow applications

Gianmarco Aversano[a,b,c], Aurelie Bellemans[a,b], Zhiyi Li[a,b], 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 91190 Gif-sur-Yvette, France*

Abstract

Detailed numerical simulations of detailed combustion systems require substantial computational resources, which limit their use for optimization and uncertainty quantification studies. Starting from a limited number of CFD simulations, reduced-order models can be derived using a few detailed function evaluations. In this work, the combination of Principal Component Analysis (PCA) with Kriging is considered to identify accurate low-order models. PCA is used to identify and separate invariants of the system, the PCA modes, from the coefficients that are instead related to the characteristic operating conditions. Kriging is then used to find a response surface for these coefficients. This leads to a surrogate model that allows performing parameter exploration with reduced computational cost. Variations of the classical PCA approach, namely Local and Constrained PCA, are also presented. This methodology is demonstrated on 1D and 2D flames produced by OpenSmoke++ and OpenFoam, respectively, for which accurate surrogate models have been developed.

Keywords: Principal Component Analysis, Kriging, Surrogate Models

1. Introduction

In many engineering applications, complex physical systems can only be described by high-fidelity expensive simulations. Due to the non-linearity of these problems, changing the operating conditions, namely the model's input parameters, can drastically change the state of the considered system. Complete knowledge about the investigated system's behavior for a full range of operating conditions can therefore only be achieved by running these expensive simulations several times with different inputs, until enough observations of the system's state are obtained.

In this study, we focus on combustion systems that fall in this category as they are characterized by very complex physical interactions, between chemistry, fluid-dynamics and heat transfer processes. Our objective is to develop advanced Surrogate Models (SMs) that can accurately represent the behavior of complex reacting systems in a wide range of conditions, without the need for expensive CFD simulations. This is particularly attractive for the development of digital counterparts of real systems, with

application in monitoring, diagnostics and prognostics [1, 2]. To this purpose, we derive techniques from the Machine Learning community.

In our approach a specific computationally-expensive CFD simulation or computer code, referred to as Full-Order Model (FOM) [3, 4], is treated as a black box that generates a certain output $\mathbf{y}$ (e.g. the temperature field) given a set of input parameters $\mathbf{x}$ (e.g. the equivalence ratio) and indicated by $\mathcal{F}(\cdot)$:

$$\mathbf{y} = \mathcal{F}(\mathbf{x}). \tag{1}$$

The evaluation of the function $\mathcal{F}(\cdot)$ usually requires many hours of computational time. After enough observations of the FOM's output are available, $\mathbf{y}(\mathbf{x}_i) \quad \forall i = 1, ..., M$, a SM can be trained and the output $\mathbf{y}^*$ for a particular set of unexplored inputs $\mathbf{x}^*$ can be predicted without the need to evaluate $\mathcal{F}(\mathbf{x}^*)$ and, thus, no simulation is run. The function $\mathcal{F}(\cdot)$ is therefore approximated by a new function $\mathcal{M}(\cdot)$ whose evaluation is very cheap compared to $\mathcal{F}(\cdot)$:

$$\mathbf{y}^* = \mathcal{F}(\mathbf{x}^*) \approx \mathcal{M}(\mathbf{x}^*). \tag{2}$$

SMs are mathematical models based on available data that try to approximate the underlying *hidden* relationship between input and output. A very simple example of a SM is a linear regression of available data. SMs are useful when this relationship is either not known or comes in the form of a computationally expensive computer code. This is the case of a CFD simulation. SMs are constructed or *trained* from a relatively small set of *training* observations of the modelâĂŹs output, which correspond to a set of training locations or points in the model parameter space. Once trained, SMs allow for a fast evaluation of the system's state over a wide range of their input parameters. Therefore, they are very appealing in the context of optimization studies as well as for Uncertainty Quantification (UQ) [5] and global optimization problems [6, 7]. In [8], SMs are used to optimize the performance of chemical kinetics with respect to MILD combustion. In [9], SMs are employed in a Bayesian approach to calibrate various neutronics and thermal-hydraulics parameters. SMs are also used in the AlgoRithms for Global Optimization of coNstrAined grey-box compUTational problems (ARGONAUT) framework [10], which was also utilised in [11] for the optimization of the operation of an oilfield using water-flooding. Ideally, the trained SMs should preserve the physics of the investigated phenomena, and be developed from a limited number of expensive function evaluations, i.e. CFD simulations. Examples of SMs are Radial Basis Functions and Polynomial Chaos Expansion [12]. Examples of SMs used in combustion applications can be found in [13].

SMs are generally constructed directly on the analyzed system's output, i.e. directly on the variables of interest like the velocity and temperature fields. For each individual output variable a SM is trained and a response surface is found, indicating the relationship between the variable and the input parameters. If the number of variables of interest is high, many SMs need to be trained. Besides, any correlation between these variables of interest might be lost in the process of training individual SMs: the information about the physics of the phenomena involved is lost. Reducing the number of SMs to train is possible if the original set of variables can be represented by a new set of fewer scalars. This

corresponds to the idea that the original variables are actually realization of unknown *latent* variables [14].

Principal Component Analysis (PCA) [15] offer the potential of preserving the physics of the system while reducing the size of the problem. PCA is a statistical technique used to find a set of orthogonal low-dimensional basis functions, called Principal Components (PCs), to represent an ensemble of high-dimensional data. PCA finds a new, smaller set of uncorrelated variables, often referred to as *PCA scores*, which is representative of the original variables of interest. PCA is also used for data interpretation, usually combined with rotation methods [16]. Once these PCA scores are found, a SM can be built for each one of them. They are indicated as Reduced-Order (Surrogate) Models (ROMs).

SMs usually include interpolation or regression techniques which depend on the choice of some particular design functions. These design functions are defined by a set of so-called hyper-parameters (or also length-scales) whose values 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. As shown in [17], ROMs are less sensitive to the particular design functions chosen for their construction, which is desirable. ROMs also have a reduced number of variables for which a SM needs to be trained. This means that fewer optimization problems are solved in order to estimate feasible values for the hyper-parameters of the design functions. In addition, in [17] it is also shown how ROMs usually scale better than classic SMs for parallel computing. These features are what makes PCA-based ROMs very attractive candidates for the development of physics-preserving SMs.

Combustion problems are well-known for being characterized by a set of strongly inter-dependent variables. In fact, PCA has been employed in [18, 19] to re-parameterize the thermo-chemical state of a reacting system by a small number of progress variables, drastically reducing the number of transport equations to solve, and in the process showing the intrinsic lower-dimensionality of these systems, which will be exploited in the present work. PCA has also been employed in the context of turbulent combustion in [20], for the a-posteriori validation of a turbulent combustion model based on the solution of transport equations for the principal components [21] and for on-line process monitoring and fault diagnostics [22].

The objective is to develop advanced SMs, trained on a reduced number of full simulations, able to predict the full system state in unexplored conditions, without running a new simulation. To this end, an approach based on the combination of PCA and Kriging was chosen. PCA was used for dimensionality reduction, thus to extract the invariant (w.r.t. the input parameters) physics-related information of an investigated combustion system and identify the system's coefficients which instead depend on the operating conditions, the PCA scores. The Kriging interpolation method was then able to find a response surface for these scores. With this strategy it was possible to build a ROM that granted the possibility of parameter exploration with reduced computational cost. The Kriging interpolation method was chosen over other regression techniques not only because Kriging provides a distribution for the prediction value, rather than just one value as the prediction, thus also capturing the model's uncertainty, but especially because it allows the user to add prior knowledge on the model by selecting different kernel functions. The use of Kriging for Computational Fluid Dynamics (CFD) data has also produced encour-

aging results. In fact, Kriging was employed for the shape optimization of a car engine intake port in [23] and for aerodynamical shape optimization problems as shown in [24, 25]. However, the application was limited to non-reacting flow problems.

In the present work, the Kriging-PCA approach is extended to combustion applications, to develop a ROM that can faithfully reproduce the temperature and chemical species fields in a reacting flow simulation. The methodology is demonstrated on a methane laminar premixed flame, increasing the complexity of the problem gradually, increasing the number of input parameters (equivalence ratio, inlet temperature and fuel composition), and the problem dimensionality, going from one to two-dimensional flames. The objective of the present work is to demonstrate the applicability of the proposed methodology for the development of reduced-order models of multi-scale and multi-physics computer models. In this perspective, this work paves the way for the development of digital twins [26] of realistic engineering systems. The methodology outlined in the present work shows the advantages of the PCA+Kriging formulation in terms of predictive capabilities and computational efficiency. Indeed, these features are necessary for the development of predictive models of engineering systems, which can be employed for visualization, real-time control, optimization and troubleshooting.

From the application perspective, this paper presents the first application of the Kriging-PCA methodology to combustion problems. From the methodology point of view, this paper presents the first application of Local PCA and of Constrained PCA in combination with Kriging.

This paper is organised as follows. Section 2 will cover the employed methodology in details, while in section 3, PCA and its variations, Kriging and a sampling strategy referred to as Adaptive Sampling are shown. Results are presented and discussed in section 4. In section 5 conclusions are drawn.

2. Methodology

The methodology used in the present work is sketched in figure 2. Consider that a certain high-fidelity simulation model or Full Order Model (FOM) $\mathbf{y}(\mathbf{x}) = F(\mathbf{x}) \in \mathbb{R}^N$ is available, such as a CFD-combustion solver. For one value of the input parameter(s) $\mathbf{x}$, the solver returns a vector $\mathbf{y}^{(j)}$ of observations of all the involved physical variables at every grid point:

$$\mathbf{y}^{(j)} = [T(r_1, \mathbf{x}_j), \ldots, T(r_L, \mathbf{x}_j), Y_{CH_4}(r_1, \mathbf{x}_j), \ldots, Y_{CH_4}(r_L, \mathbf{x}_j), \ldots]^T, \tag{3}$$

where L is the total number of grid points, r_i is the i-th spatial location and $\mathbf{x}_j$ is the j-th point in the input parameter space. This FOM is solved for a limited amount $M < N$ of training points in the input parameter space $\mathbf{X} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, ..., \mathbf{x}^{(M)}\} \in \mathcal{D}$, where $\mathcal{D}$ is the region spanned by the training points. Thus, only M simulations are available, one for each of those points: $\mathbf{Y} = \{\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, ..., \mathbf{y}^{(M)}\}$. The full exploration of the region $\mathcal{D}$ is possible only by running the expensive CFD-combustion solver $F(\cdot)$ for every $\mathbf{x} \in \mathcal{D}$. From the data-set $\mathbf{Y}$ of run simulations, PCA is able to extract a set of basis functions $\mathbf{\Phi} = \{\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, ..., \boldsymbol{\phi}_q\}$, with $q < N$ usually, called PCA modes that are invariant with respect to the input parameters $\mathbf{x}$. A set of coefficients $\mathbf{a}(\mathbf{x}) = \{a_1(\mathbf{x}), a_2(\mathbf{x}), ..., a_q(\mathbf{x})\}$, called PCA scores and depending on $\mathbf{x}$, is consequently found. An illustrative example is reported in Figure 1, where a temperature spatial

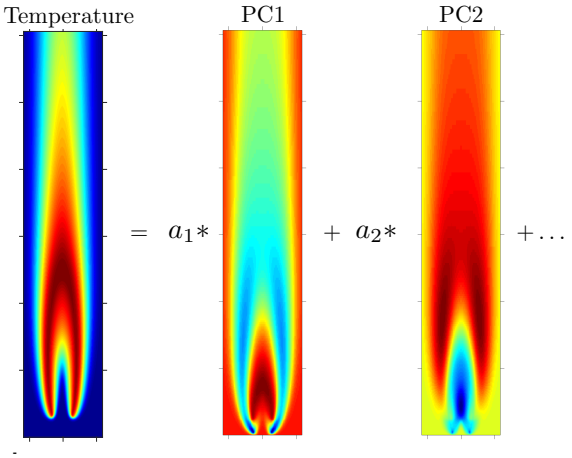


Figure 1: Illustrative example of PCA applied to combustion data: one particular temperature spatial field is represented by a set of coordinates (the coefficients a_i, called PCA scores) on the basis functions or Principal Components (PCs) found by PCA.

$$\begin{bmatrix} \mathbf{y}^{(1)} = \mathcal{F}(\mathbf{x}^{(1)}) \\ \mathbf{y}^{(2)} = \mathcal{F}(\mathbf{x}^{(2)}) \\ \cdots \\ \mathbf{y}^{(M)} = \mathcal{F}(\mathbf{x}^{(M)}) \end{bmatrix} \rightarrow \text{PCA} \rightarrow \begin{bmatrix} a_1^{(1)},\ \ldots,\ a_q^{(1)} \\ a_1^{(2)},\ \ldots,\ a_q^{(2)} \\ \cdots \\ a_1^{(M)},\ \ldots,\ a_q^{(M)} \\ \boldsymbol{\Phi} = (\boldsymbol{\phi}_1,\ \boldsymbol{\phi}_2,\ \ldots,\ \boldsymbol{\phi}_q) \end{bmatrix} \longrightarrow \begin{array}{c} \text{Kriging response} \\ \text{surface for the} \\ \text{PCA scores} \end{array}$$

Figure 2: PCA finds the set of Principal Directions $\boldsymbol{\Phi} = (\boldsymbol{\phi}_1,\ \boldsymbol{\phi}_2,\ ...,\ \boldsymbol{\phi}_q)$ and encodes each observation $\mathbf{y}^{(i)} \in \mathbb{R}^N$ into a small set of scalars $a_1^{(i)},\ ...,\ a_q^{(i)}$ for $q < N$. A Kriging response surface is then found for these scalars.

field is represented as a set of coefficients that weight a set of basis functions, i.e. the PCs. These coefficients are less in number than the original number of variables as $q < N$ and can be interpolated in order to acquire knowledge about the system's state for any unexplored point $\mathbf{x}^* \in \mathcal{D}$.

One advantage of this approach is that a much smaller number of variables, namely q PCA scores, are interpolated instead of N. Another advantage is that the N original variables might be correlated. The application of PCA for the detection of latent variables, the PCA scores, preserves this correlation, which might be lost if each original variable is interpolated independently. One additional remark is that considering, for example, $T(r_i, \mathbf{x}_j)$ and $T(r_j, \mathbf{x}_j)$ as two separate variables (rather than using the spatial locations r_i as additional input parameters) also reduces the computational costs, as we shall see.

3. Theory

3.1. PCA

The key idea of Principal Component Analysis (PCA) is to reduce (compress) a large number of interdependent variables (i.e. independent up to the second-order statistical moments) to a smaller number of uncorrelated variables while retaining as much of the original data variance as possible [27, 28, 29, 30, 31].

For a data-set $\mathbf{Y}(N \times M)$, containing M observations of N original variables, namely temperature and species mass fractions measured at each spatial location of a considered geometrical domain, as described in Section 2, 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. In general, $q \leq min(N, M)$. Thus, the vector $\mathbf{y} \in \mathbb{R}^N$ of observed temperature and species mass fractions can be encoded into a lower dimensional vector, $\mathbf{a} \in \mathbb{R}^q$.

Data are usually centered and scaled before PCA is carried out. Here we report six different choices for the scaling of the data:

1. Auto-scaling (STD), each variable is normalized by its standard deviation;
2. RANGE, each variable is normalized by its range;
3. PARETO, each variable is scaled by the square root of its standard deviation;
4. VAST, each variable is scaled by the standard deviation and coefficient of variation;
5. LEVEL, each variable is normalized by the mean of the data;
6. MAX, each variable is scaled by its maximum value.

The 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. A matrix containing the mean of each of the N variables, namely $[T(r_1, \mathbf{x}_j),\ \ldots,\ T(r_L, \mathbf{x}_j),\ Y_{CH_4}(r_1, \mathbf{x}_j),\ \ldots,\ Y_{CH_4}(r_L, \mathbf{x}_j),\ \ldots]$ over the M observations is indicated by $\bar{\mathbf{Y}}(N \times M)$.

$$\bar{\mathbf{Y}} = \begin{bmatrix} \bar{y}_1 & \bar{y}_1 & \bar{y}_1 & \ldots & \bar{y}_1 \\ \bar{y}_2 & \bar{y}_2 & \bar{y}_2 & \ldots & \bar{y}_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \bar{y}_N & \bar{y}_N & \bar{y}_N & \ldots & \bar{y}_N \end{bmatrix} \quad \mathbf{D} = \begin{bmatrix} d_1 & 0 & 0 & \ldots & 0 \\ 0 & d_2 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & d_N \end{bmatrix}. \tag{5}$$

After centering and scaling the data, the covariance matrix $\mathbf{C}$ is evaluated as:

$$\underset{(N \times N)}{\mathbf{C}} = \frac{1}{M-1}\mathbf{Y}_0\mathbf{Y}_0^T. \tag{6}$$

This matrix is symmetric and its rank, $rank(\mathbf{C}) = rank(\mathbf{Y}) = min(N, M)$. The set of PCA directions, the Principal Components (PCs) or modes, is found by solving the following set of eigenproblems: $\mathbf{C}\boldsymbol{\phi}_i =$

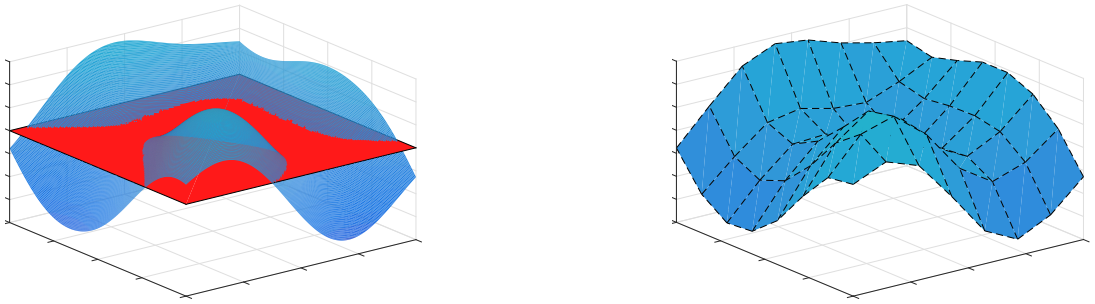


Figure 3: (*left*) A non-linear hyper-surface is approximated by only one hyper-plane in the data space. (*right*) The same hyper-surface is approximated by a set of local hyper-planes. The application of PCA can lead to better performances if local regions in the data space are detected and PCA is applied locally and independently in each region.

$\lambda_i \boldsymbol{\phi}_i \quad \forall i = 1, 2, ..., q$. Each PCA mode $\boldsymbol{\phi}_i$ has an associated eigenvalue, λ_i, which represents the variance of the original data taken into account by that mode [32]. The PCA modes can be collected in a $N \times q$ matrix $\boldsymbol{\Phi} = \{\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, ..., \boldsymbol{\phi}_q\}$, sorted in descending order of importance.

The number q of PCA modes that are retained is usually much smaller than the dimension N. The PCA modes with the highest eigenvalues are the ones that are kept. Once the PCA modes are found, the data can be encoded in a set of q scalars called PCA scores. The PCA scores corresponding to the realization $\mathbf{y}(\mathbf{x}_j)$ are given by the projection:

$$a_i(\mathbf{x}_j) = \boldsymbol{\phi}_i^T \, \mathbf{y}(\mathbf{x}_j) \tag{7}$$

with $\forall i = 1, \ldots, q$ and $\forall j = 1, \ldots, M$. The PCA reduction can be expressed in matrix form as:

$$\mathbf{Y} = \bar{\mathbf{Y}} + \mathbf{D}\mathbf{Y}_0 \approx \bar{\mathbf{Y}} + \mathbf{D}\boldsymbol{\Phi}\mathbf{A} = \tilde{\mathbf{Y}}, \tag{8}$$

where $\mathbf{Y}$ is the data matrix as described in Section 2, $\mathbf{A} = \{\mathbf{a}(\mathbf{x}_1), \mathbf{a}(\mathbf{x}_2), ..., \mathbf{a}(\mathbf{x}_M)\}$ with $\mathbf{a} \in \mathbb{R}^q$ is the matrix where all the PCA scores for different values of the input parameters are stored and $\tilde{\mathbf{Y}}$ is the reconstruction of the data matrix $\mathbf{Y}$ after the compression achieved by PCA.

3.2. Local PCA

PCA is a linear combination of basis functions. A large number of PCs may be required when applying PCA on highly non-linear systems [31, 30]. Local PCA (LPCA) constructs local models, each pertaining to a different disjoint region of the data space [33]. Within each region, the model complexity is limited, and thus it is possible to construct linear models using PCA [34, 33]. Figure 3 provides a general representation for a set of 3-dimensional observations forming a curved surface. Each axis shows the co-domain for each of the three scalar components that identify the 3-dimensional observations. The figure shows how a local representation of the curved surface can provide a better representation with respect to a single hyper-plane.

The partition in local clusters, where PCA is carried out, is accomplished using a Vector Quantization (VQ) algorithm that minimizes the reconstruction error. The reconstruction error is the squared Euclidean distance from one point or observation in the data-space to the linear manifold that is found by applying PCA in the local region. Mathematically, it can be expressed in a general fashion as:

$$d(\mathbf{z},\ \mathbf{r}^{(i)}) = \left(\mathbf{z} - \mathbf{r}^{(i)}\right)^T \mathbf{\Phi}^{(i)\ T}\mathbf{\Phi}^{(i)} \left(\mathbf{z} - \mathbf{r}^{(i)}\right), \tag{9}$$

where $\mathbf{z}$ is the object to be assigned to the cluster $\mathcal{R}^{(i)}$, represented by the reference vector $\mathbf{r}^{(i)}$, defined as the centroid of the i-th region: $\mathbf{r}^{(i)} = \mathrm{E}[\mathbf{z} \in \mathcal{R}^{(i)}]$. The cluster i is defined as:

$$\mathcal{R}^{(i)} = \{\mathbf{z} \mid d(\mathbf{z}, \mathbf{r}^{(i)}) \leq d(\mathbf{z}, \mathbf{r}^{(j)});\ \forall\, j \neq i\}. \tag{10}$$

In this work, the objects to be assigned to different clusters are chosen to be the vectors of observations of one single variable, namely the rows of the data-matrix:

$$\mathbf{Y} = \begin{bmatrix} T(r_1, \mathbf{x}_1) & T(r_1, \mathbf{x}_2) & T(r_1, \mathbf{x}_3) & \dots & T(r_1, \mathbf{x}_M) \\ \vdots & \vdots & \vdots & & \vdots \\ T(r_L, \mathbf{x}_1) & T(r_L, \mathbf{x}_2) & T(r_L, \mathbf{x}_3) & \dots & T(r_L, \mathbf{x}_M) \\ Y_{CH_4}(r_1, \mathbf{x}_1) & Y_{CH_4}(r_1, \mathbf{x}_2) & Y_{CH_4}(r_1, \mathbf{x}_3) & \dots & Y_{CH_4}(r_1, \mathbf{x}_M) \\ \vdots & \vdots & \vdots & & \vdots \\ Y_{CH_4}(r_L, \mathbf{x}_1) & Y_{CH_4}(r_L, \mathbf{x}_2) & Y_{CH_4}(r_L, \mathbf{x}_3) & \dots & Y_{CH_4}(r_L, \mathbf{x}_M) \\ \vdots & \vdots & \vdots & & \vdots \end{bmatrix} \tag{11}$$

3.3. Constrained PCA

The truncation of the PCA basis may inevitably involve the violation of important physical laws such as the conservation of mass when the observations $\mathbf{y}(\mathbf{x}_j)\ \forall j = 1, \dots, M$ are reconstructed from the PCA scores. To avoid that, the PCA scores can be evaluated by solving a constrained minimization problem, where the functional to be minimized is the PCA reconstruction error [17]. This approach is usually referred to as Constrained PCA (CPCA). The constraints are the physical laws which are intended not to be violated. This minimization problem can be mathematically expressed as:

$$\begin{aligned} minimize: \quad & \mathcal{J}\left(\boldsymbol{\gamma}^{(i)}\right) = \frac{1}{2}\|\mathbf{y}^{(i)} - \left(\overline{\mathbf{y}} + \mathbf{\Phi}_k\boldsymbol{\gamma}^{(i)}\right)\|^2 \\ s.t.: \quad & l_j\left(\overline{\mathbf{y}} + \mathbf{\Phi}_k\boldsymbol{\gamma}^{(i)}\right) = l_j\left(\mathbf{y}^{(i)}\right) = 0 \qquad \forall j = 1, ..., N_c \end{aligned} \tag{12}$$

where $\mathbf{y}^{(i)}$ represent the vector $\mathbf{y}(\mathbf{x}_i)\ \forall i = 1, \dots, M$, also introduced in Section 2, 3.1 and 3.2, $\boldsymbol{\gamma}$ are the CPCA scores (they have been indicated with a different symbol to differentiate them from the PCA scores), $l_j(\cdot)$ is the function related to the j-th constraint and N_c is the number of constraints, which can

also be inequality constraints. Minimizing the functional $\mathcal{J}(\cdot)$ when no constraints are enforced leads to the PCA scores $\mathbf{a}$.

It is preferable that the solution of this system be not too computationally expensive. In [17], the constrained optimization problem has a straightforward solution due to the linearity of the imposed constraints, which allows for a fast evaluation of the CPCA coefficients. If more complex constraints are imposed, the solution of the constrained optimization problem for the evaluation of the CPCA scores might involve the reconstruction of the considered physical fields (via eq. (8)) and the use of more expensive optimization algorithms, making the evaluation of the aforementioned coefficients unfeasible.

3.4. Kriging

Accurate prediction of the PCA scores at unexplored points $\mathbf{x}^* \in \mathcal{D}$ in the input parameter space (e.g. inlet temperature, equivalence ratio, etc.) translate into accurate estimation of the original variables as the mapping from $\mathbf{a}(\mathbf{x}^*)$ to $\mathbf{y}(\mathbf{x}^*)$ is known and explained in section 3.1. The data-set $\mathbf{A} = \{\mathbf{a}(\mathbf{x}_1),\ \mathbf{a}(\mathbf{x}_2),\ ...,\ \mathbf{a}(\mathbf{x}_M)\}$ of PCA scores evaluated at different training points, with $\mathbf{a}(\mathbf{x}) = [a_1(\mathbf{x}), a_2(\mathbf{x}), \ldots, a_q(\mathbf{x})]$, is used to build a response surface in the region $\mathcal{D}$ spanned by $\mathbf{X}$.

Kriging is an interpolation method in which every realization $a(\mathbf{x})$, where a is one PCA score indicated with no subscript for brevity, is expressed as a combination of a trend function and a residual [35]:

$$a(\mathbf{x}) = \mu(\mathbf{x}) + z(\mathbf{x}) = \sum_{i=0}^{p} \beta_i f_i(\mathbf{x}) + z(\mathbf{x}) = \mathbf{f}^T(\mathbf{x})\boldsymbol{\beta} + z(\mathbf{x}) \tag{13}$$

The trend function $\mu(\mathbf{x})$ is a low-order polynomial regression and provides a global model in the input space. The term $z(\mathbf{x})$ creates a localized deviation weighting the points in the training set that are closer to the target point $\mathbf{x}$. The trend function $\mu(\mathbf{x})$ is expressed as a weighted linear combination of $p+1$ polynomials $\mathbf{f}(\mathbf{x}) = [f_0(\mathbf{x}),\ ...,\ f_p(\mathbf{x})]^T$ with the weights $\boldsymbol{\beta} = [\beta_0,\ ...,\ \beta_p]^T$ 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). Many possible correlation functions are available: linear, quadratic, exponential, Gaussian, Matern 3/2, Matern 5/2, just to name a few [35, 36]. A detailed discussion about these functions can be found in [37]. One of the main differences among these kernels is their smoothness.

In the definition of both the trend function and the residual, it is up to the designer to choose the polynomials $\mathbf{f}(\mathbf{x})$ and the correlation model or kernel. In this way, the designer has the possibility to add prior into the problem and subsequently let the data speak for themselves by estimating the hyper-parameters $\boldsymbol{\theta}$.

3.5. Adaptive Sampling

The sampling strategy employed to explore the region $\mathcal{D}$ of the input space can affect the construction of a PCA+Kriging-based ROM. The construction of high-performing ROMs with a very limited

number of samples is possible if an effective sampling strategy is developed. As a PCA-based model strongly depends on its modal basis, a first step towards the improvement of this kind of model consists in improving the basis [38]. Given a set of (centered-scaled) observations $\mathbf{Y} = \{\mathbf{y}^{(1)}, ..., \mathbf{y}^{(M)}\}$ and its corresponding PCA-based model, we want to choose a new sample point, $\mathbf{x}_{\text{new}}$, that will meet the trade-off between the modal basis improvement and the parametric space exploration. Firstly, the influence of each observation on the modal basis is computed. The influence of the j-th observation on the i-th mode is defined by:

$$\text{Infl}_{\phi_i}(\mathbf{x}_j) = \frac{1}{|\boldsymbol{\phi}_i^T \boldsymbol{\phi}_i^{-j}|} - 1, \tag{14}$$

where $\boldsymbol{\phi}_i^{-j}$ is the i-th basis function evaluated from a data set:

$$\mathbf{Y} = \{\mathbf{y}^{(1)}, ..., \mathbf{y}^{(j-1)}, \mathbf{0}, \mathbf{y}^{(j+1)}, ..., \mathbf{y}^{(M)}\}. \tag{15}$$

The influence of the observation $\mathbf{y}(\mathbf{x}_j)$ on the modal basis is defined by:

$$\text{Infl}_{\text{Basis}}(\mathbf{x}_j) = \sum_{i=1}^{K} s_i \text{Infl}_{\phi_i}(\mathbf{x}_j), \tag{16}$$

where s_i is the singular value of the i-th mode. The relative influence of the j-th observation on the modal basis is given by:

$$\text{Infl}_{\text{Basis}}^{\text{Rel}}(\mathbf{x}_j) = \frac{\text{Infl}_{\text{Basis}}(\mathbf{x}_j)}{\sum_{l=1}^{M} \text{Infl}_{\text{Basis}}(\mathbf{x}_l)}. \tag{17}$$

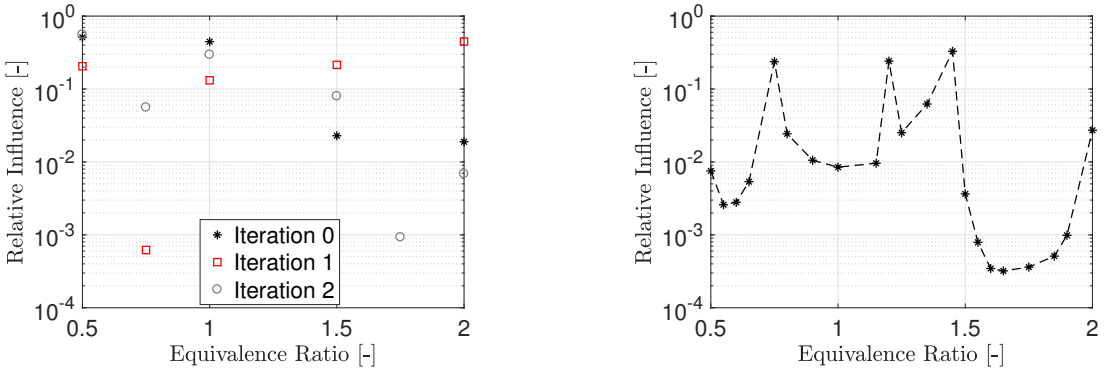
After the computation of this equation for each $\mathbf{x}_j$, the parametric space is heavily sampled via a LHS technique and the resulting set of samples is denoted by $\mathbf{Q} = \{\boldsymbol{\nu}_1, ..., \boldsymbol{\nu}_b\}$. The size b of this set of samples can be chosen as 100 times the parametric space dimension. Then the potential of enrichment $\text{PotBasis}(\boldsymbol{\nu}_i)$ of each candidate sample is computed with respect to the trade-off between the input space exploration and the improvement of the modal basis as:

$$\text{PotBasis}(\boldsymbol{\nu}_i) = d\left(\boldsymbol{\nu}_i, \mathbf{x}_j\right) \text{Infl}_{\text{Basis}}^{\text{Rel}}(\mathbf{x}_j), \tag{18}$$

where $j = \texttt{argmin}_k \, d\left(\boldsymbol{\nu}_i, \mathbf{x}_k\right)$ with $d(\cdot, \cdot)$ denoting the Euclidean distance. Finally, the new point will be selected to fulfill the following condition:

$$\mathbf{x}_{\text{new}} = \texttt{argmax}_{\boldsymbol{\nu} \in \mathbf{Q}} \text{PotBasis}(\boldsymbol{\nu}). \tag{19}$$

A new sample is chosen as far from the other samples as possible, but at the same time as close to the samples with the highest relative influence as possible. This sampling methodology is named Adaptive Sampling (AS).



(a) (b)

Figure 4: (a) Starting from 4 initial samples, the relative influence of each sample on the modal basis is evaluated and a new training point is chosen (as explained in section 3.5). Then, the relative influence of the 5 samples is evaluated again, and a 6-th training point is chosen, and so on. (b) Relative influence of each observation on the modal basis evaluated according to section 3.5 once that a number of 21 total training observation was reached.

4. Results

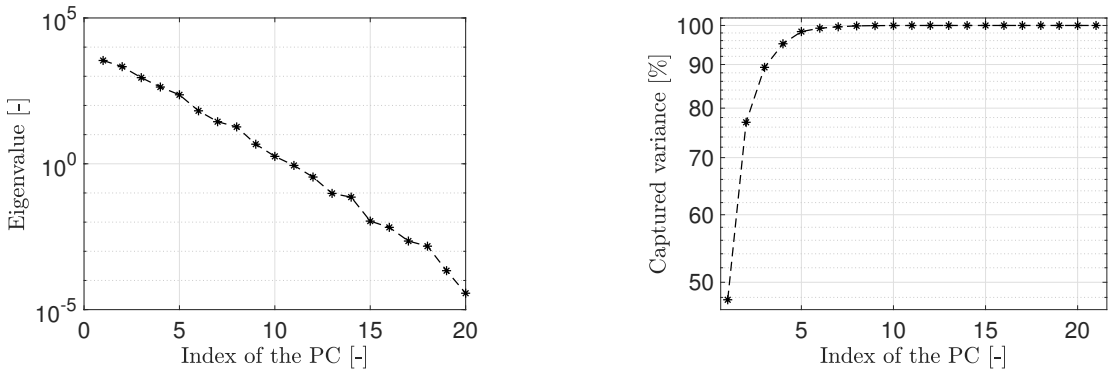
4.1. 1D flame with one input parameter

The Kriging-PCA approach was tested on a 1D methane/air laminar flame. OpenSMOKE++ [39, 40] was used to produce a data-set of 21 observations $\{\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, ..., \mathbf{y}^{(M)}\}$ of methane/air flames with GRI 3.0 mechanism for different values of the equivalence ratio ψ, in the range $0.5 \div 2$, with a step of 0.05. A subset of size $M = 21$ of the total 31 observations was used as training data-set to build a ROM and referred to as *training* observations or data. The corresponding values of ψ for the training observations were named training points. The 21 training observations were chosen using the Adaptive Sampling technique described in Section 3.5. The remaining 10 observations were used as test data to assess the accuracy of the reduced model. These 10 observations are referred to as *validation* data. The objective is to predict the validation data with acceptable accuracy.

In this paper, the term *reconstruction* will be used to indicate the reconstruction of the training data after the compression achieved by PCA, whilst the term *prediction* will be used when testing the predictive capabilities of the developed ROM on the validation data.

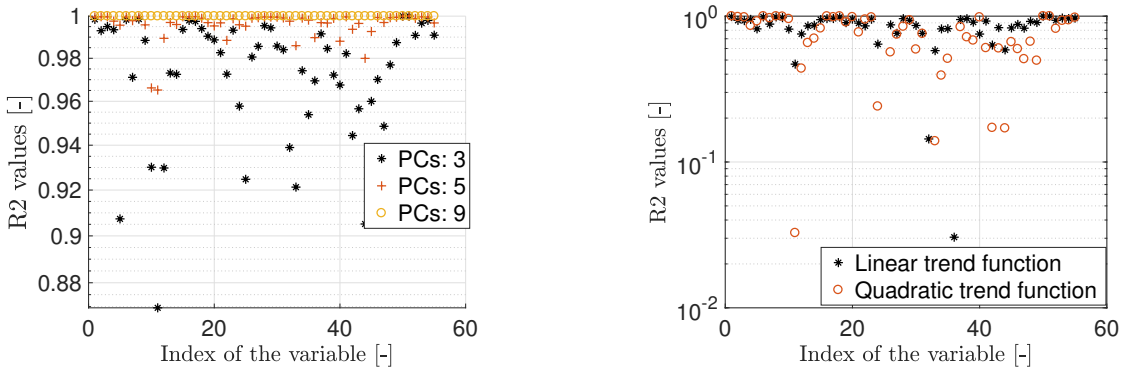
Each observation $\mathbf{y}^{(i)}$ of the data-set $\mathbf{Y} = \{\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, ..., \mathbf{y}^{(M)}\}$ is a vector of $N = 10,780$ variables: 53 chemical species plus temperature and axial-velocity evaluated at 196 grid points. The size of the matrix $\mathbf{Y}$ is $N \times M = 10,780 \times 21$. Because $N > M$, PCA shall find at most $M - 1 = 20$ PCs. Alternatively said, PCA can encode the 10 vectors $\mathbf{y}_m \in \mathbb{R}^N$ contained in the data matrix $\mathbf{Y}$ into a set of at most 20 scalars $\mathbf{a}_m \in \mathbb{R}^20$. The data-set is scaled according to the auto-scaling criterion. This criterion was chosen as other criteria tend to prioritize some variables over the others as shown in [27].

In figure 4a, an online evaluation of the relative influence of each *current* training observation on the modal basis (evaluated according to section 3.5) is reported. 4 samples are present at the start (iteration 0), and the 5-th training sample is chosen in the region of high values for the equivalence ratio as the



(a) (b)

Figure 5: (a) The spectrum of the eigenvalues associated to each PC provides a criterion for sorting the PCs in descending order as these eigenvalues can be interpreted as the importance of the PC they correspond to. Auto-scaling. (b) The cumulative original data variance that is captured when adding more PCs provides a criterion for the selection of the number of PCs when using global PCA. Auto-scaling. One-parameter case.



(a) (b)

Figure 6: (a) R-squared values for the reconstruction of the original data by PCA, for different numbers of retained PCs. (b) R-squared values for the prediction of the original data for two different choices of the trend function for the Kriging. 5 PCs retained. Matern52 correlation function. One-parameter case.

already present training observations in that region have a higher relative influence. Once this observation is added to the training data-set, the relative influence of each sample is re-evaluated (iteration 1) and a 6-th training observation is chosen. Thus, the relative influence of the new set of samples is re-evaluated (iteration 2). And so on, until the desired number of training samples is reached. In figure 4b, the relative influence of each observation on the modal basis is shown, once that 21 training observations have been chosen. Notice that the relative influence of each snapshot changes when new observations are introduced.

Figure 5a shows the eigenvalue spectrum of the PCs found by PCA. Figure 5b shows the cumulative original data variance that is captured when more PCs are retained. Interestingly, a few number of PCs are enough to recover most of this variance. As the maximum number of PCs that can be found is limited by the number of training observations, the cumulative recovered original data variance, reported in figure 5b, can also be used to determine if the number of available training observations is enough for the development of the ROM. In this case, 5 PCs could recover 98.4% of the original data variance and 10 PCs could recover 99.8%, thus suggesting that a number of 21 training observations can be considered sufficient. In figure 6a, the R^2 values for the reconstruction of the training data are reported for different numbers of retained PCs, q. These values, logically, increase with the approximation order, q, because more of the original data variance was accounted for by the reduced basis. Analysis on the R^2 values and thus of the reconstruction errors is a good estimation about the amount of information that is lost due to compression and can thus provide a good criterion for the choice of the number of PCs that is required.

Figure 6b shows the R^2 values for the prediction of the validation data by means of PCA+Kriging. These values are reported for two PCA+Kriging models with a number of 5 PCs retained: with a linear and with a quadratic trend function. Both models used a Matern52 kernel. From these results, we can conclude that the PCA+Kriging model with a linear trend function performed better in terms of predicting capabilities for this data-set.

In figures 7a - 7d the effects of the choice for the approximation order q and the choice of different trend functions and kernels for the Kriging interpolation are investigated. A reduction on prediction errors can be observed when increasing the approximation order q. Differences with the data (prediction errors) are still present due to the Kriging interpolation process and the fact that the data used for validation were not included in the data used to find the PCA modes. For the evaluation of these errors, the validation data-set and the one predicted by the model are scaled (range-scaling) and their difference is evaluated and stored in a new matrix. A mean value for each observation in this matrix is then used as prediction error. The formula for this error is:

$$Err(\mathbf{x}_i) = \frac{1}{N} \sum_{j=1}^{N} \frac{y_j(\mathbf{x}_i) - \tilde{y}_j(\mathbf{x}_i)}{y_j^{max} - y_j^{min}}, \tag{20}$$

where $\mathbf{x}_i$ is the i-th prediction point, N is the number of variables, $y_j(\mathbf{x}_i)$ is the true value of the j-th variable for the i-ith prediction point and $\tilde{y}_j(\mathbf{x}_i)$ is the ROM's prediction for the same variable, for

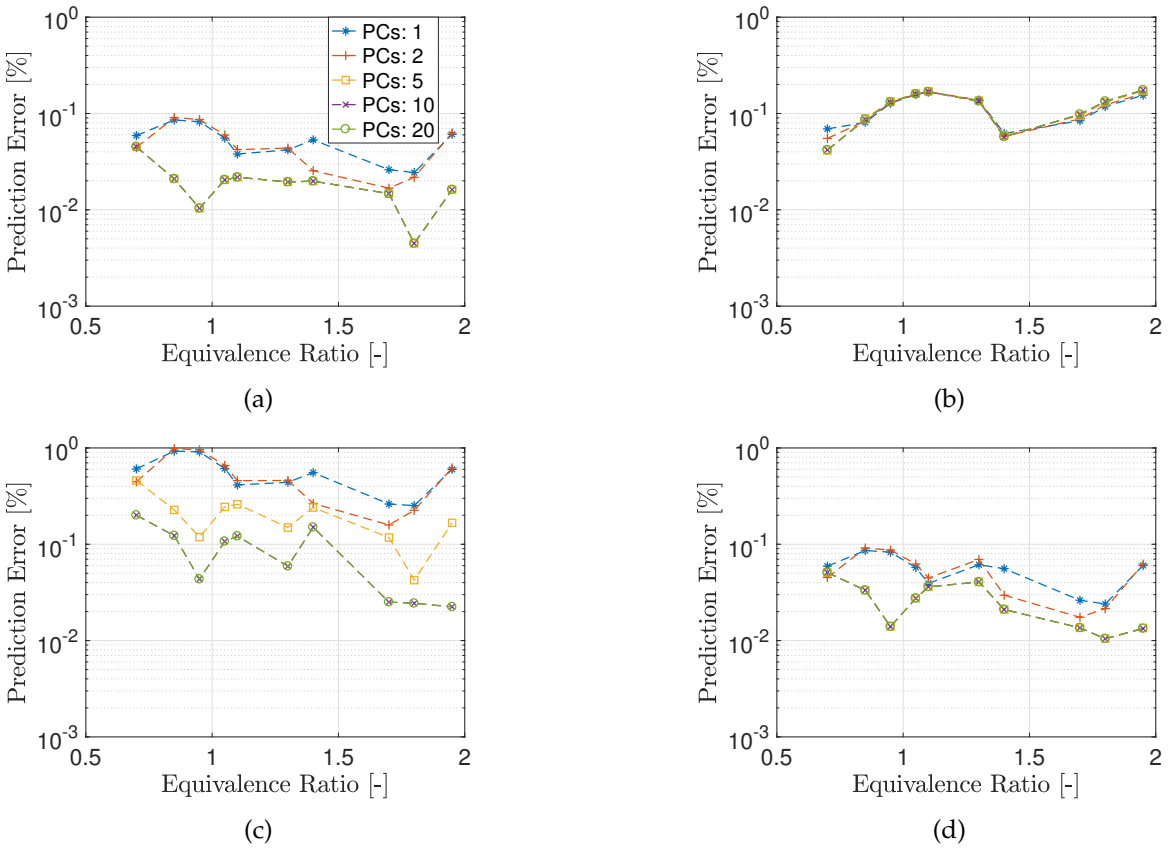
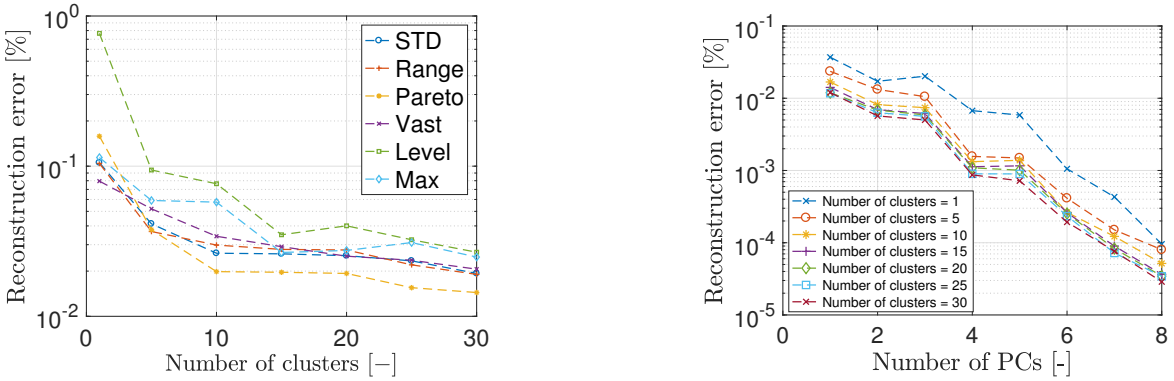


Figure 7: Prediction errors for different numbers of retained PCs. Global PCA+Kriging. (a) Constant trend function and exponential kernel. (b) Linear trend function and Gaussian kernel. (c) Quadratic trend function and Matern12 kernel. (d) Linear trend function and Matern52 kernel. One-parameter case.



(a) (b)

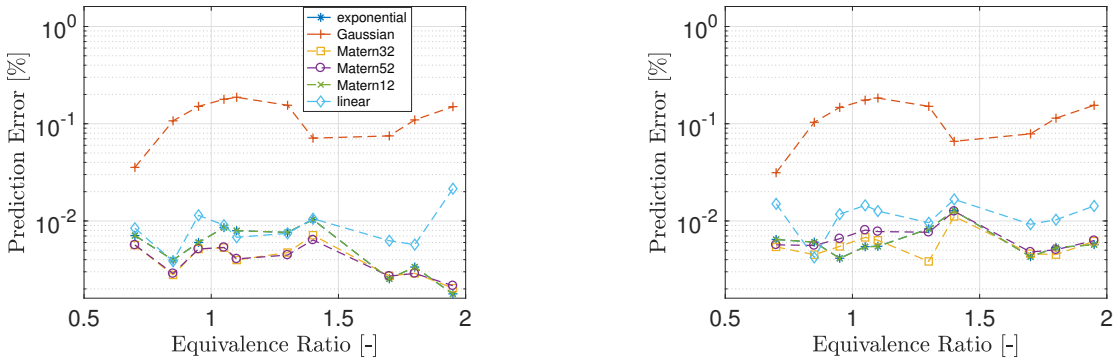
Figure 8: Local PCA reconstruction errors. (a) Reconstruction errors decreases as the number of clusters used is increased. (b) Reconstruction errors decrease as more PCs are retained (reported for the auto-scaling criterion). One-parameter case.

the same prediction point. These errors were below a value of 4% for all observations when $q > 1$. Different choices of trend and correlation function for the Kriging interpolation led to different results. For example, the choice of a Gaussian kernel (see figure 7b) for a PCA+Kriging model for this data-set could not reduce further its prediction errors when increasing q from a value of 2 to a value of 5 or 9. Indeed, although increasing q means accounting for more of the original data variance, it also means that more scalars have to be interpolated.

Next, the effects of performing PCA in separated clusters (Local PCA) on the quality of the reduced model is investigated. The effects of the choice of the scaling criterion are also examined. Figure 8a reports the data reconstruction errors when increasing the number of clusters used for the Local PCA formulation, for different scaling criteria. Increasing the number of clusters translated into better data reconstructions. The auto-scaling and the Pareto scaling criteria emerged as the ones with the lowest reconstruction errors for the Local PCA formulation. The general trend was a decreasing reconstruction error as the number of clusters is increased. Figure 8b reports the Local PCA reconstruction error when the number of retained PCs is increased (auto-scaling), for different number of clusters used for the Local PCA formulation.

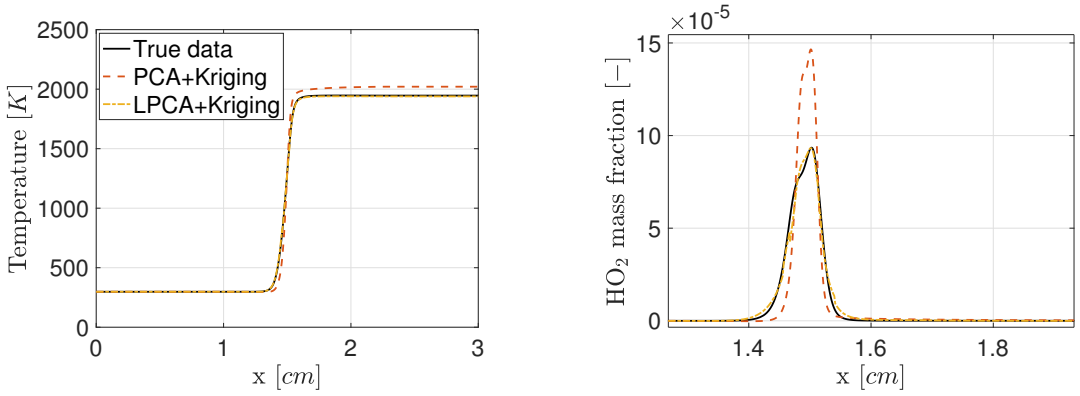
Prediction errors also decreased when a higher number of clusters was used for the Local PCA formulation. This can be observed in figures 9a and 9b. Local PCA, in combination with Kriging, resulted in better performances, in terms of both for reconstruction at the training points and predictions. The gain in accuracy for the prediction points is a clear proof that also the Kriging interpolation method benefited from the local formulation. Figure 10 reports the predictions of the spatial profiles of temperature and HO_2 mass fraction by PCA+Kriging and Local PCA+Kriging, where the gain in accuracy for Local PCA formulation is evident.

The clustering strategy explained in section 3.2 was used to group the rows of the data matrix $\mathbf{Y} \in \mathbb{R}^{N \times M}$ into clusters. The general encoding process for one observation in the data matrix $\mathbf{Y}$, which



(a) (b)

Figure 9: Local PCA prediction errors for $q = 2$. (a) 5 clusters, auto-scaling, constant trend function. (b) 10 clusters, auto-scaling, constant trend function. Reported for different kernel functions. One-parameter case.



(a) (b)

Figure 10: Prediction of the spatial profile of temperature and HO_2 mass fraction. Comparison between PCA and LPCA (15 clusters) combined with Kriging. One-parameter case.

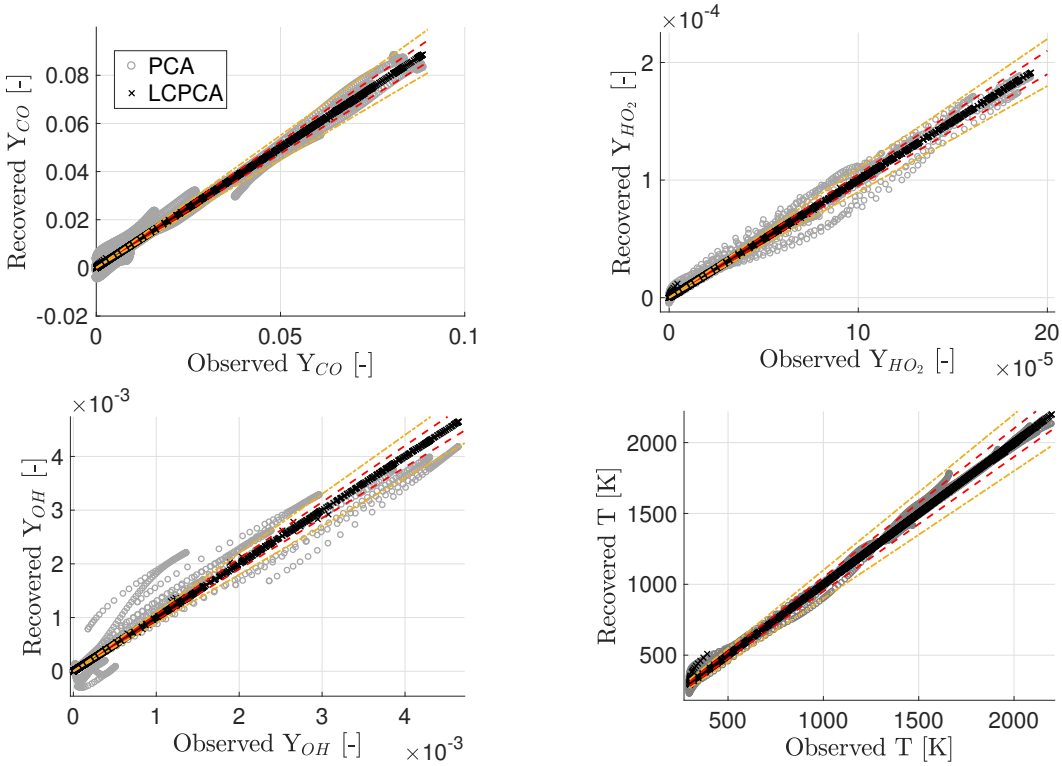


Figure 11: Parity plots for the reconstruction of the CO, HO_2, OH mass fractions and temperature. PCA: 3 PCs retained. 20 clusters for Local CPCA. Dashed: 5% error line. Dotted: 10% error line. One-parameter case.

can be stated as $\mathbf{y} \in \mathbb{R}^N \rightarrow \mathbf{a} \in \mathbb{R}^q$ with $q \ll N$, is possible via the detection of a lower-dimensional manifold spanned by the $M = 21$ observations $\mathbf{y}_m \in \mathbb{R}^N \; \forall \; m = 1, ..., M$. The dimension of this lower-dimensional manifold is thus $M - 1 = 20$. A new observation $\mathbf{y}^*$ not already present in the data-set will be approximated by a point on the lower-dimensional manifold even if there is no guarantee that this new observation lies *close* to this manifold. The data-partitioning process reduces the number of variables for each cluster. This means we are moving closer to a situation where the M samples are enough for the detection of a manifold which can accurately approximate unseen observations.

The PCA-reconstructed mass fraction spatial profile can include negative values as shown in the parity plots (figure 11). This is a known issue with PCA [41, 42] since the method itself cannot guarantee that physical constraints are verified, such as the positivity of mass fractions and the conservation of mass ($\sum Y_i = 1$). An increase in the approximation could indirectly lead to a solution to this issue, but it is here interesting to investigate if the implementation of a constrained version of PCA (CPCA) can guarantee the non-violation of the aforementioned physical laws. As discussed in section 3.3, the CPCA implementation used in the present work consisted in forcing each rebuilt mass fraction, on every grid point, to be positive.

However, one of the challenges of using CPCA in combination with Kriging is that the chemical species fields recovered from the interpolated CPCA scores might still violate the set of imposed con-

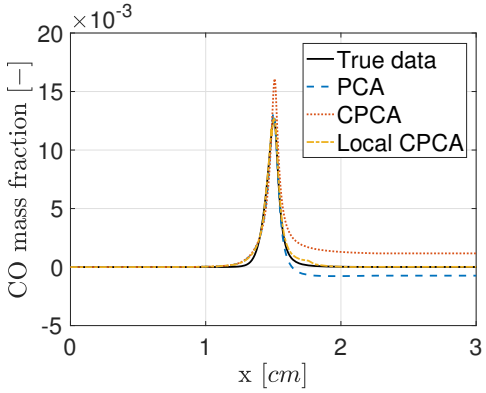


Figure 12: Reconstruction of the spatial profile for the CO mass fraction. Comparison between PCA, CPCA and LCPCA (15 clusters). One-parameter case.

straints. Even if for each training point a constrained optimization problem is solved for the evaluation of the CPCA scores, no constraint is imposed when these coefficients are interpolated by means of Kriging. Thus, there is no guarantee that the predicted CPCA scores belong to the feasible region delimited by the constraints $l_j(\boldsymbol{\gamma}) = 0$ of (12). CPCA in combination with Local PCA guarantees the satisfaction of the physical constraint, since the reconstruction accuracy improves, as visible in figures 11 and 12. Not only the imposed constraint was satisfied by the LCPCA+Kriging model's predictions, but a satisfying level of accuracy was also achieved as the parity plots in figure 13 indicate, with most predicted values being within the 5% or 10% error lines. The R^2 values for the prediction of the validation data as the number of training samples increases are reported in figure 14. These values increase as more observations are employed for the training of the model, as expected.

The results presented in this section show that different choices for the Kriging kernels, PCA formulation (standard, local and constrained) affect the ROM's predictive capabilities.

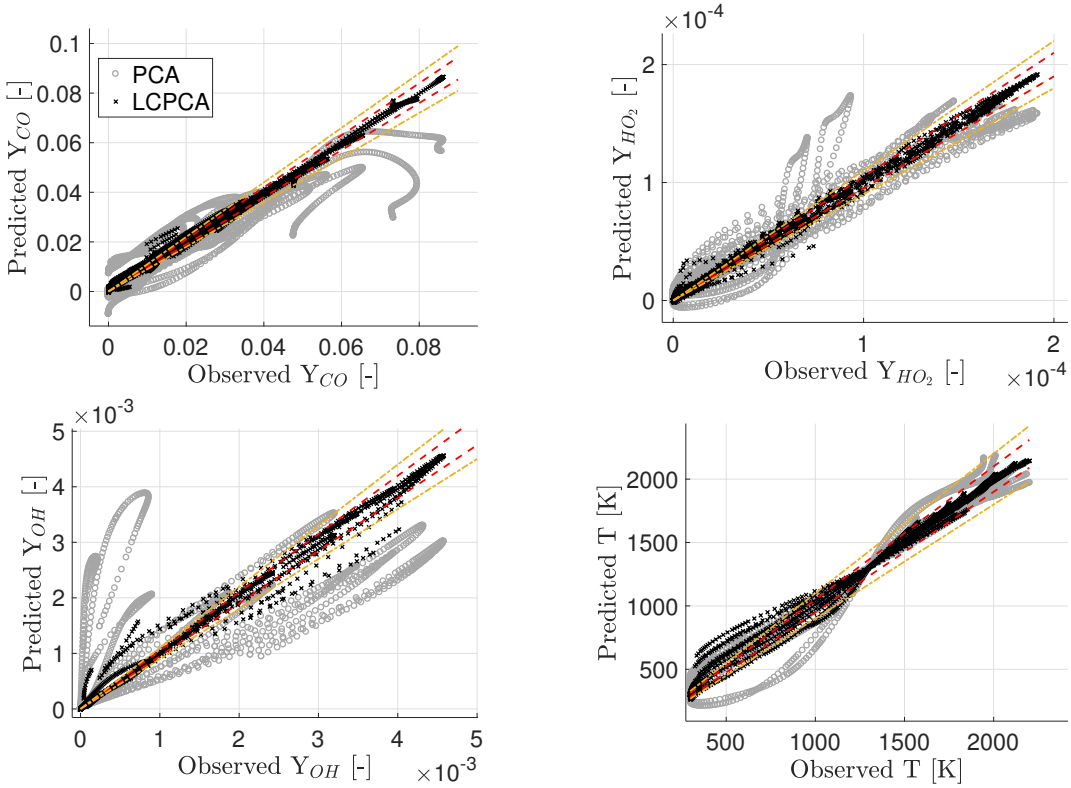


Figure 13: Parity plots for the predictions of the CO, HO_2, OH mass fractions and temperature. PCA: 3 PCs retained. 50 clusters for Local PCA. Kriging: linear trend function, Matern52 kernel. Dashed: 5% error line. Dotted: 10% error line. One-parameter case.

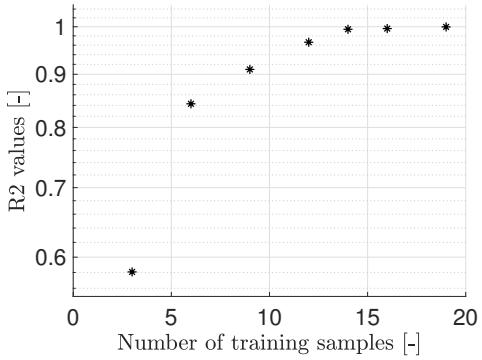
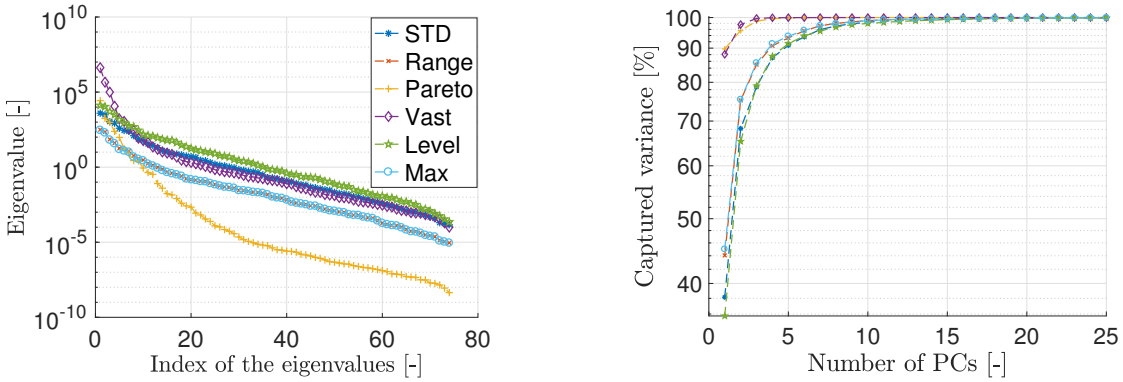


Figure 14: Global R^2 values for the validation data as the number of training observations increases. Local PCA+Kriging with linear trend function and Matern12 kernel.



(a) (b)

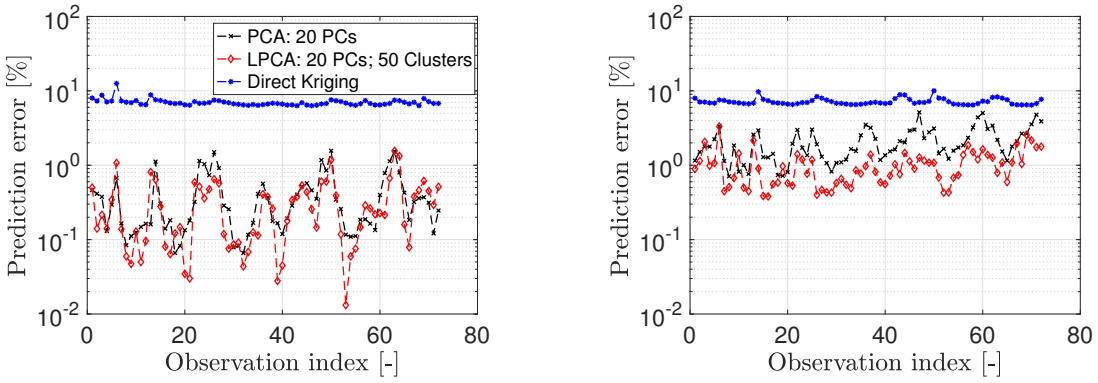
Figure 15: (a) The spectrum of the eigenvalues associated to each PC provides a criterion for sorting the PCs in descending order as these eigenvalues can be interpreted as the importance of the PC they correspond to. (b) The cumulative original data variance that is captured when adding more PCs provides a criterion for the selection of the number of PCs when using global PCA. Two-parameters case.

4.2. 1D flame with two input parameters

The methodology explained in section 2 was tested on the same system presented in Section 4.1, increasing the number of parameters to two. A molar fraction of N_2 was introduced in the inlet stream. A total of 147 observations were generated by OpenSMOKE++ in the range $0.5 \div 1.5$ for the equivalence ratio and $0.4 \div 1.0$ for the molar fraction of N_2. A set of 75 observations was selected with the AS strategy explained in Section 3.5 for the training of the SM. The remaining observations were used as test data.

Figure 15a reports the eigenvalue spectrum of the PCs, for different scaling criteria used. Figure 15b shows the cumulative original data variance that is captured when more PCs are retained in the reduced basis. This is reported for different scaling criteria applied to the data-set. As for the one-parameter case, most of the total data variance is concentrated on the first few PCs, indicating that PCA is a valid strategy for the reduction the problem size. Besides, considered that 75 training observations were used, a number of 15 PCs could recover over 99% of the original data variance, indicating that the chosen number of training observations is sufficient, as the prediction errors for this case will show.

Figure 16a shows the prediction errors for a PCA+Kriging model with 20 PCs, a constant trend function and exponential kernel. Prediction errors for a Local PCA+Kriging model with 20 PCs, 50 clusters and again constant trend function and exponential kernel and prediction errors for a direct Kriging model (Kriging applied directly on the original variables without the use of PCA) with constant trend function and exponential kernel are also reported. The prediction errors are shown again for the three models in figure 16b, with the only difference being that a Gaussian kernel for the Kriging was chosen. These errors are reported for the 72 observations used for validation. From these results, it can be concluded that the clustering of the data improved the ROM's predictive capabilities. In figure 17, the parity plots for the predicted values of temperature and CH_4, OH, CO mass fractions are reported. The number of retained PCs was 20, corresponding to a recovered original data variance of 99.78% with



(a) (b)

Figure 16: Prediction errors reported for 3 different SMs. Auto-scaling. Kriging: (a) constant trend function and exponential kernel; (b) constant trend function and Gaussian kernel. Two-parameters case.

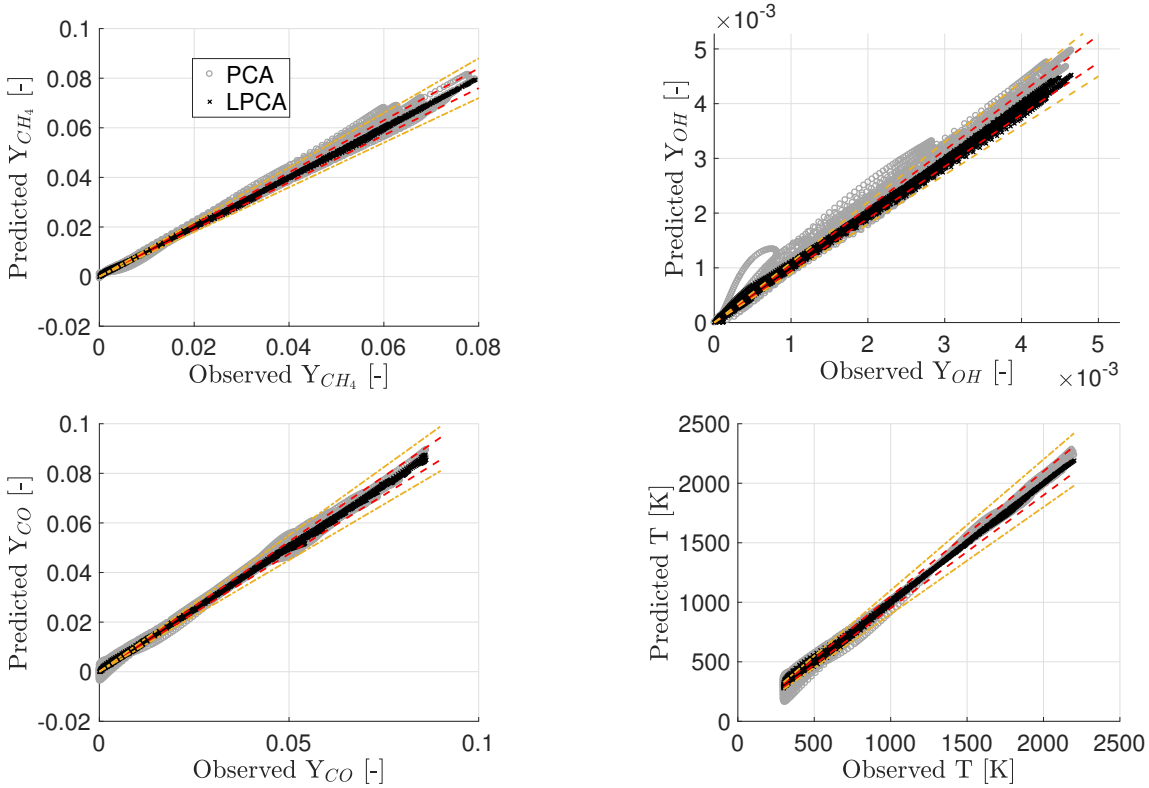
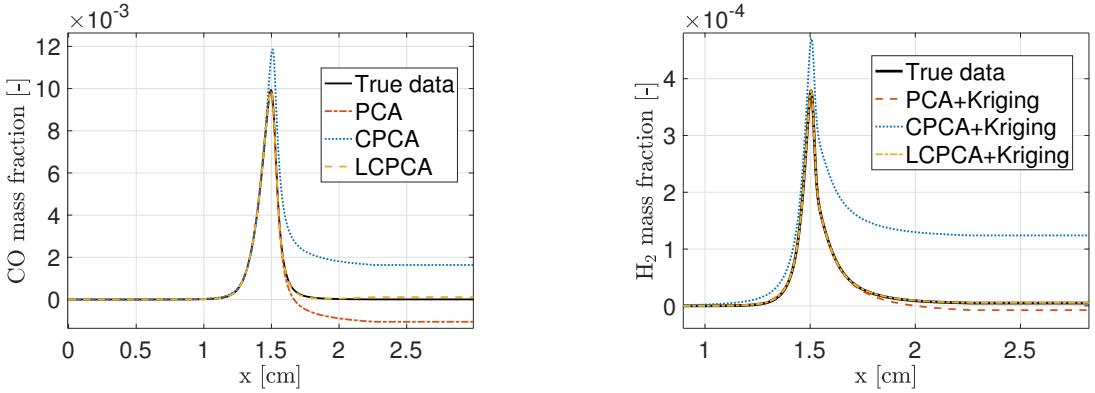


Figure 17: Parity plots for the prediction of temperature and of the CH_4, OH, CO mass fractions. PCA: 20 PCs retained. Local PCA: 50 clusters. Kriging: linear trend function, Gaussian kernel. Dashed: 5% error line. Dotted: 10% error line. Two-parameter case.



(a) (b)

Figure 18: (a) Reconstruction of the spatial profile for CO. Comparison between PCA, CPCA and LCPCA (50 clusters). (b) Prediction of the spatial profile for H2. Comparison between PCA, CPCA and LCPCA (50 clusters) combined with Kriging. Two-parameter case.

the auto-scaling criterion. The number of clusters used for Local PCA was 50. The figure shows that the PCA+Kriging methodology performed efficiently, with clear improvements when the data are clustered. The sum of mass fractions of all chemical species added up to 1 with a maximum error of 4×10^{-3}.

The use of CPCA was also investigated for this case. In figure 18a, the rebuilt spatial profiles for CO by means of PCA, CPCA and LCPCA are reported and compared. While the reconstructed CO spatial profile showed negative values when rebuilt by PCA, CPCA provided a spatial profile that did not violate this specific physical constraint. Figure 18b shows that even after the Kriging interpolation, the physical constraint imposed in the application of CPCA was not violated, meaning that the interpolated values for the CPCA scores remained within the feasible region delimited by the constraints indicated in (12). Besides, the spatial profiles in the post-reaction zone were wrongly reconstructed and predicted by the CPCA+Kriging model. This aspect was corrected by the Local PCA formulation.

As explained, most of the data variance was recovered by a small number of PCs, 15, in comparison to the number of available training samples, which was 75. In order to further test the proposed methodology, and be sure that the high accuracy of the model's predictions were not to be attributed mainly to the high number of samples used for its training, another ROM was developed based on the same approach with a smaller number of training observations, 46. The parity plots for the prediction of temperature and OH, CO, CH_4 mass fractions are reported in figure 19. As visible, a high level of accuracy was achieved also with less observations, indicating that the analysis based on the cumulative data variance recovered by the PCs can be an efficient indicator for the amount of information fed to the model by the training samples.

The promising results obtained from the application of the proposed methodology for the one-parameter case, were confirmed with two input parameters. The results confirm that the combination of PCA, or its variations, with Kriging is a valid choice for the development of reduced-order surrogate models. In the next section, the complexity of the case is increased by adding a third input parameter.

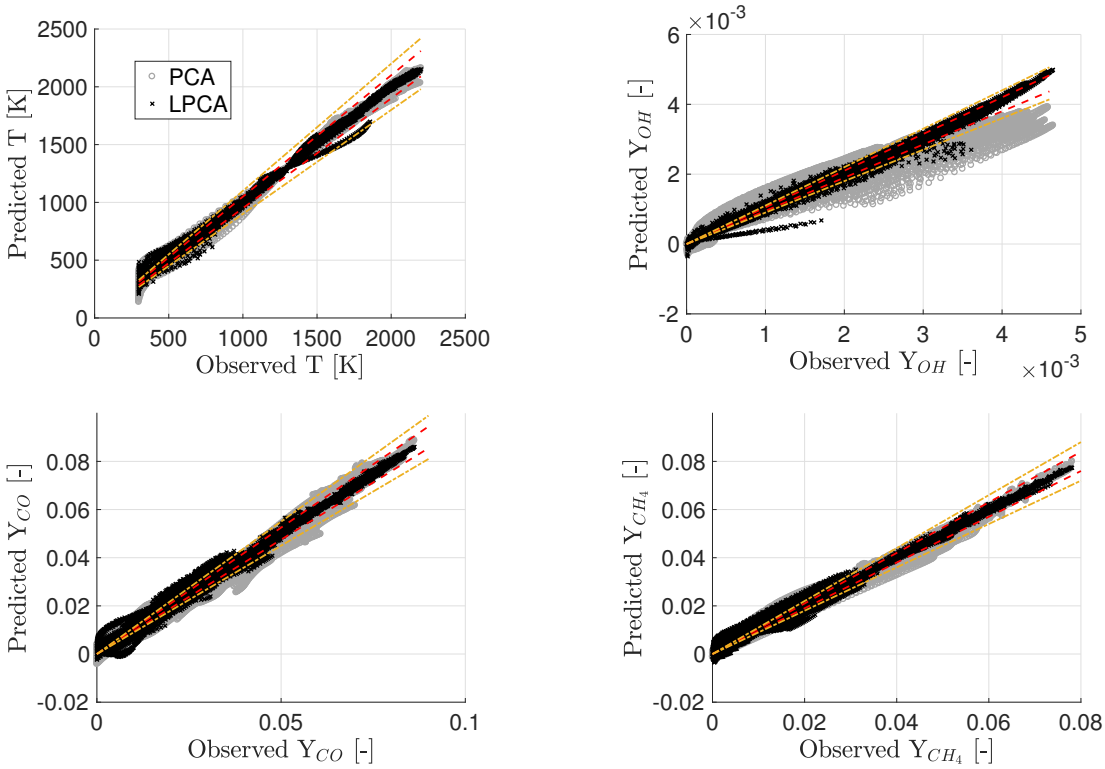
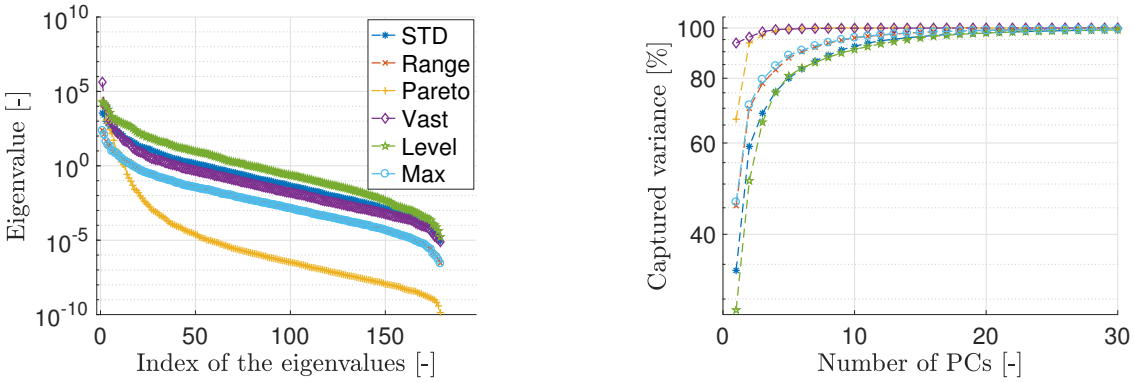


Figure 19: Parity plots for the predictions of temperature and the OH, CO, CH_4 mass fractions, using less training observations. PCA: 30 PCs retained. Local PCA: 90 clusters. Kriging: linear trend function, Gaussian kernel. Dashed: 5% error line. Dotted: 10% error line. Two-parameter case with less training observations.



(a) (b)

Figure 20: (a) The spectrum of the eigenvalues associated to each PC provides a criterion for sorting the PCs in descending order as these eigenvalues can be interpreted as the importance of the PC they correspond to. Reported for different scaling criteria. (b) The cumulative original data variance that is captured when adding more PCs provides a criterion for the selection of the number of PCs when using global PCA. Reported for different scaling criteria. Three-parameters case.

4.3. 1D flame with three input parameters

After validating the predictive capabilities of a ROM developed on the combination of PCA and Kriging on systems with one or two input parameters, one more parameter is added to the system, namely the inlet stream temperature in the range $298 \div 798$ K. Two sets of observations were again generated by OpenSMOKE++. A set of 180 solutions, selected using the AS strategy, was used to train the model. The other set (702 solutions) was used for the validation of the ROM's predictions.

Figures 20a and 20b show the eigenvalue spectra and original data variance recovery for the different scaling criteria, respectively. Over 99% of the original data variance was captured by 28 PCs.

Figure 21 shows predictions for T, CO, OH and NO for a PCA+Kriging (48 PCs) and LPCA+Kriging (20 clusters) model with a quadratic trend function and Matern52 kernel. The application of Local PCA+Kriging notably improve the predictions for some species like OH and NO whereas good predictions were already obtained by PCA+Kriging for T and CO.

The application on the 1D flame allowed to test the potential and limitations of the PCA+Kriging approach. For this reason, it is interesting to test the proposed methodology on a 2D flame with detailed chemistry and transport phenomena.

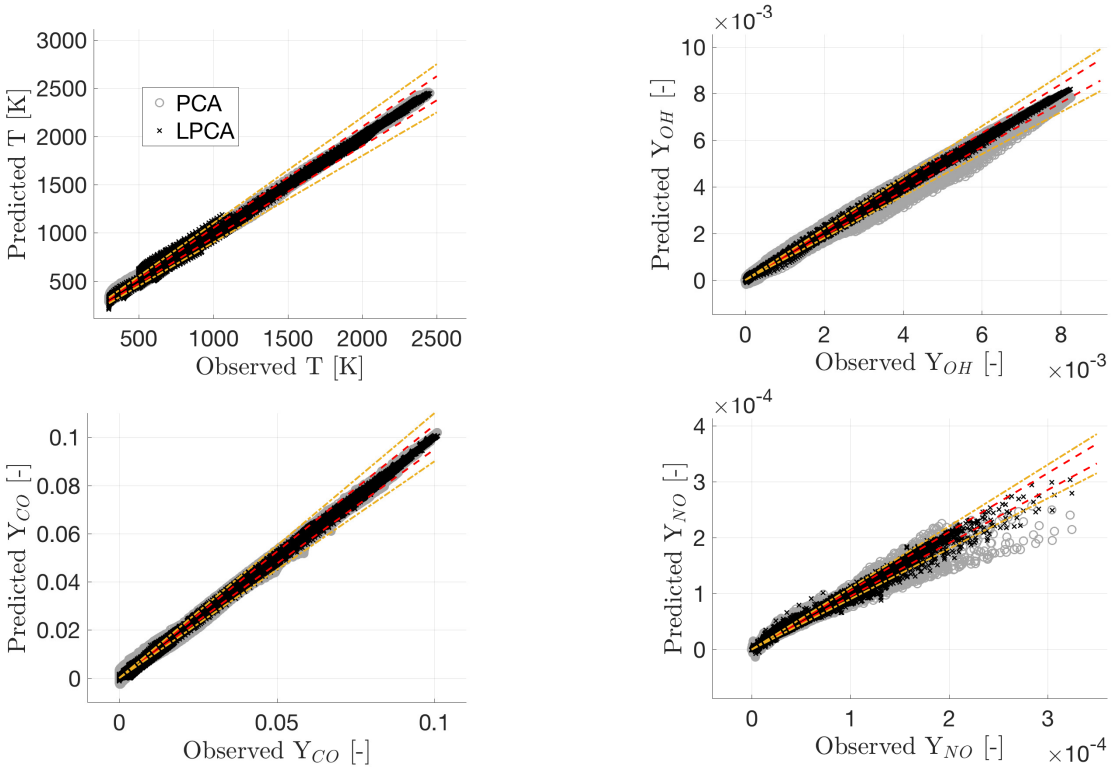
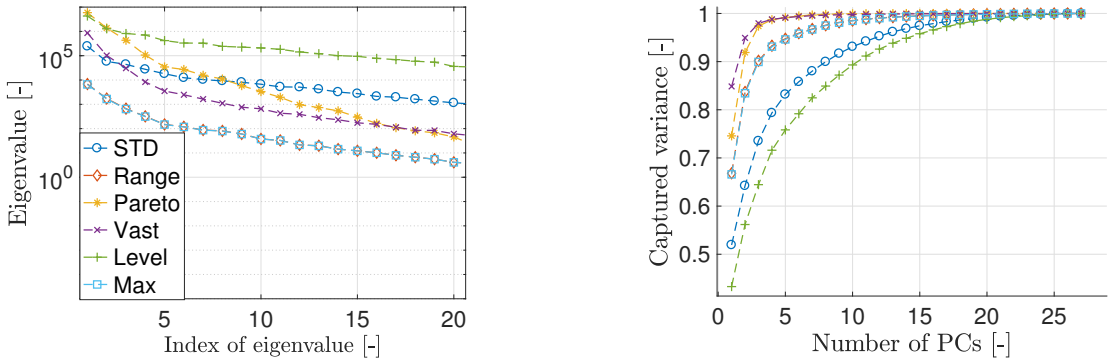


Figure 21: Parity plots for the prediction of temperature and OH, CO, NO mass fractions. PCA: 48 PCs. LPCA: 20 clusters, 48 PCs. Kriging: quadratic trend function, Matern52 kernel. Dashed: 5% error line. Dotted: 10% error line. Three-parameter case.



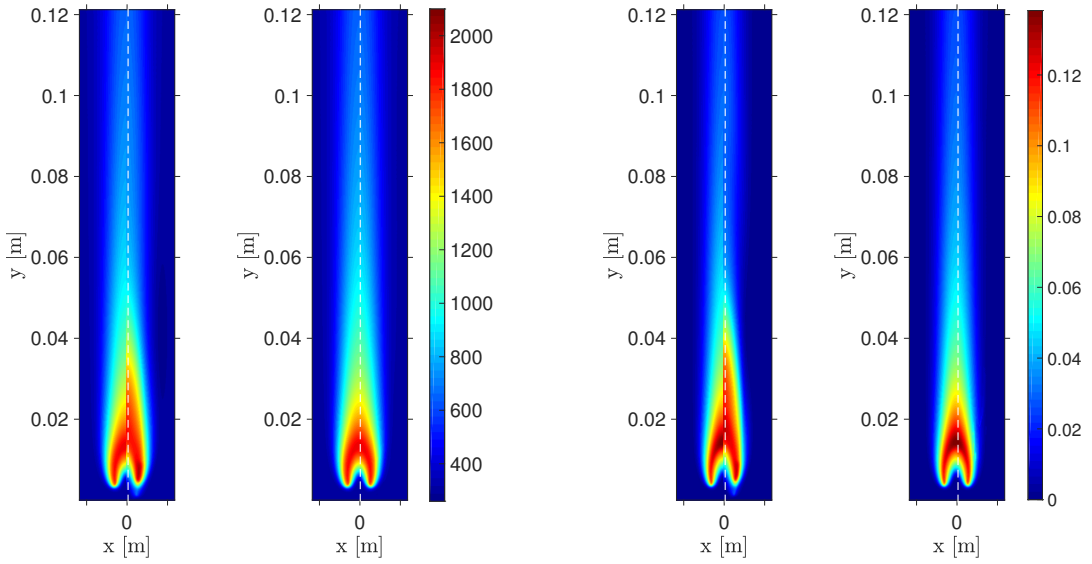
(a) (b)

Figure 22: (a) The spectrum of the eigenvalues associated to each PC provides a criterion for sorting the PCs in descending order as these eigenvalues can be interpreted as the importance of the PC they correspond to. Reported for different scaling criteria. (b) The cumulative original data variance that is captured when adding more PCs provides a criterion for the selection of the number of PCs when using global PCA. Reported for different scaling criteria. 2D flame.

4.4. 2D flame with two input parameters

The methodology described in section 2 is applied to a multidimensional flame in more complex configurations in order to test its potential. The configuration of the simulated flame is described in [43] and in [44]. The computational domain starts from the exit of the nozzle and extends 122 mm further downstream. The radial direction is expanded to 42.88 mm. A 2D structured axi-symmetric mesh with around 25600 cells is used and the nozzle radius is resolved with 12 cells. The laminarBuoyantSimpleSMOKE solver is applied. Gravity is turned on. The multi-component diffusion model is adopted to consider molecular diffusion. The GRI3.0 mechanism without NO_x (35 species and 219 reactions) is applied. For the velocity boundary condition, the profile provided from [43] and in [44] is used. The input parameters are two, namely the inlet velocity and the molar fraction of CH_4 in the inlet stream, which is a mixture of CH_4 and N_2. A total of 30 samples was produced by OpenFOAM, spanning the two input parameters in the range $24 \div 89$ cm/s and $40 \div 100$ $\%$ for the inlet velocity and inlet molar fraction of CH_4, respectively. A total of 36 physical variables was considered: 35 chemical species and temperature. 25 observations were chosen to train a PCA+Kriging model using the adaptive strategy described in section 3.5. Figure 22a reports the eigenvalue spectrum of a PCA performed on this data-set, for different scaling criteria. Figure 22b shows the cumulative original data variance that is recovered when more PCs are retained. The recovery of 99% of the original data variance was achieved with 20 PCs for all scaling criteria. This indicates that despite the high dimensionality of the system, recurrent structures could be found in the data. At the same time, the fact that 20 PCs were needed for the 99% data variance recovery, considered that 25 observations were used for the training of the model, also suggests that more training observations might be needed, or a more narrow parameter region should be explored, as we shall see.

Figure 23 shows the temperature field for an inlet velocity of 24 cm/s and an inlet CH_4 molar fraction



(a) (b) (c) (d)

Figure 23: Left halves: true data. Right halves: (a) reconstruction of the temperature field by PCA; (b) reconstruction of the temperature field by Local PCA; (c) reconstruction of the CO_2 mass fraction field by PCA; (d) reconstruction of the CO_2 mass fraction field by Local PCA. Inlet velocity: $24\ cm/s$; inlet CH_4 molar fraction: 40%. 2D flame. PCA: 5 PCs. Local PCA: 100 clusters.

of 40%. The original field is shown on the left half of the figures, while PCA and Local PCA reconstructions are shown on the right half of the figure. PCA was performed by retaining 5 PCs while a number of 100 clusters was used for Local PCA. The data-set was centered and scaled according to the VAST scaling criterion. The reconstruction performed by PCA could clearly be improved by the Local PCA formulation. The mean reconstruction errors were 7% for PCA and 3% for Local PCA. The overall performance of the model for the reconstruction of the original data can be analyzed from figure 25. These figures report the parity plots for the reconstructed temperature and CO_2, CH_4, CO mass fraction fields. The dashed lines correspond to the 5% error. The dotted lines correspond to the 10% error. It is easily noticeable that the grey points corresponding to the PCA reconstruction are more scattered, while the the black points corresponding to the Local PCA reconstruction lie within the reported 5% or 10% error lines more frequently.

Next, the predictive capabilities of a PCA+Kriging and Local PCA+Kriging ROM for the 2D system are investigated. Figure 26 reports the true temperature field for an inlet velocity of $64\ cm/s$ and an inlet CH_4 molar fraction of 70%, as well as the predictions provided by PCA with 5 PCs, and Local PCA with 100 clusters and 5 PCs, both in combination with a Kriging model that used a quadratic trend function and exponential kernel. The prediction by the PCA+Kriging model for this temperature field had a mean error of 8%. This error was 4% for the LPCA+Kriging prediction. Similarly, figure 27 reports the CO and OH mass fractions fields for an inlet velocity of $64\ cm/s$ and an inlet CH_4 molar fraction of 70%. Figure 28 reports the R^2 values for the prediction of all the 36 fields, namely temperature and chemical

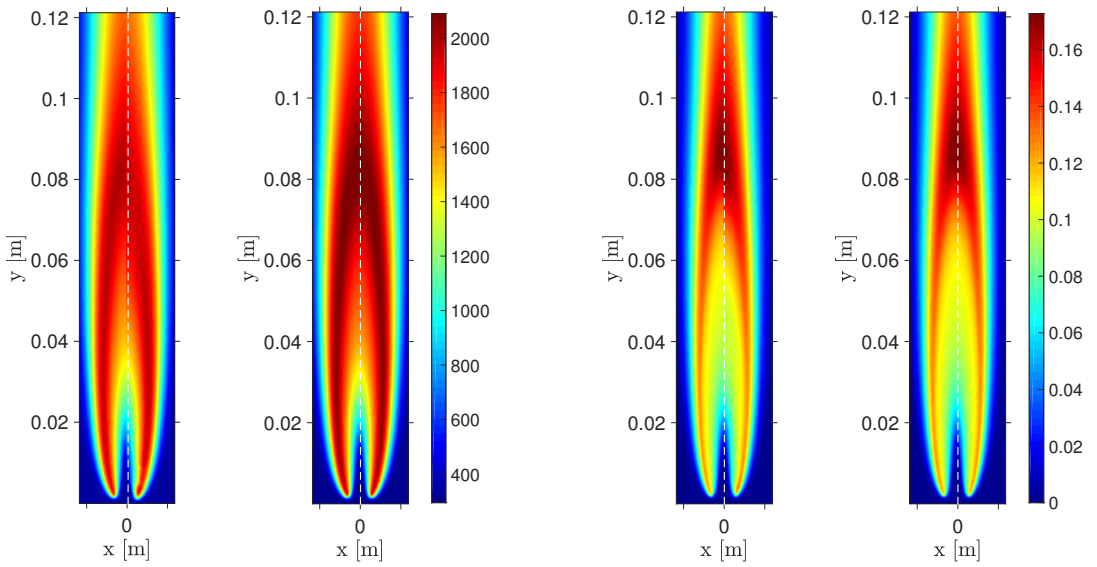


Figure 24: Left halves: true data. Right halves: (a) reconstruction of the temperature field by PCA; (b) reconstruction of the temperature field by Local PCA; (c) reconstruction of the CO_2 mass fraction field by PCA; (d) reconstruction of the CO_2 mass fraction field by Local PCA. Inlet velocity: $89\ cm/s$; inlet CH_4 molar fraction: 85%. 2D flame. PCA: 10 PCs. Local PCA: 100 clusters.

species. Interestingly, these values are higher when the Local PCA formulation is employed, confirming the importance of the clustering process for the proposed methodology. Parity plots for the prediction of temperature and CO_2, CH_4, CO mass fractions are reported in figure 29. The 5% and 10% error lines are reporter again in dashed red and yellow, respectively. The predictions of the Local PCA+Kriging model, reported in black, lie within the 10% error lines more than the PCA+Kriging predictions, indicating an overall better performance by the Local PCA+Kriging model.

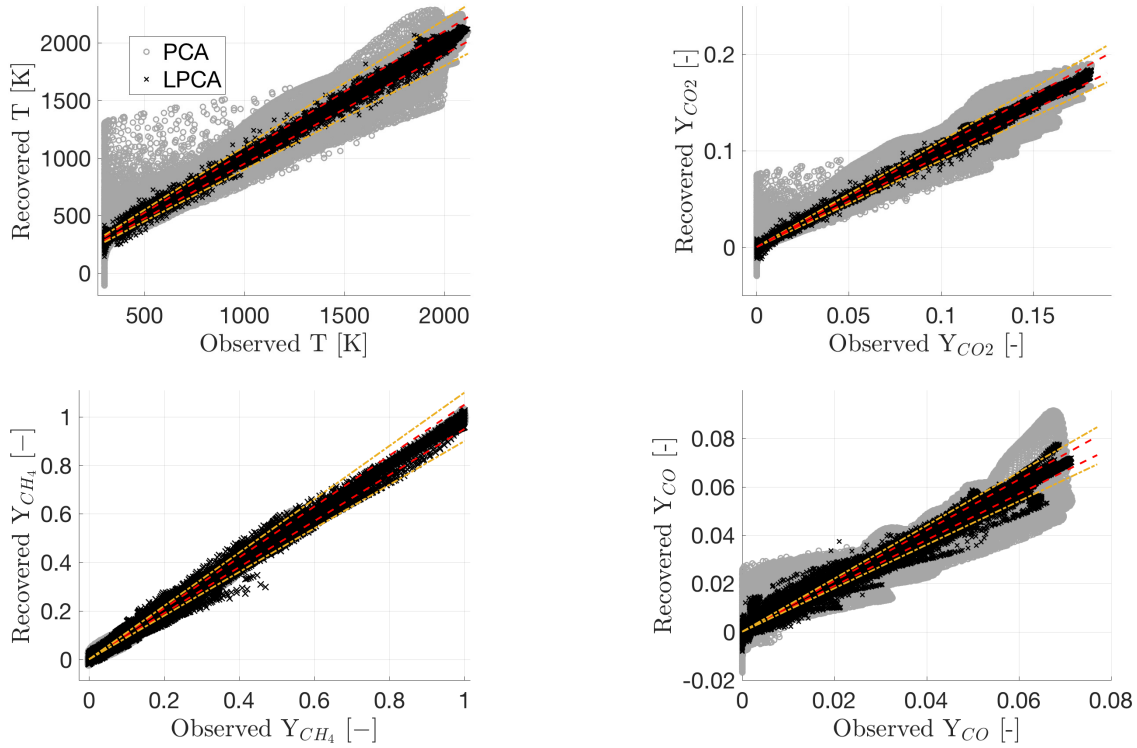


Figure 25: Parity plot for the reconstruction of temperature and CO_2, CH_4, CO mass fractions (using 25 training samples): comparison between PCA and LPCA (5 PCs, 100 clusters). Dashed: 5% error line. Dotted: 10% error line. 2D flame.

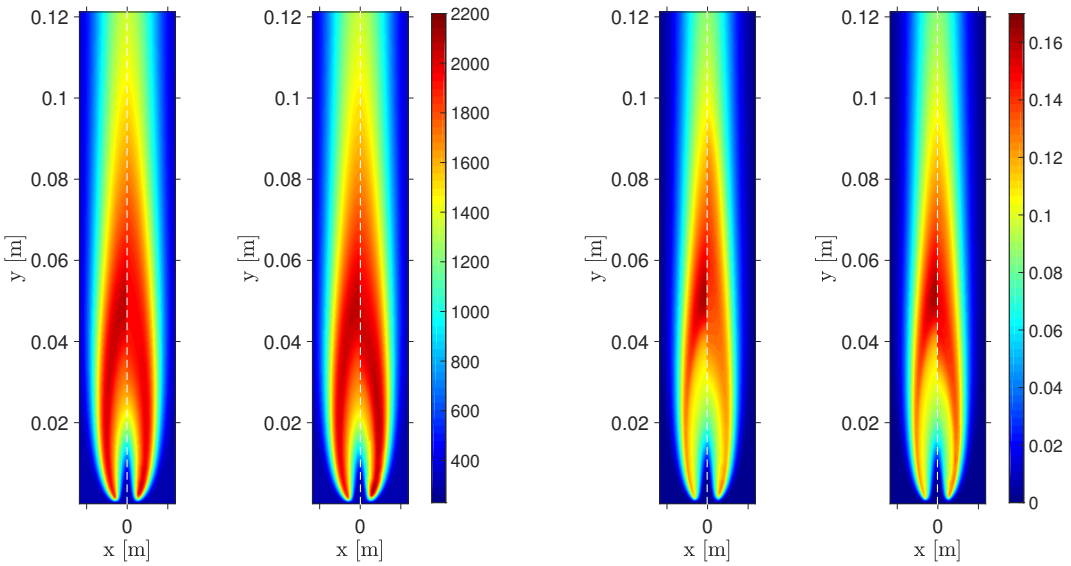


Figure 26: Left halves: true data. Right halves: (a) prediction of the temperature field by PCA; (b) prediction of the temperature field by Local PCA; (c) prediction of the CO_2 mass fraction field by PCA; (d) prediction of the CO_2 mass fraction field by Local PCA. Inlet velocity: $64\ cm/s$; inlet CH_4 molar fraction: 70%. 2D flame. PCA: 5 PCs. Local PCA: 100 clusters. Kriging: quadratic trend function, exponential kernel.

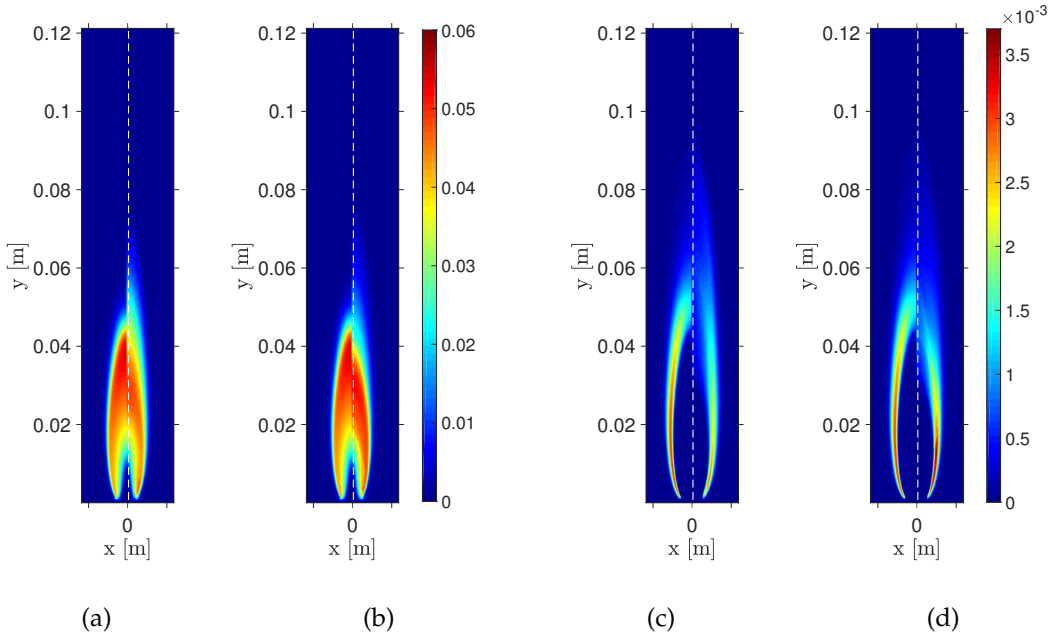


Figure 27: Left halves: true data. Right halves: (a) prediction of the CO mass fraction field by PCA; (b) prediction of the CO mass fraction field by Local PCA; (c) prediction of the OH mass fraction field by PCA; (d) prediction of the OH mass fraction field by Local PCA. Inlet velocity: $64\ cm/s$; inlet CH_4 molar fraction: 70%. PCA: 5 PCs. Local PCA: 100 clusters. Kriging: quadratic trend function, exponential kernel. 2D flame.

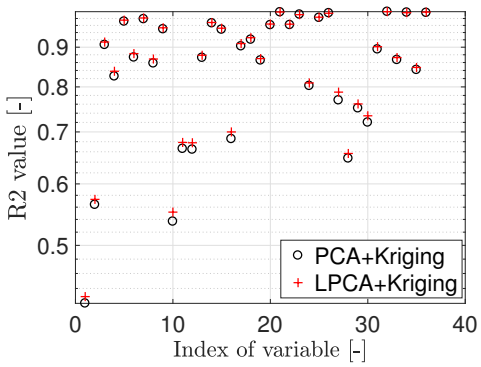


Figure 28: R2 values for all the 36 physical variables: comparison between PCA and LPCA (5 PCs, 100 clusters). Quadratic trend function, Matern52 kernel. 2D flame.

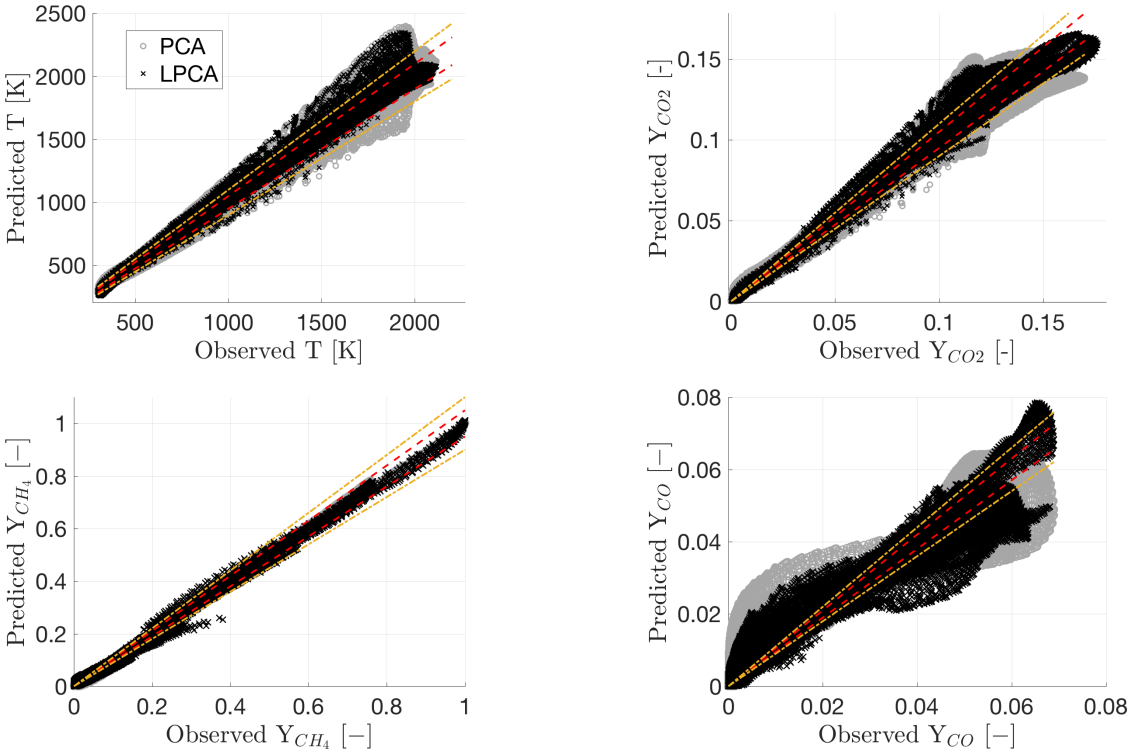


Figure 29: Parity plot for the predictions of temperature and CO_2, CH_4, CO mass fractions (using 25 training samples): comparison between PCA and LPCA (5 PCs, 100 clusters). Kriging: quadratic trend function, Matern52 kernel. Dashed: 5% error line. Dotted: 10% error line. 2D flame.

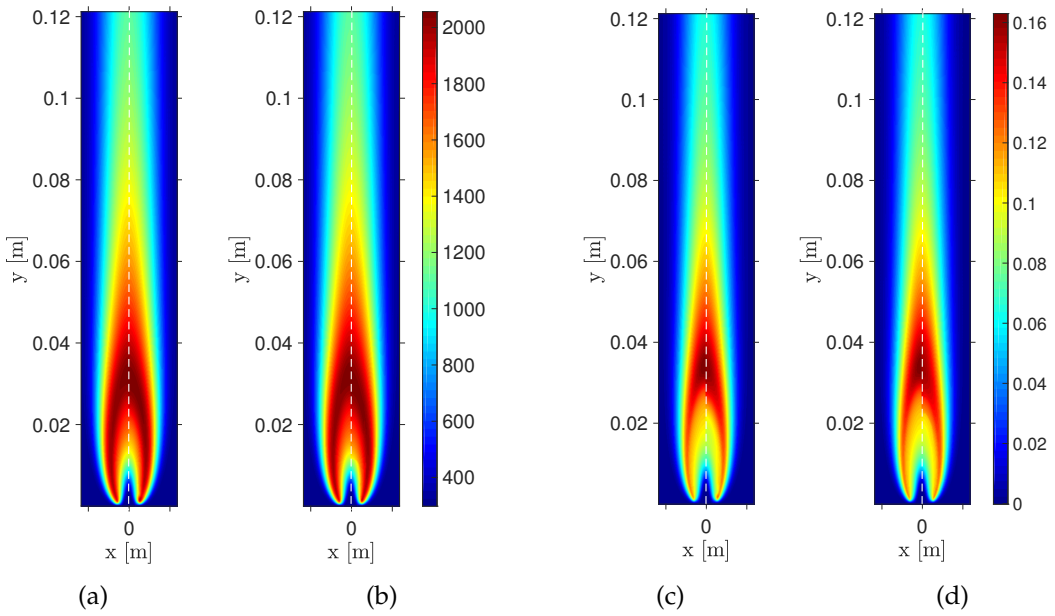


Figure 30: Left halves: true data. Right halves: (a) prediction of the temperature field by PCA; (b) prediction of the temperature field by Local PCA; (c) prediction of the CO_2 mass fraction field by PCA; (d) prediction of the CO_2 mass fraction field by Local PCA. Inlet velocity: $40\ cm/s$; inlet CH_4 molar fraction: 75%. PCA: 5 PCs. Local PCA: 100 clusters. Kriging: quadratic trend function, exponential kernel. 2D flame with a smaller range of velocities.

4.5. 2D flame with two input parameters using a smaller parameter region

The presented 2D data-set included observations for a high range of velocities. The parameter region which was intended to be explored included regions of strong non-linearities which were difficult to model with *only* 30 observations, given the high dimensionality of the data. Thus, the predictive capabilities of the developed ROM were limited by this fact. In order to confirm that, new observations were produced for a more narrow range of velocities ($24 \div 55\ m/s$). A total of 23 observations were used for the training of a new ROM based on PCA and Kriging in this region. Predictions were validated for 4 combinations of input parameters: $45\ \%$ - $35\ m/s$, $65\ \%$ - $30\ m/s$, $75\ \%$ - $40\ m/s$, $95\ \%$ - $40\ m/s$.

Figure 30 and 31 report the prediction of the temperature field and of the CO_2, CO, OH mass fraction fields by means of PCA+Kriging and Local PCA+Kriging for an inlet velocity of $40\ cm/s$ and an inlet CH_4 molar fraction of 75%. PCA was performed with a number of 5 PCs, Local PCA with a number of 100 clusters. Kriging was performed with a quadratic trend function and an exponential kernel. Figure 32 reports the parity plots for the predicted values of temperature and CO_2, CH_4, CO mass fractions. These figures clearly indicate that better predictions were obtained in the region of low inlet velocities in comparison to figure 29, where the data spread was higher. In particular, it is observable that most predictions by the Local PCA+Kriging model fall within the region delimited by the 5% error lines (dashed red). The limitations of the predictive model observed in the previous case were due to the wide range of explored conditions: the inlet velocity was indeed allowed to change roughly by a factor of 4, which led to substantial modifications in the flame topology and structure. Limiting the ratio of velocity

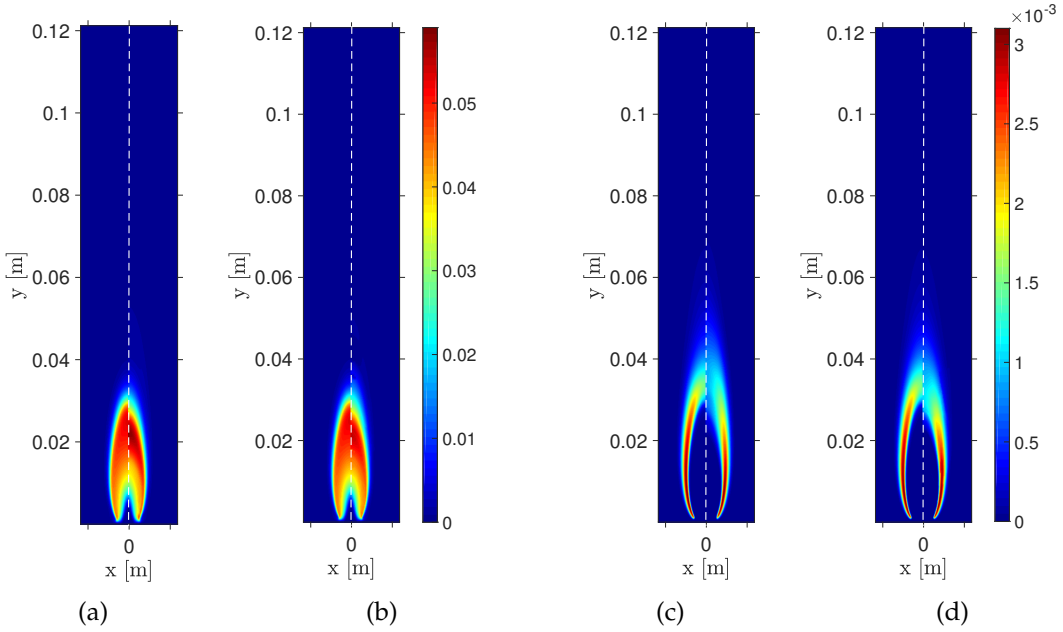


Figure 31: Left halves: true data. Right halves: (a) prediction of the CO mass fraction field by PCA; (b) prediction of the CO mass fraction field by Local PCA; (c) prediction of the OH mass fraction field by PCA; (d) prediction of the OH mass fraction field by Local PCA. Inlet velocity: $40\ cm/s$; inlet CH_4 molar fraction: 75%. PCA: 5 PCs. Local PCA: 100 clusters. Kriging: quadratic trend function, exponential kernel. 2D flame with a smaller range of velocities.

to a factor of 2 allows to develop a ROM whose reliability in the range of conditions is significantly higher.

Figure 33a reports the cumulative data variance captured by the PCs. In the previous case, the number of training samples was 25 and 20 PCs were needed to capture over 99% of the data variance. In this case, 15 PCs were sufficient to recover over 99% of data variance for 23 training observations, thus confirming that for this case the explored parameter region has been more efficiently sampled. Figure 33b reports the global R^2 values for the prediction of the validation data as the number of samples used for the training of a Local PCA+Kriging model increases. As expected, these values go to unity as more training samples are used.

4.6. Performance evaluation

The computational cost of the two-dimensional CFD with two input parameters, described in Section 4.4, is over 30 CPU-hours per simulation. By developing a ROM, as shown throughout this paper, the outcome of such simulations can be predicted instantaneously. Besides, the costs associated to the training process of a ROM based on PCA and Kriging are also smaller if compared to the computational burden needed for direct Kriging (no PCA compression is performed and one Kriging model is trained per original number of variables). Table 1 summarizes the training costs associated to different models, namely: a direct Kriging model, a PCA+Kriging model with 10 PCs, and a Local PCA+Kriging model with 10 PCs and 100 clusters. As clearly shown in Table 1, the training costs and the number of hyperparameters associated to a Direct Kriging model are very high when compared to the two ROMs based

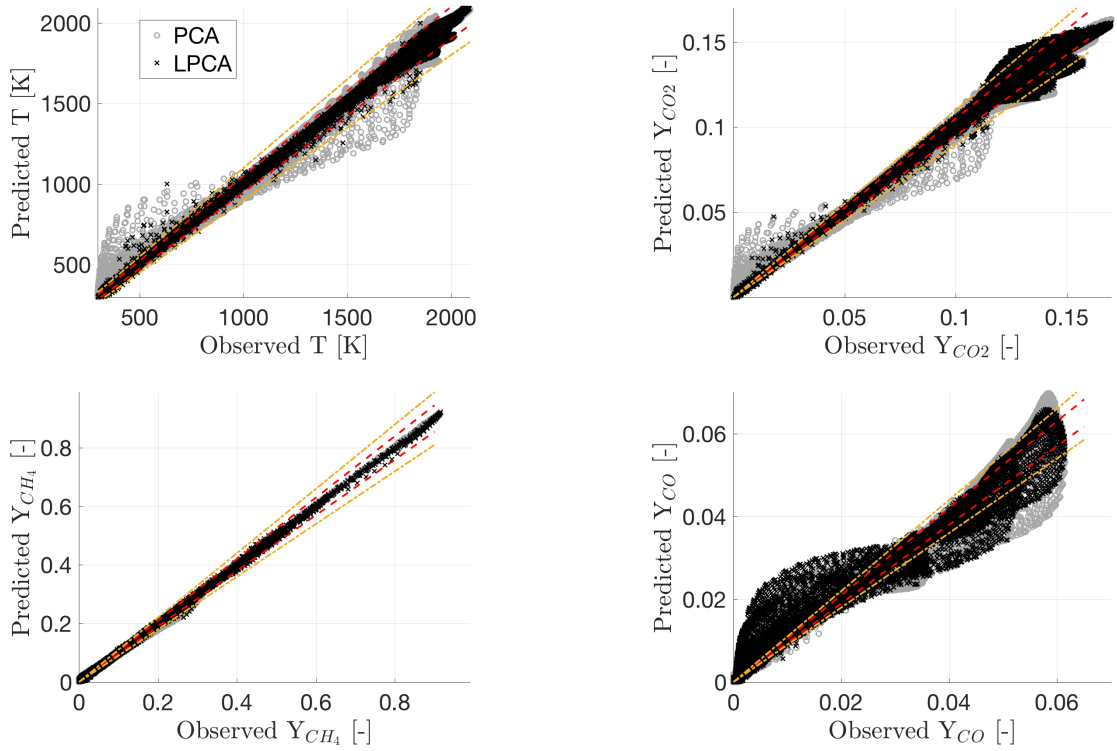
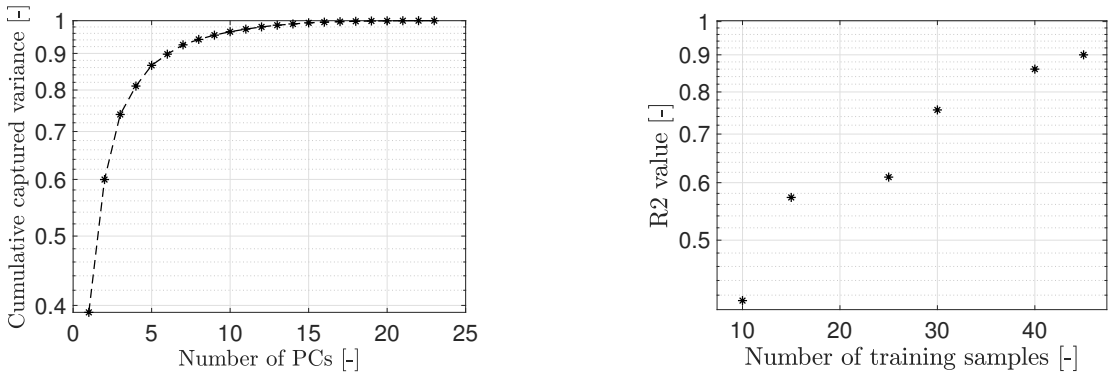


Figure 32: Parity plot for the predictions of temperature and CO_2, CH_4, CO mass fractions: comparison between PCA and LPCA (5 PCs, 100 clusters). Kriging: quadratic trend function, Matern52 kernel. Dashed: 5% error line. Dotted: 10% error line. 2D flame with a smaller range of velocities.



(a) (b)

Figure 33: (a) The cumulative original data variance that is captured when adding more PCs provides a criterion for the selection of the number of PCs when using global PCA. VAST-scaling. 2D flame with a smaller range of velocities. (b) Global R^2 values for the validation data as the number of training observations increases. Local PCA+Kriging with quadratic trend function and Matern52 kernel.

	Kriging	10 **PCs**	100 **clusters**
Training time	9.7 hrs	4 s	9 s
Clustering time	− s	− s	53 min
Speed-up	1	8 712	11
# Hyper-parameters	945 360	30	3 000
# Coefficients	0	5 672 160	5 672 160

Table 1: Comparison between the computational performances of different Kriging models: a direct Kriging model, a PCA+Kriging model with 10 PCs, and a Local PCA+Kriging model with 10 PCs and 100 clusters.

on PCA or Local PCA. On the other hand, after training a (Local) PCA+Kriging model, there is the need for additional coefficients to be stored in memory, due to the centering and scaling procedure and the PCA compression. For a PCA+Kriging model with 10 PCs, 2 vectors of mean values and scaling factors have to be stored as well as 10 PCA modes, for a total of 12 vectors of size 472 680, as the data-set was composed of 36 physical variables evaluated at 13 130 spatial locations (half of the original mesh was analyzed exploiting the problem's symmetry). The same number of coefficients needed to be stored for the Local PCA+Kriging model with 10 PCs and 100 clusters.

5. Conclusions

In the present work, a methodology for the development of accurate and robust ROM generation using a combination of PCA and Kriging was presented. The methodology was demonstrated for a 1D laminar flame with an increasing number of input parameters (equivalence ratio, composition of the fuel, inlet temperature), and for a 2D flame with two input parameters (inlet velocity and inlet fuel composition). In all three cases and for the 2D flame as well, both training and test data were available. The training data was employed to generate the ROM. The test data was used to assess the ROM's predictive capabilities.

The results showed that the combination of PCA with Kriging can be a valid solution for the development of physics-based SMs that can perform accurately with reduced computational cost. Mainly, we want to remark the following points:

- The model performed parameter exploration with low prediction errors: $< 10\%$.
- The Local PCA formulation provided an improvement over PCA as it better deals with the non-linearities of the original system.
- CPCA guarantees that the imposed physical constraints are not violated when the data are reconstructed. In this work, the imposed constraint was that all variables be positive in value when reconstructed. This was not guaranteed by PCA, although Local PCA alleviated this issue by simply improving the accuracy of the data reconstruction.
- Once the model was correctly trained, instantaneous predictions were possible, making parameter exploration much easier, even for very CPU-intensive systems.

The present work represents the first application of the PCA+Kriging methodology to combustion problems. As such, it is intended to be a proof of concept that will pave the way for the application of this methodology to more complex systems. In fact, as 3D simulations of practical combustion systems usually require significant amount of CPU hours, having a low-order model that can reliably and instantaneously predict the outcome of these simulations is precious. Moreover, the promptness of the ROM's predictions is paramount for the development of digital twins for real systems which can be employed for system control and visualization. A correctly trained ROM also grants the possibility of performing sensitivity analysis of the investigated system w.r.t. its input parameters and can be employed to solve optimization problems in the context of system design, where the evaluation of the objective function is the computational burden. The training costs of the PCA+Kriging ROM are also lower in comparison to a SM with no compression. This is very useful when new training processes are continuously needed in order to update the developed ROM in the event of new available data. The predictive capabilities of the ROM can also be employed for the initialization of complex simulation, reducing the time needed by the solver to converge. In the application of the proposed methodology to 2D flames, relevant computational savings were present as one 2D simulation needed over 30 CPU hours to run. Despite the simplicity of the test cases, the present work allowed to investigate the advantages and limitations of the method, as well as its potential for applications to more complex combustion systems. A substantial reduction in the system dimensionality was accomplished via PCA (e.g. from $10,780$ to 5 scalars in the one-parameter case), while the use of Kriging allowed to capture the non-linear relation between the reduced set of coefficients and the input parameters, enabling the prediction of non-observed system states.

Future work will involve the application of the presented methodology to more complex systems, e.g. 3D systems.

Acknowledgements

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 from the European Research Council, Starting Grant No 714605.

References

[1] B. Schleich, N. Anwer, L. Mathieu, and S. Wartzack, "Shaping the digital twin for design and production engineering," *CIRP Annals - Manufacturing Technology*, vol. 66, no. 1, pp. 141–144, 2017.

[2] T. H. Uhlemann, C. Schock, C. Lehmann, S. Freiberger, and R. Steinhilper, "The Digital Twin: Demonstrating the Potential of Real Time Data Acquisition in Production Systems," *Procedia Manufacturing*, vol. 9, pp. 113–120, 2017.

[3] K. Bizon and G. Continillo, "Reduced order modelling of chemical reactors with recycle by means of POD-penalty method," *Computers and Chemical Engineering*, vol. 39, pp. 22–32, 2012.

[4] K. Bizon, G. Continillo, M. Berezowski, and J. Smua-Ostaszewska, "Optimal model reduction by empirical spectral methods via sampling of chaotic orbits," *Physica D: Nonlinear Phenomena*, vol. 241, no. 17, pp. 1441–1449, 2012.

[5] G. Lin, "On the Bayesian calibration of expensive computer models with input dependent parameters," *Spatial Statistics*, 2017.

[6] J. Müller, C. A. Shoemaker, and R. Piché, "SO-MI: A surrogate model algorithm for computationally expensive nonlinear mixed-integer black-box global optimization problems," *Computers and Operations Research*, vol. 40, no. 5, pp. 1383–1400, 2013.

[7] R. G. Regis and C. A. Shoemaker, "Constrained global optimization of expensive black box functions using radial basis functions," *Journal of Global Optimization*, vol. 31, no. 1, pp. 153–171, 2005.

[8] M. Fürst, P. Sabia, M. Lubrano Lavadera, G. Aversano, M. de Joannon, A. Frassoldati, and A. Parente, "Optimization of Chemical Kinetics for Methane and Biomass Pyrolysis Products in Moderate or Intense Low-Oxygen Dilution Combustion," *Energy & Fuels*, p. acs.energyfuels.8b01022, 2018.

[9] B. A. Khuwaileh and P. J. Turinsky, "Surrogate Based Model Calibration for Pressurized Water Reactor Physics Calculations," *Nuclear Engineering and Technology*, vol. 49, no. 6, pp. 1219–1225, 2017.

[10] B. Beykal, F. Boukouvala, C. A. Floudas, and E. N. Pistikopoulos, "Optimal design of energy systems using constrained grey-box multi-objective optimization," *Computers and Chemical Engineering*, vol. 0, pp. 1–15, 2018.

[11] B. Beykal, F. Boukouvala, C. A. Floudas, N. Sorek, H. Zalavadia, and E. Gildin, "Global optimization of grey-box computational systems using surrogate functions and application to highly constrained oil-field operations," *Computers and Chemical Engineering*, vol. 114, pp. 99–110, 2018.

[12] T. Crestaux, O. Le Maître, and J. M. Martinez, "Polynomial chaos expansion for sensitivity analysis," *Reliability Engineering and System Safety*, vol. 94, no. 7, pp. 1161–1172, 2009.

[13] T. Lancien, N. Dumont, K. Prieur, D. Durox, S. Candel, O. Gicquel, and R. Vicquelin, "Uncertainty quantification of injected droplet size in mono-dispersed Eulerian simulations," 2016.

[14] C. M. Bishop, *Pattern Recognition and Machine Learning*, vol. 53. 2013.

[15] I. T. Jolliffe, "Principal Component Analysis, Second Edition," 2002.

[16] A. Bellemans, G. Aversano, A. Coussement, and A. Parente, "Feature extraction from reduced order models based on principal component analysis," *Computers & Chemical Engineering*, no. March, 2018.

[17] M. Xiao, P. Breitkopf, R. Filomeno Coelho, C. Knopf-Lenoir, M. Sidorkiewicz, and P. Villon, "Model reduction by CPOD and Kriging," *Structural and Multidisciplinary Optimization*, vol. 41, no. 4, pp. 555–574, 2010.

[18] B. J. Isaac, A. Coussement, O. Gicquel, P. J. Smith, and A. Parente, "Reduced-order PCA models for chemical reacting flows," *Combustion and Flame*, vol. 161, no. 11, pp. 2785–2800, 2014.

[19] B. J. Isaac, J. N. Thornock, J. Sutherland, P. J. Smith, and A. Parente, "Advanced regression methods for combustion modelling using principal components," *Combustion and Flame*, vol. 162, no. 6, pp. 2592–2601, 2015.

[20] H. Mirgolbabaei, T. Echekki, and N. Smaoui, "A nonlinear principal component analysis approach for turbulent combustion composition space," *International Journal of Hydrogen Energy*, vol. 39, no. 9, pp. 4622–4633, 2014.

[21] T. Echekki and H. Mirgolbabaei, "Principal component transport in turbulent combustion: A posteriori analysis," *Combustion and Flame*, vol. 162, no. 5, pp. 1919–1933, 2015.

[22] J. Yu, "Local and global principal component analysis for process monitoring," *Journal of Process Control*, vol. 22, no. 7, pp. 1358–1373, 2012.

[23] M. Xiao, P. Breitkopf, R. Filomeno Coelho, C. Knopf-Lenoir, and P. Villon, "Enhanced POD projection basis with application to shape optimization of car engine intake port," *Structural and Multidisciplinary Optimization*, vol. 46, no. 1, pp. 129–136, 2012.

[24] M. Xiao, P. Breitkopf, R. Filomeno Coelho, C. Knopf-Lenoir, P. Villon, and W. Zhang, "Constrained Proper Orthogonal Decomposition based on QR-factorization for aerodynamical shape optimization," *Applied Mathematics and Computation*, vol. 223, pp. 254–263, 2013.

[25] M. Xiao, P. Breitkopf, R. F. Coelho, P. Villon, and W. Zhang, "Proper orthogonal decomposition with high number of linear constraints for aerodynamical shape optimization," *Applied Mathematics and Computation*, vol. 247, pp. 1096–1112, 2014.

[26] S. Haag and R. Anderl, "Digital twin âĂŞ Proof of concept," *Manufacturing Letters*, pp. 10–12, 2018.

[27] A. Parente and J. C. Sutherland, "Principal component analysis of turbulent combustion data: Data pre-processing and manifold sensitivity," *Combustion and Flame*, vol. 160, no. 2, pp. 340–350, 2013.

[28] A. Coussement, B. J. Isaac, O. Gicquel, and A. Parente, "Assessment of different chemistry reduction methods based on principal component analysis: Comparison of the MG-PCA and score-PCA approaches," *Combustion and Flame*, vol. 168, pp. 83–97, 2016.

[29] K. Bizon, G. Continillo, E. Mancaruso, S. S. Merola, and B. M. Vaglieco, "POD-based analysis of combustion images in optically accessible engines," *Combustion and Flame*, vol. 157, no. 4, pp. 632–640, 2010.

[30] J. C. Sutherland and A. Parente, "Combustion modeling using principal component analysis," *Proceedings of the Combustion Institute*, vol. 32 I, no. 1, pp. 1563–1570, 2009.

[31] A. Parente, J. C. Sutherland, L. Tognotti, and P. J. Smith, "Identification of low-dimensional manifolds in turbulent flames," *Proceedings of the Combustion Institute*, vol. 32 I, no. 1, pp. 1579–1586, 2009.

[32] L. J. Williams, "Principal component analysis Âŧ," vol. 2, no. August, 2010.

[33] N. Kambhatla and T. Leen, "Dimension reduction by local principal component analysis," *Neural Computation*, vol. 9, no. 7, pp. 1493–1516, 1997.

[34] S. Sahyoun and S. Djouadi, "Local Proper Orthogonal Decomposition based on space vectors clustering," *Systems and Control (ICSC), 2013 3rd International Conference on*, pp. 665–670, 2013.

[35] P. G. Constantine, E. Dow, and Q. Wang, "Active Subspace Methods in Theory and Practice," *SIAM Journal of Scientific Computation*, vol. 36, no. 4, pp. 1500–1524, 2014.

[36] S. N. Lophaven, J. Søndergaard, and H. B. Nielsen, "Kriging Toolbox," pp. 1–28, 2002.

[37] M. Seeger, *Gaussian processes for machine learning.*, vol. 14. 2004.

[38] M. Guenot, I. Lepot, C. Sainvitu, J. Goblet, and R. F. Coelho, "Adaptive sampling strategies for non-intrusive POD-based surrogates," *Engineering Computations*, vol. 30, no. 4, pp. 521–547, 2013.

[39] A. Cuoci, A. Frassoldati, T. Faravelli, and E. Ranzi, "A computational tool for the detailed kinetic modeling of laminar flames: Application to C2H4/CH4 coflow flames," *Combustion and Flame*, vol. 160, no. 5, pp. 870–886, 2013.

[40] A. Cuoci, A. Frassoldati, T. Faravelli, and E. Ranzi, "Numerical modeling of laminar flames with detailed kinetics based on the operator-splitting method," *Energy and Fuels*, vol. 27, no. 12, pp. 7730–7753, 2013.

[41] A. Parente, J. C. Sutherland, B. B. Dally, L. Tognotti, and P. J. Smith, "Investigation of the MILD combustion regime via Principal Component Analysis," *Proceedings of the Combustion Institute*, vol. 33, no. 2, pp. 3333–3341, 2011.

[42] B. J. Isaac, J. N. Thornock, J. Sutherland, P. J. Smith, and A. Parente, "Advanced regression methods for combustion modelling using principal components," *Combustion and Flame*, vol. 162, no. 6, pp. 2592–2601, 2015.

[43] S. Cao, B. Bennett, B. Ma, and D. Giassi, "Effects of Fuel Dilution and Gravity on Laminar Coflow Methane-Air Diffusion Flames: A Computational and Experimental Investigation," pp. 1–9, 2013.

[44] S. Cao, B. Ma, B. A. Bennett, D. Giassi, D. P. Stocker, F. Takahashi, M. B. Long, and M. D. Smooke, "A computational and experimental study of coflow laminar methane/air diffusion flames: Effects of fuel dilution, inlet velocity, and gravity," *Proceedings of the Combustion Institute*, vol. 35, no. 1, pp. 897–903, 2015.

5.2 PCA and kriging for the efficient exploration of consistency regions in uncertainty quantification

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

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

[a] *Université Libre de Bruxelles, Aero-Thermo-Mechanics Departement, Avenue F.D. Roosevelt 51, CP 165/41, 1050 Brussels, 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 91190 Gif-sur-Yvette, France*
[d] *University of Utah, Department of Chemical Engineering, Salt Lake City, UT, USA*

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 that tries to simulate the behavior of Alstom's 15 $\mathrm{MW_{th}}$ PC tangentially fired oxy-pilot Boiler Simulation Facility (BSF) situated in CT, USA. In practice, experiments cannot provide full coverage of the pulverized-coal utility boiler due to both practicality and expense. Values of the model's parameters which guarantee consistency with the experimental data of this test facility for 121 QOIs are found by training a SM based on Kriging for only 5 PCA scores.

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 that are based on the deep understanding of the underlying processes. These mathematical models, even if deterministic, often include uncertainties

*Corresponding author:
Email addresses: Gianmarco.Aversano@ulb.ac.be (Gianmarco Aversano), jcparraa@gmail.com (Javier C. Parra-Alvarez), benjamin.j.isaac@utah.edu (Benjamin Isaac), sean.t.smith@utah.edu (Sean T. Smith), axel.coussement@ulb.ac.be (Axel Coussement), olivier.gicquel@centralesupelec.fr (Olivier Gicquel), Alessandro.Parente@ulb.ac.be (Alessandro Parente)

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 this function, i.e. the CFD model, is not feasible. Having a Surrogate Model (SM) that can approximate the costly model's prediction of the QOIs but with 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 Expasion (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 in combination of no other 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 to train is that the training can still be costly. 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 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, consequently the number of SMs to train, and thus develop reduced-order SMs, 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 values for the parameters of a CFD model [11] for the behavior of Alstom's $15\,\mathrm{MW_{th}}$ PC 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 these new set of 5 scalars and the region in the 3-dimensional parameter space for which consistency is guaranteed with the reference data 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 $y_i\ \forall i = 1, ..., N$ 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 $\mathbf{y}$ with 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 $\mathbf{\Sigma}$ is a diagonal matrix such that $\mathbf{\Sigma} = diag(\boldsymbol{\sigma})$, we can express (1) in matrix form:

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

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:

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

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 for $\mathbf{x}$ for which the predictive capabilities of the model $M(\mathbf{x})$ 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. 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 $\mathbf{V_q} = \{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_q\}$ thus the data can be encoded in a set of q scalars called PCA scores or Principal Components (PCs) as follows:

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

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 $\mathbf{A_q} = \{\mathbf{a}(\mathbf{x}_1), \mathbf{a}(\mathbf{x}_2), ..., \mathbf{a}(\mathbf{x}_M)\}$ is the matrix in which all the PCA scores for different values of the input parameters are stored; $\mathbf{a}(\mathbf{x}_j) = \{a_1(\mathbf{x}_j), ..., a_q(\mathbf{x}_j)\}^T$ 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 $\mathbf{Y}$ achieved by PCA if q PCs are retained. Equivalently, one observation $\mathbf{y}(\mathbf{x}_j) \in \mathbb{R}^N$ contained in the data-set matrix $\mathbf{Y}$ can be approximated as: $\mathbf{y}(\mathbf{x}_j) \approx \bar{\mathbf{y}} + \mathbf{D}\mathbf{V}_\mathbf{q}\mathbf{a}(\mathbf{x}_j)$.

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 $\mathbf{f}(\mathbf{x}) = [f_0(\mathbf{x}), ..., f_p(\mathbf{x})]^T$ with the weights $\boldsymbol{\beta} = [\beta_0, ..., \beta_p]^T$ 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:

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

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 of the output.

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:

$$\mathbf{a}(\mathbf{x}_m) = {\mathbf{V}_\mathbf{q}}^T \mathbf{D}^{-1} \left(\mathbf{y}(\mathbf{x}_m) - \bar{\mathbf{y}}\right) \tag{10}$$

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

The subscript $m = 1, ..., M$ 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 the predicted PCA scores in (6)) and consistency is achieved if (2) or (3) is satisfied. Computational savings are achieved because less SMs are trained ($q \ll N$).

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, and thus evaluating (13) involves less operations than evaluating (3).

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 Figure 1: dimension

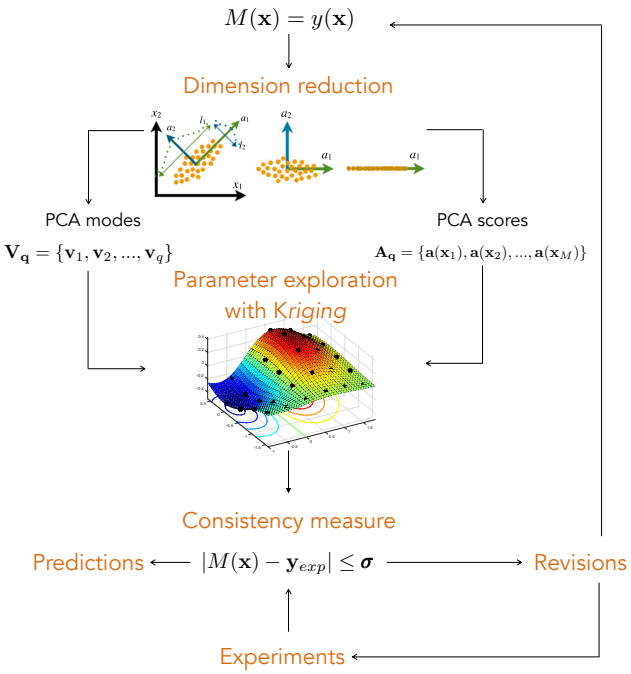


Figure 1: Flowchart for the Reduced-Order Bound-to-Bound Data Collaboration procedure. Dimension reduction is employed on a set of available observations of the model's output and combined with an interpolation technique to build a surrogate model for parameter exploration and thus find consistency with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

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

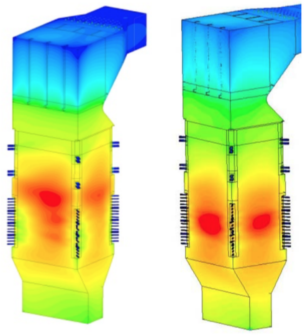


Figure 2: CFD simulations for Alstom's BSF Heat Flux profiles. Figure from [12].

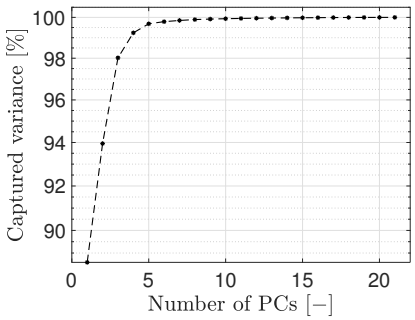


Figure 3: The cumulative original data variance that is recovered when using more principal component directions provides a criterion for the selection of the number of PCs.

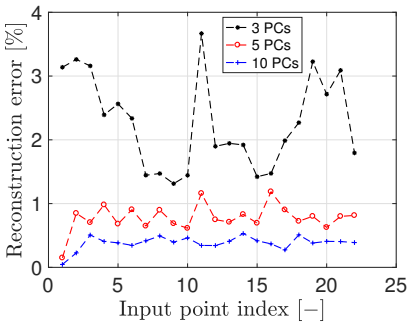


Figure 4: Error for the reconstruction of the original data by PCA when using 3 PCs (*star*), 5 PCs (*circle*) and 10 PCs (*cross*). This error is zero if all the PCs are kept.

Alstom's Boiler Simulation Facility (BSF) is a 15 MW_{th} capacity tangential-fired pilot facility located at Alstom's 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]. Alstom has applied CFD tools to modeling and design improvement of boilers since the early 1990s. Information about the CFD model can be found in [11, 22]. The computational cost of this model for the BSF is 740.000 CPU hours per simulation. Figure 2 shows wall Heat Flux profiles from CFD simulations of the BSF.

An UQ analysis is carried out in order to identify the parameters which have the highest impact, with the latter defined as $uncertainty \times sensitivity$. This study included mesh resolution, the CFL number, spatial and temporal schemes, devolatilization parameters such as the swelling factor, char oxidation

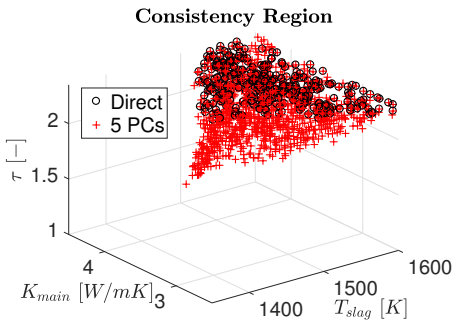


Figure 5: Consistency region found by Kriging on the original variables (*circles*) and Kriging on the first 5 PCs (*crosses*). Using the constraint (2).

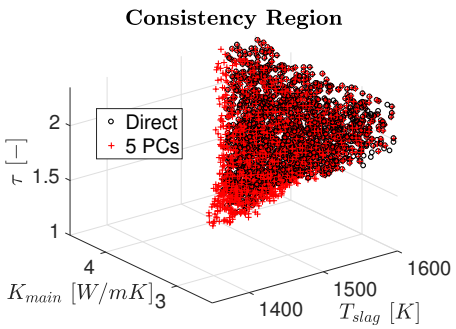


Figure 6: Consistency regions as in Figure 5 but using the constraint (3) with $\alpha = 15$.

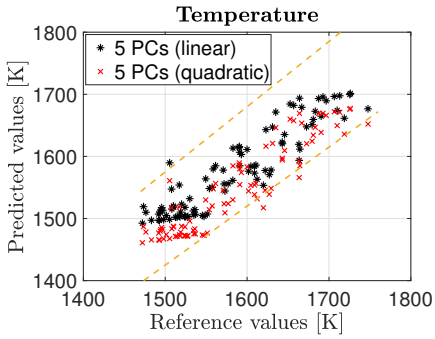


Figure 7: Parity plot for Temperature when using 5 PCs. Comparison between the reference experimental data and one consistent model's prediction according to the constraint (2) (*linear*) and (3) (*quadratic*). The dashed lines represent the 5% error.

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.

Experimental data are available as probes are present in the BSF at specific locations. A number of 22 simulations are run for different values of the described parameters, identified by means of Latin Hypercube Sampling. The parameter region or hypercube $\mathcal{H} = [1350, 1600]\ K \times [2.5, 4.5]\ W/(m \cdot K) \times [1, 2.5]$ 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

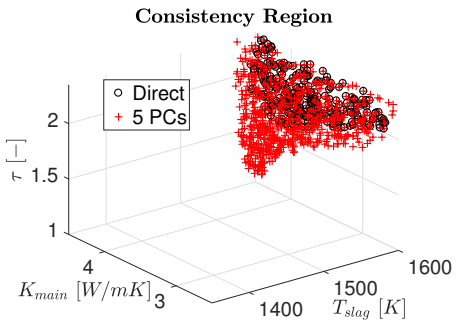


Figure 8: Consistency regions as in Figure 5 but using the constraint (3) with $\alpha = 11$.

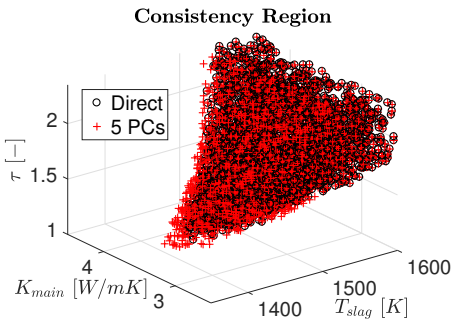


Figure 9: Consistency regions as in Figure 5 but using the constraint (3) with $\alpha = 20$.

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 the 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 explained 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 a number of 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 suggest consistency (Figure 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 difference in volume of these regions is about 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 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, predictions from SMs trained on the original variables can lie outside of it. Figure 7 is a parity plot between the reference experimental data and one SM's prediction belonging to the consistency regions shown in Figure 5 and Figure 6. 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 is still found. The choice for the value of α depends on how strict a consistency between the model's predictions and the experimental data is wanted. 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 with 5 PCs) both shrink or grow larger together. In conclusion, there is no need to train 121 SMs, because 5 PCs are enough

	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

Table 1: Comparison between the computational performances of 4 different Kriging models.

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 computational performances of the two methodologies, in the form of training times, number of hyper-parameters to train and number of coefficients to store in memory. 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 Alstom's 15 MW_{th} BSF. The BSF is a test facility for the development of Alstom's combustion systems. A CFD model of the BSF is also available [11, 22] but not fully defined as the value of 3 of its parameters is not known. 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. Suitable values for these parameters can be found using the B2B-DC approach and thus 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 simulations is 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 or PCs, 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 PCs or, more in general, of latent variables and not in terms of the original physical variables involved. Results obtained from the Reduced-Order B2B-DC approach are compared with those of a consistency analysis carried out without the use of PCA. The results show that the Reduced-Order B2B-DC approach is able to find the consistency region with a difference in volume of about 5% if the input space is standardized and with a smaller set of variables. The advantages of the approach include computational savings as less SMs need to be trained: less hyper-parameters have to be found for the construction of the needed 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.

References

[1] D. R. Yeates, W. Li, P. R. Westmoreland, W. Speight, T. Russi, A. Packard, M. Frenklach, Integrated data-model analysis facilitated by an Instrumental Model, Proceedings of the Combustion Institute 35 (1) (2015) 597–605.
URL http://dx.doi.org/10.1016/j.proci.2014.05.090

[2] G. Lin, On the Bayesian calibration of expensive computer models with input dependent parameters, Spatial Statistics.
URL http://dx.doi.org/10.1016/j.spasta.2017.08.002

[3] B. A. Khuwaileh, P. J. Turinsky, Surrogate Based Model Calibration for Pressurized Water Reactor Physics Calculations, Nuclear Engineering and Technology 49 (6) (2017) 1219–1225.
URL http://linkinghub.elsevier.com/retrieve/pii/S1738573317303182

[4] J. Jatnieks, M. De Lucia, D. Dransch, M. Sips, Data-driven Surrogate Model Approach for Improving the Performance of Reactive Transport Simulations, Energy Procedia 97 (2016) 447–453.
URL http://dx.doi.org/10.1016/j.egypro.2016.10.047

[5] R. E. Edwards, J. New, L. E. Parker, B. Cui, J. Dong, Constructing large scale surrogate models from big data and artificial intelligence, Applied Energy 202 (2017) 685–699.
URL http://dx.doi.org/10.1016/j.apenergy.2017.05.155

[6] N. E. Owen, P. Challenor, P. P. Menon, S. Bennani, Comparison of surrogate-based uncertainty quantification methods for computationally expensive simulators (2015) 1–39.
URL http://arxiv.org/abs/1511.00926

[7] M. De Lozzo, A. Marrel, Sensitivity analysis with dependence and variance-based measures for spatio-temporal numerical simulators, Stochastic Environmental Research and Risk Assessment 31 (6) (2017) 1437–1453.

[8] S. Dubreuil, M. Berveiller, F. Petitjean, M. Salaün, Construction of bootstrap confidence intervals on sensitivity indices computed by polynomial chaos expansion, Reliability Engineering and System Safety 121 (2014) 263–275.
URL http://dx.doi.org/10.1016/j.ress.2013.09.011

[9] A. Marrel, B. Iooss, B. Laurent, O. Roustant, Calculations of Sobol indices for the Gaussian process metamodel, Reliability Engineering and System Safety 94 (3) (2009) 742–751.

[10] I. T. Jolliffe, Principal Component Analysis, Second Edition.

[11] J. Pedel, J. N. Thornock, S. T. Smith, P. J. Smith, International Journal of Multiphase Flow Large eddy simulation of polydisperse particles in turbulent coaxial jets using the direct quadrature method of moments, International Journal of Multiphase Flow 63 (2014) 23–38.
URL http://dx.doi.org/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.

[13] R. Feeley, P. Seiler, A. Packard, M. Frenklach, Consistency of a reaction dataset, Journal of Physical Chemistry A 108 (44) (2004) 9573–9583.

[14] M. Frenklach, A. Packard, P. Seiler, Prediction uncertainty from models and data, Proceedings of the American Control Conference 5 (2002) 4135–4140.

[15] M. Frenklach, A. Packard, P. Seiler, R. Feeley, Collaborative data processing in developing predictive models of complex reaction systems, International Journal of Chemical Kinetics 36 (1) (2004) 57–66.

[16] A. Parente, J. C. Sutherland, Principal component analysis of turbulent combustion data: Data pre-processing and manifold sensitivity, Combustion and Flame 160 (2) (2013) 340–350.
URL http://dx.doi.org/10.1016/j.combustflame.2012.09.016

[17] A. Coussement, B. J. Isaac, O. Gicquel, A. Parente, Assessment of different chemistry reduction methods based on principal component analysis: Comparison of the MG-PCA and score-PCA approaches, Combustion and Flame 168 (2016) 83–97.
URL http://www.sciencedirect.com/science/article/pii/S0010218016300384

[18] K. Bizon, G. Continillo, E. Mancaruso, S. S. Merola, B. M. Vaglieco, POD-based analysis of combustion images in optically accessible engines, Combustion and Flame 157 (4) (2010) 632–640.
URL http://dx.doi.org/10.1016/j.combustflame.2009.12.013

[19] P. G. Constantine, E. Dow, Q. Wang, Active Subspace Methods in Theory and Practice, SIAM Journal of Scientific Computation 36 (4) (2014) 1500–1524.
URL http://inside.mines.edu/{~}pconstan/docs/constantine-asm.pdf

[20] S. N. Lophaven, J. Søndergaard, H. B. Nielsen, Kriging Toolbox (2002) 1–28.

[21] M. Seeger, Gaussian processes for machine learning., Vol. 14, 2004.

[22] J. Pedel, J. N. Thornock, P. J. Smith, Ignition of co-axial turbulent diffusion oxy-coal jet flames : Experiments and simulations collaboration, Combustion and Flame 160 (6) (2013) 1112–1128.
URL http://dx.doi.org/10.1016/j.combustflame.2013.01.022

[23] G. R. Hadley, Thermal conductivity of packed metal powders, International Journal of Heat and Mass Transfer 29 (6) (1986) 909–920.

[24] N. Akkiraju, H. Edelsbrunner, M. Facello, P. Fu, E. P. Mucke, C. Varela, Alpha Shapes: Definition and Software, Proc. Internat. Comput. Geom. Software Workshop (1995) 1–8.

5.3 COMBINATION OF POLYNOMIAL CHAOS AND KRIGING FOR REDUCED-ORDER MODEL OF REACTING FLOW APPLICATIONS

Combination of polynomial chaos and Kriging for reduced-order model of reacting flow applications

Gianmarco Aversano[a,c], Giuseppe D'Alessio[a,c], Zhiyi Li[a,c], Axel Coussement[a,c], Francesco Contino[b,c], Alessandro Parente[a,c]

[a] *Université Libre de Bruxelles, Aero-Thermo-Mechanics Departement, Avenue F.D. Roosevelt 51, CP 165/41, 1050 Brussels, Belgium*
[b] *Fluid and Thermal Dynamics (FLOW), Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium*
[c] *Université Libre de Bruxelles and Vrije Universiteit Brussel, Combustion and Robust Optimization Group (BURN), Brussels, Belgium*

Abstract

The combination of Proper Orthogonal Decomposition (POD) with Kriging has been shown to be a reliable choice for the development of Reduced-Order Models (ROMs) for the prediction of combustion data at unexplored operating conditions. In this study, the authors combine POD with Polynomial Chaos Expansion (PCE), with a combination of PCE and Kriging (PC-Kriging) and with Artificial Neural Networks (ANN) for the development of a ROM that can predict 2D combustion data for unexplored operating conditions. The choice of Non-negative Matrix Factorization (NMF) instead of POD as compression method is also investigated. This method is chosen because it can intrinsically guarantee the non-violation of physical constraints such as positivity of chemical species mass fractions, although POD's data reconstruction errors are lower. The performances of the POD and NMF in combination with the proposed supervised methods are compared, with prediction normalized root mean squared errors (NRMSE) being less than 10% for spatial fields of temperature, CH_4 and O_2 for all approaches.

Keywords: PCA, Surrogate models, Polynomial Chaos, Kriging

1. Introduction

High-fidelity expensive computer simulation are necessary for the description of many complex physical systems. This is the case with computational fluid dynamics (CFD) and combustion, where design optimization studies are limited by the computational cost of running a large number of simulations. Due to the non-linearity of CFD and combustion systems, changing operating conditions can lead to drastic changes in the state of the considered system. Thus, complete knowledge about the investigated system's behavior for a full range of operating conditions can only be achieved by running these expensive simulations several times with different inputs, until enough observations of the system's state are obtained.

Surrogate modeling reduces the computational cost of these studies by evaluating only a small subset of the proposed simulations and fitting a computationally cheap model to them. This model can then be quickly evaluated to further guide designs instead of simply running additional CFD simulations.

In a previous study [1], the authors showed that the combination of an unsupervised method for data compression such as Principal Component Analysis (PCA) or Proper Orthogonal Decomposition (POD), with a supervised technique such as Kriging for the development of Reduced-Order Models (ROMs) is a reliable choice for the prediction of combustion data, namely spatial profiles of temperature and chemical species mass fractions, for unseen operating conditions, e.g. inlet velocity and fuel composition of a 2D laminar methane flame. POD could compress the size of the high-dimensional data whose prediction is of interest, leading to a high reduction in the number of surrogate models (SMs) to be trained.

Because of the reconstruction errors that are involved in low-rank approximations such as the one provided by POD, important physical laws such as positivity of the chemical species mass fractions might be violated. The approach of finding reduced rank non-negative factors to approximate a given non-negative data matrix thus becomes a natural choice. Even though Constrained PCA (CPCA) has shown to be capable to alleviate or solve this problem [1, 2], in the present work the authors investigate the choice of Non-negative Matrix Factorization (NMF) as a data size reduction instead of POD. Thus, the ability of NMF to reconstruct training and test data is assessed, in comparison to POD, as well as the predictive capabilities of a ROM that combines NMF instead of POD with a supervised technique, such as PC-Kriging.

In the present study, the authors also investigate the effects of combining Kriging with Polynomial Chaos Expansion (PCE) for the supervised part of their approach. As shown in [3], PCE can act as Kriging trend function, increasing the model's predictive capabilities especially in absence of data, while Kriging helps to interpolate the training data. The combination of the two methods, here referred to as PC-Kriging or PCK, have already produced encouraging results as shown in [3, 4]. However, in the present study the approach is applied to reacting systems, which are multi-physics, multi-scale problems, so to pave the way for the use of this approach to real industrial applications. Besides, no compression method was used before training a SM based on PC-Kriging in [3, 4].

A compression method grants the possibility to encode high-dimensional vectors (such as a spatial field) into a few scalars, as explained in [1] in the context of combustion data, thus reducing the number of predictive models to train. However, although the size of the reduced data can be even 10 times smaller than the original data size, the reduced size can still be large if strong non-linearities are inherent to the considered problem, meaning that a large number of predictive models still needs to be trained. Thus, in the present work, the authors also investigate the combination of a compression method such as POD or NMF with Artificial Neural Networks (ANNs) as one advantage of ANN is that it makes it possible to train only one net that is able to predict the values of all the targets, namely the POD or NMF scores, for new input parameters. Whereas with the other techniques, it is necessary to train one model per score.

The discussed approaches are tested on 2 designs of experiment (DoE) made out of a total of 64 simulations produced by OpenFOAM, spanning the two input parameters in the range $24 \div 89\ cm/s$ and $40 \div 100\ \%$ for the inlet velocity and inlet molar fraction of CH_4, respectively. A DoE with 24 training simulations (DoE-A) was firstly used to compare the performances of POD+Kriging, POD+PCE and POD+PC-Kriging. The small number of training simulations was chosen as results from [3] indicate that the PC-Kriging approach works better for small experimental designs. Then, the approaches are also compared on a bigger experimental design, namely DoE-B, which included 55

training simulations. In both cases, the training samples were randomly chosen. This part of the SM training was not optimized as the main objective of the present study was to compare the methodologies, more than effectively train a ROM. Besides, by not optimizing the sampling procedure, the authors guaranteed that similarities in performances, in particular in high predictive capabilities, were not due to abundance of available data. ROMs with low prediction errors are nonetheless constructed in the present work and a leave-k-out analysis is also carried out [5].

The paper is organized as follows: the methodology employed for the ROM development is explained briefly in Section 2, as well as the chosen unsupervised and supervised techniques, such as POD, NMF, PCE, Kriging, PC-Kriging and ANN. Section 3 introduces the combustion data-set chosen to validate the developed ROMs and shows the results obtained from data compression performed by means of POD and NMF. Section 4 assesses the predictive capabilities of PCE, Kriging, PC-Kriging and ANN when combined with POD, while Section 5 discusses the performances of an NMF-based ROM using PC-Kriging. Finally, conclusions are drawn in Section 6.

2. Theoretical background

2.1. Methodology

In the present work, a data matrix containing m output of a certain computer model or CFD simulation, indicated by $\mathcal{F}$, will be indicated by a matrix $\mathbf{Y}$ of size $(m \times n)$, where n is the size of the output of the considered model. The m combinations of values for the d input parameters to the simulations that produced $\mathbf{Y}$ are collected in the matrix $\mathbf{X}$ of size $(m \times d)$. One particular simulation or output of the considered computer code for one combination of the input parameters $\mathbf{x}$ is indicated by $\mathbf{y}(\mathbf{x}) = \mathcal{F}(\mathbf{x})$.

In the present work, a low-rank approximation of $\mathbf{Y}$ will be attained, so that each particular simulation output $\mathbf{y}(\mathbf{x})$ can be expressed as a weighted sum of basis functions, as shown in Figure 1. By doing so, it is subsequently possible to train SM for unexplored combinations of the input

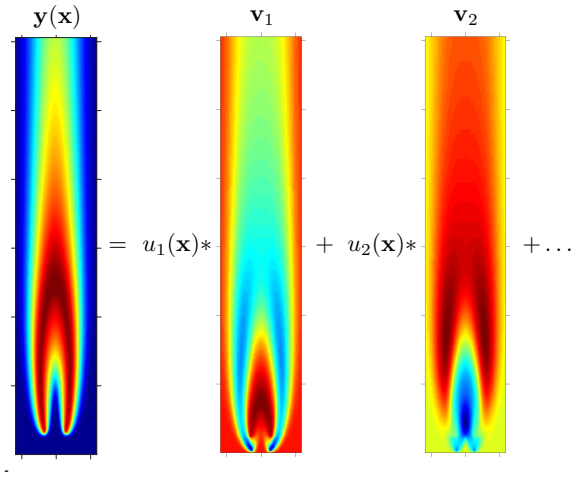


Figure 1: Illustrative example of low-rank approximation of a spatial field $\mathbf{y}$ observed for a certain operating condition $\mathbf{x}$, represented by a set of coordinates (the coefficients u_i) on some basis functions $\mathbf{v}_i$ [1]: $\mathbf{y}(\mathbf{x}) = \sum_{i=1}^{k} \mathbf{v}_i u_i(\mathbf{x})$, where k is the total number of used basis functions.

parameter only for the coefficients of the basis functions [1]. Two different methods to perform low-rank approximation of a given data matrix and thus find the basis functions and their coefficients are explained in the next Section. Each method requires the data matrix to be pre-processed in a certain way [6, 7]. In the following Section, it is implied that the data matrix is appropriately pre-processed (centered and scaled) for each method.

2.2. Low-rank approximations

Data compression or low-rank approximation is a lower dimensional representation of a higher-dimensional data-set $\mathbf{Y}$ of size $(m \times n)$. This matrix is compressed to or represented by a lower dimensional matrix of size $(m \times k)$, with $k < n$ and k being called the approximation order.

2.2.1. Proper Orthogonal Decomposition

The POD problem can be stated as follows: given a matrix $\mathbf{Y}$ of size $(m \times n)$, find $\mathbf{Z}$ of size $(m \times k)$ and $\mathbf{A}$ of size $(n \times k)$ with $k < n$ such that the functional $f(\mathbf{Z}, \mathbf{A}) = \frac{1}{2}||\mathbf{Y} - \mathbf{Z}\mathbf{A}^T||^2$ is minimized, subject to $\mathbf{A}^T\mathbf{A} = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix. This problem can be solved by computing the singular value decomposition (SVD) of the matrix $\mathbf{Y}$, which corresponds to finding the eigenvectors and eigenvalues of the matrix $\mathbf{C} = \frac{1}{m-1}\mathbf{Y}^T\mathbf{Y}$ [8]. The eigenvectors of $\mathbf{C}$ are the POD modes and the associated eigenvalues represent their relevance for the low-rank approximation of $\mathbf{Y}$. The POD modes are thus found all at once, and by ordering them in descending order according to their corresponding eigenvalue and retaining only a subset $k < n$ of them, a low-rank approximation of $\mathbf{Y}$ is possible as follows $\mathbf{Y} \approx \mathbf{Z}\mathbf{A}^T = \mathbf{Y}\mathbf{A}\mathbf{A}^T$, where the columns of $\mathbf{A}$ of size $(n \times k)$ are the POD modes and $\mathbf{Z}$ of size $(m \times k)$ is the matrix of POD coefficients. Each row of $\mathbf{Z}$ are the k coefficients for the retained k POD modes so that one particular simulation, or row of $\mathbf{Y}$, can be expressed as $\mathbf{y}(\mathbf{x}_j) = \sum_{i=1}^{k} \mathbf{a}_i z_i(\mathbf{x}_j)$. As the solution of the POD problem only leads to the evaluation of $\mathbf{A}$, the matrix $\mathbf{Z}$ can be computed only after $\mathbf{A}$ is known, as follows: $\mathbf{Z} = \mathbf{Y}\mathbf{A}$. This holds for new, unseen data as well: $\mathbf{Z}' = \mathbf{Y}'\mathbf{A}$.

2.2.2. Non-negative matrix factorization

The NMF problem can be stated as follows: given a non-negative matrix $\mathbf{Y}$ of size $(m \times n)$, find non-negative matrix factors $\mathbf{U}$ of size $(m \times k)$ and $\mathbf{V}$ of size $(n \times k)$ with $k < n$ such that the functional $f(\mathbf{U}, \mathbf{V}) = \frac{1}{2}||\mathbf{Y} - \mathbf{U}\mathbf{V}^T||^2$ is minimized, subject to $u_{ij}, v_{ij} > 0 \quad \forall i, j$ [9, 10]. Differently from POD, the non-negative factors or NMF modes are found for a given approximation order k. For a different value of k, the NMF problem needs to be solved again. As for POD, once the NMF problem is solved and thus the NMF modes are found, one particular simulation, or row of $\mathbf{Y}$, can be expressed as $\mathbf{y}(\mathbf{x}_j) = \sum_{i=1}^{k} \mathbf{v}_i u_i(\mathbf{x}_j)$. Differently from POD, both matrices $\mathbf{U}$ and $\mathbf{V}$ are determined by solving the NMF problem. The determination of the NMF scores $\mathbf{U}'$ for new, unseen data $\mathbf{Y}'$ is possible only by solving the least-squares error minimization problem: $\mathbf{U}' = \underset{\mathbf{U}'}{\operatorname{argmin}} \frac{1}{2}||\mathbf{Y}' - \mathbf{U}'\mathbf{V}^T||^2$, whose solution is taken as $\mathbf{U}' = \mathbf{Y}' \left(\mathbf{V}^T\mathbf{V}\right)^{-1} \mathbf{V}^T$.

2.3. Surrogate modeling

After a low-rank approximation of the data matrix is found, by either POD or NMF, the lower-dimensional data, namely the POD or NMF scores, need to be predicted for new values of the input

parameters to the computational model, as explained in Section 2.1. To this purpose, the following supervised methods will be used in the present work, so that a response surface for the compressed data is found.

The methods for surrogate modeling presented next require the matrix of the input parameter values $\mathbf{X}$ to be pre-processed, namely centered and scaled, before the surrogate models can be trained, thus this will be implied.

2.3.1. Polynomial Chaos Expansion

The theoretical background for Polynomial Chaos Expansion (PCE) is reported from [3, 4]. Consider a system whose behavior is represented by a computational model $\mathcal{M}$ which maps the d-dimensional input parameter space to the 1-dimensional output space $\mathcal{M} : \mathbf{x} \in \mathbb{R}^d \rightarrow y \in \mathbb{R}$. In Section 2.1, the computational model was indicated by $\mathcal{F}$ and its output was multi-dimensional. For this reason, in this section the symbol $\mathcal{M}$ will be employed. In the present work, the components of the input vector $\mathbf{x} = x_1, \ldots, x_d$ are assumed independent. The case of dependent input variables can easily be addressed as explained in [11]. In the present work, the authors consider that the computational model $\mathcal{M}$ is a deterministic mapping from the input to the output space, i.e. repeated evaluations with the same input values lead to the same output value. As explained in [3], the computational model $\mathcal{M}$ can be approximated by a finite, truncated set of polynomials:

$$y(\mathbf{x}) = \mathcal{M}(\mathbf{x}) \approx \sum_{\boldsymbol{\alpha} \in \mathbb{N}^d} a_{\boldsymbol{\alpha}} \boldsymbol{\psi}_{\boldsymbol{\alpha}}(\mathbf{x}), \tag{1}$$

where $a_{\boldsymbol{\alpha}}$ are the expansion coefficients of the multivariate polynomials $\boldsymbol{\psi}_{\boldsymbol{\alpha}}(\mathbf{x})$ and $\boldsymbol{\alpha}$ is the multi-index. Because of the statistical independence of the input variables, the multivariate polynomials are evaluated as product of uni-variate polynomials $\boldsymbol{\psi}_{\boldsymbol{\alpha}}(\mathbf{x}) = \prod_{i=1}^{d} \boldsymbol{\psi}_{\alpha_i}^{(i)}(x_i)$, where $\boldsymbol{\psi}_{\alpha_i}^{(i)}$ is the polynomial of degree α_i for the i-th variable. The total degree of the (multivariate) polynomials is defined by $|\boldsymbol{\alpha}| = \sum_{i=1}^{d} \alpha_i$. The total number of (multivariate) polynomials depends on the adopted truncation scheme. In the present work, the adopted truncation scheme is the *hyperbolic truncation set* [3]: $\boldsymbol{\alpha} \in \mathbb{N} : ||\boldsymbol{\alpha}||_q \leq N$, where N is the total order of the polynomials and the norm $|| \cdot ||_q$ is defined as $||\boldsymbol{\alpha}||_q = \left(\sum_{i=1}^{d} \alpha_i^q\right)^{1/q}$.

The maximum number of terms in the polynomial basis, attainable for $q = 1$, is given in Eq. (2):

$$p + 1 = \frac{(d + N)!}{d!\, N!}. \tag{2}$$

The expansion coefficients $a_{\boldsymbol{\alpha}}$ can be estimated by a least-square minimization method, as explained in [3].

2.3.2. Kriging

Kriging is an interpolation method in which every realization $y(\mathbf{x})$ is expressed as a combination of a trend function and a residual [12]:

$$y(\mathbf{x}) = \mu(\mathbf{x}) + s(\mathbf{x}) = \sum_{i=0}^{p} \beta_i f_i(\mathbf{x}) + s(\mathbf{x}) = \mathbf{f}^T(\mathbf{x})\boldsymbol{\beta} + s(\mathbf{x}) \tag{3}$$

The trend function $\mu(\mathbf{x})$ is a low-order polynomial regression and provides a global model in the input space. The term $s(\mathbf{x})$ creates a localized deviation weighting the points in the training set that

are closer to the target point $\mathbf{x}$. The trend function $\mu(\mathbf{x})$ is expressed as a weighted linear combination of $p+1$ polynomials $\mathbf{f}(\mathbf{x}) = [f_0(\mathbf{x}),\ ...,\ f_p(\mathbf{x})]^T$ with the weights $\boldsymbol{\beta} = [\beta_0,\ ...,\ \beta_p]^T$ determined by generalized least squares (GLS). The subscript p also indicates the degree of the polynomial. The residuals $s(\mathbf{x})$ are modeled by a Gaussian process with a kernel or correlation function that depends on a set of hyper-parameters $\mathbf{l}$ to be evaluated by Maximum Likelihood Estimation (MLE) [13, 14].

In the definition of both the trend function and the residual, it is up to the designer to choose the polynomials $\mathbf{f}(\mathbf{x})$ and the correlation model or kernel. In this way, prior knowledge can be added into the problem.

2.3.3. PC-Kriging

As explained in [3], Kriging is able to interpolate local variations of the output of the computational model. In contrast, polynomial chaos expansions (PCE) are generally used for approximating global behaviors of computational models. The two techniques can be combined if PCE is used as trend function for the Kriging interpolation method. This approach is referred to as PC-Kriging and its formulation is as follows:

$$y(\mathbf{x}) = \mathcal{M}(\mathbf{x}) \approx \sum_{\boldsymbol{\alpha} \in \mathcal{A}} a_{\boldsymbol{\alpha}} \boldsymbol{\psi}_{\boldsymbol{\alpha}}(\mathbf{x}) + s(\mathbf{x}). \tag{4}$$

Building a PC-Kriging meta-model consists of determining the optimal set of polynomials for PCE first and then calibrating the Kriging hyper-parameters $\mathbf{l}$.

2.3.4. Artificial Neural Networks

Artificial Neural Networks (ANN) [15, 16, 17] are a supervised method that have certain characteristics in common with biological neural networks and have been developed as generalization of mathematical models. An ANN is characterized by its patterns of connections between neurons (architecture), its method on determining suitable values for the weights which its architecture is composed of (training or learning algorithm), its activation functions. Figure 2 is an illustrative example of an Artificial Neural Network with 2-dimensional input layer, 4-dimensional hidden layer

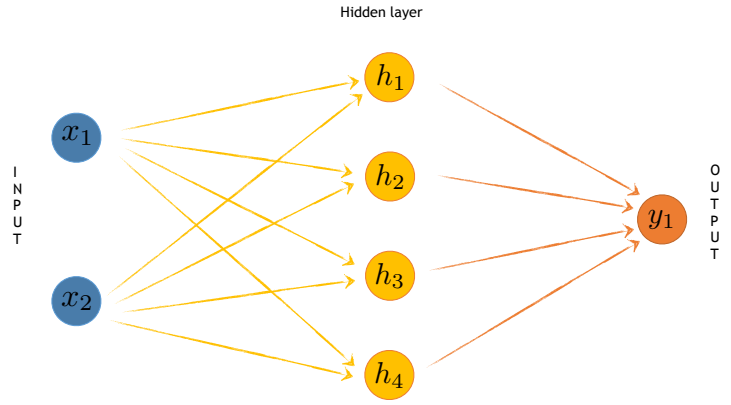


Figure 2: Illustrative example of an Artificial Neural Network with 2-dimensional input layer, 4-dimensional hidden layer and 1-dimensional output layer. The value of h_i is given by $h_i = \sum_{j=1}^{d} f_l(x_j w_{ji}^{(l)} + b_j^{(l)})$, where d is the size of the input layer, $f_l()$ is the activation function for l-th layer, $w_{ji}^{(l)}$ is the weight from the j-th input to h_i (in layer l) and $b_j^{(l)}$ is the bias for the j-th input (in layer l).

and 1-dimensional output layer. The size of the output layer can even be greater than 1. In such a case, one ANN can be trained for multiple outputs. For instance, only one net can be trained to predict the values of the POD or NMF scores altogether. In the case of Kriging, PCE or PC-Kriging, this is not possible and the models have to be trained for each score separately.

3. Data-set description and data compression

The configuration of the simulated flame is described in [18] and in [19]. The computational domain starts from the exit of the nozzle and extends 122 mm further downstream. The radial direction is expanded to 42.88 mm. A 2D structured axi-symmetric mesh with around 25600 cells is used and the nozzle radius is resolved with 12 cells. The laminarBuoyantSimpleSMOKE solver is applied. Gravity is turned on. The multi-component diffusion model is adopted to consider molecular diffusion. The GRI3.0 mechanism without NO_x (35 species and 219 reactions) is applied. For the velocity boundary condition, the profile provided from [18] and in [19] is used. The input parameters are two, namely the inlet velocity and the molar fraction of CH_4 in the inlet stream, which is a mixture of CH_4 and N_2. 64 samples were produced by OpenFOAM, spanning the two input parameters in the range $24 \div 55$ cm/s and $40 \div 100$ % for the inlet velocity and inlet molar fraction of CH_4, respectively. The approaches investigated in the present work will only be performed on the following spatial profiles: CH_4, CO, CO_2, H_2, H_2O, N_2, O_2, OH and temperature.

3.1. Results from data compression

One of the advantages of POD is that the POD modes can all be estimated beforehand. Besides, the POD modes come with an eigenvalue associated to them representing the portion of the original data variance that they account for. This made it possible to solve the POD problem once and then choose the number of modes to retain by looking at the cumulative data variance data that they recovered, differently from NMF, where a different choice of the approximation order, namely the number of modes or non-negative factors to find, means that the NMF problem has to be solved again.

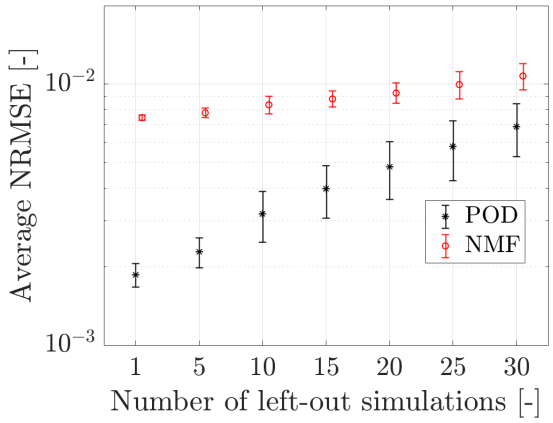
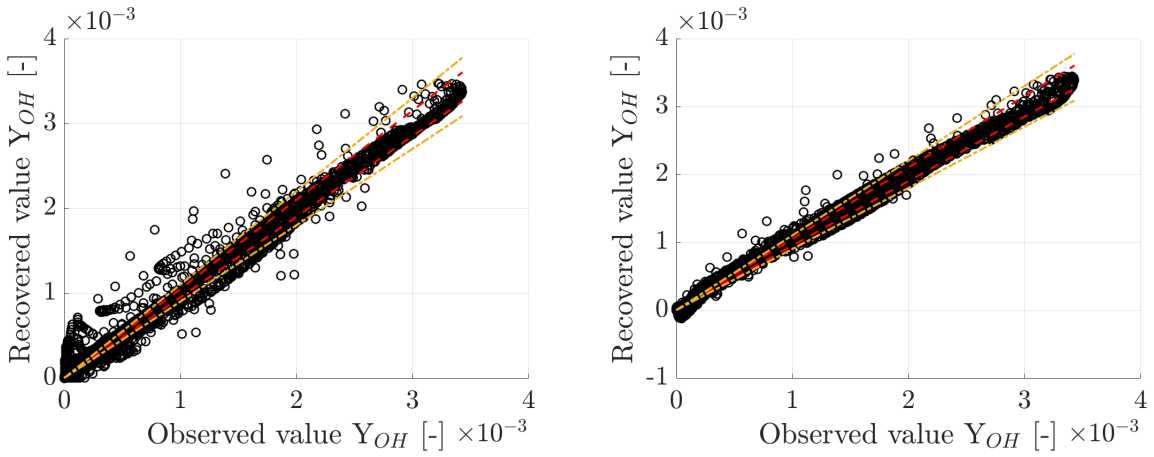


Figure 3: Average NRMSE for the reconstruction of the test data by POD and NMF (both with a number of 30 modes) for an increasing number of left-out simulations.

A leave-k-out analysis was carried out as well in order to assess the influence of the number of training simulations on POD and NMF. Leave-k-out analysis is generally a more robust way to assess how a model will generalize to unseen data. k was the number of simulations left out from the training data set employed to find the POD and NMF reduced basis. Each time, k simulations were left out and the error to reconstruct the left-out simulations from the POD and NMF basis was estimated. Figure 3 reports the normalized root mean squared errors (NRMSE) for the reconstruction of the test data by POD and NMF with 30 modes for an increasing number of left-out simulations. In terms of reconstruction, POD was clearly superior to NMF. The gap between the two methods in average NRMSE for the test data reconstruction decreased for smaller training sizes. Figure 4 is a



(a) NMF. (b) POD.

Figure 4: Reconstruction of OH mass fraction. Although more accurate, POD recovered negative values which are physically impossible. The inner and out dashed lines represent the 5% and 10% error, respectively.

parity plot for the true, observed values for the OH mass fraction and the ones recovered by (a) NMF and (b) POD, for a case where 10 random simulations were left out. Although NMF's higher data reconstruction errors can be observed here as well, POD recovered negative values of the OH mass fractions, which is physically impossible, while NMF did not. A physical interpretation of the data-driven basis functions found by POD and NMF can be accomplished by analyzing Figures 5 and 6, which report the eigenflames found by the two methods when compressing the available simulations to a 5-dimensional reduced manifold. These eigenflames were different. POD's first mode (Figure 5a) represents the main direction of variation. The second mode (Figure 5b) was representative of the temperature high gradients due to the cold inlet fuel jet entering the flame zone. The third mode, reported in Figure 5c, was representative of the reactive zone, where the maximum values of temperature were encountered. NMF's first mode (Figure 6a) was representative of the temperature field for higher values of inlet velocities, while NMF's second mode was for lower values, with the following modes representing different features to be added. In both cases, the different flame positions, which depended more on the inlet temperature than on the fuel composition input parameter, played a major role on the shapes of the eigenflames, whereas the maximum values of the different spatial fields, which depended on the fuel composition mainly, were accounted for by the eigenflame coefficients or scores.

In conclusion, despite NMF's additive nature, POD's lower reconstruction errors indicated that

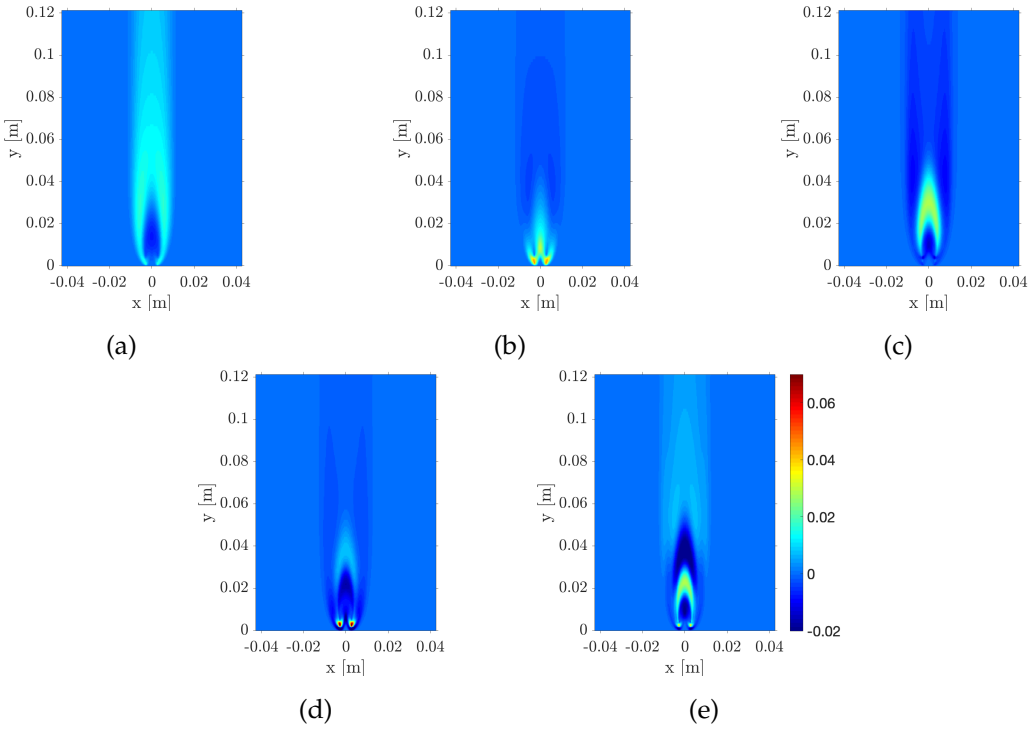


Figure 5: Data-driven orthogonal basis functions for the temperature field found by POD with a number of 5 retained modes.

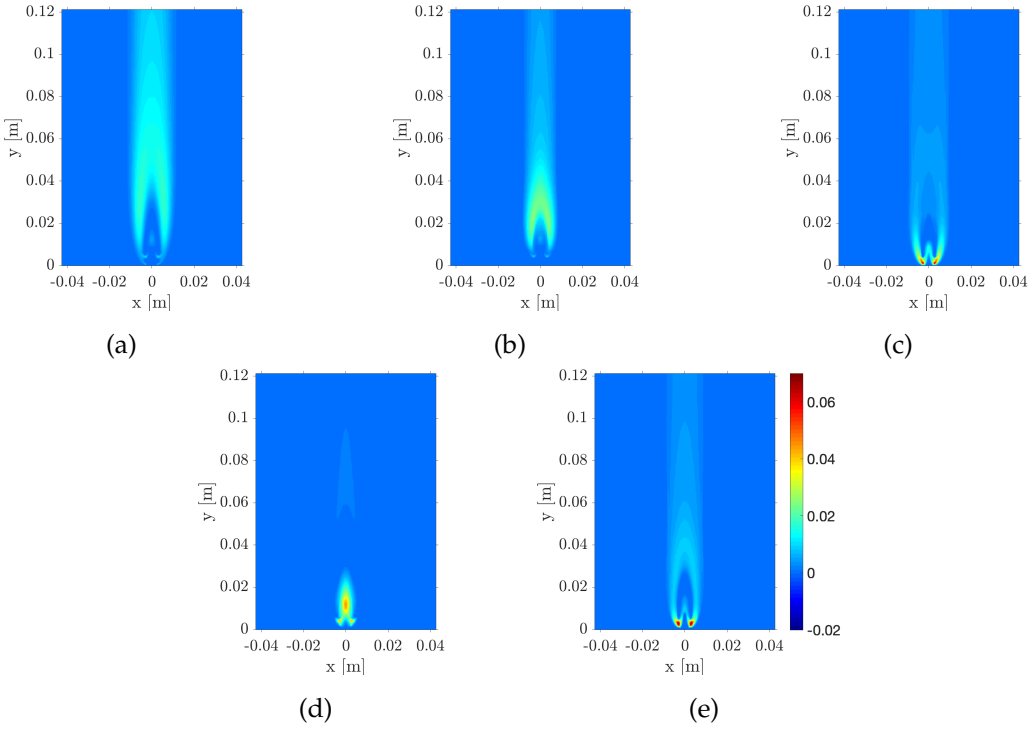
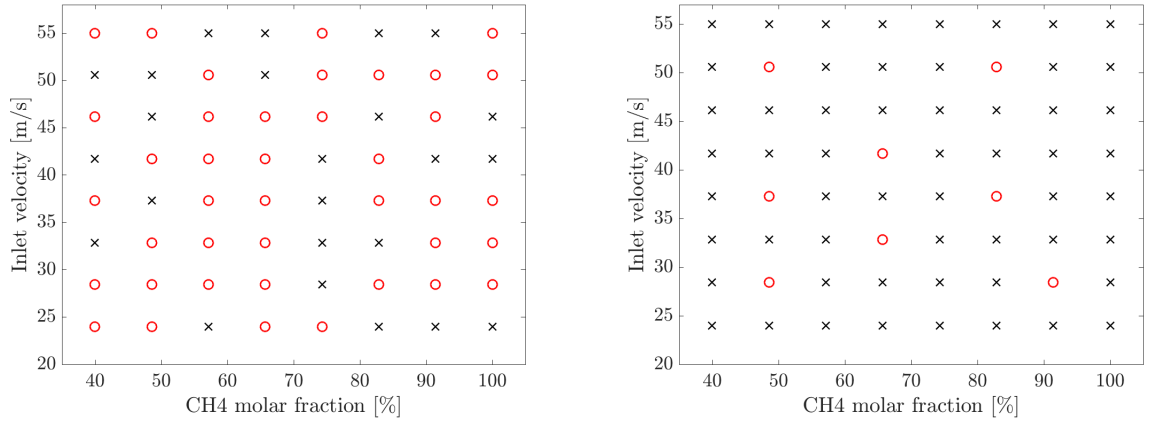


Figure 6: Data-driven non-negative basis functions for the temperature field found by NMF with a number of 5 retained modes. Same color map as of Figure 5.

this method was a better choice over NMF for the compression of the data of the present study.

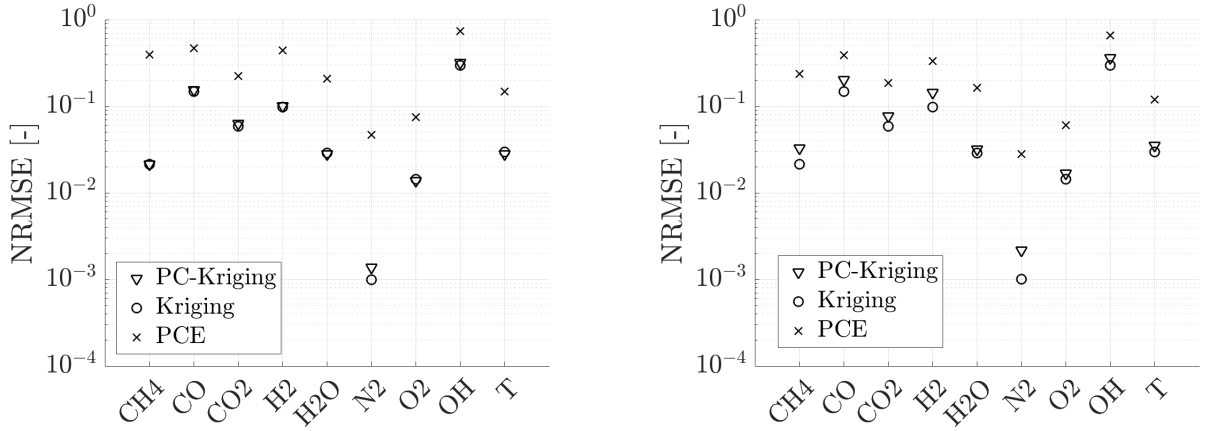
4. Application of the POD-based ROM

The performance of the ROM developed by combining POD with PCE, Ordinary Kriging (OK) and PC-Kriging were investigated in the present work. Specifically, the effects of changing the values of the PCE parameters such as the total polynomial order N and degree of polynomial interaction q was investigated, as well as the effects of the training data-set's size. The leave-k-out strategy was employed for the analysis, but also two different reference DoEs (randomly chosen) of different sizes were employed when a leave-k-out analysis was computationally prohibitive. These DoEs are reported in Figure 7.



(a) DoE-A. (b) DoE-B.

Figure 7: Random design of experiment for the surrogate model. Black cross: training points; red circle: test points.



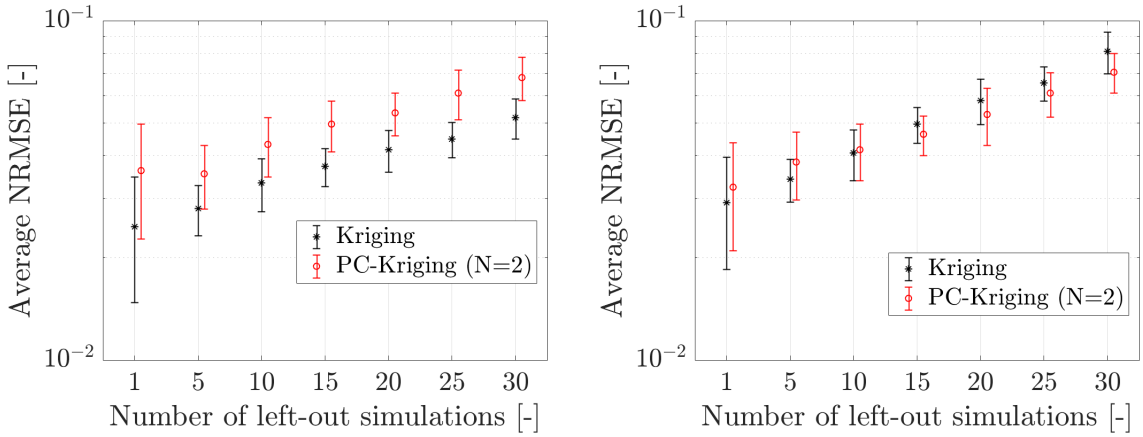
(a) $q = 0.5$ (b) $q = 1$

Figure 8: NRMSE for the prediction of the test data of DoE-A with Matern52 as kernel for the Kriging, with $N = 2$ and with (a) $q = 0.5$ and (b) $q = 1$ for the Polynomial Chaos Expansion.

Figure 8 reports the NRMSE for the prediction of the considered system's variables for DoE-A, for two different values of the parameter q, for $N = 2$. In both Figure 8a and 8b, the NRMSE for the Kriging model are the same as q is a parameter of PCE. It can be observed that decreasing the value of q led to an improvement for the PC-Kriging model, indicating that the interaction between

polynomials of high degrees can be detrimental and meaning that a trade-off in terms of polynomial complexity (optimal values for N and q) should be found when using PC-Kriging.

A leave-k-out analysis was carried out in order to determine the influence of the number of training simulations on ROM's performances. This analysis was repeated for different values of N, so to determine the sensitivity of the model to this parameter as well. k was the number of simulations left out from the training data set employed to find the POD reduced basis and train the predictive models. Each time, k simulations were left out and the error to predict the left-out simulations was estimated.



(a) All modes.

(b) 30 modes.

(c) All modes.

(d) 30 modes.

Figure 9: Average NRMSE among all the considered variables for the prediction of the test data for an increasing number of training simulations by means of POD + Kriging and POD + PC-Kriging with $N = 2$ and $N = 6$ and (a, c) all modes and (b, d) only 30 modes. $q = 0.5$. The height of the bars represents the standard deviation of the NRMSE.

Figure 9 reprots the average prediction errors by means of leave-k-out for the POD+Kriging and POD+PC-Kriging models with $N = 2$ and $N = 6$, for a different number of retained POD modes. In particular, all POD modes were retained for the results of Figure 9a and 9c, while only 30 POD modes were retained for Figure 9b and 9d. POD+Kriging's prediction errors were lower when more modes were retained, while the opposite happened for POD+PC-Kriging. Specifically, when less modes were retained, the POD+PC-Kriging ROM performed better than the POD+Kriging ROM, for

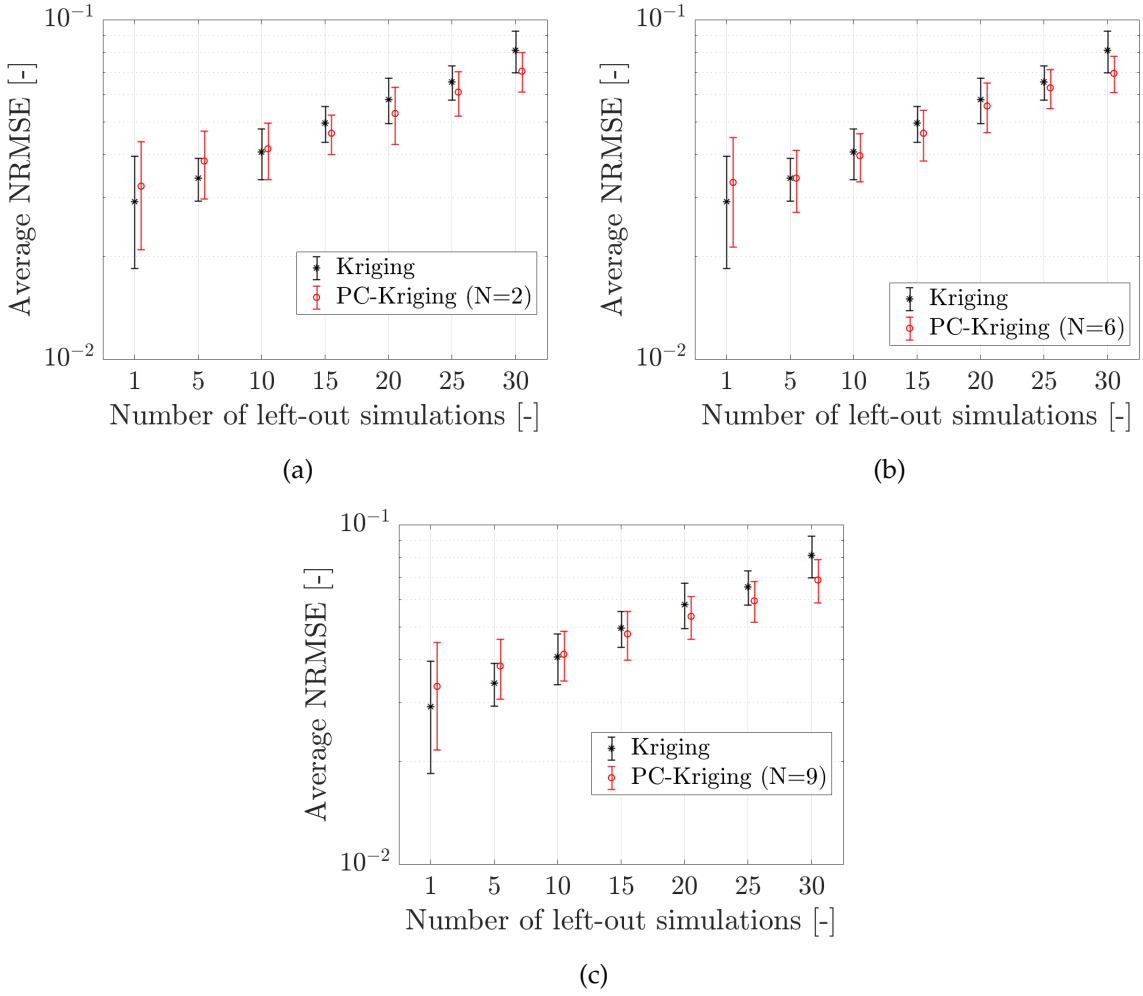


Figure 10: Average NRMSE among all the considered variables for the prediction of the test data for an increasing number of training simulations by means of POD + Kriging and POD + PC-Kriging with (a) $N = 2$, (b) $N = 6$ and (c) $N = 9$. $q = 0.5$ and 30 modes retained. The height of the bars represents the standard deviation of the NRMSE.

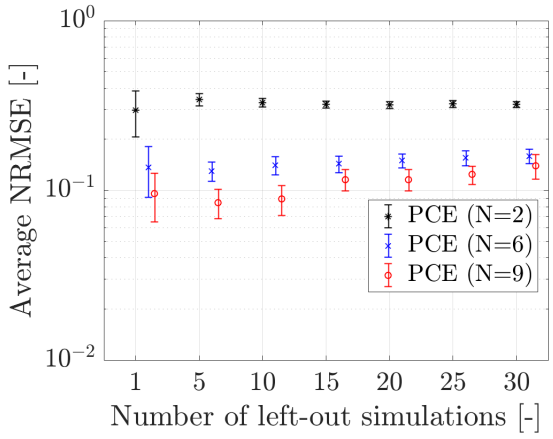
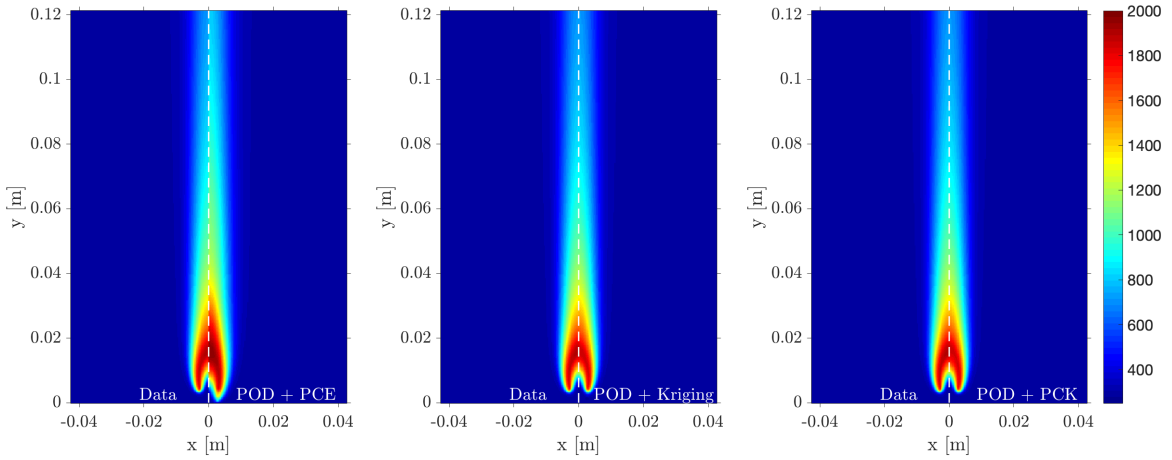


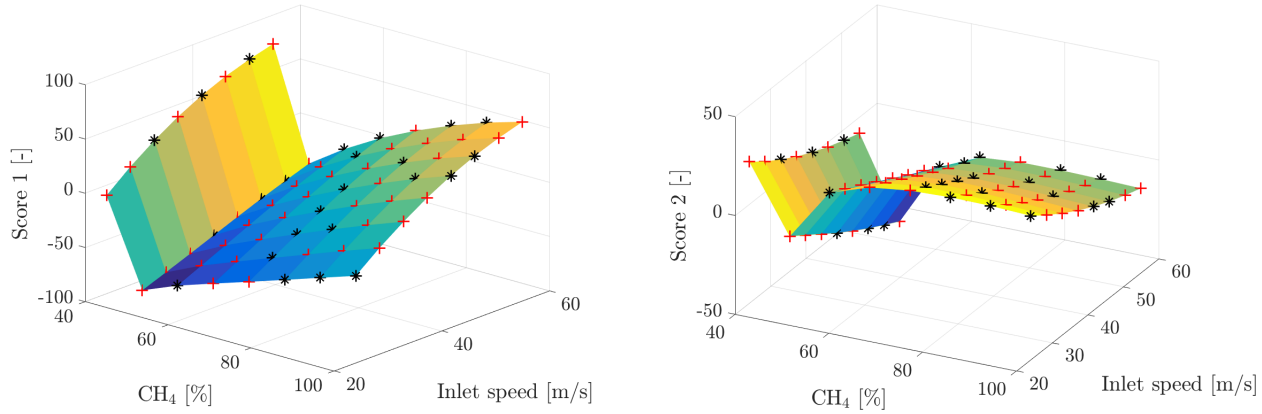
Figure 11: Average NRMSE among all the considered variables for the prediction of the test data for an increasing number of training simulations by means of POD + PCE with $N = 2$, $N = 6$ and $N = 9$. $q = 0.5$ and 30 modes retained. The height of the bars represents the standard deviation of the NRMSE.

smaller training sizes, indicating the PC-Kriging has the potential to be a better choice than Kriging for coarse experimental designs and low approximation orders. The estimated average prediction errors of the POD+Kriging and POD+PC-Kriging ($N = 2$, $N = 6$ and $N = 9$) models by means of leave-k-out analysis are reported in Figure 10, from which it can be observed that POD+PC-Kriging outperformed POD+Kriging for $k > 10$ when $N = 2$, and for $k > 5$ for $N = 6$. The POD+PC-Kriging model's performance declined for $N = 9$, indicating a trade-off for N has to be found. Figure 11 shows that the NRMSE for POD+PCE with $N = 2$, $N = 6$ and $N = 9$ were approximately 10 times higher with respect to the other models, further confirming that the combination of the two models led to improved performances.



(a) (b) (c)

Figure 12: Left halves: true data. Right halves: (a) prediction of the temperature field by POD + PCE; (b) prediction of the temperature field by POD + Kriging; (c) prediction of the temperature field by POD + PC-Kriging. POD: 20 PCs. Polynomial Chaos: $N = 2$, $q = 0.5$, Legendre. Kriging: Matern52 kernel.

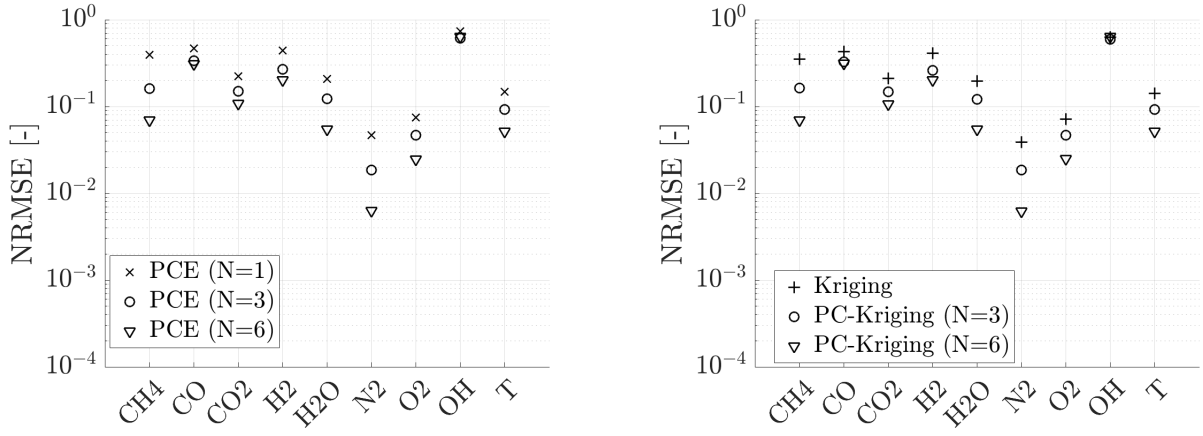


(a) (b)

Figure 13: Response surfaces of the POD score (a) 1 and (b) 2 for the training (black stars) and test (red crosses) data of DoE-A.

Figure 12 reports the contours of the true and predicted temperature fields by the 3 ROMs when trained on DoE-A, for one specific combination of the input parameters. Figure 13 reports the response surfaces to be found for the first 2 POD scores.

4.1. Fixing Kriging length-scales



(a) (b)

Figure 14: NRMSE for the prediction of the test data of DoE-A by the (a) PCE and (b) PC-Kriging model with different values for the polynomial order N for a small value of the Kriging length-scale $l = 10^{-5}$.

A POD+PC-Kriging model with small values for the Kriging length-scales was built ($l = 10^{-5}$) so to investigate on the effects of this parameter. This way, the Kriging model could influence the predictions only in the very proximity of the training data. The NRMSE for the POD+PCE and POD+PC-Kriging models for different values of N and a fixed length-scale $l = 10^{-5}$ are reported in Figure 14, from which we can observe that: for both POD+PCE and POD+PC-Kriging, lower prediction errors were obtained when N was increased; the performances of POD+PCE and POD+PC-Kriging for the prediction of the test data coincided, indeed indicating that the predictive capabilities of the PC-Kriging model came from the PCE part when small values for l were set. However, the POD+PC-Kriging model still offered an advantage over the plain POD+PCE model as, thanks to Kriging, it interpolated the training data, differently from PCE which is a regression method. The combination of PCE with Kriging thus offers the possibility of training only the PCE part of the model (estimating the polynomial coefficients) which is computationally cheaper than training a Kriging model, and of manually setting the values of the Kriging kernel length-scales to small values so that its predictions are forced by the Kriging part of the model to be closer to the training values in proximity of the training data and thus to interpolate them.

4.2. ANN as supervised method

The use of ANN as supervised technique in combination with POD for the prediction of two random DoEs of different sizes, DoE-A and DoE-B, was also investigated in the present work. These DoEs are reported in Figure 7. In this case, a leave-k-out analysis was not performed because of its computational cost. As explained in Section 2.3.4, ANN offers the possibility of training one model, namely one neural network, for the prediction of all the POD scores, simultaneously. In the present

work, the architecture for the ANN was chosen with one input layer of dimensions $d = 2$ (number of input parameters) and 4 hidden layers of dimensions 8, 32, 128, 256, respectively. The dimension of the output layer was 20 for DoE-A and for DoE-B, given that a total of 20 modes was retained. The activation functions for all the hidden layers were Leaky ReLU with different slopes, specifically the first 2 hidden layers had a slope of 10^{-4} and the last two had a slope of 10^{-3}. Figure 15 reports the

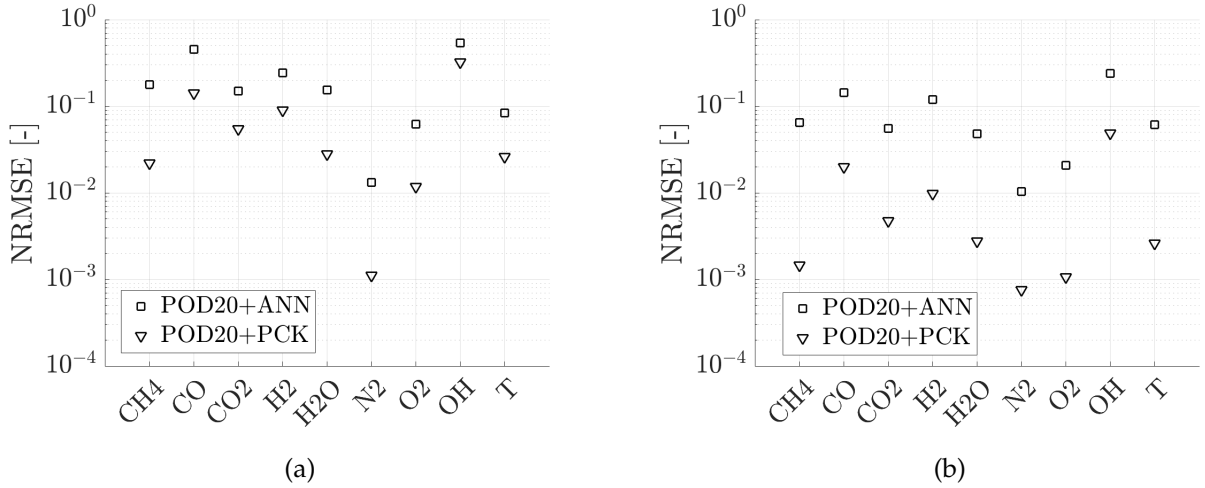


Figure 15: NRMSE for the prediction of (a) DoE-A and (b) DoE-B by a ROM which combines POD with 20 modes and ANN. The NRMSE by POD+PC-Kriging are also reported for comparison. Kriging: Matern52. Polynomial Chaos: Legendre, $N = 2$, $q = 1.0$.

NRMSE for the prediction of the DoE-A and DoE-B by POD+ANN and by POD+PC-Kriging. Even though only one model for the prediction of the POD scores was necessary to train when using ANN in combination with POD, the resulting prediction errors were higher with respect to the POD+PC-Kriging ROM's. The prediction errors of POD+ANN decreased by $< 50\%$ when increasing the size of the training data-set, whereas the prediction errors of POD+PC-Kriging decreased by a factor of ≈ 10. The training of an ANN consists in a wide range of design choices for its architecture, such as number of hidden layers or type of activation functions, to be cross-validated. Besides, ANNs are usually employed for cases with a high number of observations available, which is not the case for computationally expensive combustion simulations. This makes ANN a more complex choice as supervised method for a ROM, which might not be preferable for the present case, where PC-Kriging was a much simpler method to set up.

In conclusion, the results presented in this Section showed that the POD+Kriging or POD+PC-Kriging ROM performed with lower prediction errors by a factor of ≈ 10 in comparison to POD+PCE. The use of PCE as Kriging trend function could improve the Kriging model's performance for smaller training sizes. In general, the POD+Kriging model performed satisfactorily. In fact, the fields of the mass fractions of CH_4, CO_2, H_2O, N_2, O_2 and temperature were predicted with NRMSE $< 1\%$ for DoE-B. The combination of POD with ANN did not lead to satisfactory results as the approach achieved higher prediction errors by a factor of ≈ 10 with respect to POD+PC-Kriging.

5. Application of the NMF-based ROM

The performance of the ROM developed by combining NMF and PC-Kriging was investigated in the present work.

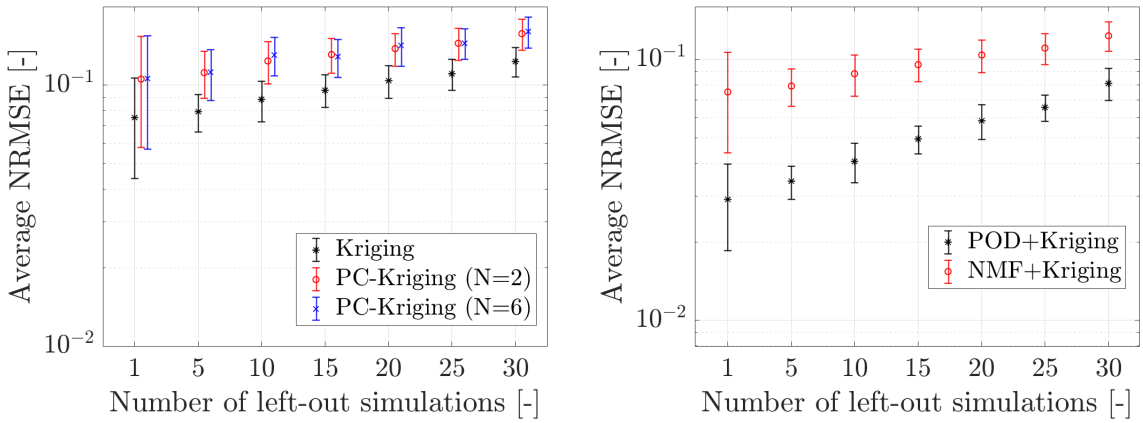
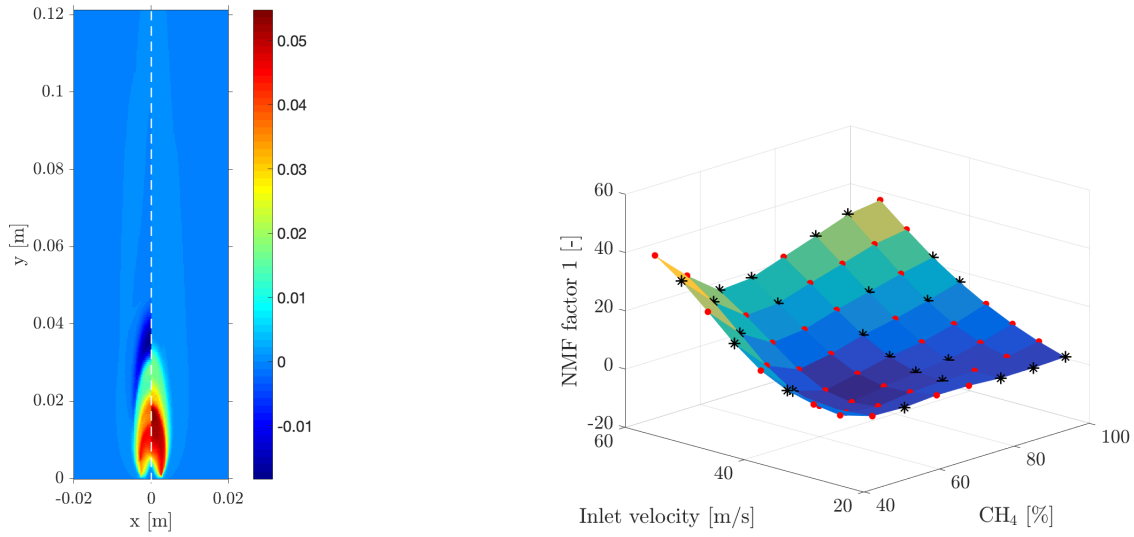


Figure 16: (a) Average NRMSE among all the considered variables for the prediction of the test data for an increasing number of training simulations by means of NMF+Kriging and NMF+PC-Kriging. The height of the bars represents the standard deviation of the NRMSE. (b) Comparison between POD+Kriging and NMF+Kriging. The POD-based ROM had lower NRMSE for all values for the number of left-out simulations k. 30 modes.

A leave-k-out analysis was carried out in order to determine the influence of the number of training simulations on performances of a ROM based on NMF as well. k was the number of simulations left out from the training data set employed to find the NMF reduced basis and train the predictive models. Each time, k simulations were left out and the error to reconstruct the left-out simulations from the POD basis was estimated. The estimated average prediction errors of the NMF+Kriging and NMF+PC-Kriging models by means of leave-k-out analysis are reported in Figure 16a. Figure 16b compares the performance of the NMF+Kriging model with the POD+Kriging one. The overall higher NRMSE in comparison to a POD-based ROM suggested that POD proved to be a better choice unsupervised method for the proposed approach.

Figure 17a reports an example of predicted CO mass fraction field by (left) the NMF-based ROM and (right) still the NMF-based ROM after setting the predicted nefative-valued scores to zero. As shown in Figure 17a, because no positivity constraint is enforced on the predicted NMF scores, the predicted CO field by the NMF-based ROM had negative values just like the one predicted by the POD-based ROM, which nullified the advantage that NMF should have provided over POD. A similar issue arose from the combination of Kriging with Constrained POD [2], as shown in [1]. In fact, as shown in Figure 17b, which reports the response surface for one NMF score, PC-Kriging predicted negative values for this quantity. However, differently from a POD-based ROM where no correction on the predicted POD scores can be made to avoid predicting negative mass fractions, such a correction is possible for NMF. In fact, manually correcting the predicted negative values for all the NMF scores to be equal to 0 led to a less accurate but physically acceptable prediction, as shown in Figure 17a, which reports the predicted CO mass fraction by the NMF-based ROM when such a correction on the NMF scores was done.



(a) (b)

Figure 17: (a) Predicted CO mass fraction by (left) the NMF-based ROM with no correction and (right) the NMF-based ROM with correction, i.e. after setting the predicted negative-valued NMF scores to zero. (b) Response surface found by PC-Kriging in a NMF-based ROM with 20 modes on DoE-A for one NMF score. PC-Kriging predicted negative values for this score as no positivity constraint is enforced on predictions. Kriging: Matern52. Polynomial Chaos: Legendre, $q = 0.5$, $N = 2$.

In conclusion, the NMF-based ROMs had higher prediction errors in comparison to the POD-based ROMs. The variables that were predicted with a NRMSE $< 10\%$ for DoE-A were mass fractions of CH_4, CO_2, H_2O, N_2, O_2 and temperature. Although the predicted NMF scores were at times negative in value, leading to predicted negative mass fractions, NMF offered the possibility to correct this by setting the predicted negative values to zero.

6. Conclusions

In the present work, inspired by [1] and [4], two data compression techniques such as POD and NMF were combined with a predictive model based on the combination of Polynomial Chaos Expansion (PCE) and Kriging for the development of a Reduced-Order Model (ROM) for the prediction of combustion data, with PCE functioning as Kriging trend. In order to compare the performance of the PC-Kriging interpolation, POD and NMF were also combined with PCE only and Kriging only.

As regards the unsupervised techniques used for data compression for the development of the ROM, the results showed that POD could reconstruct the training data with an NRMSE which was ≈ 10 times lower with respect to NMF. On the other hand, the positivity constraint imposed in the NMF formulation guaranteed that positive physical quantities were not reconstructed with negative values. Both methods generalized to unseen data with similar performances in terms of errors for data reconstruction, with POD having reconstruction errors $\approx 10\%$ lower.

As concerns the supervised part of the ROM, results from a leave-k-out analysis showed that the PC-Kriging interpolation performed with lower prediction errors than Kriging for smaller training sizes and for low approximation orders. Both PC-Kriging and Kriging outperformed PCE, when these techniques were combined with POD. PC-Kriging performed better (prediction errors lower

by $\approx$ 10%) than Kriging also when in combination with NMF. In general, the POD-based ROM performed with lower prediction errors than the NMF-based ROM, by a factor of ≈ 10 on CH_4 and N_2. The PC-Kriging method performed better than Kriging also when the Kriging length-scales were not optimized and set to low values, so that the Kriging would influence predictions only in the regions close to the training data (and interpolate those data), while PCE would contribute in the regions far from the training data. By doing so, the overall prediction errors were higher than the case where the Kriging length-scales were optimized, although the training of such a model was computationally less demanding as it did not involve the optimization procedure for the evaluation of the length-scales and thus it only involved the evaluation of the PCE coefficients.

The use of ANN as supervised part for a POD-based ROM was also investigated as ANN offers the possibility of training one model for the prediction of all the POD (or NMF) scores, simultaneously. The predictive capabilities of combining POD with ANN were compared to the ones of POD in combination with PC-Kriging. Results showed that, again, POD combined with PC-Kriging had lower prediction errors with respect to the combination of POD and ANN, by a factor of ≈ 10 for nearly all variables involved, as ANN usually requires a high number of observations to perform well, which was not the case in the present work where data from computationally expensive combustion simulations were employed. The training of an ANN also meant that a wide range of design choices for its architecture, such as number of hidden layers or type of activation functions, needed to be explored or cross-validated, which made ANN a more complex choice as supervised method.

Acknowledgements

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.

References

[1] G. Aversano, A. Bellemans, Z. Li, A. Coussement, O. Gicquel, A. Parente, Application of reduced-order models based on PCA & Kriging for the development of digital twins of reacting flow applications, Computers and Chemical Engineering 121 (2019) 422–441. doi:10.1016/j.compchemeng.2018.09.022.

[2] M. Xiao, P. Breitkopf, R. Filomeno Coelho, C. Knopf-Lenoir, P. Villon, W. Zhang, Constrained Proper Orthogonal Decomposition based on QR-factorization for aerodynamical shape optimization, Applied Mathematics and Computation 223 (2013) 254–263. doi:10.1016/j.amc.2013.07.086.

[3] R. Schöbi, B. Sudret, J. Wiart, Polynomial-Chaos-based Kriging (2015) 1–33doi : 10 . 1615 / Int . J . UncertaintyQuantification.2015012467.

[4] R. Schöbi, P. Kersaudy, B. Sudret, J. Wiart, Combining Polynomial Chaos Expansions and Kriging <hal-01432550>, Tech. rep., Orange Labs research, ETH Zurich, Switzerland (2014).

[5] G. C. Cawley, N. L. C. Talbot, On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation, Journal of Machine Learning Research 11 (2010) 2079–2107.

[6] A. Parente, J. C. Sutherland, Principal component analysis of turbulent combustion data: Data pre-processing and manifold sensitivity, Combustion and Flame 160 (2) (2013) 340–350. doi:10.1016/j.combustflame.2012.09.016.

[7] N. Gillis, Sparse and Unique Nonnegative Matrix Factorization Through Data Preprocessing, Journal of Machine Learning Research 13 (2012) 3349–3386. arXiv:arXiv:1204.2436v1.

[8] C. E. Frouzakis, Y. G. Kevrekidis, J. Lee, K. Boulouchos, A. A. Alonso, Proper orthogonal decomposition of direct numerical simulation data: data reduction and observer construction 28 (2000) 75–81.

[9] P. Paatero, U. Tapper, Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values, Environmetrics 5 (2) (1994) 111–126. doi:10.1002/env.3170050203.

[10] M. T. Daza, J. J. Orteils, C. Noguera, Negative semantic priming from consciously vs. unconsciously perceived single words, Psicologica 28 (2) (2007) 105–127. doi:10.1016/j.csda.2006.11.006.

[11] R. Lebrun, A. Dutfoy, An innovating analysis of the Nataf transformation from the copula viewpoint, Probabilistic Engineering Mechanics 24 (3) (2009) 312–320. doi:10.1016/j.probengmech.2008.08.001.
URL http://dx.doi.org/10.1016/j.probengmech.2008.08.001

[12] P. G. Constantine, E. Dow, Q. Wang, Active Subspace Methods in Theory and Practice, SIAM Journal of Scientific Computation 36 (4) (2014) 1500–1524.

[13] S. N. Lophaven, J. Søndergaard, H. B. Nielsen, Kriging Toolbox (2002) 1–28.

[14] M. Seeger, Gaussian processes for machine learning., Vol. 14, 2004. arXiv : 026218253X, doi : 10 . 1142 / S0129065704001899.

[15] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning 2016, MIT Press.

[16] C. M. Bishop, Pattern recognition and machine learning, Springer, 2006.

[17] N. Ripley, Brian D and Hjort, Pattern recognition and neural networks, Cambridge university press, 1996.

[18] S. Cao, B. Bennett, B. Ma, D. Giassi, Effects of Fuel Dilution and Gravity on Laminar Coflow Methane-Air Diffusion Flames: A Computational and Experimental Investigation, 8th US National Combustion Meeting Organized by the Western States Section of the Combustion Institute (2013) 1–9.

[19] S. Cao, B. Ma, B. A. Bennett, D. Giassi, D. P. Stocker, F. Takahashi, M. B. Long, M. D. Smooke, A computational and experimental study of coflow laminar methane/air diffusion flames: Effects of fuel dilution, inlet velocity, and gravity, Proceedings of the Combustion Institute 35 (1) (2015) 897–903. doi:10.1016/j.proci.2014.05.138.

5.4 DIGITAL TWIN FOR MILD COMBUSTION FURNACE

Digital twin for MILD combustion furnace

Gianmarco Aversano[a,b], Marco Ferrarotti[a,b], Alessandro Parente[a,b]

[a] *Université Libre de Bruxelles, Aero-Thermo-Mechanics Departement, Avenue F.D. Roosevelt 51, CP 165/41, 1050 Brussels, Belgium*

[b] *Université Libre de Bruxelles and Vrije Universiteit Brussel, Combustion and Robust Optimization Group (BURN), Brussels, Belgium*

Abstract

The combination of a data compression method such as Proper Orthogonal Decomposition (POD) with an interpolation technique such as Kriging was shown to be a reliable choice for the development of Reduced-Order Models (ROMs) for the prediction of combustion data at unexplored operating conditions. In this study, the authors report on the use of this approach to develop a digital twin for a MILD combustion furnace by training a ROM on 3D simulation data, with the objective to predict the spatial fields of temperature and chemical species mass fractions and some important scalar quantities such as the temperature values at the wall, exhaust gas composition and flame length. An analysis on the sensitivity of the ROM to size of the training data-set used for its development is also carried out. Results showed that, the developed ROM could reliably predict the fields of temperature and of the main chemical species mass fractions, such as CO_2, O_2, H_2O and CH_4, and of the mentioned scalar quantities, with an overall accuracy of less than 10%, indicating that the model can be used as digital-twin of the furnace for real-time predictions when its operating conditions change, indicated by sensory data.

Keywords: Proper orthogonal decomposition, Surrogate modeling, Kriging, Digital twins

1. Introduction

One of the objective of the modern society is to ensure a healthy environment for future generations, which is considered possible only if more and more affordable and sustainable energy is employed. However, the intermittent nature of renewable sources requires the development of novel storage solutions that can guarantee the availability of the required energy supply when renewable sources are not available. Because many applications such as air and ground transportation require high energy density and as a consequence cannot rely on the direct use of renewable energy, it is energy storage in the form of chemical compounds such as hydrogen that will lead to a true integration between renewable sources and existing infrastructure for energy conversion, such as combustion systems. Energy density is the key feature that makes the use of fuels inescapable for energy demanding applications, such as transportation and other industrial processes. This means that the transformation of excess renewable energy into energy carriers is an appealing solution for energy storage. Fuel flexibility also poses technical challenges and indicates the need for advanced combustion technologies. Such technologies have to be fuel flexible, highly efficient and non-polluting, operating in conditions which substantially differ from those characterizing traditional combustion

systems. Moderate and Intense Low-oxygen Dilution (MILD) [1] combustion represents a very attractive solution for its fuel flexibility and capability to deliver very high combustion efficiency with virtually zero pollutant emissions. MILD combustion ensures large fuel flexibility, representing an ideal technology for low-calorific value fuels, high-calorific industrial wastes and hydrogen-based fuels, as well as for liquid and solid fuels. These new technologies result from a change in perspective in the analysis of combustion problems, from purely energetic to chemical aspects. The combustion process becomes a chemical reactor, which must be optimized from the perspective of flexibility, conversion efficiency and emissions. This requires a deep understanding of the kinetic aspects related to combustion reactions, interactions between chemical kinetics and turbulence, and heat exchange phenomena, in particular radiation, meaning that both high-fidelity simulations and experimental techniques have to be used in a unified framework, to optimize the operation of existing systems and develop new combustion systems. Besides, the experiments and simulations needed for the production of informative data with the objective of acquiring usable knowledge for the development of new combustion technologies come with associated costs, which limit their production. Thus, the need for automated algorithms that can help data interpretation and speed up the process of data production becomes clear. The development of virtual models, also referred to as digital twins, of industrial systems opens up a number of opportunities, such as the use of data to anticipate the response of a system and brainstorm malfunctioning, and the use of simulations to develop new technologies, i.e. virtual prototyping. A definition of digital twin is "An integrated multi-physics, multi-scale, probabilistic simulation of an as-built system, enabled by digital thread, that uses the best available models, sensor information, and input data to mirror and predict activities/performance over the life of its corresponding physical twin" [2].

With the increase of available data coming from either experiments or simulations, the popularity of machine learning techniques is growing as these allow to fully exploit the information contained in very large data-sets [3]. In fact, the characterization of complex multi-physics problems, such as combustion, require the use of high-fidelity simulations requiring significant computational resources and produce terabytes of data that can hardly be handled by human supervision. The direct use of these simulations to predict the state of industrial components in real-time is thus not a possibility. This is the case with computational fluid dynamics (CFD) and reacting flows more in general, for which the simulation of a specific configuration may require millions of CPU hours [1]. Combining CFD simulations with real-time data coming from sensors of a real industrial system in order to foresee a change in its state is possible only if the prediction of the system's state based on the operating conditions reported by these sensors becomes instantaneous [4]. In order to do so, it is necessary to produce a set of training simulations beforehand, for a wide enough range of possible operating conditions, which is subsequently used to develop a reduced-order model (ROM) that can predict the state of the physical system in real-time.

A physics-based reduced-order model can be developed by using unsupervised learning to compress the original data size and thus extract latent features in the data, for which a response surface, for a wide range of operating conditions, is subsequently found by a supervised learning technique. Once this is accomplished, the values of these features can be predicted for the operating conditions indicated by the sensors and, because the mapping from the feature-space to the original data-space is learned by the compression method, also the full state of the real system is predicted, e.g. the

spatial fields of temperature and chemical species mass fractions.

In a previous study [5], the authors showed that the combination of an unsupervised method for data compression such as Principal Component Analysis (PCA) or Proper Orthogonal Decomposition (POD), with a supervised technique such as Kriging for the development of Reduced-Order surrogate Models (ROMs) was a reliable choice for the prediction of combustion data, namely spatial profiles of temperature and chemical species mass fractions, for unexplored operating conditions, e.g. inlet velocity and fuel composition of a 2D laminar methane flame. POD could compress the size of the high-dimensional data whose prediction was of interest, leading to a high reduction in the number of surrogate models (SMs) to be trained, which is desirable in case the predictive model needs to be continuously retrained whenever new data-sets become available.

In the present work, the objectives were twofold: to understand the factors that can affect the final performance of a digital twin and to build a digital twin [2] for a MILD combustion furnace from a set of 3D CFD simulations for the prediction of the full state of the furnace (spatial fields of temperature and main chemical species mass fractions) and important scalar quantities such as wall temperature, peak of OH mass fraction, exhaust gas composition and flame length within an accuracy of 10%, for unexplored operating conditions, namely a design parameter such as the air injector length, and two parameters whose values can be measured by sensors such as mass fraction of the inlet H_2 (fuel) and equivalence ratio. A set of simulations for different values of the mentioned parameters was produced using the commercial software Ansys Fluent. The choice of the sampling strategy described in [6] to identify the set of training simulations was investigated, as this method is computationally affordable and able to associate an importance to each available simulation based on the influence they have on the reduced basis found by POD. This method was also used in [5]. Cross validation techniques, such as the estimation of the leave-k-out errors for the prediction of the quantities of interest, were also employed to understand how the developed ROM would generalize to new data [7]. Thus, the effects of the size of the training data on the performance of the developed ROM to reconstruct and predict test data was also assessed. In the present work, for a specific case, the simulations not used for the training of the reduced-order model and that were used for its validation are referred to as test data. Thus, some guidelines for the development of a physics-based ROM of a combustion system are offered in the present work. In the end, a digital twin of the described physical system was developed.

The paper is organized as follows: the methodology employed for the development of the digital twin is explained briefly in Section 2. Section 3.1 discusses the different possibilities for data-preprocessing. In Section 3.2, the chosen method for data compression is explained in more details. Section 3.3 describes the supervised technique used in the present work. Section 4 introduces the combustion system for which the ROM is developed and results from the data compression procedure. In Section 5, an analysis on the ROM's predictive capabilities and its dependence on the size of the training data-set is carried out. Seciton 6 presents the performance of one developed ROM. Conclusions are drawn in Section 7.

2. Methodology

In the present work, a data matrix containing m output of a certain computer model or CFD simulation, indicated by $\mathcal{F}$, will be indicated by a matrix $\mathbf{Y}$ of size $(m \times n)$, where n is the size

of the output of the considered model. The m combinations of values for the d input parameters to the simulations $\mathbf{Y}$ are collected in the matrix $\mathbf{X}$ of size $(m \times d)$. One particular simulation or output of the considered computer code for one combination of the input parameters $\mathbf{x}$ is indicated by $\mathbf{y}(\mathbf{x}) = \mathcal{F}(\mathbf{x})$.

A low-rank approximation of $\mathbf{Y}$ is sought, so that each particular simulation output $\mathbf{y}(\mathbf{x})$ can be expressed as a weighted sum of basis functions, as shown in Figure 1. In the context of combustion data, the basis functions are also called *eigenflames* [5]. By doing so, it is subsequently possible to train

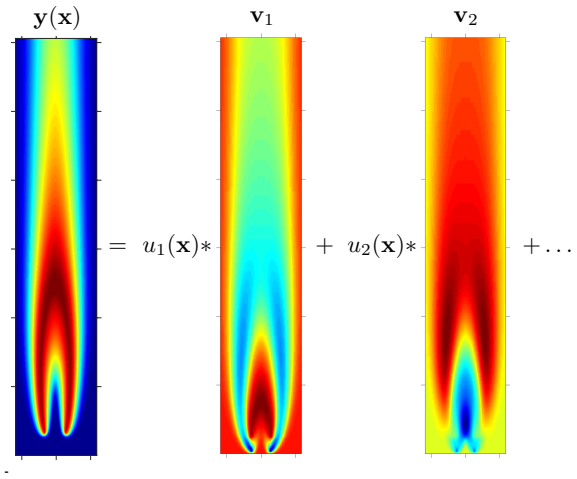


Figure 1: Illustrative example of low-rank approximation of a spatial field $\mathbf{y}$ observed for a certain operating condition $\mathbf{x}$, represented by a set of coordinates (the coefficients u_i) on some basis functions (eigenflames) $\mathbf{v}_i$ [5]: $\mathbf{y}(\mathbf{x}) = \sum_{i=1}^{q} \mathbf{v}_i u_i(\mathbf{x})$, where q is the total number of used basis functions.

predictive models for unexplored combinations of the input parameters only for the coefficients of the basis functions, as done in [5], here also referred to as *features*. In the following Section, it is implied that the data matrix is appropriately pre-processed (centered and scaled) before compressing the data [8].

3. Methods

3.1. Data pre-processing

Data are usually centered and scaled before a data compression method is carried out. Centering represents all observations as fluctuations from a chosen center, usually the mean value, leaving only the relevant variation for analysis. Scaling is a crucial operation when dealing with multivariate data-sets. In fact, in the case of combustion-related data-sets, temperature and chemical species mass fractions have different units and vary over different scales. The choice of the appropriate centering and scaling criteria also depends on the subsequent compression method that is employed. Six possible choices for the scaling of the data are here reported. AUTO scaling: each variable is normalized by its standard deviation; RANGE scaling: each variable is normalized by its range; PARETO scaling: each variable is scaled by the square root of its standard deviation; VAST: scaling each variable is scaled by the standard deviation and coefficient of variation; LEVEL scaling: each variable is normalized by the mean of the data; MAX scaling: each variable is scaled by its maximum

value. The choice of scaling usually affects the subsequent data compression process, as shown in [8] where Principal Component Analysis is applied.

3.2. Proper Orthogonal Decomposition

The POD problem can be stated as follows: given a matrix $\mathbf{Y}$ of size $(m \times n)$, find $\mathbf{Z}$ of size $(m \times q)$ and $\mathbf{A}$ of size $(n \times k)$ with $k < n$ such that the functional $f(\mathbf{Z}, \mathbf{A}) = \frac{1}{2}||\mathbf{Y} - \mathbf{Z}\mathbf{A}^T||^2$ is minimized, subject to $\mathbf{A}^T\mathbf{A} = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix. This problem can be solved by computing the singular value decomposition (SVD) of the matrix $\mathbf{Y}$, which corresponds to finding the eigenvectors and eigenvalues of the matrix $\mathbf{C} = \frac{1}{m-1}\mathbf{Y}^T\mathbf{Y}$ [9]. The eigenvectors of $\mathbf{C}$ are the POD modes and the associated eigenvalues represent their relevance for the low-rank approximation of $\mathbf{Y}$. The POD modes are thus found all at once, and by ordering them in descending order according to their corresponding eigenvalue and retaining only a subset $k < \min(n, m)$ of them, a low-rank approximation of $\mathbf{Y}$ is possible as follows $\mathbf{Y} \approx \mathbf{Z}\mathbf{A}^T = \mathbf{Y}\mathbf{A}\mathbf{A}^T$, where the columns of $\mathbf{A}$ of size $(n \times q)$ are the POD modes and $\mathbf{Z}$ of size $(m \times k)$ is the matrix of POD coefficients. Each column of $\mathbf{Z}$ are the k coefficients for the retained k POD modes so that one particular simulation, or row of $\mathbf{Y}$, can be expressed as $\mathbf{y}(\mathbf{x}) = \sum_{i=1}^{k} \mathbf{a}_i z_i(\mathbf{x})$. As the solution of the POD problem only leads to the evaluation of $\mathbf{A}$, the matrix $\mathbf{Z}$ can be computed only after $\mathbf{A}$ is known, as follows: $\mathbf{Z} = \mathbf{Y}\mathbf{A}$. This holds for new data $\mathbf{Y}'$ as well: $\mathbf{Z}' = \mathbf{Y}'\mathbf{A}$.

3.3. Kriging

Kriging is an interpolation method in which every realization $y(\mathbf{x})$ is expressed as a combination of a trend function and a residual [10]:

$$y(\mathbf{x}) = \mu(\mathbf{x}) + s(\mathbf{x}) = \sum_{i=0}^{p} \beta_i f_i(\mathbf{x}) + s(\mathbf{x}) = \mathbf{f}^T(\mathbf{x})\boldsymbol{\beta} + s(\mathbf{x}) \tag{1}$$

The trend function $\mu(\mathbf{x})$ is a low-order polynomial regression and provides a global model in the input space. The term $s(\mathbf{x})$ creates a localized deviation weighting the points in the training set that are closer to the target point $\mathbf{x}$. The trend function $\mu(\mathbf{x})$ is expressed as a weighted linear combination of $p + 1$ polynomials $\mathbf{f}(\mathbf{x}) = [f_0(\mathbf{x}),\ ...,\ f_p(\mathbf{x})]^T$ with the weights $\boldsymbol{\beta} = [\beta_0,\ ...,\ \beta_p]^T$ determined by generalized least squares (GLS). The subscript p also indicates the degree of the polynomial. The residuals $s(\mathbf{x})$ are modeled by a Gaussian process with a kernel or correlation function that depends on a set of hyper-parameters to be evaluated by Maximum Likelihood Estimation (MLE) [11, 12].

In the definition of both the trend function and the residual, it is up to the designer to choose the polynomials $\mathbf{f}(\mathbf{x})$ and the correlation model or kernel. In this way, prior knowledge can be added into the problem.

4. Data-set description and data compression

The configuration of the simulated flame is described in [13]. A 45 degrees angular sector of the 3D geometry of the furnace is considered, as a result of the symmetry of the problem. The computational grid was first created with tetrahedrons and then converted into polyhedrons, for a final number of 180 000 cells. The PaSR is used to solve the chemistry/turbulence interactions

coupled with the KEE (17 species and 58 reactions) kinetic scheme. Three input parameters are considered: fuel composition in mole fractions (mixture of methane/hydrogen), equivalence ratio and air injection geometry. A design of experiments (DoE) of 45 simulations was produced, varying the input parameters in the range 0-100 % (H_2 molar fraction), 0.7-1 (equivalence ratio) and 16-20-25 mm (air injector length). The approach investigated in the present work will be validate w.r.t. the prediction of the spatial fields of CH_4, CO, CO_2, H_2, H_2O, O_2, OH and temperature and of scalar quantities such as the flame-length and outlet composition.

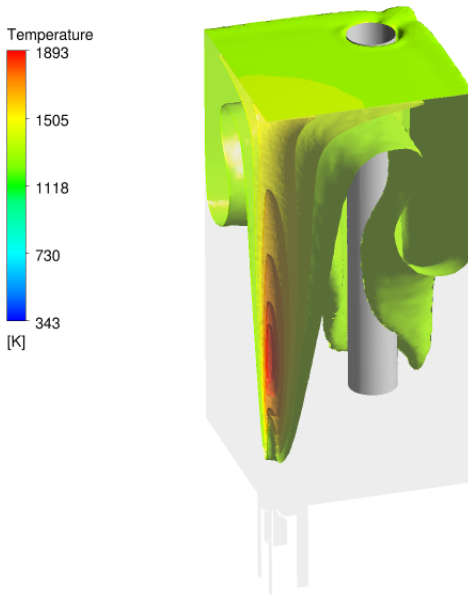


Figure 2: Temperature field for the simulation run with the following input parameters: air injector length of 25 mm, H_2 ratio of 0.70 and equivalence ratio equal to 0.94.

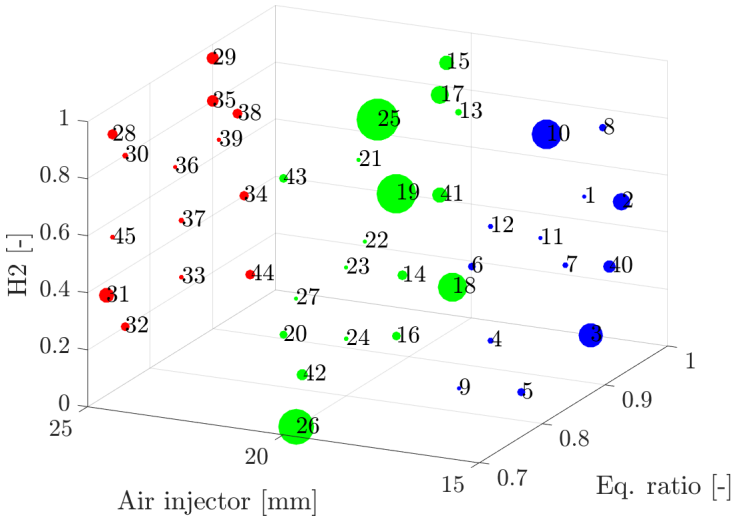


Figure 3: Input parameter values of the 45 simulations. The size of the circle represents the influence of that simulation on the POD reduced basis. Color represents the 3 different values of the air injector length.

Figure 3 reports the values of the input parameters which the 45 simulations were run for. The

relative influence of each simulation on the POD basis was evaluated as described in [6], and it is represented in Figure 3 by the size of the circles, while the color is indicative of the value of the air injector length for that simulation. Each simulation is numbered for convenience. From the values of the relative influence on the POD basis, the simulations that mostly contribute to the reduced basis can be identified and kept as training data.

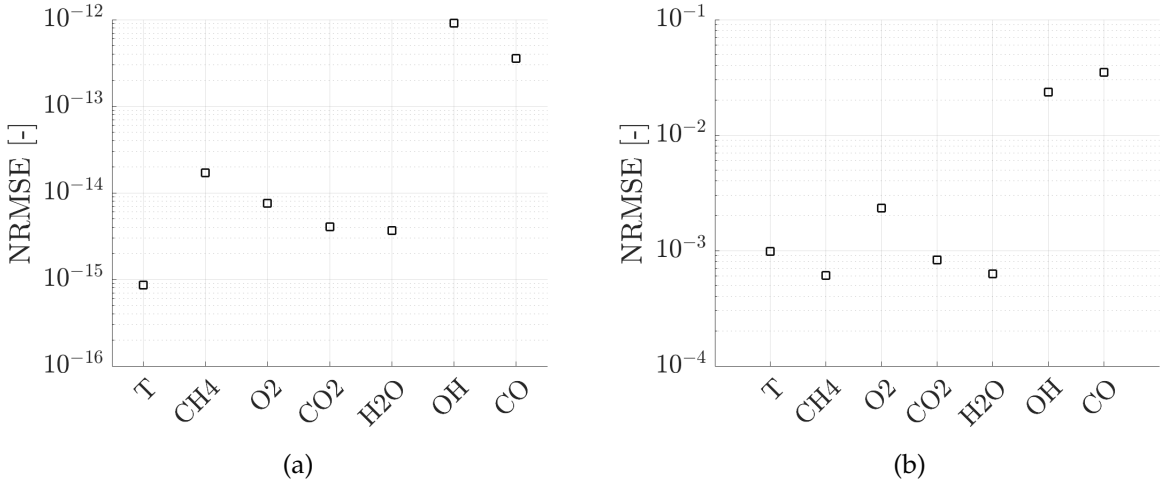


Figure 4: NRMSE for the reconstruction of the (a) training and (b) test data.

The ability of a reduced basis to compress and reconstruct unseen data (test data) could be estimated by leaving a set of simulations out of the training data-set and observe the reconstruction errors for it. To estimate the normalized root mean squared errors (NRMSE) for the reconstruction of the training and test data of Figure 4, simulations 11, 13, 23, 33 and 38, which had a low influence as shown in Figure 3, were left out of the training data-set employed to find the POD reduced basis. Visibly, the reconstruction errors associated to the training and test set were very different. As all the found POD modes were kept, the NRMSE for the training data was negligible. The NRMSE for the test data were below a value of 10^{-2} for the reconstruction of the temperature, CH_4, O_2, CO_2, H_2O fields, indicating an acceptable ability of the reduced basis to generalize to and represent unseen data. The reconstruction errors for the test data also serve as indication of the minimum prediction errors attainable by the subsequent predictive models for the prediction of the POD scores at unexplored operating conditions. These errors can be seen as the cost of the data compression part that leads to the development of a reduced-order model. As the original data were encoded into a few scalars, only a few surrogate models for these scalars needed to be trained for the prediction of the full fields. As explained in [5], having a low number of predictive models for the full system state means that only a few models have to be retrained in case new data become available. Thus, although reconstruction errors for the test data are not as low as the reconstruction errors for the training data, they are a necessary yet small cost (w.r.t. a case with no compression) for having an instantaneous predictive model that is also fast to re-train.

The influence of the scaling criteria for data pre-processing on the performance of POD for the case where the least influential simulations were left out, determined by 3, was also investigated. Figure 5 reports the reconstruction NRMSE for the recovery of the test data for six different scaling criteria. The center of each error bar represents the average value for the NRMSE among all vari-

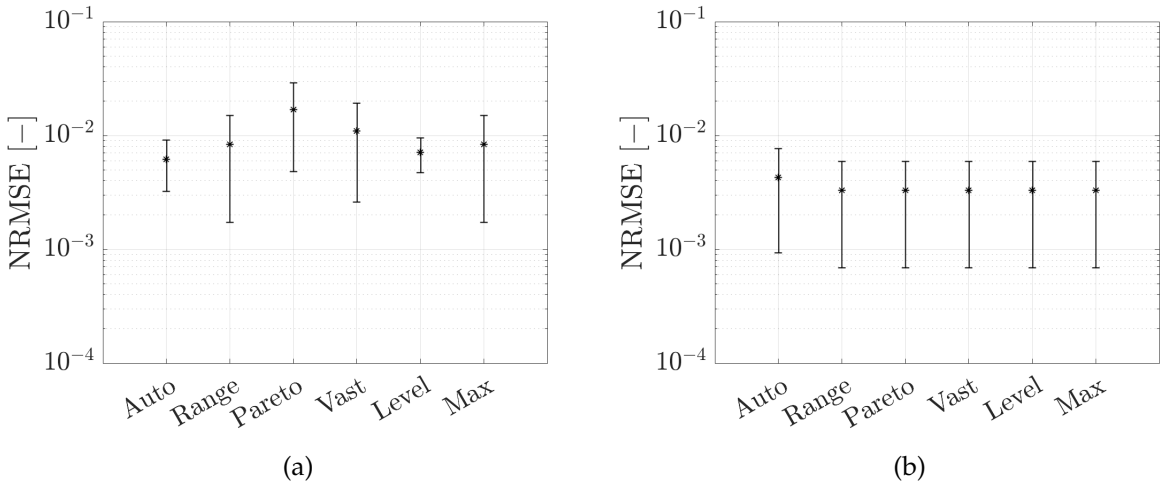
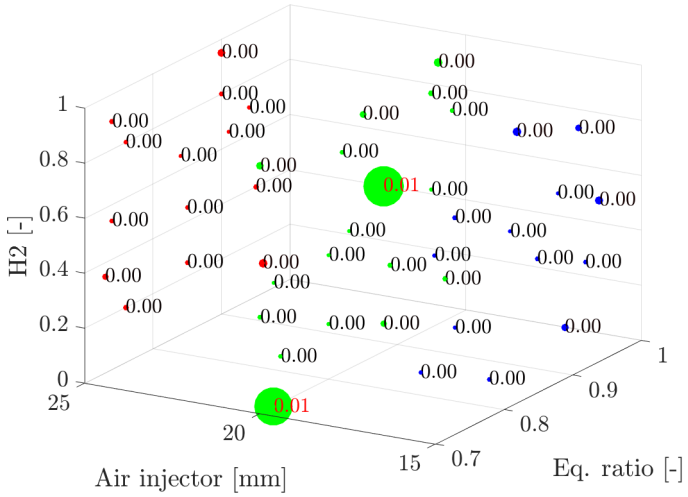


Figure 5: NRMSE for the reconstruction of the test data for different scaling criteria when (a) POD is applied to all the spatial fields altogether and when (b) POD is applied to each spatial field separately. The center of each error bar represents the average value for the NRMSE among all variables. The total height of the bar represents the standard deviation of the NRMSE among the variables.

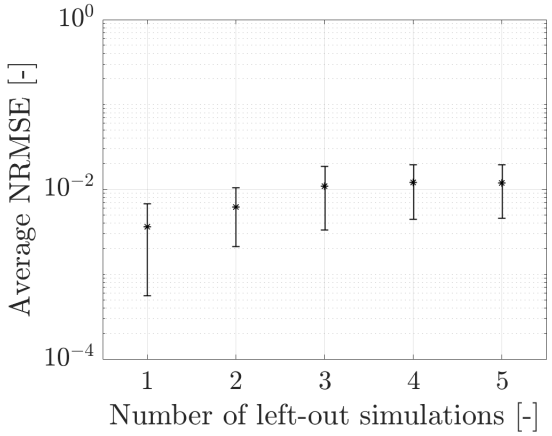
ables, for one scaling criterion. The total height of the bars represents the standard deviation of the NRMSE among the variables. In particular, Figure 5a reports the NRMSE for different scaling criteria when POD was applied to all the spatial fields of temperature and chemical species mass fractions altogether. Figure 5b reports the NRMSE for different scaling criteria when POD was applied to each spatial field of temperature and chemical species mass fractions separately. In the first case (Figure 5a), each scaling criterion led to different errors for the reconstruction of the test data. This was expected as the data matrix consisted of spatial fields with very different ranges and units, thus the scaling operation was crucial. In the second case where each field was scaled separately (Figure 5b), as visible all scaling criteria performed very similarly, with exception of the auto-scaling criterion which exhibited higher errors for OH and CO. Besides, the reconstruction NRMSE for the test data were lower in this case compared to Figure 5a. Thus, this was the chosen strategy to pre-process the data.

4.1. Sensitivity of the data compression step to the training data

In order to assess the importance of the size of the training data for the POD basis, a leave-k-out analysis was performed, where k was the number of simulations left out from the training data set employed to find the POD reduced basis. Each time, k simulations were left out and the error to reconstruct the left-out simulations from the POD basis was estimated. This was an important step to understand if enough data had been collected for the development of the ROM. Besides, leave-k-out errors are generally a more robust way to assess how a model will generalize to unseen data.



(a)



(b)

Figure 6: (a) Leave-one-out reconstruction errors, visualized in the input parameter space. The reported figures are the mean NRMSE among all variable for the reconstruction of one particular left-out simulation. The sizes of the circles also represent the this error. The color of the points represents the 3 different values of the air injector length. (b) Average NRMSE for the reconstruction of an increasing number of left-out simulations. Vertical bars represent standard deviations.

Figure 6a reports the mean reconstruction errors among all fields estimated with the leave-one-out (LOO) method. These errors reflected the influences on the POD basis of Figure 3. In the case of $k = 1$ (leave-one-out), the total number of possible design of experiments (DoE) was equal to the number of available simulations, and in this case it was also possible to visualize the results as done in Figure 6a. In the other cases, the total number of possible DoE was given by: $m!/(m-k)!\,k!$, where

m is the total number of available simulations and k is the number of simulations to leave out each time. For $k > 1$, only a random subset of all the possible combinations was considered. For a given value of k, the leave-k-out errors were estimated from random subsets of different sizes. The size of the subset was chosen to be 250, as the leave-k-out errors were converging for this value. The analysis was carried out for an increasing value of k, as reported in Figure 6b, where the average NRMSE and its standard deviation for the reconstruction of the test data are reported. Two observations can be made. Firstly, the average reconstruction error increased when more simulations were left out of the training data, as could have been expected. Secondly, the standard deviation of the reconstruction error generally decreased when k increased, indicating that for high values of k the ROM was more sensitive to the size of the training data than to the positions of the training simulations in the input parameter space.

5. Sensitivity of the ROM's predictive capabilities to the training data

After encoding the training simulations into a reduced set of POD scores or features, the values of these scores was used to train a surrogate model for the prediction of unexplored values of the POD coefficients so that the full fields of temperature and chemical species mass fractions could be predicted at unexplored operating conditions, which is vital for digital-twins that have to approximate that state of a real industrial system based on sensory data. In fact, by doing so, only a few sensors can be installed measuring a small number of physical parameters as it will be the digital-twin's task to provide an overview of what is occurring inside the system for the reported measurements.

The importance of the size of the training data for the prediction of unexplored conditions was investigated by means of leave-k-out analysis as well. This step was important to understand which and how many simulations were needed in order to develop a reliable ROM and is in general a more robust method to achieve this than the estimation of the influence on the POD basis as it considers all the steps for the construction of the ROM, instead of only one intermediate step. Figure 7a, similarly to Figure 6, reports the NRMSE associated to the prediction of each particular simulation (prediction of the spatial fields of temperature and main chemical species in this case), when that particular simulation was left out. Although relatively high prediction errors were observed for simulations that had a low influence on the POD basis, some of the highest LOO errors were observed for the prediction of the simulations that had the highest influences on the POD basis as well (as shown in Figure 3). Thus, the evaluation of these influences can be taken into consideration as a fast preliminary method to assess the quality of the training data and detect the regions in the input space were more observations are needed, even if doing so on the base of the LOO prediction errors is a more robust choice. Figure 7b reports the average NRMSE for the prediction of an increasing number of left-out simulations, similarly to what was done in Figure 6. As the LOO prediction errors of Figure 7a determined the most important simulations that as a consequence should always be included in the training set, the leave-k-out errors of Figure 7b were estimated by taking this into account and thus only the simulations whose influence was $< 15\%$ w.r.t. the most important simulation were taken into consideration as possible test data. Predictably, the prediction errors were greater for higher values of k, the number of simulations left out of the training data-set. Interestingly, as observed for the reconstruction errors, the standard deviation of the mean prediction NRMSE decreased when k

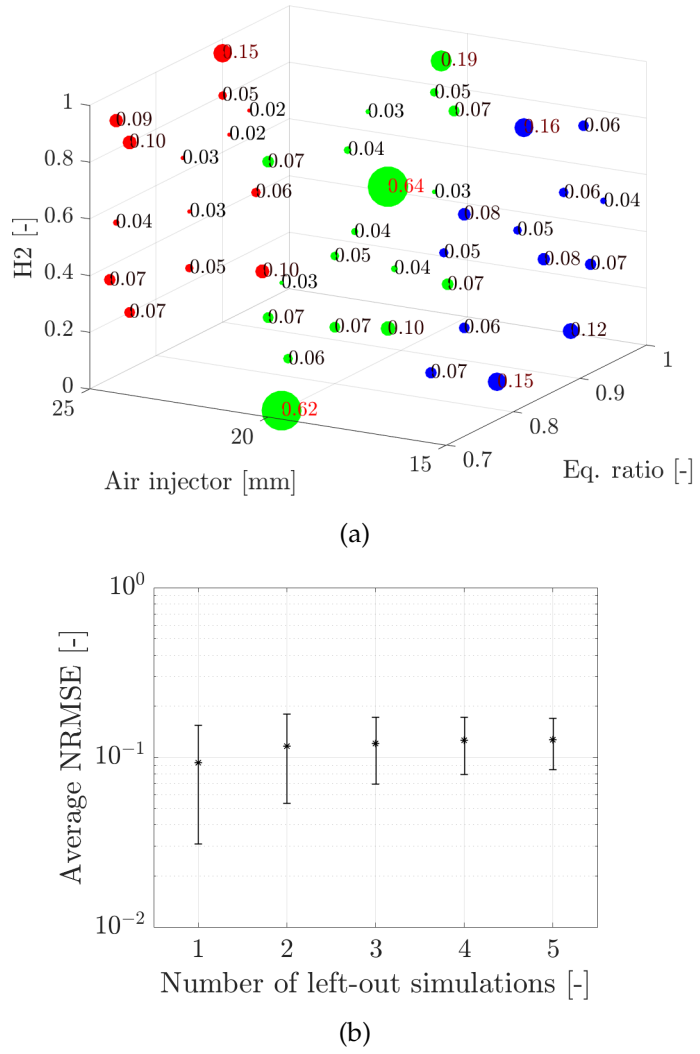


Figure 7: (a) Leave-one-out prediction errors, visualized in the input parameter space. The reported figures are the mean NRMSE among all variable for the prediction of one particular left-out simulation. The sizes of the circles also represent the this error. The color of the points represents the 3 different values of the air injector length. (b) Average NRMSE for the prediction of an increasing number of left-out simulations. Vertical bars represent standard deviations.

was increased. The leave-k-out errors for the prediction of scalar and integral quantities such as wall temperature, flame length and others are reported in Figure 8. In the context of stationary systems, it is of major interest to look at quantities that can be immediately compared to sensory data rather than at the full spatial fields, and to predict quantities such as flame-length and outlet composition. As seen in Figure 8a, a clear increase of the error, when changing the value of k from $k = 2$ to $k = 3$, was also observed for the prediction of the wall temperature. Besides, as also observed in Figure 7, also the standard deviations of Figure 8b tended to decrease for higher values of k. Low standard deviations for the prediction errors can be a preferable characteristic for a ROM when this guarantees lower upper bounds of the prediction errors, and thus gives the user of the ROM more certainties about the ROM's worst case scenarios in terms of predictive capabilities. The reliability of the ROM at predicting these quantities also means that, in case such a ROM is utilized for soft sensing [14].

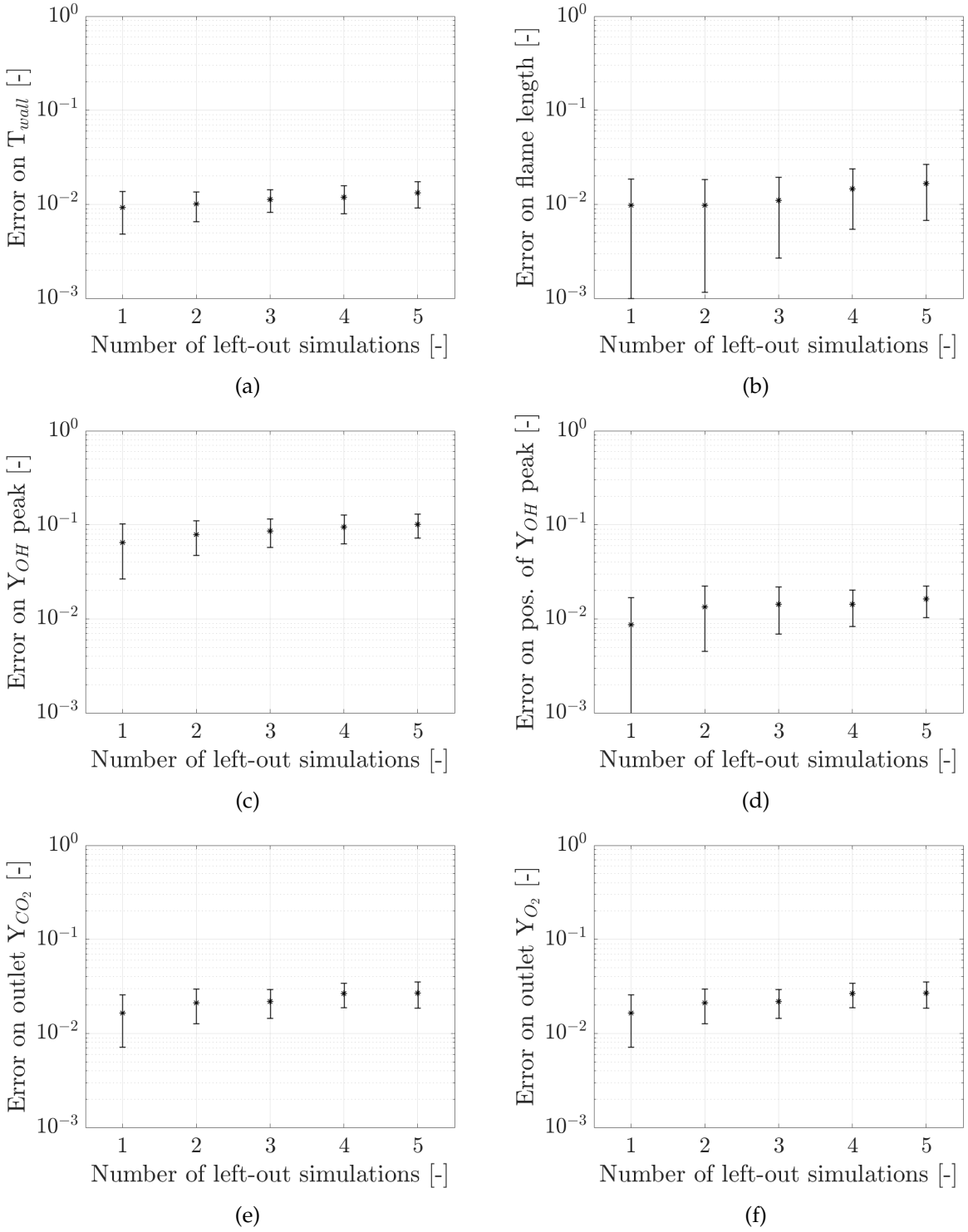
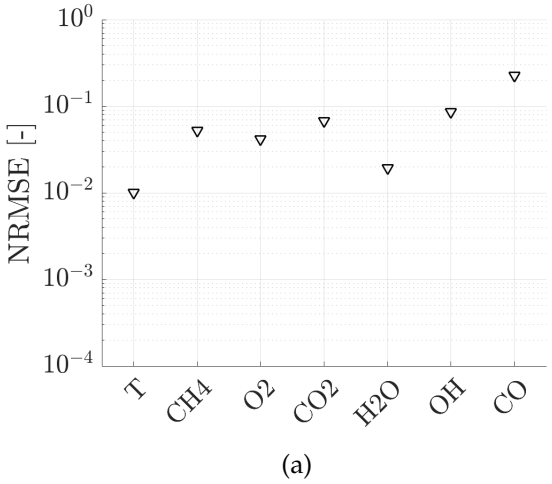


Figure 8: Leave-k-out relative errors for the prediction of scalar quantities such as (a) wall temperature, (b) flame length, (c) value of the peak of Y_{OH} and (d) its position, outlet mass fraction of (e) CO_2 and (f) O_2. Vertical bars represent standard deviations.

6. ROM developed from the training data-set determined by leave-k-out analysis

A ROM was developed based on the simulations determined by the leave-k-out analysis of the previous Section. For a given value of k, the set of left-out simulations which led to the best per-

formances in terms of predictions was chosen as test data and the remaining simulations were used to develop the ROM. The errors of Figure 8 were considered relatively low even for $k = 4$, suggesting that the use of a training set of corresponding size could lead to satisfactory performances as well, especially for the prediction of the scalar quantities. Thus, a value of $k = 4$ was chosen and all simulations except 1, 22, 28 and 39 (corresponding to the combination with minimum leave-k-out error for $k = 4$) were employed as training data-set to both find the reduced POD basis, thus the POD scores as well, and train a Kriging model for the prediction of the POD scores. The left-out simulations were used as test data, to assess the ROM's predictive capabilities.



(a)

Figure 9: NRMSE for the prediction of the test data by a ROM based on POD and Kriging.

Figure 9 reports the overall NRMSE for all the variables for the prediction of the test data. The fields of temperature and the main chemical species mass fractions were predicted with an error below 10%, whereas higher prediction errors were observed for OH and CO radicals, for which also the reconstruction errors were higher in comparison to the other variable, as seen in Figure 4. In general, the prediction errors are ≈ 10 times higher than the reconstruction errors.

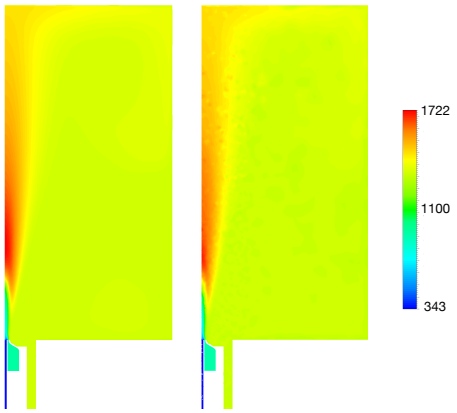


Figure 10: (*left*) True temperature field from simulation 1. (*right*) Predicted temperature field for the same operating conditions.

Figures 10, 11 and 12 compare the true temperature, CO_2, and OH mass fraction fields, respec-

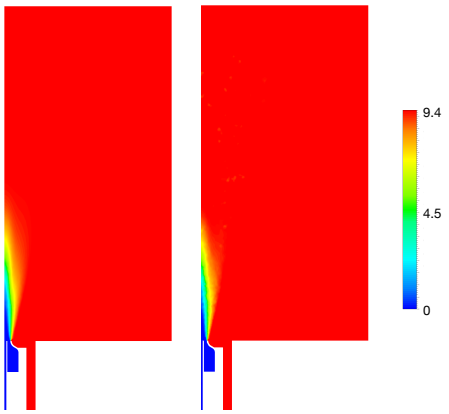


Figure 11: (*left*) True CO_2 mass fraction field from simulation 39. (*right*) Predicted CO_2 mass fraction field for the same operating conditions.

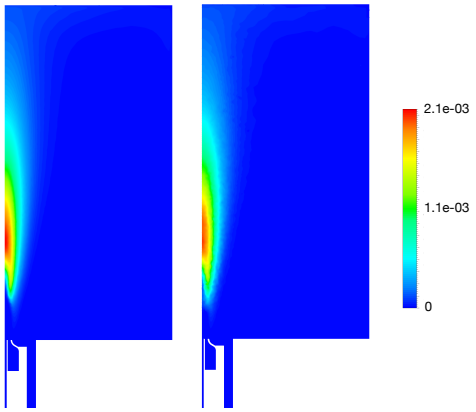


Figure 12: (*left*) True OH mass fraction field from simulation 39. (*right*) Predicted OH mass fraction field for the same operating conditions.

tively, to the predictions achieved by the developed ROM, for different unexplored operating conditions.

Figure 13 reports the true temperature values on the wall of the furnace where the sensors are installed and the values predicted by the ROM for this quantity. The discrepancy between the true and predicted values amounted to an average value of 12 K, corresponding to an error of $\approx 1\%$, which was considered acceptable. Table 1 and Figure 14 report the errors for the prediction of different scalar quantities such as flame length, position and value of the peak of OH and exhaust gas composition, for the left-out simulations. The flame length was estimated as the distance from the inlet (on a vertical axis) needed for the OH mass fraction to decrease from its maximum value to less than 5% of this value. The total height of the furnace was $0.7\ m$. The error on the flame length corresponded to less than 3% w.r.t. the true values for the left-out simulations, which was acceptable for the objective of the present work. No evident errors were observed for the prediction of the position of the peak of the OH mass fraction. The developed ROM could also predict the H_2O and CO_2 mass fractions in exhaust gas composition of the furnace with an error of up to 5% w.r.t. the true values. The errors on the exhaust gas composition for CO were also estimated, and these errors were lower

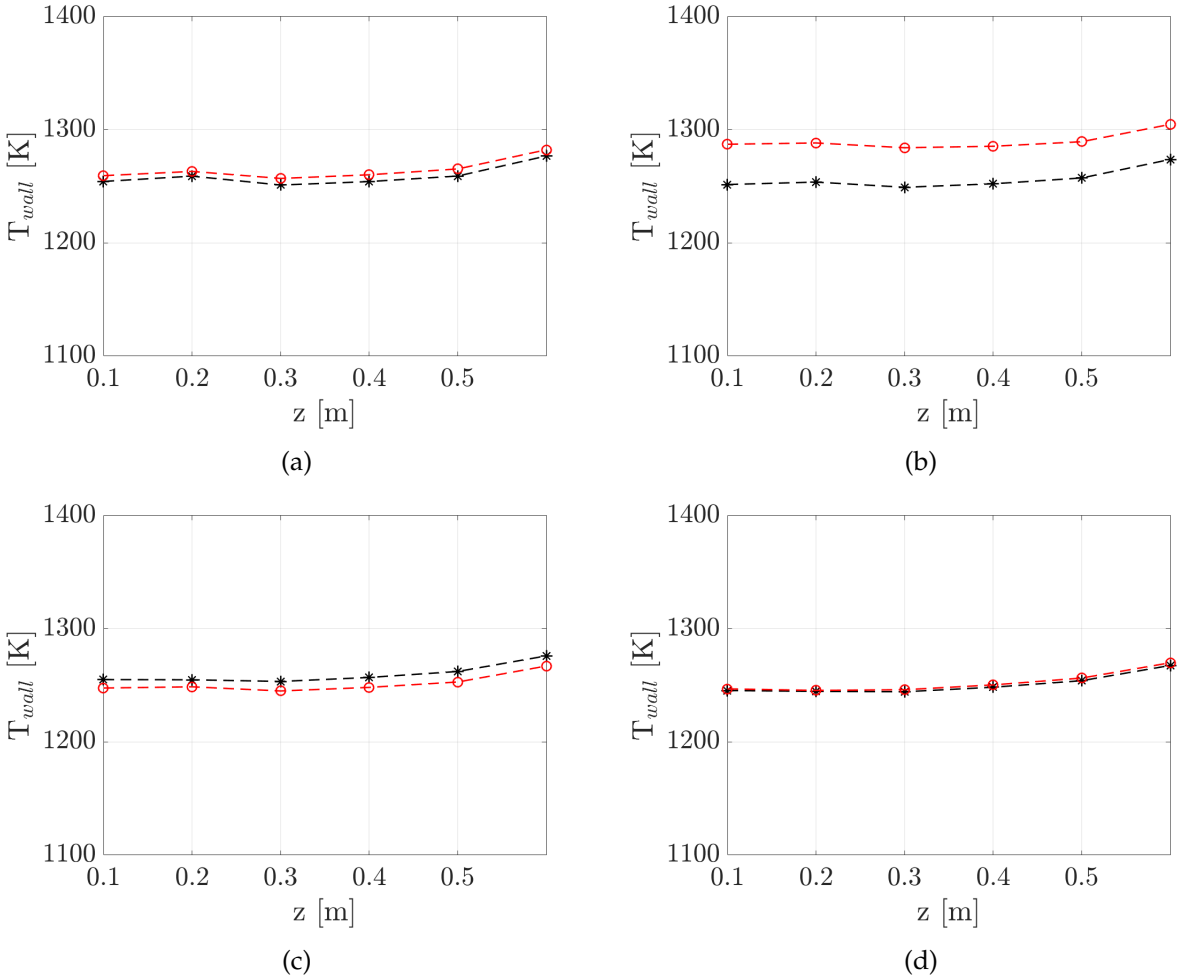
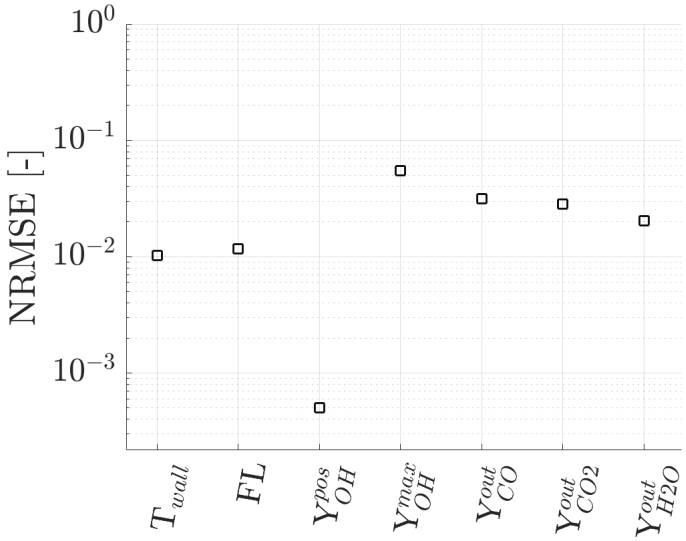


Figure 13: Prediction of wall temperature at the locations where real sensors are installed for the left-out simulations.

than 4%.

(Error on)	Sim. 1	Sim. 2	Sim. 3	Sim. 4
T_{wall}	0%	3%	1%	0%
Flame length	9%	0%	1%	0%
Position of Y_{OH} peak	2%	0%	0%	0%
Value of Y_{OH} peak	9%	5%	7%	1%
Y_{H_2O} outlet	1%	3%	4%	1%
Y_{CO_2} outlet	5%	1%	5%	1%
Y_{CO} outlet	3%	0%	0%	1%

Table 1: Digital twin's prediction errors for different scalar quantities of the furnace such as wall temperature, flame length, position of the peak of Y_{OH}, value of the peak of Y_{OH}, furnace outlet mass fractions of H_2O, CO_2 and CO.



(a)

Figure 14: Performance of the digital twin for the prediction of important scalar quantities of the furnace such as wall temperature (T_{wall}), flame length (FL), position of the peak of Y_{OH}, value of the peak of Y_{OH}, furnace outlet mass fractions of H_2O, CO_2 and CO.

7. Conclusions

In the present work, a reduced-order model (ROM) based on the combination of Proper Orthogonal Decomposition (POD) and Kriging was developed for the prediction of a 3D flame (prediction of the full fields of temperature and main chemical species) in a furnace for different operating conditions, chosen as the fuel composition, which was a mixture of methane and hydrogen, the value of the equivalence ratio and the length of the air injector, with the objective of the present work being the development of a reliable model that can predict the state of the considered combustion system and important scalar quantities such as wall temperature, exhaust gas composition and flame length, within an accuracy of 10%, in real-time and function as digital-twin for it. A set of simulations was produced and used as training data for the development of the ROM. The influence of each simulation on the reduced basis found by POD was estimated, so to identify the most important simulations to retain as training data. The influence of the number of training simulations used for the development of the ROM was also investigated. POD was used for data compression and

thus to represent the original data with a reduced number of features, the POD scores. Kriging was used to find a response surface for these scores at unexplored operating conditions. As the mapping from the reduced space to the original space was learned by POD, also the full fields of temperature and main chemical species could be predicted. A leave-k-out analysis was carried out in order to determine how many and which simulations were needed for the training of the ROM and estimate how the developed ROM would generalize to new data. Results showed that the developed ROM could predict the fields of temperature and CO_2, O_2, H_2O, CH_4 mass fractions, at unexplored operating conditions, reliably with an overall error of less than 10%. The predictions of scalar quantities such as wall temperature, position of the peak of OH mass fraction, exhaust gas composition and flame length were also obtained with an error of less than 5%. The reliability of the ROM at predicting these quantities means that the ROM can be utilized for soft sensing. As a consequence, it is concluded that the overall behavior of the considered system was well reproduced by the ROM, suggesting that the proposed methodology for ROM-development was valid and that the developed ROM could be used as a digital-twin of the furnace for real-time predictions of its state when the operating conditions change.

Acknowledgements

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.

References

[1] Z. Li, A. Cuoci, A. Parente, Large Eddy Simulation of MILD combustion using finite rate chemistry : Effect of combustion sub-grid closure 37 (x) (2019) 4519–4529. doi:10.1016/j.proci.2018.09.033.

[2] E. H. Glaessgen, D. T. Branch, D. S. Stargel, M. Sciences, The Digital Twin Paradigm for Future NASA and U . S . Air Force Vehicles (2019) 1–14.

[3] M. J. Kaur, V. P. Mishra, P. Maheshwari, The Convergence of Digital Twin, IoT, and Machine Learning: Transforming Data into Action, Springer International Publishing. doi:10.1007/978-3-030-18732-3.

[4] T. H. Uhlemann, C. Schock, C. Lehmann, S. Freiberger, R. Steinhilper, The Digital Twin: Demonstrating the Potential of Real Time Data Acquisition in Production Systems, Procedia Manufacturing 9 (2017) 113–120. doi:10.1016/j.promfg.2017.04.043.

[5] G. Aversano, A. Bellemans, Z. Li, A. Coussement, O. Gicquel, A. Parente, Application of reduced-order models based on PCA & Kriging for the development of digital twins of reacting flow applications, Computers and Chemical Engineering 121 (2019) 422–441. doi:10.1016/j.compchemeng.2018.09.022.

[6] M. Guenot, I. Lepot, C. Sainvitu, J. Goblet, R. F. Coelho, Adaptive sampling strategies for non-intrusive POD-based surrogates, Engineering Computations 30 (4) (2013) 521–547. doi:Doi10.1108/02644401311329352.

[7] G. C. Cawley, N. L. C. Talbot, On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation, Journal of Machine Learning Research 11 (2010) 2079–2107.

[8] A. Parente, J. C. Sutherland, Principal component analysis of turbulent combustion data: Data pre-processing and manifold sensitivity, Combustion and Flame 160 (2) (2013) 340–350. doi:10.1016/j.combustflame.2012.09.016. URL http://dx.doi.org/10.1016/j.combustflame.2012.09.016

[9] C. E. Frouzakis, Y. G. Kevrekidis, J. Lee, K. Boulouchos, A. A. Alonso, Proper orthogonal decomposition of direct numerical simulation data: data reduction and observer construction 28 (2000) 75–81.

[10] P. G. Constantine, E. Dow, Q. Wang, Active Subspace Methods in Theory and Practice, SIAM Journal of Scientific Computation 36 (4) (2014) 1500–1524.

[11] S. N. Lophaven, J. Søndergaard, H. B. Nielsen, Kriging Toolbox (2002) 1–28.

[12] M. Seeger, Gaussian processes for machine learning., Vol. 14, 2004. arXiv:026218253X, doi:10.1142/S0129065704001899.

[13] M. Ferrarotti, M. Fu, E. Cresci, W. D. Paepe, A. Parente, Key Modeling Aspects in the Simulation of a Quasi-industrial 20 kW Moderate or Intense Low-oxygen Dilution Combustion Chamber, Energy Fuels 32 (10) (2018) 10228–10241. doi:10.1021/acs.energyfuels.8b01064.

[14] B. Gabrys, S. Strandt, D.-d. Soft, Data-driven Soft Sensors in the Process Industry, Computers & Chemical Engineering.

5.5 FEATURE EXTRACTION IN COMBUSTION APPLICATIONS

Feature extraction in combustion applications

Giuseppe D'Alessio[a,b], Kamila Zdybał[a,b], Gianmarco Aversano[a,b], James C. Sutherland[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] *Department of Chemical Engineering, University of Utah, Salt Lake City, UT, USA*

Abstract

Interpreting turbulent combustion data sets remains a challenge due to their large dimensionality resulting from a high number of state-space variables describing the combustion process. Applying dimensionality reduction techniques can be a solution to this problem as they provide a low-rank approximation of the data which is easier to process and analyze. Recently, many such data-driven techniques from the machine learning community have been applied with success to a variety of pattern recognition and feature detection problems. In the present work, we investigate the capabilities of Principal Component Analysis (PCA), its non-linear or locally non-linear variants such as Kernel PCA (KPCA) and Local PCA (LPCA), Non-negative Matrix Factorization (NMF) and Autoencoders to extract interpretable features from a syngas jet flame numerical data set. The ability to link low-dimensional structures to physically motivated variables is shown. We present clustering methods and their impact on capturing flame regions. The quality of reconstruction using the explored methods is also investigated to find the best trade-off between detection of features and data approximation. Finally, a new method of performing clustering with artificial modes is presented.

Keywords: Principal Component Analysis, Non-negative Matrix Factorization, Autoencoders, Feature extraction

1. Introduction

High-dimensional combustion data sets are characterised by many underlying features. The successful data interpretation and parameterization of the combustion system often depends on the ability to detect the leading variables. Recently, data science techniques were applied to turbulent reacting systems to reduce the complexity of the chemical mechanisms and drive the development of combustion models. These techniques also show the potential of extracting low-dimensional structures which aids the interpretability, providing additional insight into the underlying physical phenomena. Some of the techniques that have been used in conjunction with the combustion systems are Principal Component Analysis (PCA) [1, 2, 3], Dynamic Mode Decomposition (DMD) [4, 5], Artificial Neural Networks (ANN) [6] or Merge Trees [7]. One of the major drawbacks of dimensionality reduction techniques is that the obtained manifolds are often not physically motivated. In Principal Component Analysis for instance, they are a linear combination of the original variables. There is a need for identifying these setup combi-

nations and input parameters which support the extraction of certain physical features. For flame data, the main interest could be in finding ignition or extinction regions, identifying the leading species and reactions, tracking the pollutant formation, etc.

The present work aims to uncover a set of rules that favour the feature extraction process in combustion data sets and compare the different potentials in doing so using various methods. We apply globally linear methods such as Principal Component Analysis (PCA) or Non-negative Matrix Factorization (NMF), as well as variants of those that are only locally linear. In addition, two non-linear methods are explored: Kernel PCA (KPCA) and Autoencoders. These techniques, with various constraints, are applied to capture physical phenomena from the Direct Numerical Simulation (DNS) data for a syngas flame. For each method, data was compressed by a set of retrieved *basis functions* or *modes*. All of the methods explored in this work impose their unique restrictions to the obtained modes and thus are capable of extracting different features.

We show that different modes are found by global and local techniques. In global techniques, the data reduction is applied to the entire data set and the obtained modes represented the entire set. In local algorithms, the data set is first partitioned into clusters and then a low-rank approximation is applied in each cluster separately. The main idea of the locally linear algorithm is that if the local regions are small enough the data manifold will not curve much over the extent of the region, thus improving the performance of linear encoding processes. Modes found in local clusters describe the physical phenomena associated with that particular cluster - in the reactive layer as well as the turbulent fuel jet, the identified key species were those associated to the fuel: H_2 and CO, while in the oxidizer stream O_2 and N_2 were identified. The global techniques, on the other hand, are likely to capture overall characteristics of the entire flame domain such as *mixture fraction*.

Apart from the interpretability, we also investigated the quality of reconstruction from the low-dimensional basis. It is expected that the level of reduction attainable by local algorithms in each cluster is larger compared to the use of global techniques [8]. Indeed, the greatest level of reconstruction was observed with non-linear techniques (such as Autoencoders). However, using the locally linear PCA with a high number of clusters (typicall larger than 30) was shown to approach the performance of a non-linear technique.

In addition to the restrictions inherent to the method, numerous factors can parameterize the data reduction process. Those could for instance be data pre-processing such as centering and scaling (done before applying the data reduction technique), rotation of the modes or even limiting the state-space to pre-selected variables from the original data set. In this work, we provide a guidance on how to choose the best input parameters that can extract specific combustion phenomena from a new, unseen data set. In particular, we explore the benefits of scaling criteria and show that they allow to emphasize the importance of different sets of variables.

Finally, a new method for clustering is proposed, namely Feature-Assisted Clustering (FAC), where *a-priori* user-selected basis functions (for instance ones that come from other techniques) are used as cluster identifiers.

2. Data description

In the present work, a single data set $\mathbf{X}$ consists of spatial profiles of 13 variables: temperature T and twelve mass fractions Y_i of the involved chemical species. Each data set has size $(n_s \times n_v)$, where n_s is the number of spatial points on the mesh and $n_v = 13$ is the number of variables.

The data sets chosen for the present work were provided by Direct Numerical Simulation (DNS) results of a CO/H_2 turbulent jet oxidation using a detailed mechanism consisting of 12 species and 33 reactions [9]. Additional information regarding the numerical simulation can be found in [10]. Two types of simulations were considered: the first one corresponded to a spatially-evolving jet (DNS1), the second one to a temporally-evolving jet (DNS2). For DNS1, three time steps were available: $t = 2.1 \cdot 10^{-3}$ s, $t = 2.5 \cdot 10^{-3}$ s, $t = 3.2 \cdot 10^{-3}$ s. Each time step of DNS1 employed a matrix made of approximately 1.5 million rows (observations), corresponding to the points of the mesh of the simulation, and 13 columns (variables) corresponding to temperature and the 12 chemical species considered by the mechanism. Several time steps of the numerical simulation of DNS2 were available, from $t = 1 \cdot 10^{-4}$ s to $t = 3.7 \cdot 10^{-3}$ s. Each time step was represented by matrix of approximately 700 thousand rows and 13 columns, also in this case corresponding to temperature and the 12 chemical species. Moreover, another data set corresponding to the source terms of DNS1 was available.

3. Theory

3.1. Data pre-processing

Data pre-processing operations such as centering and scaling are fundamental for the analysis of multi-variate data. Centering is achieved via subtracting a selected value from each column of the original data set. This allows to look at the data as variations from the chosen center. The centering criteria explored in this work are presented in Table 1.

Scaling is achieved by dividing each column of the data set by a chosen value. The scaling criteria explored in this work are presented in Table 1. Scaling allows to equalize variables that have different units and ranges. For instance, the Auto scaling forces all data columns to have a standard deviation equal to 1. Additionally, even for variables with the same unit, scaling allows to compare their relative importance evenly. It may for instance match the importance of major and minor species.

Centering criterion	
Mean	$\text{mean}(\mathbf{X})$
Min	$\min(\mathbf{X})$

Table 1: Centering criteria.

Scaling criterion	
Auto	$\text{std}(\mathbf{X})$
Level	$\text{mean}(\mathbf{X})$
Max	$\max(\mathbf{X})$
Pareto	$\sqrt{\text{std}(\mathbf{X})}$
Range	$\max(\mathbf{X}) - \min(\mathbf{X})$
Vast	$\text{std}(\mathbf{X}) \frac{\text{std}(\mathbf{X})}{\text{mean}(\mathbf{X})}$

Table 2: Scaling criteria.

3.2. Principal Component Analysis

Principal Component Analysis (PCA) [11] exploits the fact that the original basis in which the raw data set is represented is in general not an optimal basis. PCA thus finds a new, orthogonal basis to represent the data set: the original data set is approximated by retaining only a subset of fewer directions, namely the directions that account for most of the data variance in a transformed data set.

Prior to performing PCA, the raw data set $\mathbf{X}$ needs to be centered and scaled. We perform the data set pre-processing as follows:

$$\mathbf{X}' = (\mathbf{X} - \bar{\mathbf{X}})\mathbf{D}^{-1} \tag{1}$$

where $\bar{\mathbf{X}}$ is the centering matrix and $\mathbf{D}$ is the scaling matrix. For simplicity, we will from now on assume that $\mathbf{X}$ represents the already scaled and centered data set. In the present work, various centering and scaling options were explored in their capability to extract different features of the original data sets.

We compute the covariance matrix from the data set:

$$\mathbf{S} = \frac{1}{n_s - 1}\mathbf{X}^T\mathbf{X} \tag{2}$$

The eigenvalue decomposition of the covariance matrix yields the eigenvectors, the Principal Components (PCs), and their corresponding eigenvalues. We may write the eigendecomposition as:

$$\mathbf{S} = \mathbf{A}\mathbf{\Lambda}\mathbf{A}^T \tag{3}$$

where $\mathbf{A}$ is the matrix whose columns are the eigenvectors that form a new basis in which we can represent our original data set. The eigenvectors are orthonormal and the following property holds: $\mathbf{A}^{-1} = \mathbf{A}^T$. The diagonal elements of the matrix $\mathbf{\Lambda}$ are the corresponding eigenvalues. The eigenvalues are ordered decreasingly and they represent the amount of variance explained by the corresponding eigenvector. Since the convariance matrix is size $(n_v \times n_v)$, we obtain n_v Principal Components.

The data set can be transformed to the basis associated with the PCs:

$$\mathbf{Z} = \mathbf{X}\mathbf{A} \tag{4}$$

The matrix $\mathbf{Z}$ of size $(n_s \times n_v)$ is the matrix of Principal Component scores (PC-scores). The covariance matrix $\mathbf{S_Z} = \frac{1}{n_s - 1}\mathbf{Z}\mathbf{Z}^T = \mathbf{\Lambda}$ is a diagonal matrix, thus the columns of $\mathbf{Z}$ are linearly independent.

The data set can then be approximated by retaining only the first q PCs and q PC-scores:

$$\mathbf{X} \approx \mathbf{X_q} = \mathbf{Z_q}\mathbf{A_q}^T \tag{5}$$

the truncated matrix $\mathbf{Z_q}$ is size $(n_s \times q)$ and the matrix $\mathbf{A_q}$ is size $(n_v \times q)$. The matrix of differences $\mathbf{X} - \mathbf{X_q}$ between the original data and its low-rank approximation achieved by PCA can be used to estimate the associated reconstruction error to the data-compression process.

3.3. Rotation of the Principal Components

PCA can be coupled to rotation methods in order to increase the physical interpretation of the scores [12]. Many rotation methods exist, they can be either be orthogonal or oblique. The first class of techniques includes orthogonal transformations, as described by the varimax rotation by Kaiser [13], where the PCs are rotated rigidly over a fixed angle while ensuring orthogonality between the scores. The varimax rotation developed by Kaiser has generally been accepted as the most accurate orthogonal rotation and has been used widely in combination with PCA [14]. The varimax rotation criterion maximizes the sum of the variances of the squared coefficients within each eigenvector. The axes in the new system are rotated to maximize the rotation criterion given by the following expression

$$v = \frac{n_v \sum_{i=1}^{n_v} (a_{ij}/h_i)^4 - (\sum_{i=1}^{n_v} (a_{ij}/h_i)^2)^2}{n_v^2}, \tag{6}$$

where a stands for the principal component loadings, n_v for the number of variables, with $i \in 1, \ldots, n_v$ and $i \in 1, \ldots, q$ with q the number of PCs, and h_i are the communalities.

3.4. Local Principal Component Analysis

In combustion data sets the relationship among variables is often highly nonlinear. Since PCA is a linear technique, many PCs might be needed in order to approximate the original data set with low reconstruction error, thus combinations with other techniques are needed in order to overcome the limitations of global PCA. In local PCA (LPCA), the data are partitioned into clusters by either using an *a-priori* chosen quantity such as mixture fraction (supervised clustering), or using a Vector Quantization (VQ) technique, and then PCA is performed in each cluster separately (unsupervised clustering). This is based on the assumption that the non-linear manifold can be locally approximated by a linear one as the data manifold will not curve too much over the extent of the local region if these are small enough. This technique is also abbreviated to VQPCA. A more thorough description of LPCA can be found in [8].

In LPCA, the centroids of the clusters are found based on the minimization of the global reconstruction error. The reconstruction error of LPCA associated to one particular observation of the original data set is:

$$d(\mathbf{x}_i,\ \mathbf{r}^{(k)}) = \left(\mathbf{x}_i - \mathbf{r}^{(k)}\right)^T \mathbf{A_q}^{(k)\,T} \mathbf{A_q}^{(k)} \left(\mathbf{x} - \mathbf{r}^{(k)}\right), \tag{7}$$

where $\mathbf{x}_i$ is the i^{th} observation taken from the original data set and $\mathbf{r}^{(k)}$ is the centroid of the k^{th} cluster $\mathcal{R}^{(k)}$ to which this observation belongs, defined as $\mathbf{r}^{(k)} = \mathrm{E}[\mathbf{x} \in \mathcal{R}^{(k)}]$. The cluster k is defined as:

$$\mathcal{R}^{(k)} = \{\mathbf{x} \mid d(\mathbf{x}, \mathbf{r}^{(k)}) \leq d(\mathbf{x}, \mathbf{r}^{(j)});\ \forall\, j \neq k\}. \tag{8}$$

The reconstruction with q PCs of the i^{th} observation is given by:

$$\mathbf{x}_{i,q} = \mathbf{z}_{i,q} \mathbf{A}_q^{(k)\,T} + \mathbf{r}^{(k)} \tag{9}$$

The algorithm partitions the data into k clusters, computes the centroids for every cluster and associates every observation $\mathbf{x}_i$ to the cluster for which the error will be the least.

3.5. Kernel PCA

Kernel PCA (KPCA) [15], [16] is a non-linear dimensionality reduction technique that makes use of the kernel methods. In KPCA, the original data-set $\mathbf{X} \in \mathbb{R}^{m \times n}$ where $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m]^T$ is transformed to an arbitrarily high-dimensional feature space in the following way:

$$\mathbf{x}_i \in \mathbb{R}^n \longmapsto \boldsymbol{\phi}(\mathbf{x}_i) \in \mathbb{R}^N \qquad \forall i = 1, \ldots, m \tag{10}$$

with $N >> n$, and then PCA is carried out in this new feature space. A non-trivial, arbitrary function $\boldsymbol{\phi}$ is chosen but is never calculated explicitly, which allows for the possibility to use very high-dimensional feature spaces. In order to avoid working in the feature-space, a m-by-m kernel representing the inner product space (Gramian matrix) can be created:

$$K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j) = \boldsymbol{\phi}(\mathbf{x}_i)^T \boldsymbol{\phi}(\mathbf{x}_j). \tag{11}$$

The non-linearity of the KPCA method comes from the use a non-linear kernel function that populates the covariance matrix (known as the *kernel trick*). Thus, the eigenvectors and the eigenvalues of the covariance matrix in the feature space are never actually solved. The choice of the kernel function $k(\mathbf{x}_i, \mathbf{x}_j)$ modeling the covariance between two different points in the feature space is up to the designer. An example is the choice of a Gaussian kernel:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{\sigma} e^{\frac{|\mathbf{x}_i - \mathbf{x}_j|^2}{h}} \tag{12}$$

where h is the kernel width and $\frac{1}{\sigma}$ is the kernel scaling factor. Because no operations are directly carried out in the feature space, KPCA does not compute the PCs themselves, but the projections of our data onto those components, namely the scores.

PCA in the feature space would require the solution of the following eigenproblem:

$$\lambda \mathbf{a}_k = \mathbf{K} \mathbf{a}_k, \tag{13}$$

where $\mathbf{a}_k$ is the k-th PC in the feature space. All solutions must lie in the span of $\boldsymbol{\phi}$-images of the training data, meaning that there exist some coefficients $\boldsymbol{\alpha^k} = (\alpha_1^k, \alpha_2^k, \ldots, \alpha_m^k)$ such that:

$$\mathbf{a}_k = \sum_{i=1}^{m} \alpha_i^k \boldsymbol{\phi}(\mathbf{x}_i). \tag{14}$$

Substituting 14 into 13, and multiplying by $\boldsymbol{\phi}(\mathbf{x})$, leads to:

$$m \lambda_k \boldsymbol{\alpha}^k = \mathbf{K} \boldsymbol{\alpha}^k. \tag{15}$$

Normalizing the solutions $\mathbf{a}$ translates into $\lambda(\boldsymbol{\alpha}^T \cdot \boldsymbol{\alpha}) = 1$. The projection of the $\boldsymbol{\phi}$-image of a point $\mathbf{x}_i$ onto the k-th component, the k-th KPCA score for point i, is given by:

$$z_i^k = \boldsymbol{\phi}(\mathbf{x}_i)\mathbf{a}_k = \sum_{j=1}^{m} \alpha_i^k k(\mathbf{x}_i, \mathbf{x}_j). \tag{16}$$

One of the main drawbacks of the kernel PCA is that PCA is performed on a covariance matrix that scales with the number of observations m and not the number of variables n as in the PCA which increases the computational demands of the method. In addition, no straightforward operation is available to reconstruct the data from the KPCA scores, since the function $\boldsymbol{\phi}(\mathbf{x}_i)$ is never computed explicitly. This is in contrast to linear PCA, where the reconstruction of $\mathbf{x}_i$ from its compressed representation $\mathbf{z}_i$ is given by the linear operation: $\mathbf{x}_i \approx \sum_{k=1}^{q} \mathbf{z}_i \mathbf{a}_k = \tilde{\mathbf{x}}_i$, where q is the dimension of the low-dimensional space found by PCA. In order to reconstruct the data in Kernel PCA, the following functional needs to be minimized for $\tilde{\mathbf{x}}_i$:

$$\rho(\tilde{\mathbf{x}}_i) = k(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}) - 2 \sum_{k=1}^{q} z_k \sum_{j=1}^{m} \alpha_i^k k(\tilde{\mathbf{x}}_i, \mathbf{x}_j), \tag{17}$$

which makes the process computationally more expensive in comparison to PCA.

3.6. Non-negative Matrix Factorization (NMF)

Non-negative Matrix Factorization (NMF) [17] is a matrix factorization algorithm that focuses on the analysis of data matrices whose elements are nonnegative [18]. Given a non-negative data matrix $\mathbf{X} \in \mathbb{R}^{n_s \times n_v}$, NMF aims to find two non-negative matrices $\mathbf{U} \in \mathbb{R}^{n_s \times q}$ and $\mathbf{V} \in \mathbb{R}^{n_s \times q}$ which minimize the following objective function:

$$O = \left\| \mathbf{X} - \mathbf{U}\mathbf{V}^T \right\|. \tag{18}$$

In reality, we have $q < n_s$ and $q < n_v$. Thus, NMF essentially tries to find a compressed approximation of the original data matrix, $\mathbf{X} \approx \mathbf{U}\mathbf{V}^T$. The matrix $\mathbf{V}$ of non-negative factors can be regarded as the containing a basis that is optimized for the linear approximation of the data in $\mathbf{X}$, similarly to the matrix $\mathbf{A}$ of basis functions found by PCA. The matrix $\mathbf{U}$ is the compressed data, namely the NMF scores, similar to the matrix $\mathbf{Z}$ for PCA. Since relatively fewer basis vectors are used to represent the data, good approximations can only be achieved if the basis vectors discover structure that is latent in the data [19]. The non-negative constraints on $\mathbf{U}$ and $\mathbf{V}$ require the combination coefficients among different basis only be positive. This is the most significant difference between NMF and other matrix factorization methods, e.g., SVD. Unlike SVD, no subtractions can occur in NMF. For this reason, it is believed that NMF can learn a parts-based representation [17].

3.7. Feature-assisted clustering

In the present work, the reduced basis determined by PCA and NMF were used to perform data clustering, i.e. partitioning the rows of $\mathbf{X}$ into different groups. NMF for data clustering has already been used, as explained in [20]. To do so, imagine that a reduced basis for $\mathbf{X} \in \mathbb{R}^{n_s \times n_v}$, indicated by $\mathbf{A} \in \mathbb{R}^{q \times q}$, is available, with $q \leq n_v$. As said, this reduced basis could have been previously determined by either PCA, NMF, or any other method for low-rank approximation. The reduced basis is then used for clustering by evaluating the projection of the data on this basis as:

$$\mathbf{Z} = \mathbf{X}\mathbf{A}. \tag{19}$$

In case PCA is applied to obtain the reduced basis, the matrices $\mathbf{Z}$ and $\mathbf{A}$ are the PCA scores and PCA modes, respectively. The rows of $\mathbf{Z}$ are the projections of the rows of $\mathbf{X}$ on the columns of $\mathbf{A}$. The i-th row of X is assigned to cluster k if $|z_{ik}| > |z_{ij}| \, \forall j = 1, \ldots, q$ and $k \neq j$. The advantage of this procedure is that the clustering algorithm is usually faster than VQPCA, although the number of clusters that can be sought is limited by the maximum number of features that can be extracted from $\mathbf{X}$ as $q \leq min(n_s, n_v)$.

3.8. Autoencoder

An Autoencoder is a type of unsupervised artificial neural network (ANN). The aim of an Autoencoder is to learn a representation (encoding) for a set of data, typically for dimensionality reduction. Along with the reduction side, a reconstructing side is learnt, where the autoencoder tries to generate from the reduced encoding a representation as close as possible to its original input.

The simplest form of an Autoencoder is a feedforward, non-recurrent neural network having an input layer, an output layer and one or more hidden layers connecting them, where the output layer has the same number of neurons as the input layer, and with the purpose of reconstructing its inputs (minimizing the difference between the input and the output). Therefore, Autoencoders are unsupervised learning models.

Given one hidden layer, the encoder process of an Autoencoder takes to input $\mathbf{x} \in \mathcal{R}^{n_v}$ and maps it to $\mathbf{h} \in \mathcal{R}^{q}$, with $q < n_v$:

$$\mathbf{h} = f(\mathbf{W}\mathbf{x} + \mathbf{b}). \tag{20}$$

This image $\mathbf{h}$ is usually referred to as *code* or latent representation. Here, $f()$ is an element-wise activation function such as a sigmoid function or a rectified linear unit. $\mathbf{W}$ is a weight matrix and $\mathbf{b}$ is a bias vector. The decoder stage of the Autoencoder maps $\mathbf{h}$ to the reconstruction $\mathbf{x}'$ of the same shape as $\mathbf{x}$:

$$\mathbf{x}' = f'(\mathbf{W}'\mathbf{h} + \mathbf{b}'). \tag{21}$$

where f', $\mathbf{W}'$ and $\mathbf{b}'$ may be unrelated to f, $\mathbf{W}$ and $\mathbf{b}$.

Autoencoders are trained by minimizing the loss function $\mathcal{L}(\mathbf{x}, \mathbf{x}') = ||\mathbf{x} - \mathbf{x}'||^2$. This loss function can be modified to accommodate a regularization and a sparsity term.

3.9. Procrustes analysis

Procrustes analysis is a form of statistical shape analysis used to analyze the distribution of a set of shapes. An orthogonal Procrustes problem is a method which can be used to find out the optimal rotation and/or reflection (i.e., the optimal orthogonal linear transformation) for the Procrustes Superimposition (PS) of an object with respect to another. Given two shapes, determined by a finite set of points in a high-dimensional space, Procrustes analysis aims at finding the optimal linear operations such as translation, rotation and scaling so that the two shapes can be superimposed, with the square root of the difference between corresponding points used as a statistical measure to be minimized. This analysis can be useful to compare the data compressed by Autoencoder with the low-rank approximation found by LPCA. As the LPCA scores belong to different low-dimensional manifolds (corresponding to the different clusters), Procrustes analysis offers a way to transform the LPCA scores into the manifold found by the Autoencoder.

Given two set of m points in two different n-dimensional spaces, $\mathbf{X}_1$ and $\mathbf{X}_2$, both of size $(m \times n)$, the transformation to superimpose $\mathbf{X}_2$ over $\mathbf{X}_1$ is as follows: $\mathbf{X}_2' = \mathbf{b} \cdot \mathbf{X}_2 \cdot \mathbf{T} + \mathbf{c}$; where $\mathbf{b}$ is a scaling component, $\mathbf{T}$ is a rotation component and $\mathbf{c}$ is a translation component. If $\mathbf{X}_1$ and $\mathbf{X}_2$ are taken as the Autoencoder scores (corresponding to one LPCA cluster) and the LPCA scores from one cluster, respectively, Procrustes analysis lets $\mathbf{X}_1$ and $\mathbf{X}_2$ belong to the same space and thus allows to visualize the low-rank approximations found by the two methods.

4. Results

4.1. Global Principal Component Analysis

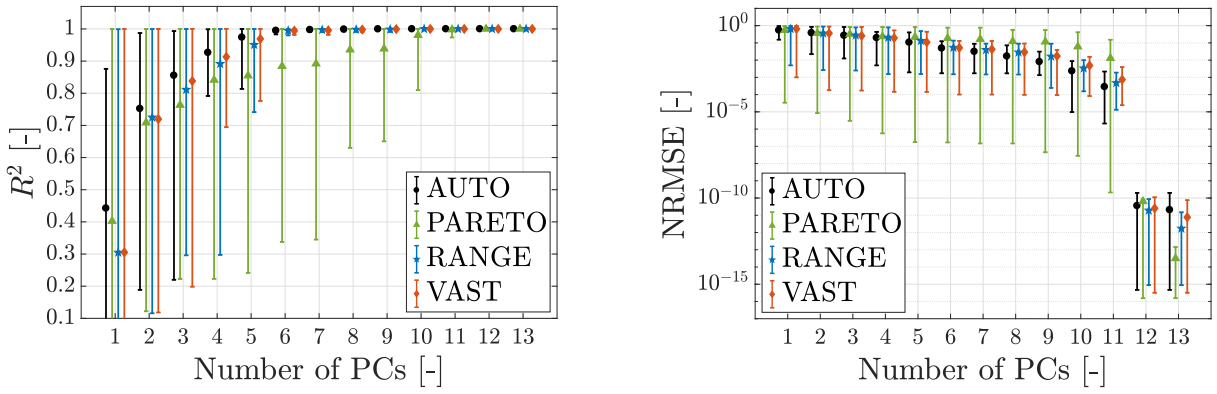
In section 3, it was explained that one of the free parameters of PCA is the number of Principal Components to retain. This parameter may be chosen by examining the percentage of explained variance associated with each PC, and retaining as many PCs as necessary to recover at least 99% of the total data variance so that only a limited amount of information is lost due to the dimensionality reduction. However, we show that the choice of the number of PCs based on this criterion is not always optimal. This can be understood analysing the data set corresponding to the time step $t = 2.5 \cdot 10^{-3}$ s of the spatially-evolving jet DNS1. We performed the recovery of the original data from the reduced representation provided by PCA, with an increasing number of retained PCs. In Table 3, the number of PCs necessary to recover at least 99% of the data variance is reported for several scaling criteria. In Figure 1, the average coefficients of determination R^2 and the average Normalized Root Mean Squared Errors (NRMSE) are computed taking the reconstruction of all (13) variables into account. As shown in Figure 1(a), if the 99% variance criterion was used, the average coefficient of determination for the recovery of the original data would be less than 50% for Pareto and less than 75% for Vast scaling. Even so, according to explained variance for these two scaling criteria, over 99% of the original data variance was recovered with not more than two dimensions, as indicated in Table 3.

In order to extract the features of the system, it is important that a sufficient amount of information is preserved in the compressed data representation. The reduced basis shall be able to recover the original data with a low approximation error. This issue plays a key role as the choice of the number of PCs based only on the explained variance could clearly undermine the effectiveness of the data recovery process and at the same time the reduced dimensionality could be too low to effectively gain any physical insight into the data.

Scaling criterion	*Number of PCs*
Auto	6
Pareto	1
Range	5
Vast	2

Table 3: Number of Principal Components necessary to recover at least 99% of the data variance with several scalings, for the time step $t = 2.5 \cdot 10^{-3}$ s.

As also discussed in [2], the way in which data is normalized before applying PCA has a very strong influence on the variables' reconstruction error and on the amount of information possible to extract from a lower-dimensional data representation. In particular, Pareto scaling criterion favours variables with high numerical values, which in the case of DNS1 data set is the temperature. Vast scaling, on the other hand, focuses on variables that do not show strong variation and thus focuses on stable variables such as mass fraction of the species N_2 which is not reacting. Figure 2 reports the weights on the first Principal Component for different scaling criteria, for two time steps of the spatially-evolving simulation: $t = 2.5 \cdot 10^{-3}$ s and $t = 3.2 \cdot 10^{-3}$ s. After scaling with Pareto, temperature was the only



(a) (b)

Figure 1: Average quality of reconstruction from all variables in the DNS1 data set when increasing the number of PCs in PCA. Average (a) coefficient of determination R^2 and (b) NRMSE, reported for different scaling criteria with maximum and minimum values indicated by bars. Time step: $t = 2.50 \cdot 10^{-3}$ s.

variable with acceptable values of R^2 and NRMSE after the original data reconstruction from the one-dimensional reduced space. Thus, no additional information on the features of the system could be extracted using only one PC apart from the temperature, which is an already known variable. However, the Pareto scaling criterion can be exploited knowing that it will not change the original data structure dramatically while decreasing the relative importance of large values, a property that can be useful for univariate data sets.

For Auto and Range scalings, a number of six PCs was necessary to attain values of the coefficients of determination close to 100% for all variables, which indicated that the considered data set could globally be compressed to a six-dimensional reduced manifold with negligible loss of information. This result was also exploited when considering local approaches for data analysis, where local clusters of observations were sought.

Each Principal Component is a linear combination of the original variables in the data set. Thus, all variables are characterized by a weight on a certain PC specifying how much they are represented by that PC. Analyzing the distribution of the variables' weights, it is possible to gain physical insight into the underlying process and also give a physical interpretation to the PCs. From the PC weight distribution on data scaled with Range, shown in Figure 2(c), we may observe that the first PC represents the mixture fraction. Similar interpretation is also true for the first PC after scaling with Vast (Figure 2(d)), despite the very high weight associated with the mass fraction of N_2. In Figure 3, scatter plots show how mixture fraction was dependent on the first PCA score for all four scaling criteria. A linear dependence can be observed when data was scaled with Range or Vast scaling. This is a particularly interesting aspect since mixture fraction was not among the variables in the original data set used for PCA.

The interpretation of the PCs can be further improved if the Varimax factor rotation is exploited. With this technique each PC is rotated and aligned with a fewer number of variables, making the physical interpretation easier. In Figure 4 the comparison between the first six original and rotated modes is shown for the Range scaling only. Examining the Varimax-rotated PC-3 to PC-6, instead of a mixture of

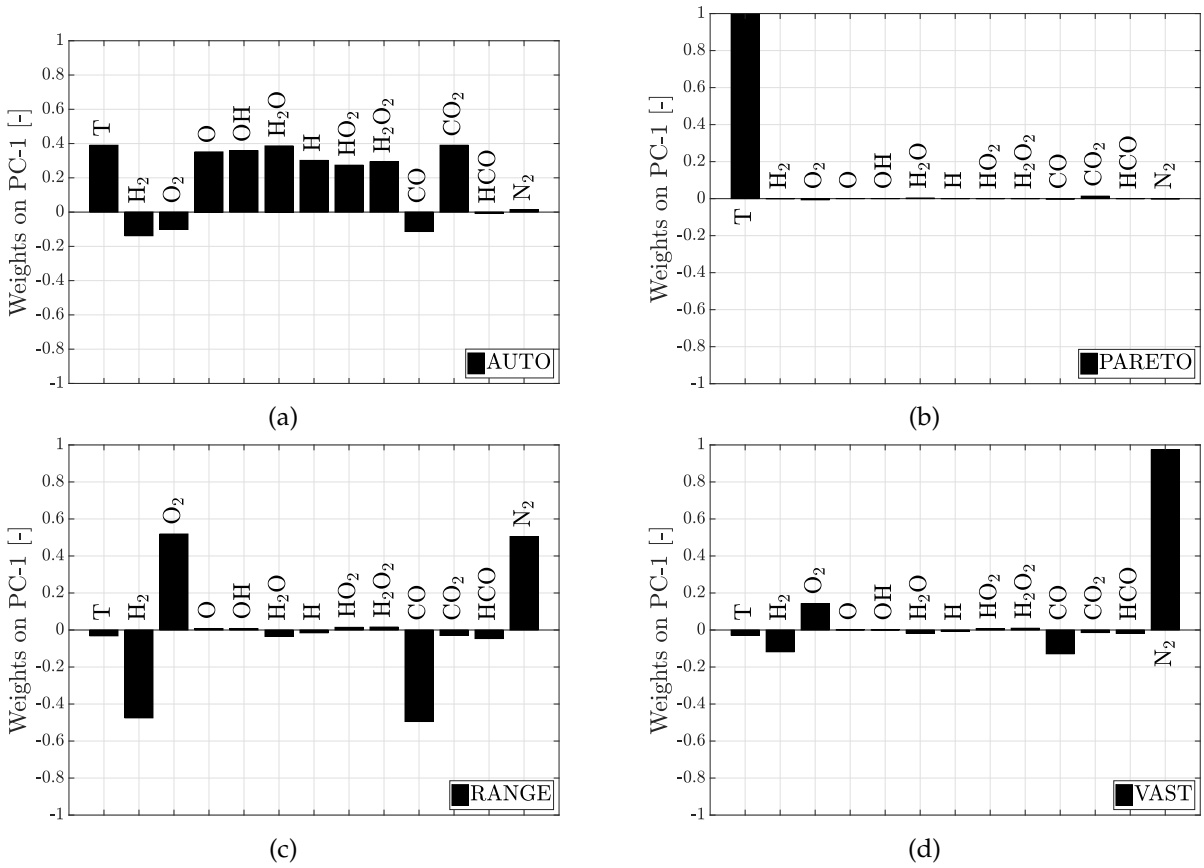


Figure 2: First PC determined by PCA with various scaling methods on the time steps $t = 2.50 \cdot 10^{-3}$ s and $t = 3.20 \cdot 10^{-3}$ s of DNS1. (a) Auto; (b) Pareto; (c) Range; (d) Vast.

variables, only one or two global variables have high weights.

4.2. Local PCA

With global PCA, a number of six PCs was necessary to achieve an accurate data reconstruction. By partitioning the data into regions and applying PCA locally, the same accuracy for the original data reconstruction could be achieved using fewer PCs per region. To partition the state-space into clusters, two approaches were investigated. As explained in section 3.4 a supervised approach (FPCA) or an unsupervised approach (VQPCA) was explored. The VQPCA algorithm is based on the unsupervised partitioning of the data into clusters minimizing the reconstruction error, while the FPCA partitioning is based on the a-priori knowledge of the data and a variable is used to condition the partition. In the context of turbulent non-premixed combustion, a valid choice as a conditioning variable could be represented by mixture fraction. Thus, PCA was applied in the local cluster after data has been partitioned into bins of mixture fraction. However, even if FPCA allowed for a very fast clustering with respect to the VQPCA approach, the choice of the mixture fraction as conditioning variable should be assessed on a case-by-case basis, as it could not always be optimal. As shown in Figure 5, with all the scaling criteria except for Pareto, it was possible to reconstruct all the variables at values of $R^2 \approx 100\%$ and at

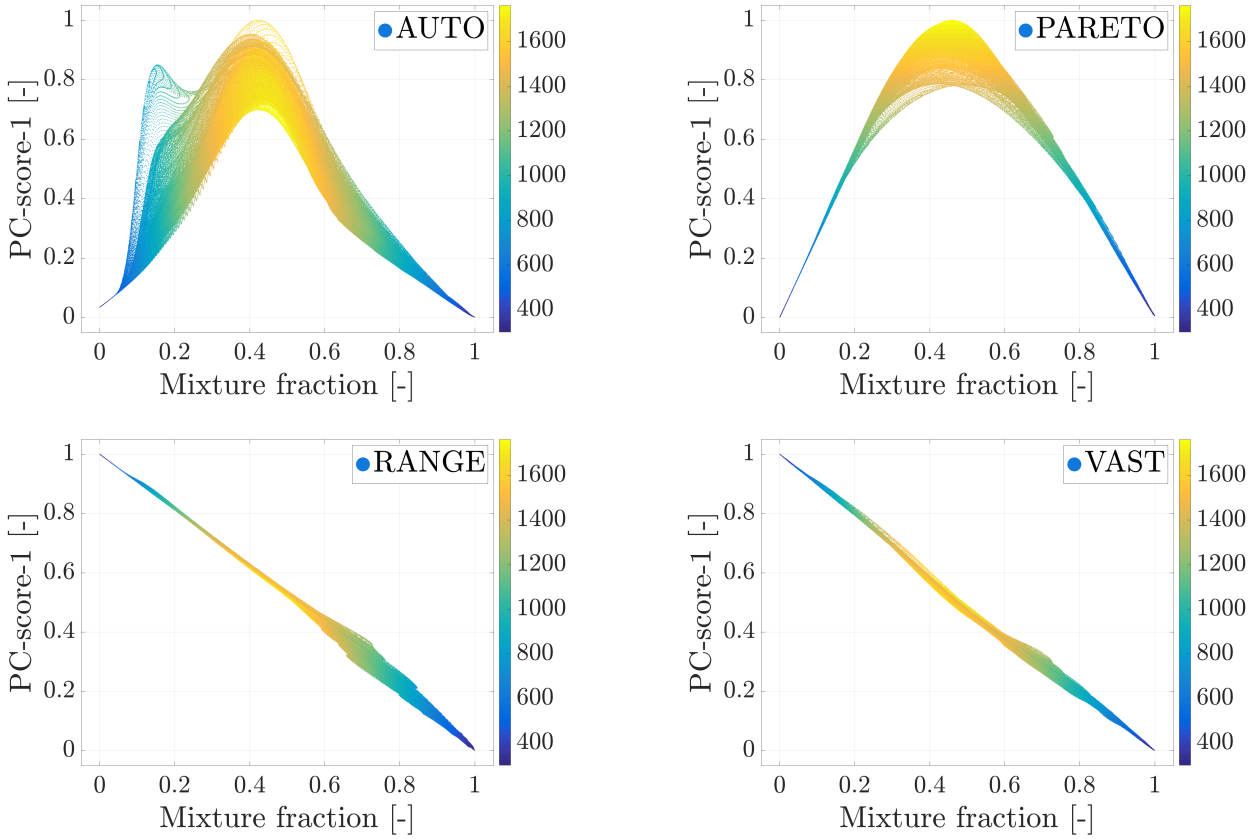
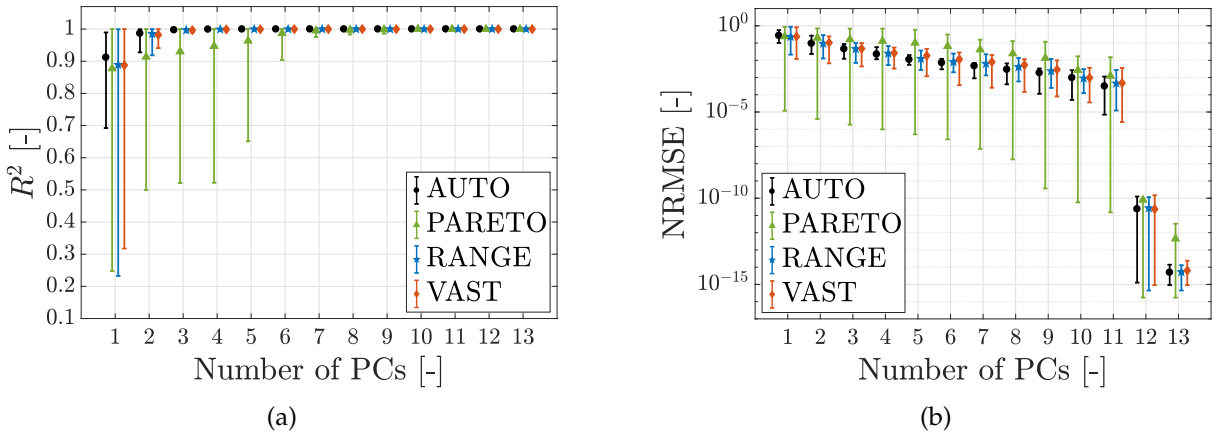


Figure 3: Scatter plot of mixture fractions against the first PCA score for different scaling criteria for DNS1. The PC-scores were normalized to attain values between 0 and 1. Time step: $t = 2.5 \cdot 10^{-3}$ s.

low values of NRMSE with only three PCs.



(a) (b)

Figure 5: Average quality of reconstruction from all variables in the DNS1 data set when increasing the number of PCs in LPCA with $k = 5$. Average (a) coefficient of determination R^2 and (b) NRMSE, reported for different scaling criteria with maximum and minimum values indicated by bars. Time step: $2.50 \cdot 10^{-3}$ s.

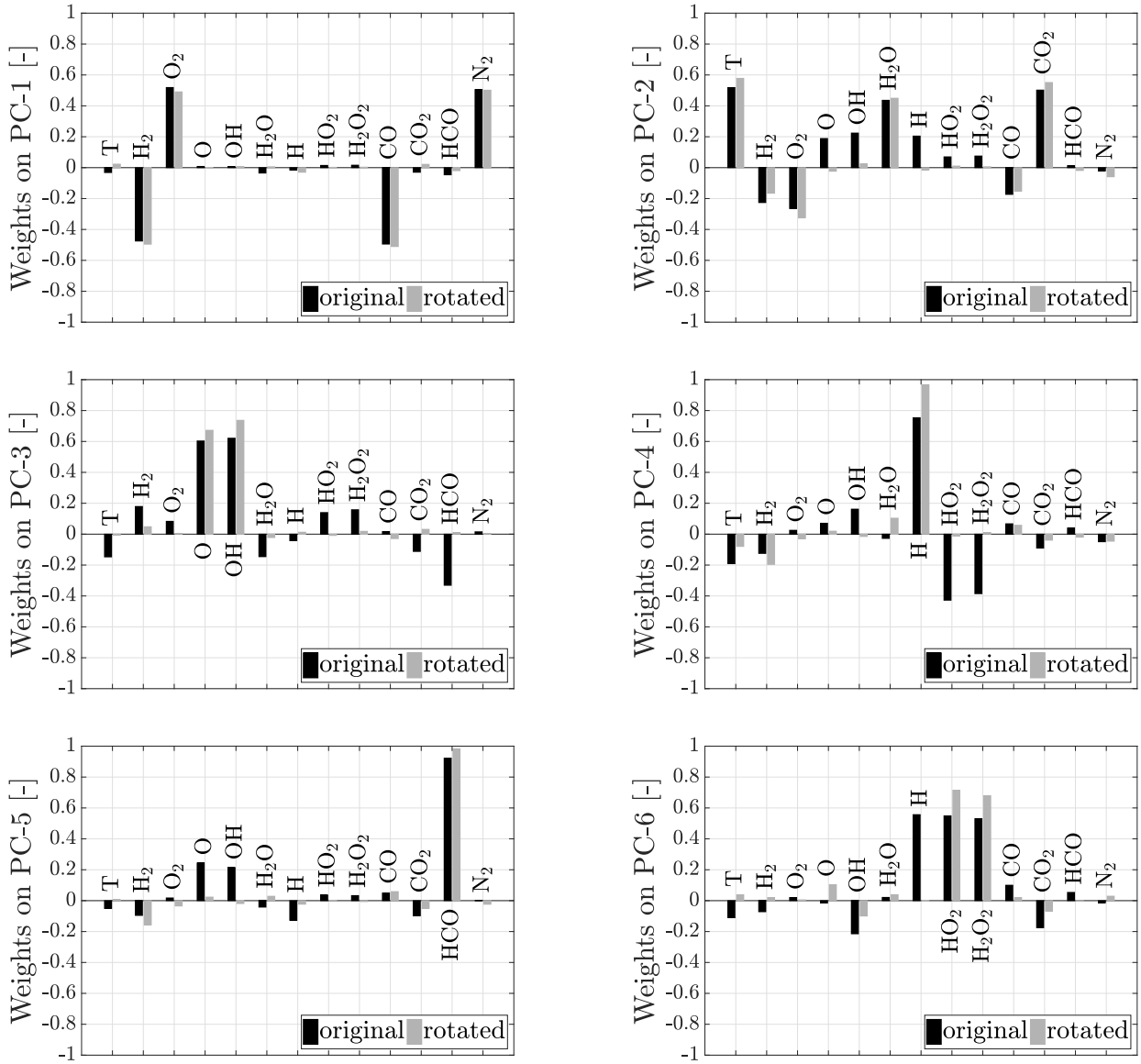


Figure 4: The original first six PCs found by PCA and rotated with the Varimax criterion for Range scaling. Time step: $t = 2.5 \cdot 10^{-3}$ s of DNS1.

As explained in section 3, the number of clusters k is a free parameter of the VQPCA partitioning algorithm and there is no explicit rule in the literature to choose a suitable value for feature extraction purposes.

In order to compare the potentials of global and local PCA for feature extraction, LPCA was performed using the VQPCA algorithm on the time step $2.5 \cdot 10^{-3}$ s of DNS1, for an increasing number of clusters after scaling the data set with Range. When VQPCA was applied to the data set, the $k = 4$ clusters represented distinctive zones of the flame characterized by rich, lean, and quasi-stoichiometric

mixture fraction values, as shown in Figure 6. As visible in Figure 6b, cluster k_1 corresponds to zones with low Z values, cluster k_3 corresponds to zones with high Z values, and clusters k_2 and k_4 to zones with values of Z close to the stoichiometric mixture fraction Z_{st}. The algorithm separated the non-reacting zones with different fuel-air ratio from the reacting ones in an unsupervised manner.

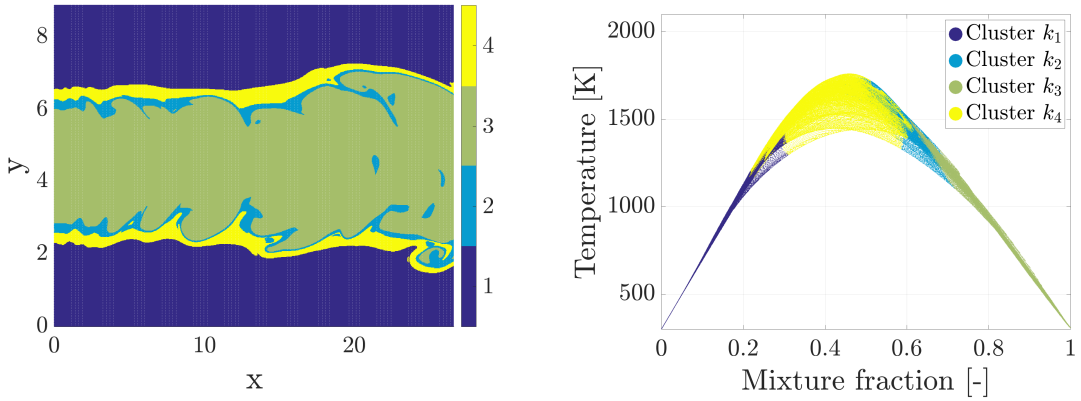


Figure 6: VQPCA performed with $k = 4$ clusters and four PCs in each cluster. Time step: $2.50 \cdot 10^{-3}$ s of DNS1 with the Range scaling criterion. (a) Results of the clustering process of LPCA with $k = 4$; (b) Temperature vs. mixture fraction plot - unsupervised division into clusters.

Figure 6a shows the data partitioninig achieved by LPCA with $k = 4$. Figures 7a, 7c, 7e, 7g show the weights on the first modes in each corresponding cluster, while Figure 7b, 7d, 7f, 7h show the cluster locations on the mesh. In case of $k = 4$, cluster k_1 was representative of the oxidizer stream, while cluster k_3 was representative of the syngas fuel jet. Clusters k_2 and k_4, instead, were representative of the reactive part, as visible from the high weights on the PCs for the radical species. Also in this case, Varimax rotation can be exploited to make the PCs interpretation in each cluster easier. It can be for instance observed, that after Varimax rotation of the first Principcal Component in the first cluster k_1, the three highest weights are on temperature, oxygen and carbon dioxide. There is also a smaller weight on N_2 and negligible weights on other species, which were significant in the unrotated PC. Thus, the rotated PC allows to better classify and intepret the cluster k_1 - it is mostly composed of oxidizer (O_2 and N_2) and one of the complete products of combustion reaction CO_2. The presence of CO_2 could for instance be attributed to the diffusion of that species into the oxidizer layer or the fact that the first cluster captures an outer portion of the reactive layer where CO_2 is largely present. Analysing Figure 7(b) where the temperature field is plotted in that cluster, we can indeed say that the latter is the case, especially that the temperature variable also maintains a high weight after Varimax rotation.

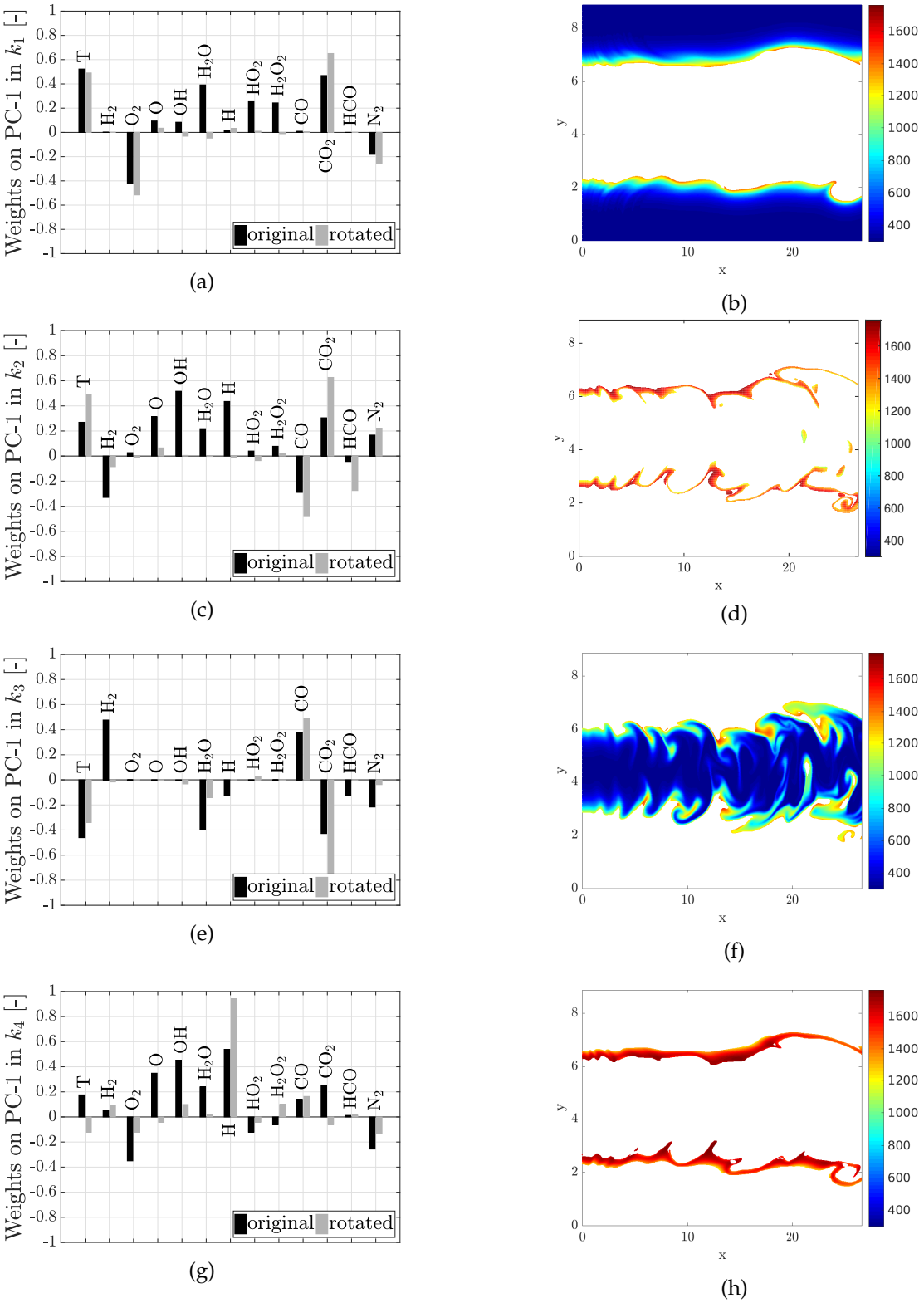


Figure 7: VQPCA performed with a number of $k = 4$ clusters and four PCs, with the Range scaling criterion. (a), (c), (e), (g) Weights on the first PC in each cluster; (b), (d), (f), (h) cluster position on the mesh - temperature field.

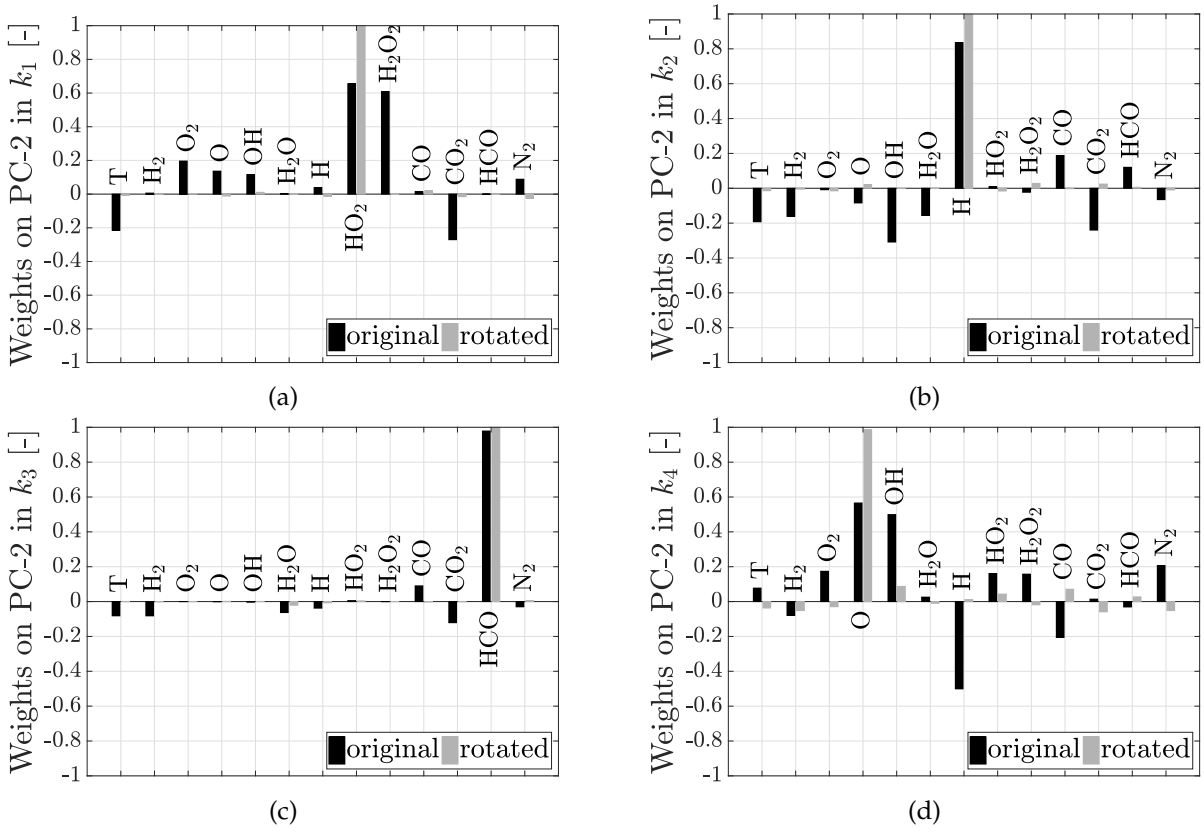


Figure 8: VQPCA performed with a number of $k = 4$ clusters and four PCs, with the Range scaling criterion. (a) PC-2 in cluster k_1; (b) PC-2 in cluster k_2; (c) PC-2 in cluster k_3; (c) PC-2 in cluster k_4.

A particularly interesting behaviour of VQPCA is that as the number of clusters grow, the amount of PCs per cluster needed to reconstruct the data decreases. From the data analysis point of view, this aspect may be essential as analyzing the data with a high number of clusters would be unfeasible in practice, especially in conjunction with a high number of variables. When increasing the number of clusters, the eigenvalue spectrum in each cluster started to have a positive skewness. This aspect is clearly visible in Figure 9a where the normalized eigenvalue spectra with respect to the first eigenvalue are plotted for the cluster representative of the syngas jet. Starting from a global PCA approach (which can be considered as VQPCA with $k = 1$) and increasing the number of clusters up to $k = 12$, the ratio between two consecutive eigenvalues started to increase. In global PCA, the ratio between the first and the second eigenvalue is not as high as with $k = 12$, where the second eigenvalue is two orders of magnitude smaller than the first. As shown in paragraph 4, the percentage of explained variance (i.e. the magnitude of the associated eigenvalues) was not a sufficient criterion to choose the number of PCs. Therefore, the reconstruction error analysis for different combinations of number of clusters k and the number of PCs was done and reported in terms of mean coefficient of determination $\bar{R^2}$ in Figure 9b. The reconstruction error analysis confirmed the results obtained from the examination of the eigenvalue spectrum. The $\bar{R^2}$ exhibited a clear relation not only to the number of PCs, but also to the

number of clusters. The $\bar{R^2}$ was increasing when the original data were partitioned in more bins. For this application, one PC could recover all the original variables with $\bar{R^2} > 95\%$ using $k = 12$ (or more) clusters. The increase in the ratio of consecutive eigenvalues shown in Figure 9 could also be explained

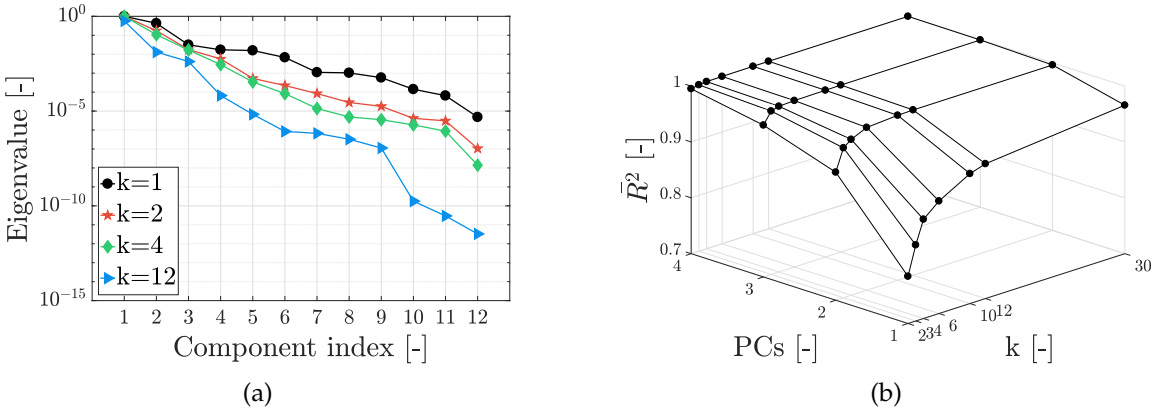


Figure 9: (a) Eigenvalues magnitude distribution as function of the number of clusters; (b) Mean coefficient of determination for the variables' reconstruction $\bar{R^2}$ as function of number of clusters k and number of PCs. DNS1, time step $2.50\ 10^{-3}$ s.

from both mathematical and physical point of view. As the number of clusters grew, the clusters started to have smaller extensions in the high-dimensional space. Mathematically speaking their geometries were also simpler, so a lower dimensional local manifold (i.e. a lower number of PCs) was required to correctly project the original data with a limited information loss. This was also physically confirmed by the fact that, as k was increased, the homogeneity in terms of chemical species in each cluster also increased (i.e. there were smaller high-dimensional groups of points, limited to only a small range of concentrations). This is shown in Figure 10, where the hydrogen concentration in the fuel jet cluster is reported for an increasing number of clusters.

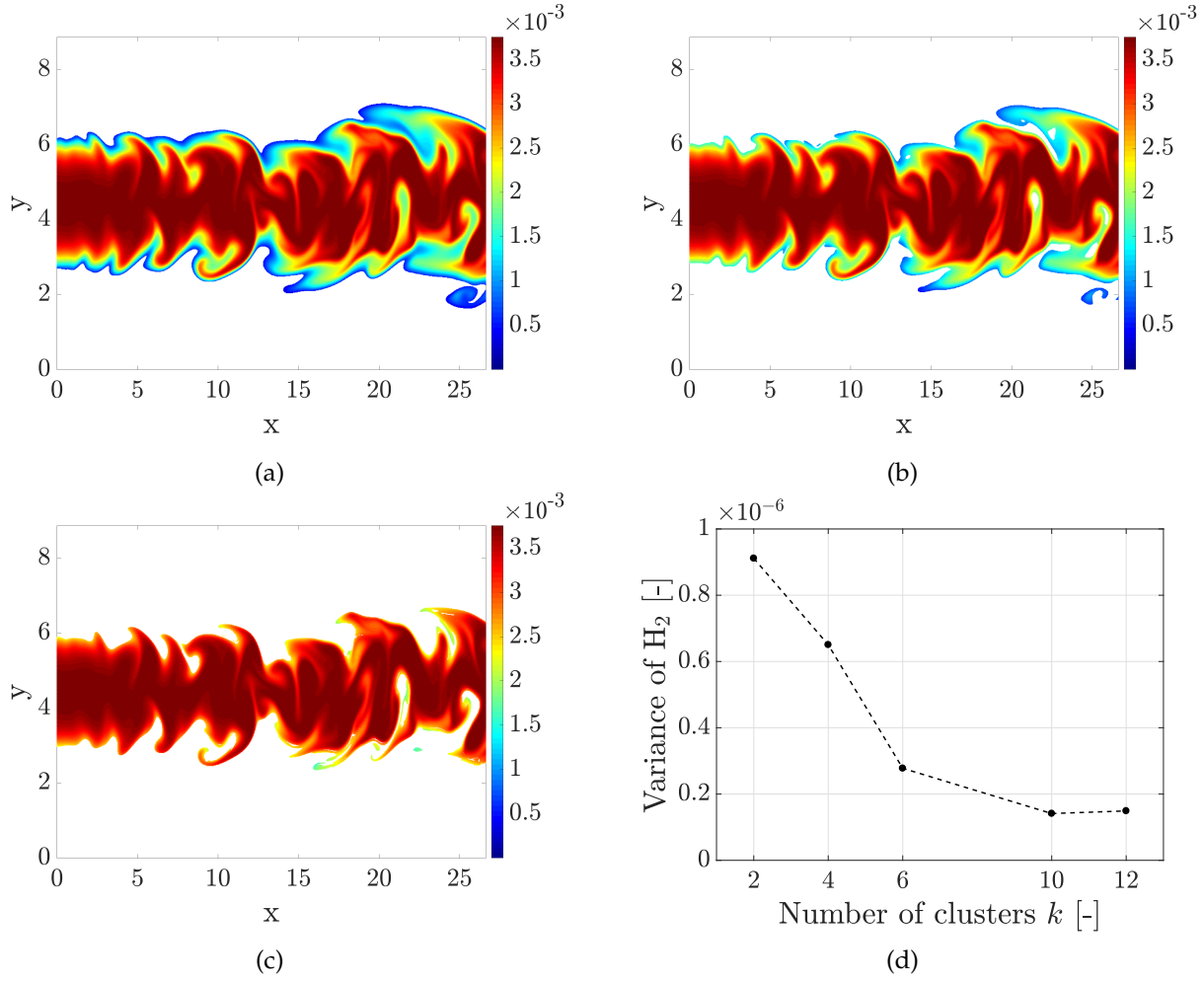
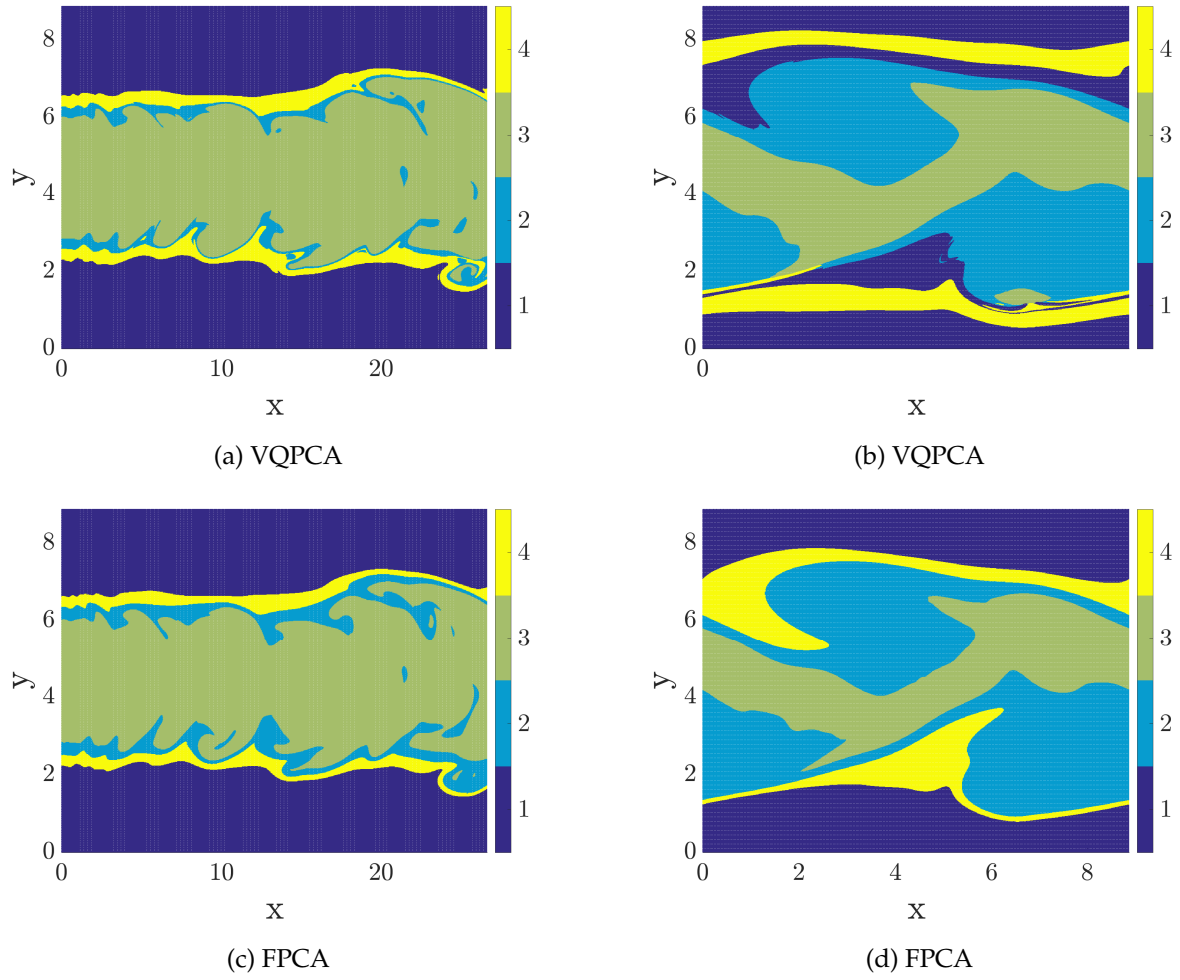


Figure 10: Cluster representing the fuel jet for a number of clusters $k = 2$ (a), $k = 4$ (b), $k = 12$ (c). 4 PCs retained. Colormap shows the values of the H_2 mass fraction. (d) Variance of H_2 concentration in the fuel jet cluster with the increasing number of clusters in input k. DNS1, time step $2.50 \cdot 10^{-3}$ s.

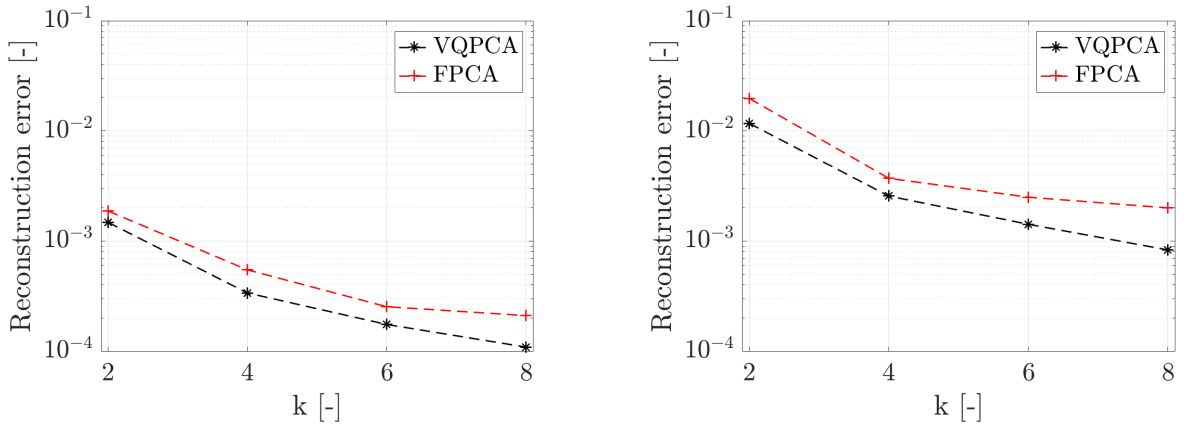
Applying FPCA and conditioning with mixture fraction to find clusters of the data set corresponding to time step $2.5 \cdot 10^{-3}$ s of DNS1 led to very similar data partitioning and consequently to comparable reconstruction errors as can be observed from Figures 11a and 11c. As far as DNS2 is concerned, the two approaches led to a different partitioning of the flame, as shown in Figures 11b and 11d where the partition for time step $3.7 \cdot 10^{-3}$ s is examined. Moreover, as shown in Figure 12, the VQPCA algorithm outperformed FPCA, providing values of the reconstruction error up to 58% lower than those obtained by means of FPCA, using the same scaling criteria, number of PCs and number of clusters k.



(a) VQPCA (b) VQPCA

(c) FPCA (d) FPCA

Figure 11: (a) VQPCA clustering with 4 PCs and $k = 4$ for DNS1, time step $t = 2.5 \cdot 10^{-3}$. (b) VQPCA clustering with 4 PCs and $k = 4$ for DNS2, time step $t = 3.7 \cdot 10^{-3}$. (c) FPCA clustering using 4 bins of mixture fraction for DNS1, time step $t = 2.5 \cdot 10^{-3}$. (d) FPCA clustering using 4 bins of mixture fraction for DNS2, time step $t = 3.7 \cdot 10^{-3}$.



(a) (b)

Figure 12: Reconstruction errors obtained with VQPCA and FPCA for different numbers of clusters, with 4 PCs and scaling (Range scaling criterion): (a) time step $t = 2.5 \cdot 10^{-3}$ s of DNS1; (b) time step $t = 3.7 \cdot 10^{-3}$ s of DNS2.

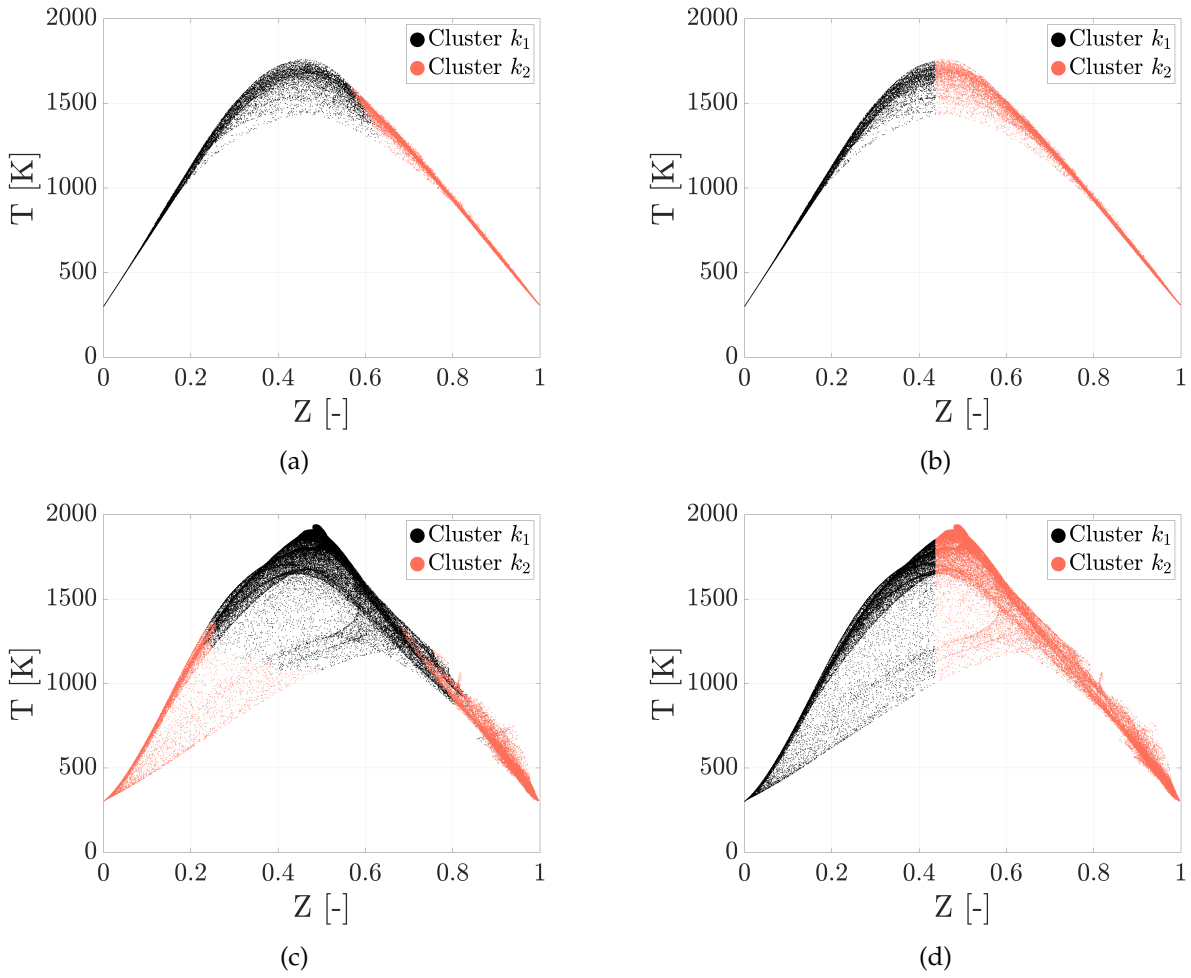


Figure 13: Data partitioning with $k = 2$ and Range scaling criterion: (a) VQPCA for DNS1; (b) FPCA for DNS1; (c) VQPCA for DNS2; (d) FPCA for DNS2.

Figure 13 shows how temperature in function of mixture fraction Z was partitioned into $k = 2$ clusters for the DNS1 (a) and (b), and DNS2 (c) and (d). For DNS1, the data is clustered into two regions for both VQPCA and FPCA, almost corresponding to rich and lean zones of the flame. This result suggests that the mixture fraction can be considered an optimal variable for the parameterization of the thermochemical state of the system, as it is generally assumed in many models for non-premixed combustion. For DNS2 instead, figure 13c shows that the first cluster is characterized by both lean and rich branches of the flame. This result is due to the complexity of the flame, characterized by significant local extinctions and re-ignitions. The VQPCA and FPCA reductions appeared comparable for the DNS1 data set, while VQPCA more clearly outperformed FPCA for DNS2. This confirmed that mixture fraction was not optimal from the point of view of error minimization when the physics under investigation became too complex. This was expected, as mixture fraction is only a measure of the local system stoichiometry and it can only cover relatively fast scales. The small discrepancies between FPCA and VQPCA for DNS1

suggest that VQPCA actually tended to FPCA when dealing with relatively simple systems, characterized by fast chemistry and a small degree of extinction. In the above discussion, VQPCA was shown to be in general superior to FPCA from the point of view of reconstruction error minimization. However, it should be reminded that VQPCA is an iterative algorithm, whereas FPCA is based on the supervised partition of the data into bins of mixture fraction: the CPU time associated with the first one is certainly higher than that of FPCA. Moreover, the computational cost of the iterative algorithm is proportional to the number of observations and to the number of clusters, k. The CPU time associated with VQPCA can reach values of order of minutes or hours, depending on k and the data set size, whereas the CPU time associated to FPCA is of the order of seconds and minutes. Thus, FPCA represents certainly a valid solutions for applications similar to the DNS1 one as it is a good trade-off for both CPU time and error.

4.3. Cluster homogeneity metrics

In order to quantify the goodness of a clustering technique, and compare all the clustering techniques which have been tested on the DNS data, three homogeneity metrics were introduced. The different clustering techniques described in the previous sections are summarized in the Table 4 by means of mean Normalized Root Mean Squared Error (NRMSE) and three different metrics to assess the clusters' homogeneity. Metric 1 is the mean silhouette value, which is a measure of how each point in the cluster is similar to other points in the same cluster. Metric 1 is defined as:

$$\text{Metric 1} = \frac{1}{k}\sum_{i=1}^{k}\frac{\bar{d}_{b,i} - \bar{d}_{a,i}}{max(\bar{d}_{a,i}, \bar{d}_{b,i})} \tag{22}$$

where $\bar{d}_{a,i}$ is the average distance from the i^{th} point to other points in the same cluster and $\bar{d}_{b,i}$ is the minimum average distance from the i^{th} point to other points in a different cluster. Metric 2 starts from the computation of the quantities g_{ij}, with the index i and j referring to a particular cluster and variable, respectively, as follows:

$$g_{ij} = \frac{a_{ij} - b_{ij}}{\mu_{ij}} \tag{23}$$

where a_{ij} and b_{ij} are the maximum and minimum values of the variable j in cluster i, and μ_{ij} is the mean of variable j in cluster i. Thus, Metric 2 is the global mean expressed as follows:

$$\text{Metric 2} = \frac{1}{k}\sum_{i=1}^{k}\left(\frac{1}{N_s}\sum_{j=1}^{N_s} g_{ij}\right), \tag{24}$$

while Metric 3 is defined as:

$$\text{Metric 3} = \frac{1}{k}\sum_{i=1}^{k}(\sigma_i) \tag{25}$$

where σ_i are the standard deviations of g_{ij} $\forall j = 1, \ldots, N_s$ for cluster i. The effects of the distance metrics on the clustering algorithm were also investigated, to see if this parameter can impact on the feature extraction process. As explained in Section 3, in VQPCA this distance is the distance between

a particular point or observation and the low-dimensional manifold determined by PCA in a certain cluster. Here, a modification of VQPCA was also used, evaluating the distances from the centroids of the clusters to a particular observation. The distance metrics used are: Euclidean, Minkowski and Cityblock.

No.	*Data set*	*Log*	*Clustering meth.*	*Centering/Scaling*	NRMSE	Metric 1	Metric 2	Metric 3
1	DNS1		VQPCA	Mean/Range	0.02482	0.7693	36.0	41.6
2	DNS1		FPCA	Mean/Range	0.03540	0.7985	58.9	65.1
3	DNS1	✓	VQPCA	Min/Range	0.08775	0.7135	428.6	441.0
4	DNS1		Euclidean	Mean/Range	0.04448	0.8318	98.2	119.8
5	DNS1		Minkowski	Mean/Range	0.05421	0.8138	72.3	80.9
6	DNS1		Cityblock	Mean/Range	0.03980	0.8362	88.5	106.5
7	DNS1		feature-assisted	Mean/Auto	0.03421	0.5057	19.0	16.6
8	DNS1		feature-assisted	Mean/Range	0.05514	0.0718	20.5	36.2
9	DNS1		source	Mean/Auto	0.02922	0.6297	27.8	27.9
10	DNS1		source	Mean/Range	0.05024	0.4946	62.2	98.8
11	DNS1	✓	source	Mean/Auto	0.11275	0.4636	626.7	1200.7
12	DNS1	✓	source	Mean/Range	0.12323	0.5024	626.7	1200.7
13	DNS2		VQPCA	Mean/Range	0.05575	0.3517	9.9	4.8
14	DNS2		FPCA	Mean/Range	0.06637	0.6131	12.4	8.6

Table 4: Summary of clustering methods in terms of mean NRMSE and three cluster homogeneity metrics. If data was transformed to log-space before clustering it is indicated by a tick in the *Log* column.

In terms of mean NRMSE errors, the best performance was achieved on the DNS1 data set when the data were clustered with VQPCA (No. 1 in Table 4) and a similar result was obtained with source terms clustering (No. 9). A still close value of NRMSE is achieved for FPCA with mixture fraction as the conditioning variable, however VQPCA outperforms FPCA on both DNS1 and DNS2 data sets. Clustering based on source terms had the second lowest NRMSE, just after VQPCA on DNS1. The analysis showed that the log space transformation significantly increased the NRMSE. In fact, the worst NRMSE out of all clustering methods tested, was achieved for clustering with source terms in the log space (No. 11 and 12). A high silhouette value indicates that a point is well matched to its own cluster. The three highest silhouette values are achieved for clustering with three distance metrics: Cityblock, Euclidean and Minkowski. The highest homogeneity indicated by small values of Metric 2 was achieved on the DNS2 data set for VQPCA and FPCA, while on the DNS1 data set it was achieved for feature-assisted clustering. This can be explained by the fact that feature-assisted clustering employs a-priori known features when looking for clusters.

4.4. Kernel PCA

As explained in section 3.5, the KPCA projects the data into an arbitrarily high-dimensional space through kernel transformation. In the present work, three kernels were explored: polynomial of 2^{nd} and 3^{rd} order, Gaussian (RBF) and sigmoid. In each case, the data set was scaled with the Range criterion and

the dimensionality was reduced to $q = 2$ or $q = 4$ target dimensions. Figure 14 presents the selected four lowest NRMSE reconstruction errors. It can be found that KPCA performed with $q = 4$ and a polynomial kernel of 3^{rd} order led to the lowest reconstruction errors for the original variables and this case is presented next for the feature detection capabilities.

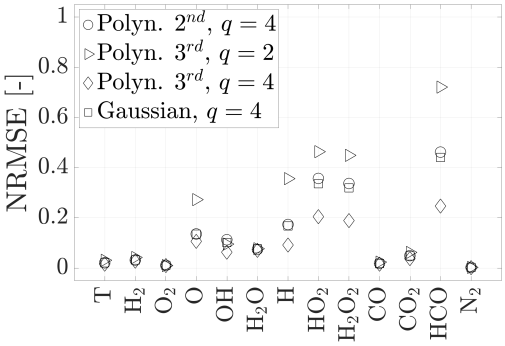


Figure 14: NRMSE for the reconstruction of the original variables from KPCA performed with a polynomial kernel of 3^{rd} order and 4 dimensions. Time step: $2.50 \cdot 10^{-3}$ s of DNS1 with Range scaling.

Figure 15 shows the correlations of kernel PCA scores with each of the original variables. Similarly to global PCA, the first PC-score from KPCA represents the mixture fraction (with the highest correlation with variables H_2, CO, O_2, N_2 and the total correlation with mixture fraction of 99.93%). Figures 16-17 report the first four KPCA scores obtained. Figure 16(a) shows the linear dependence of the mixture fraction and it can be seen in Figure 16(b) that the first score well separates the fuel stream from the oxidizer stream.

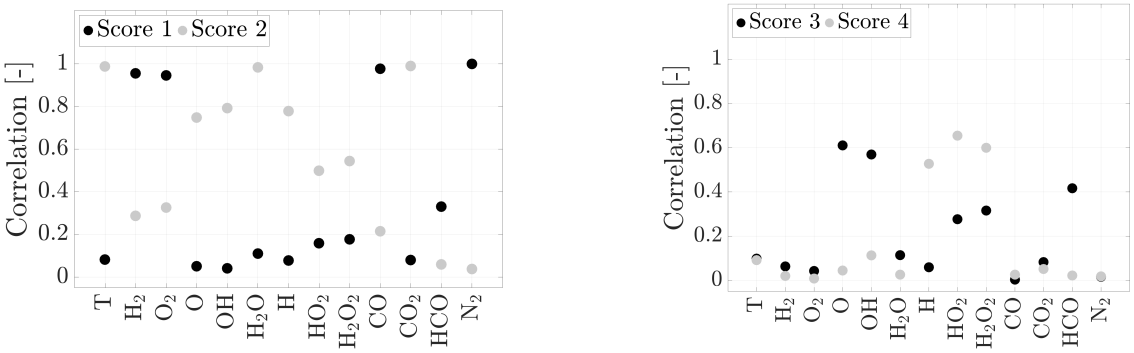


Figure 15: Absolute correlation of kernel PCA scores with each of the data set variables. KPCA performed with a polynomial kernel of 3^{rd} order and 4 dimensions. Time step: $2.50 \cdot 10^{-3}$ s of DNS1 with Range scaling.

The second PC-score correlates highly with temperature (the total correlation of 98.68%) and thus can identify the reactive layer, shown also in Figure 16(d). Additionally, from Figure 15 we can observe that the second score is highly correlated with the fully oxidized species H_2O and CO_2 and therefore it identifies the complete products of reaction in the reaction zone. The third score, plotted in Figure 17(b), represents the radical species. This can be better observed analyzing Figure 15, where the third

PC-score correlates highly with O and OH and also with HCO, HO_2 and H_2O_2. Similarly, the fourth score, plotted in Figure 17(d), represents species H, HO_2 and H_2O_2.

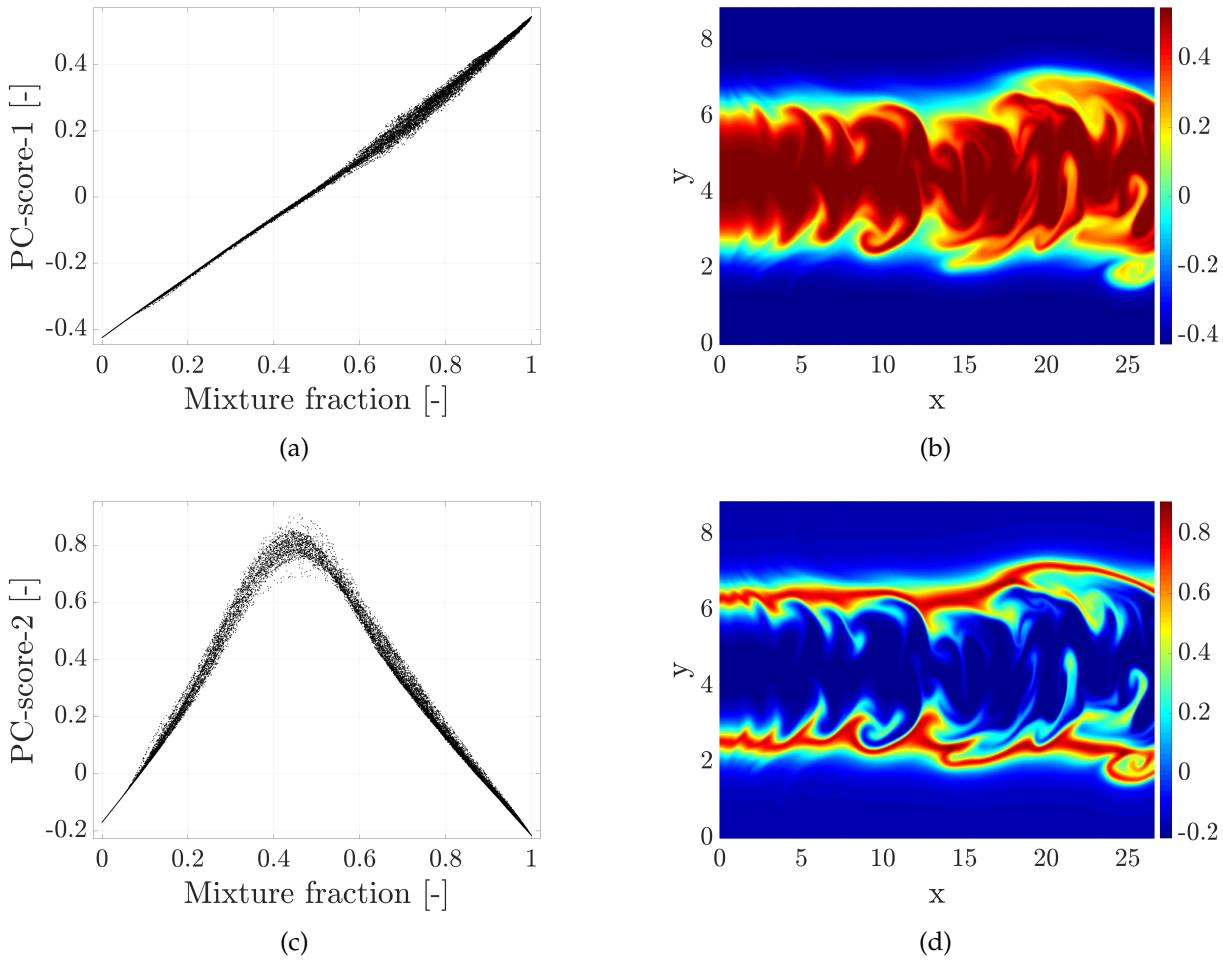


Figure 16: KPCA scores obtained with polynomial kernel of 3^{rd} order using 4 dimensions. (a) First PC-score in function of mixture fraction; (b) field of the first PC-score. (c) Second PC-score in function of mixture fraction; (d) field of the second PC-score. Time step: $2.50 \cdot 10^{-3}$ s of DNS1 with Range scaling.

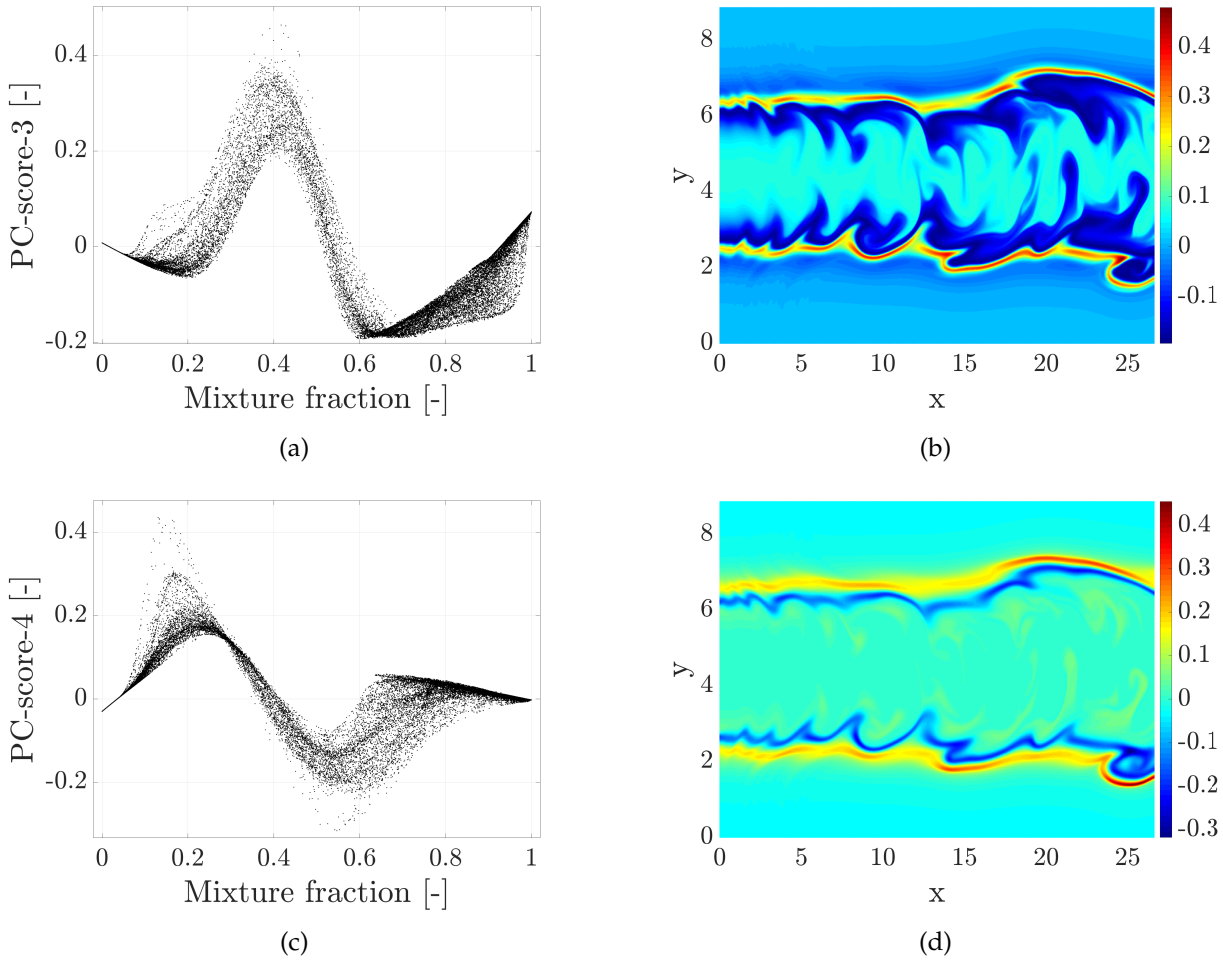


Figure 17: KPCA scores obtained with polynomial kernel of 3^{rd} order using 4 dimensions. (a) Third PC-score in function of mixture fraction; (b) field of the third PC-score. (c) Fourth PC-score in function of mixture fraction; (d) field of the fourth PC-score. Time step: $2.50 \cdot 10^{-3}$ s of DNS1 with Range scaling.

Observing the two-dimensional manifold obtained from KPCA in Figure 18 we can attribute regions of this manifold to certain flame regions and the phenomena characterizing them. In Figures 18(a)-(b) the manifold is coloured with the fuel species CO and the oxidizer O_2 respectively. These variables change smoothly on the manifold and can be viewed as progress variables. The leftmost region of the manifold can thus be attributed to the fuel stream and the rightmost one to the oxidizer stream. More so, the reaction marking variables OH and temperature, which are plotted in Figures 18(c)-(d), have the highest values close to the central region of the 2D manifold. The manifold of KPCA can thus adequately parameterize the cross section of the flame with the combustion reaction taking place where the fuel and oxidizer meet at the stoichiometric conditions. This is yet another interesting point to make since we know from the previous analysis that the first PC-score represents the mixture fraction and the second one the temperature. The region of the reaction zone in Figure 18 also corresponds to the peak where the stoichiometric conditions arise. This result could be particularly useful in the context of counter-

flow flames, where fuel and oxidizer come from two opposing regions and the reaction is taking place in-between, where the mixing conditions are stoichiometric.

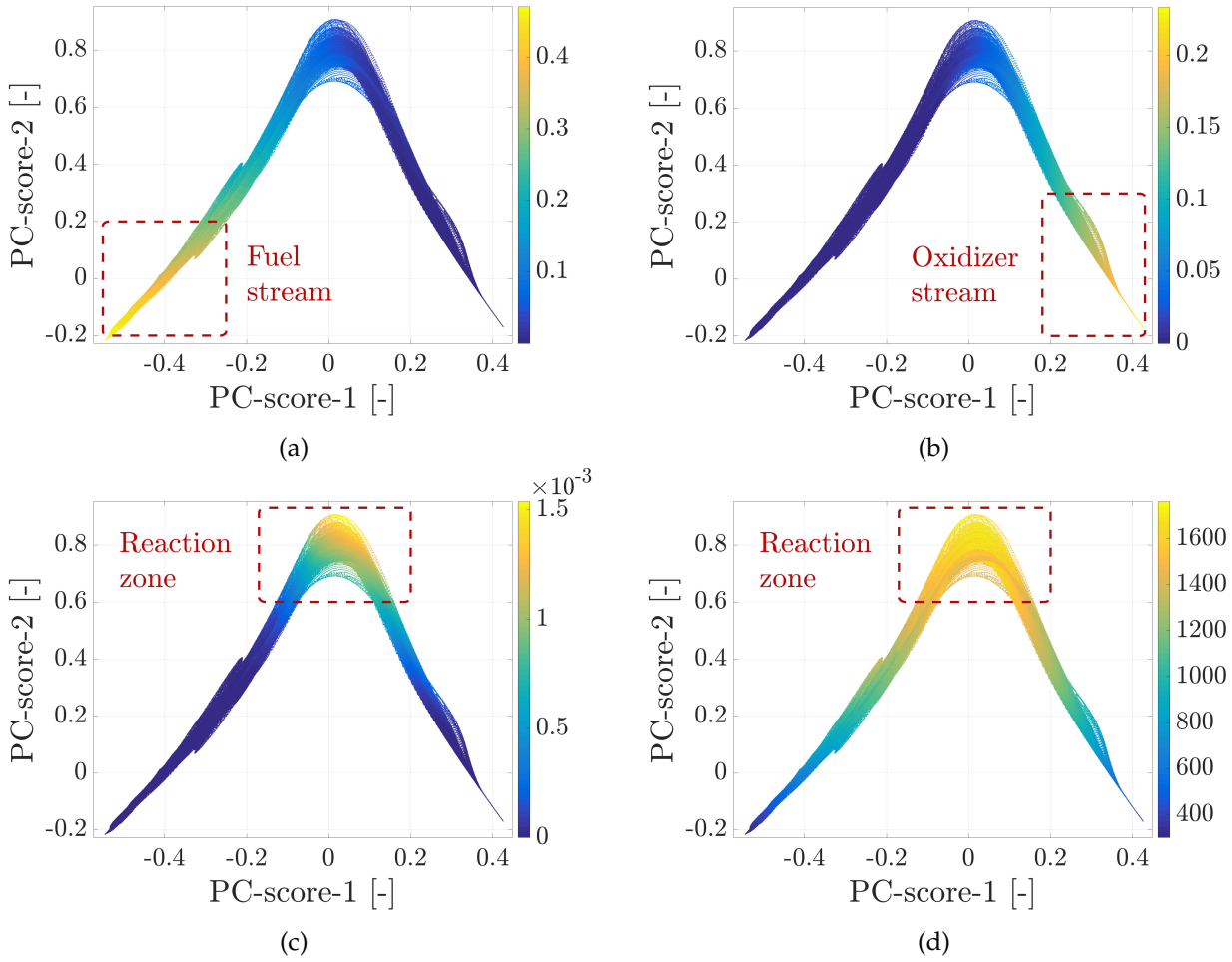
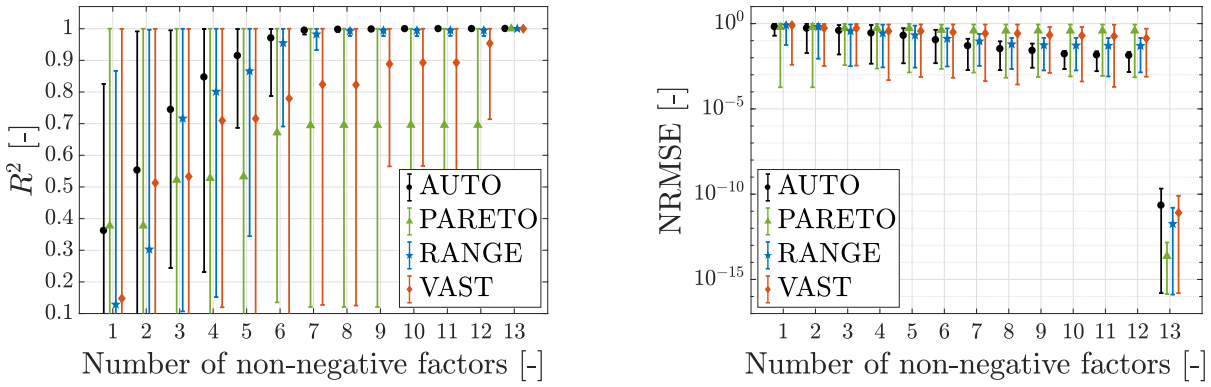


Figure 18: Manifold found by KPCA obtained with polynomial kernel of 3^{rd} order using 4 dimensions colored with (a) CO concetration profile (b) O_2 concetration profile (c) OH concentration profile (d) temperature profile. Time step: $2.50 \cdot 10^{-3}$ s of DNS1 with Range scaling.

4.5. Non-negative Matrix Factorization (NMF)

For NMF, similarly to PCA, a critical point is the assessment of the number of factors to retain. This quantity depends, as in PCA, on the physics of the problem and on the scaling criteria used for the data set. In Figure 19a and Figure 19b the statistics for the R^2 and the NRMSE are reported for a different number of retained factors and several scaling criteria. In this case, only Auto and Range seemed to be able to reconstruct the original data with acceptable accuracy, while Vast and Pareto scalings were characterized by large errors for every number of retained factors. Thus, NMF was performed with a lower-rank approximation of six factors. This value allowed for an acceptable recovery of the original data.



(a) (b)

Figure 19: (a) Average coefficient of determination R^2 with maximum and minimum values as error bars and (b) average NRMSE when recovering the original data with a given number of non-negative factors, reported for different scaling criteria. DNS1, time step: $2.50 \cdot 10^{-3}$ s.

As was done with PCA in section 4, the analysis of the weights on the non-negative factors was carried out. In case of Auto scaling, the first two non-negative factors appeared to identify the different reactants of the system: the highest weights on the first one corresponded to the oxidizer species O_2 and N_2, while on the second factor to the fuel H_2 and CO. On the remaining factors, other features that were identified appeared similar to the ones extracted by means of PCA. On the 4^{th} non-negative factor the highest weights corresponded to the main radicals and the fully oxidized species H_2O and CO_2, on the 5^{th} non-negative factor they corresponded to the hydroperoxil radical and the hydrogen peroxide and on the 6^{th} and last mode to HCO.

If compared to PCA on equal terms of number of modes, NMF seemed to be able to extract less information because of the constraint of the non-negative weights. The first PC alone was indeed characterized by the same amount of information (positive weights for oxidizer, negative weights for the fuel) as the first two non negative factors. NMF was extracting more information, if compared to PCA, only if data were previously normalized using the Pareto scaling criterion.

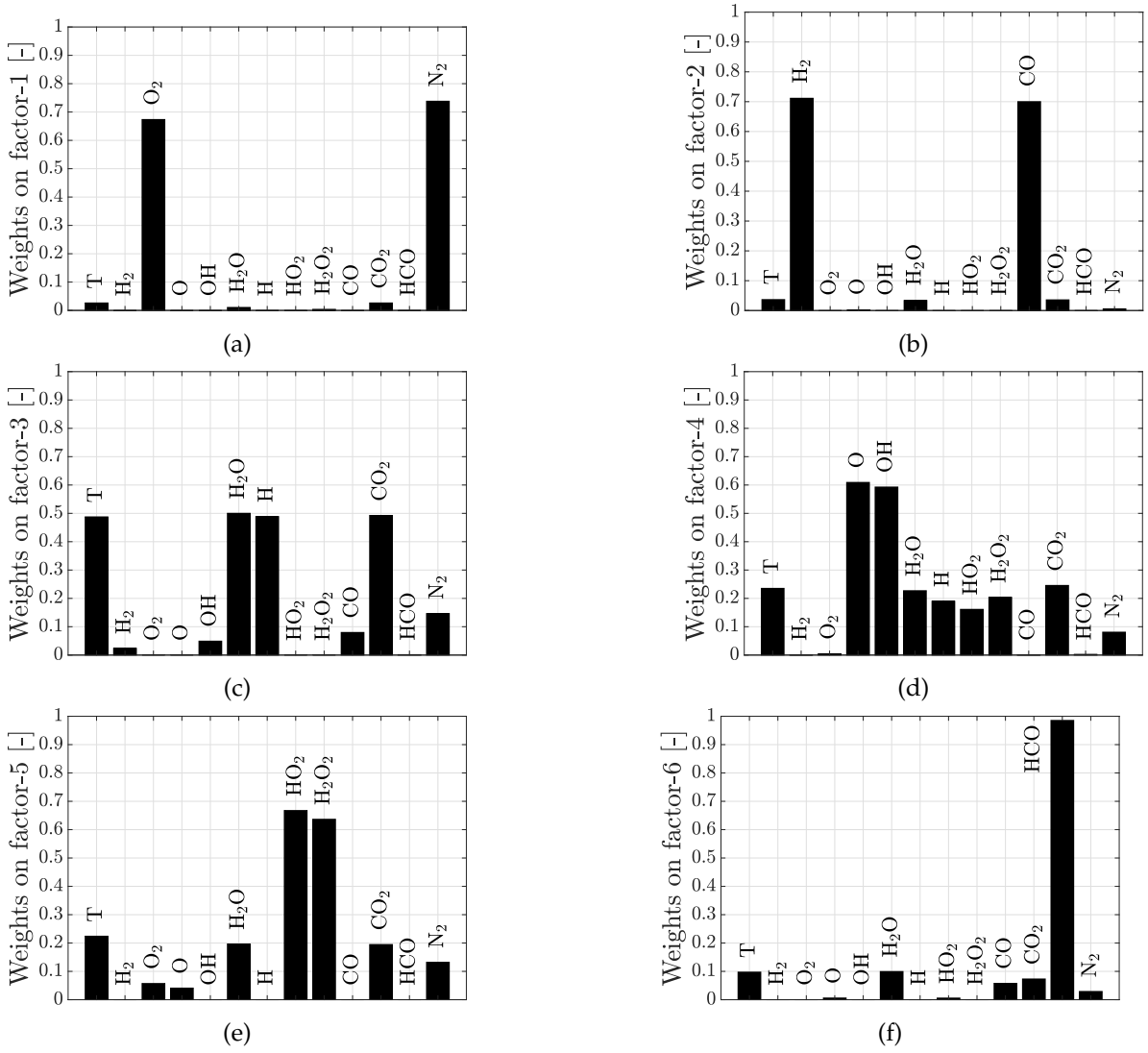


Figure 20: Non-negative factors determined by NMF with lower-rank 6 and with auto-scaling on the time step $2.5 \cdot 10^{-3}$ s.

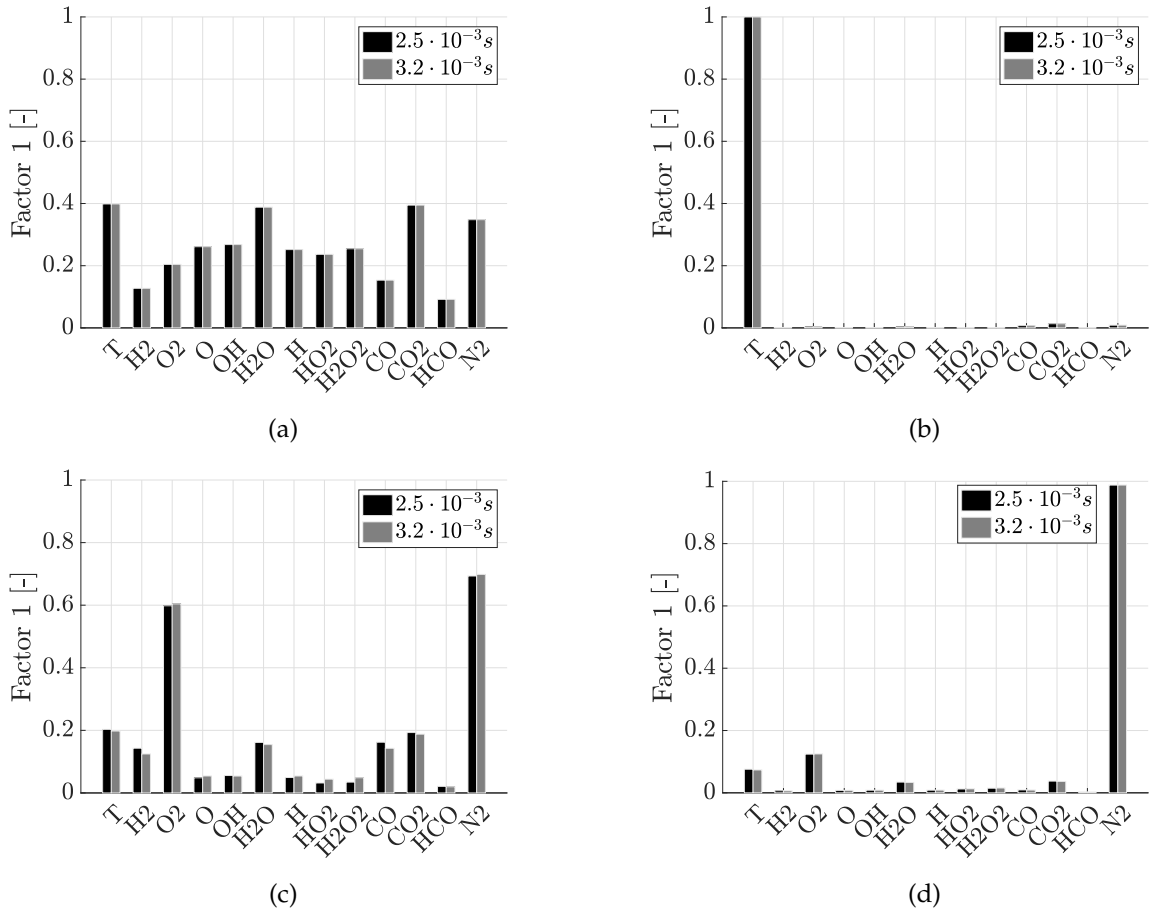


Figure 21: Non-negative factors determined by NMF ($q = 1$) with various scaling methods on the time steps $t = 2.50 \cdot 10^{-3}$ s and $t = 3.20 \cdot 10^{-3}$ s of DNS1. (a) Auto; (b) Pareto; (c) Range; (d) Vast.

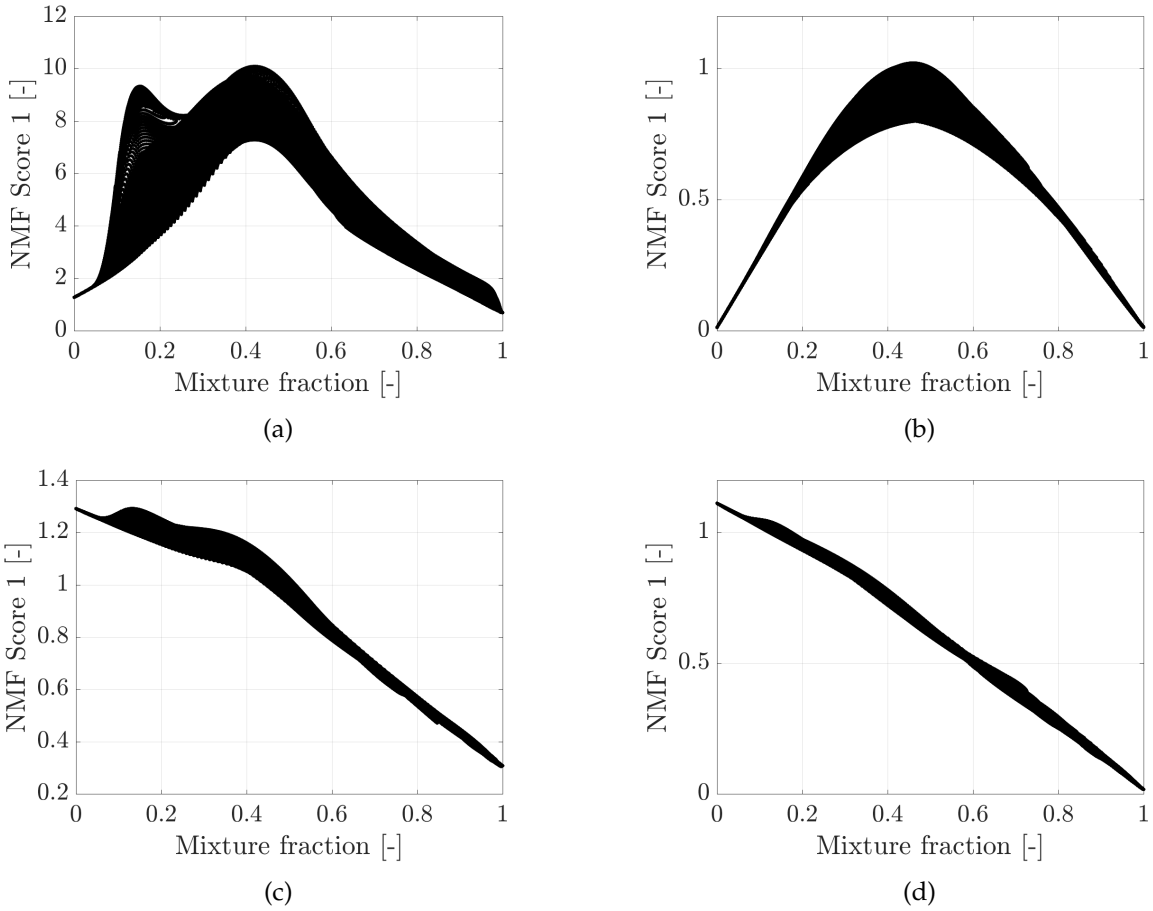


Figure 22: Mixture fraction against non-negative factor determined by NMF ($q = 1$) with various scaling methods on the time steps $t = 2.50 \cdot 10^{-3}$ s of DNS1. (a) Auto; (b) Pareto; (c) Range; (d) Vast.

Figure 21 reports the non-negative factor found by NMF for $q = 1$, for four different scaling criteria, for two time steps of DNS1. No relevant changes in the lower-dimensional manifold determined by NMF were detected, similarly to PCA (Figure 2). Mixture fraction is plotted against the corresponding NMF score for time step $t = 2.50 \cdot 10^{-3}$ s of DNS1 in Figure 22. As visible in Figure 22c and 22d, when the approximation order $q = 1$, the NMF score is representative of the mixture fraction for the Range and Vast scaling criterion. However, if NMF is run for $q > 1$, the NMF factors reported in Figure 21 are not found anymore. For example, the NMF factor found by NMF with $q = 1$ of Figure 21 is not present among the NMF factors found by NMF with $q = 6$ reported in Figure 20. This makes NMF a less straightforward problem to be solved in comparison to PCA, where the PCs are found all together and then only the first q are kept. For NMF, the choice of q is also affected by the factors that are actually found by the method.

4.6. Autoencoders

The use of Autoencoders was investigated as a nonlinear technique for feature extraction purposes. A deep architecture with five hidden layers was adopted to reduce the dimensionality from 13 to 4 and 2, respectively. Five activation functions were employed: linear, sigmoidal, hyperbolic tangent, Scaled Exponential Linear Unit (SELU) [21] and Exponential Linear Unit (ELU) [22].

The performance in terms of reconstruction of the original variables was compared for different activation functions and with a fixed architecture. In Figure 23 the comparison of the performances with the different activations is reported. Except for linear and sigmoid functions which had, as expected, error values consistently above all the others, the remaining three showed similar performances. The ELU activation function attained the lowest NRMSE values for all reconstructed variables.

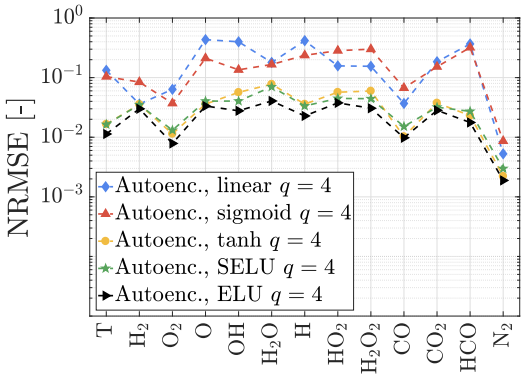


Figure 23: NRMSE for the reconstruction of the DNS1 data set after low-rank approximation achieved by Autoencoder with 4 layers, 4 dimensions and various activation functions: linear, sigmoidal, hyperbolic tangent, SELU, ELU.

The possibility to perform a dimensionality reduction in a non-linear fashion had, with respect to global and local PCA, the main advantage to be characterized by a lower reconstruction error for the original variables. The comparison in terms of NRMSE between the original and the reconstructed variables for the three methods is reported in Figures 24a and 24b: The methods were compared on equal terms of q (up to 2 for Figure 24a and up to 4 for Figure 24b), respectively. In both cases, the Autoencoder was characterized by a lowest average NRMSE for the reconstruction of the original variables.

The possibility to have a good reconstruction with only 2 dimensions is a strong attribute for the use of Autoencoders, as it can also provide a reliable tool for high-dimensional data visualization. As shown in Figure 25, the Autoencoder manifold is representative of a reaction progress variable. In Figures 25a and 25b, the initial reactants CO and O_2 are separated on two opposite sides of the manifold. In Figures 25c and 25d the manifold is marked with the OH concentration profile and temperature showing the reaction zone which is localized near the center of the manifold. This is consistent with the flame physics of a non-premixed jet. Similarly, in Figure 26, the manifold obtained from the dimensionality reduction via Autoencoder is coloured with the LPCA partitioning using $k = 4$ clusters. The separation of reactants (the fuel jet and the air co-flow) and reaction layers on the Autoencoder manifold is consistent with the above analysis.

The low recontruction errors obtained by the Autoencoder with respect to PCA can be better ex-

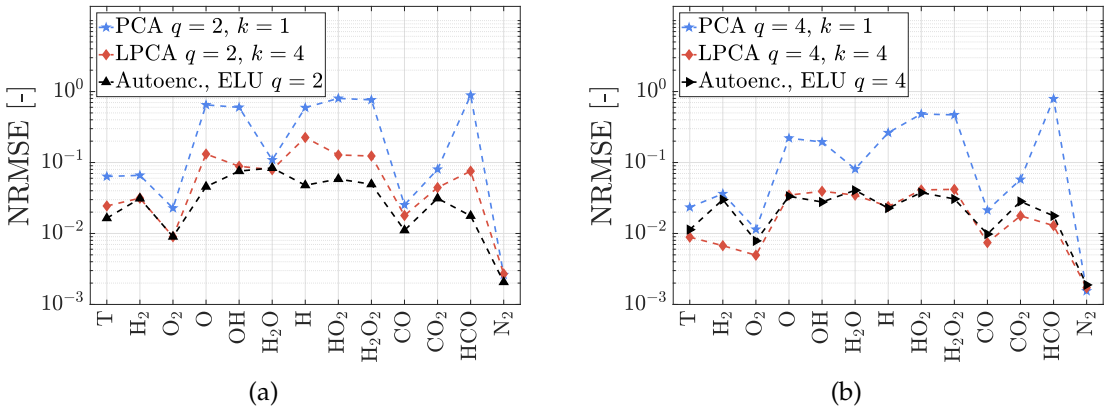


Figure 24: NRMSE for the reconstruction of the DNS1 data set after low-rank approximation achieved by Autoencoder with 5 layers, PCA and LPCA with 4 clusters for (a) $q = 2$ and (b) $q = 4$.

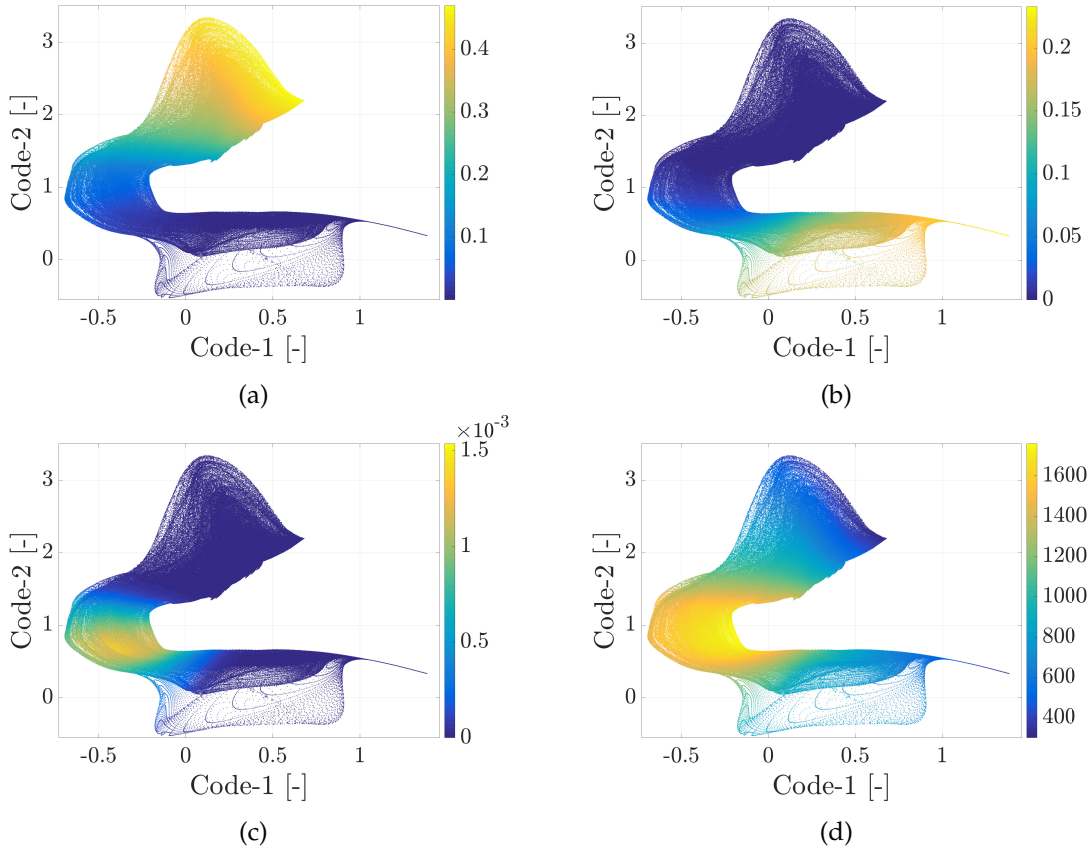


Figure 25: Manifold found by a 5-layers Autoencoder with $q = 2$ colored with (a) CO concentration profile; (b) O_2 concentration profile; (c) OH concentration profile; (d) temperature profile.

plained by comparing the low-order manifolds in case of $q = 2$ obtained by means of these two methods. In fact, as shown in Figure 27, even if the PCA manifold was also separating the reactants on the two opposite sides of the manifold as done by Autoencoder, discontinuity in OH mass fraction and temper-

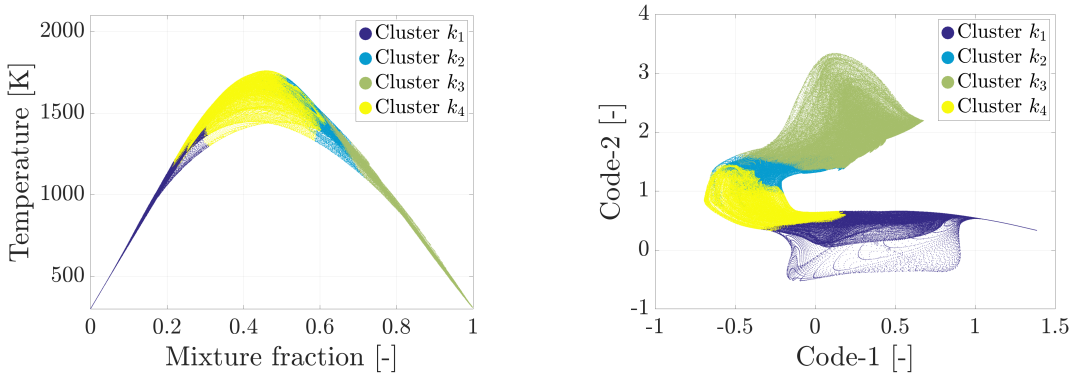


Figure 26: (a) Mixture fraction against temperature, colored with LPCA 4 clusters. (b) Manifold found by a 5-layers Autoencoder with $q = 2$ colored with the LPCA solution with $k = 4$ and 4 PCs.

ature values are present in the reaction zone, as its intrinsic linearity was not capable to follow the space curvature.

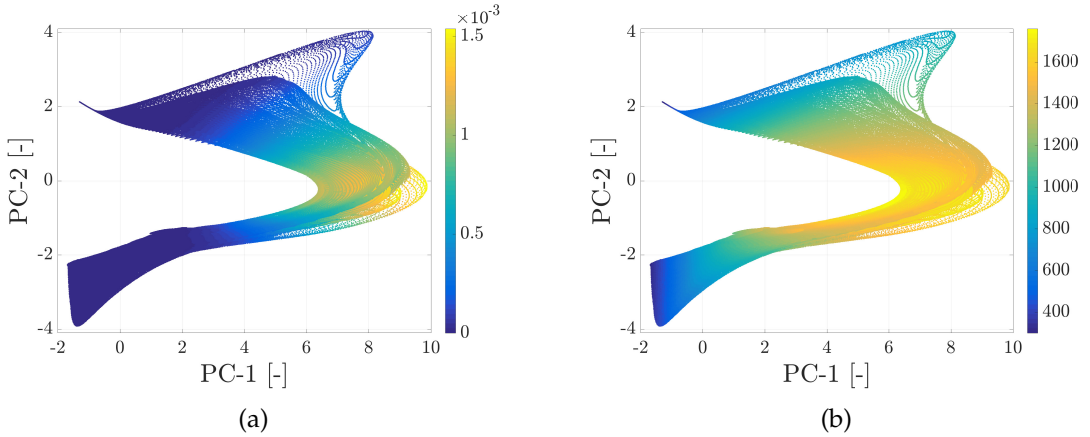
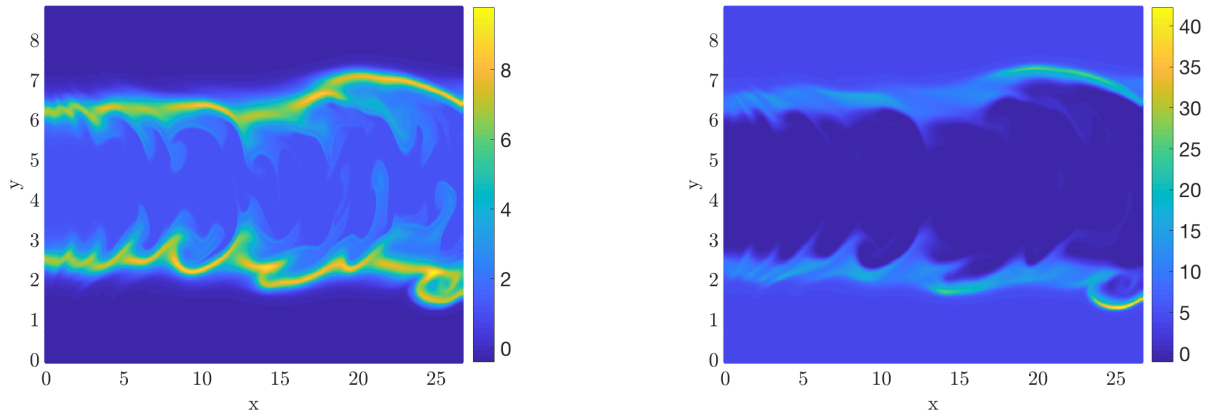


Figure 27: Manifold found by PCA with $q = 2$ when scaling with Auto, colored with (a) OH concentration profile; (b) temperature profile.

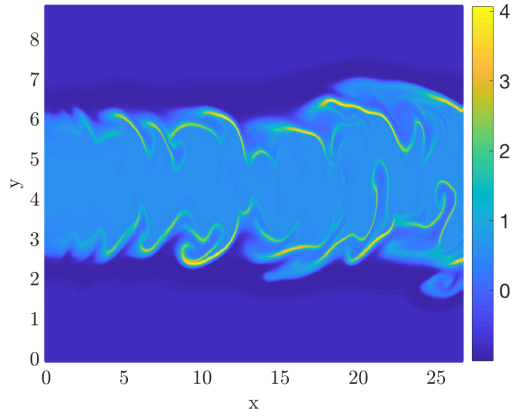
Differently from PCA where the (linear) mapping from the original state-space to the PC-space is explicitly available in the matrix $\mathbf{A}$, the relation between the original variables and the reduced set of scalars is not explicit with the Autoencoders. The codes shown in Figure 28 (Autoencoder with 5 hidden layers, ELU as activation function, $q = 4$), were correlated with the original variables, and the maximum values found for the correlations are reported in Table 5. For all the variables it was possible to find high correlations with the one of the 4 codes found by the Autoencoder. In particular, the main reactants (i.e. H_2, O_2, CO, N_2 and also HCO) were highly correlated with the third code, while the fully oxidized species (i.e. CO_2 and H_2O) and the more reactive radicals (i.e. O, OH, H) with the first one. Other radicals such as HO_2 and H_2O_2 were highly correlated with the second code. Anyway, the impossibility to explicitly access the mapping from the original to the reduced space constitutes a

limit for the Autoencoder for feature extraction purposes. In this case, indeed, not a single variable was correlated with the fourth code and also, in contrast to PCA, in this case it was not possible to correlate the mixture fraction at a significant level with any of the codes.

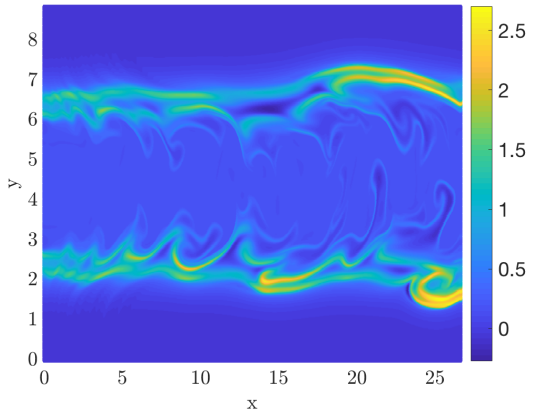


(a) Code 1

(b) Code 2



(c) Code 3



(d) Code 4

Figure 28: Codes found by Autoencoder with 5 layers and $q = 4$, with the ELU activation function.

Variable	T	H_2	O_2	O	OH	H_2O	H	HO_2	H_2O_2	CO	CO_2	HCO	N_2
Code	1	3	3	1	1	1	1	2	2	3	1	3	3
Corr. (%)	87.9	86.7	79.8	71.6	75.2	90.1	82.7	83.21	85.7	90.4	88.3	67.7	88.6

Table 5: Maximum correlations of each variable with one of the Autoencoder codes.

Moreover, by increasing the number of clusters, it was possible for LPCA to outperform the Autoencoder in terms of reconstruction error for all the variables on equal terms of q, as shown in Figure 29 where the NRMSE for the reconstruction of the DNS1 data set is compared for the 5 layers Autoencoder and LPCA with 128 clusters after compressing the data to 2 dimensions. Because LPCA is characterized by only two hyper-parameters (i.e. number of PCs and number of clusters), it is straightforward to un-

derstand how to improve a given LPCA set-up. On the contrary, the Autoencoder's performance are not as straightforward to be improved because of the higher number of hyper-parameters (i.e.: number of layers, number of neurons per layer, activation function, batch size) to be chosen. Anyway, partitioning large data sets with a high number of clusters could not always be feasible as the iterative algorithm could be very computationally expensive. Figure 30 reports the Auto-encode code 1 against code 2

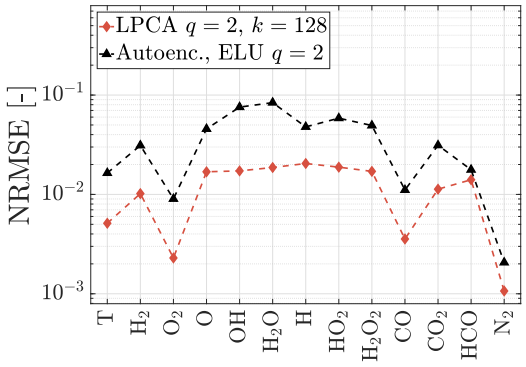


Figure 29: NRMSE for the reconstruction of the DNS1 data set after low-rank approximation achieved by Autoencoder with 5 layers and LPCA with 128 clusters for $q = 2$.

(black crosses), and LPCA score 1 against score 2 (red dots), for 4 different values of the number of clusters k (4, 16, 64, 128). Procrustes analysis offers a way to compare the Autoencoder and LPCA solution. In each cluster, a Procrustes analysis was performed in order to best conform the values of the LPCA scores of each cluster to the corresponding values of the Autoencoder codes. A given matrix $\mathbf{Z}_k$ (matrix of LPCA scores from one cluster) is transformed in order to conform to the values of a target matrix, $\mathbf{H}_k$ (matrix of corresponding values of the Autoencoder codes). The goodness-of-fit criterion is the sum of squared errors between $\mathbf{H}_k$ and the transformed $\mathbf{Z}_k$. The transformation is as follows: $\mathbf{Z}'_k = \mathbf{b} \cdot \mathbf{Z}_k \cdot \mathbf{T} + \mathbf{c}$; where $\mathbf{b}$ is a scaling component, $\mathbf{T}$ is a rotation component and $\mathbf{c}$ is a translation component. Interestingly, the manifolds found by Autoencoder and LPCA overlap, even more so for high values of k, indicating that the methods both converged to very similar solutions, which is also confirmed by Figure 31, where the parity plots of one Autoencoder code and one LPCA score, from 2 clusters, are reported, from the LPCA solution with $k = 4$ and $q = 2$.

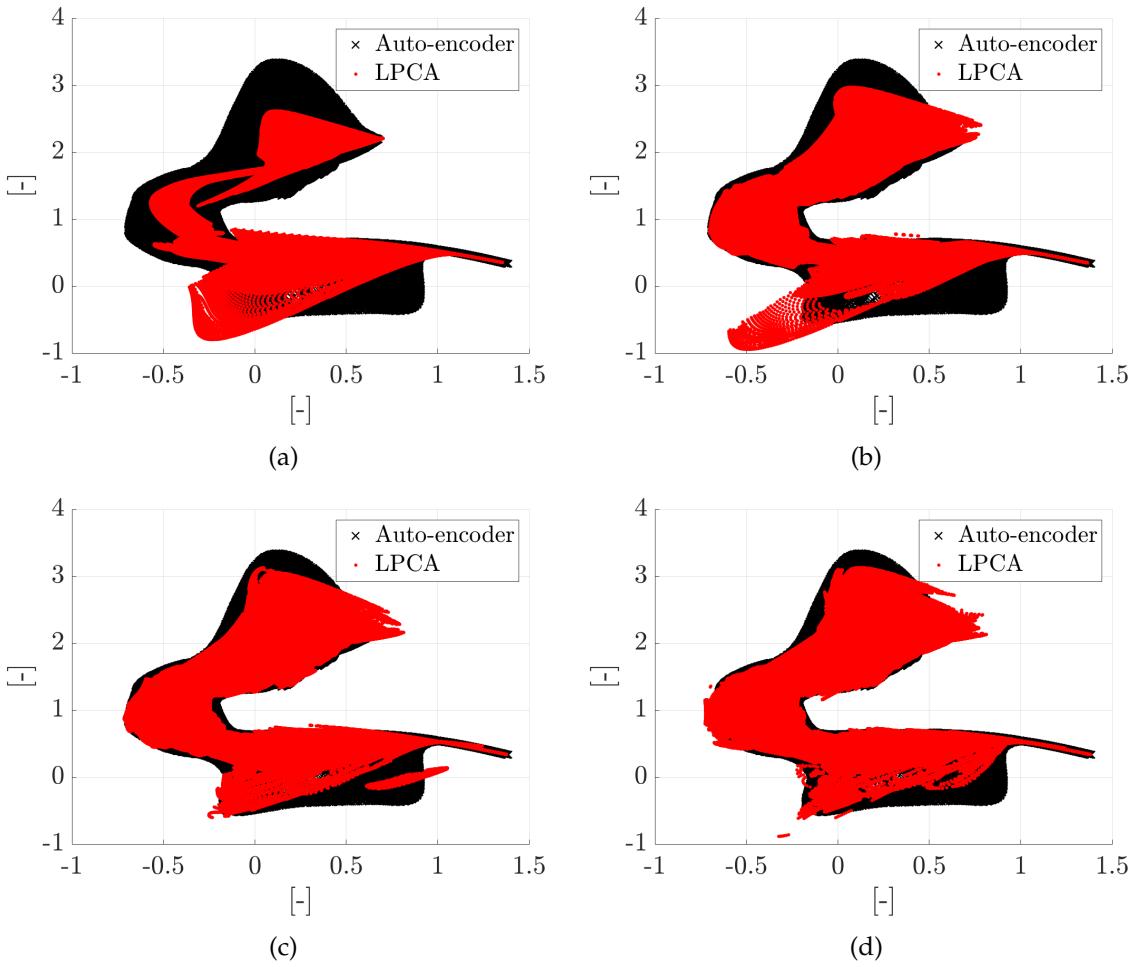


Figure 30: Procrustes analysis offers a way to compare the Autoencoder and LPCA solution. Autoencoder code 1 against Autoencoder code 2 (black crosses); score 1 against score 2 (red dots) from LPCA with $q = 2$ and (a) $k = 4$, (b) $k = 16$, (c) $k = 64$, (d) $k = 128$.

4.7. Feature-assisted clustering

From the analysis of the PCA and NMF modes extracted for DNS1 at time step $t = 2.5 \cdot 10^{-3}$ s, seven groups of recurrent features were isolated by the authors. Thus, 7 feature vectors were created, representative of 7 different chemical or physical phenomena that are reported in Table 6, and were used to partition the data into 7 clusters, reported in Figure 32, as explained in section 3.7. The first two clusters represented the two groups of reactants, fuel and oxidizer, while other four clusters were devoted to the radicals and the last one to the fully oxidized products. Two separate clusters were dedicated to H and HCO, since for both of them there was one PC with almost unitary weight for these variables. Other two groups of variables were one for O and OH and another for HO_2 and H_2O_2, as both in PCA and NMF there were modes with high weights for these variables. The clusters found by the feature-assisted clustering procedure are reported in Figure 32, and a comparison between these clusters and the one found by LPCA (VQPCA unsupervised algorithm) is reported in Figure 33.

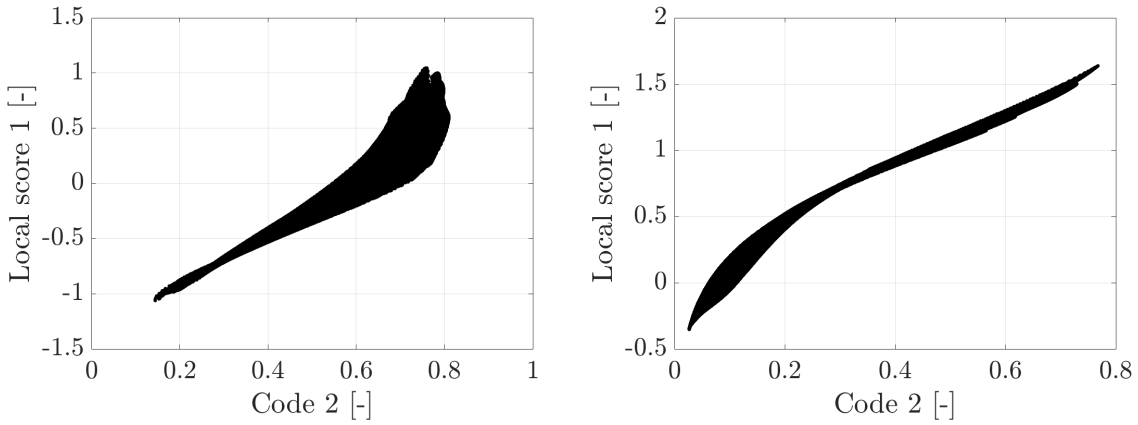


Figure 31: Parity plot of the Autoencoder score 2 vs. score 1 of LPCA with $k = 4$ in cluster (a) 2 and (b) 3. DNS1. $q = 2$.

Cluster	Features	Source
1	H_2, CO	First NMF factor
2	O_2, N_2	Second NMF factor
3	O, OH	Second LPCA mode in cluster 4, fourth NMF factor
4	H	First LPCA mode in cluster 4, second LPCA mode in cluster 2
5	HO_2, H_2O_2	Second LPCA mode in cluster 1, fifth NMF factor
6	HCO	HCO Second LPCA mode in cluster 3, sixth NMF factor
7	CO_2, H_2O	First LPCA mode in cluster 2 and 3, third NMF factor

Table 6: Main variables and features, previously extracted by means of PCA and NMF, for the cluster division reported in figure 32 imposed via artificial modes.

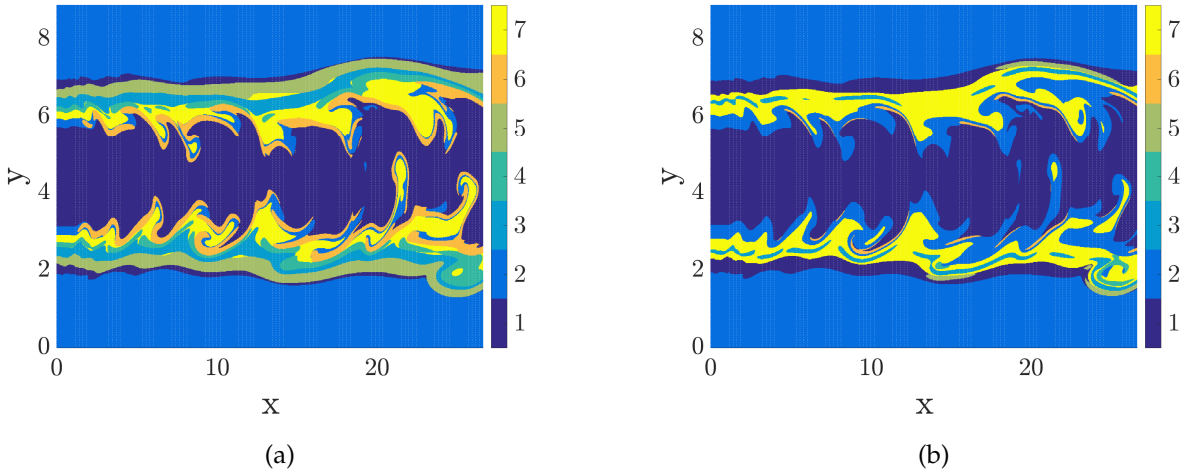


Figure 32: Clustering with artificial modes built using the features which were extracted by means of PCA and NMF. (a) Centering with Mean, scaling with Auto. (b) Centering with Mean, scaling with Range.

As visible in this figure, the clusters found by means of feature-assisted clustering were very similar

to the ones found by LPCA. In spite of this, these two methods are based on two completely different approaches. VQPCA is totally unsupervised: k groups are found with no prior knowledge on the data, from which the main features are extracted. Thus, it can be considered as a top-down approach: from the original data set, clusters are found by means of a mathematical criterion and subsequently the main physical and chemical features in each of them are retrieved. On the contrary, feature-assisted clustering is in fact a bottom-up approach: the main features of the system are already known (from previous feature extraction, as in this application, or from human expertise) and a cluster partition based on these features is achieved. It is remarkable that LPCA could find very similar clusters to the ones found by means of feature-assisted clustering, indicating that the clustering procedure implemented by VQPCA can converge to a clustering solution that is based on the physics of the problem even if no prior knowledge is available as input to the algorithm.

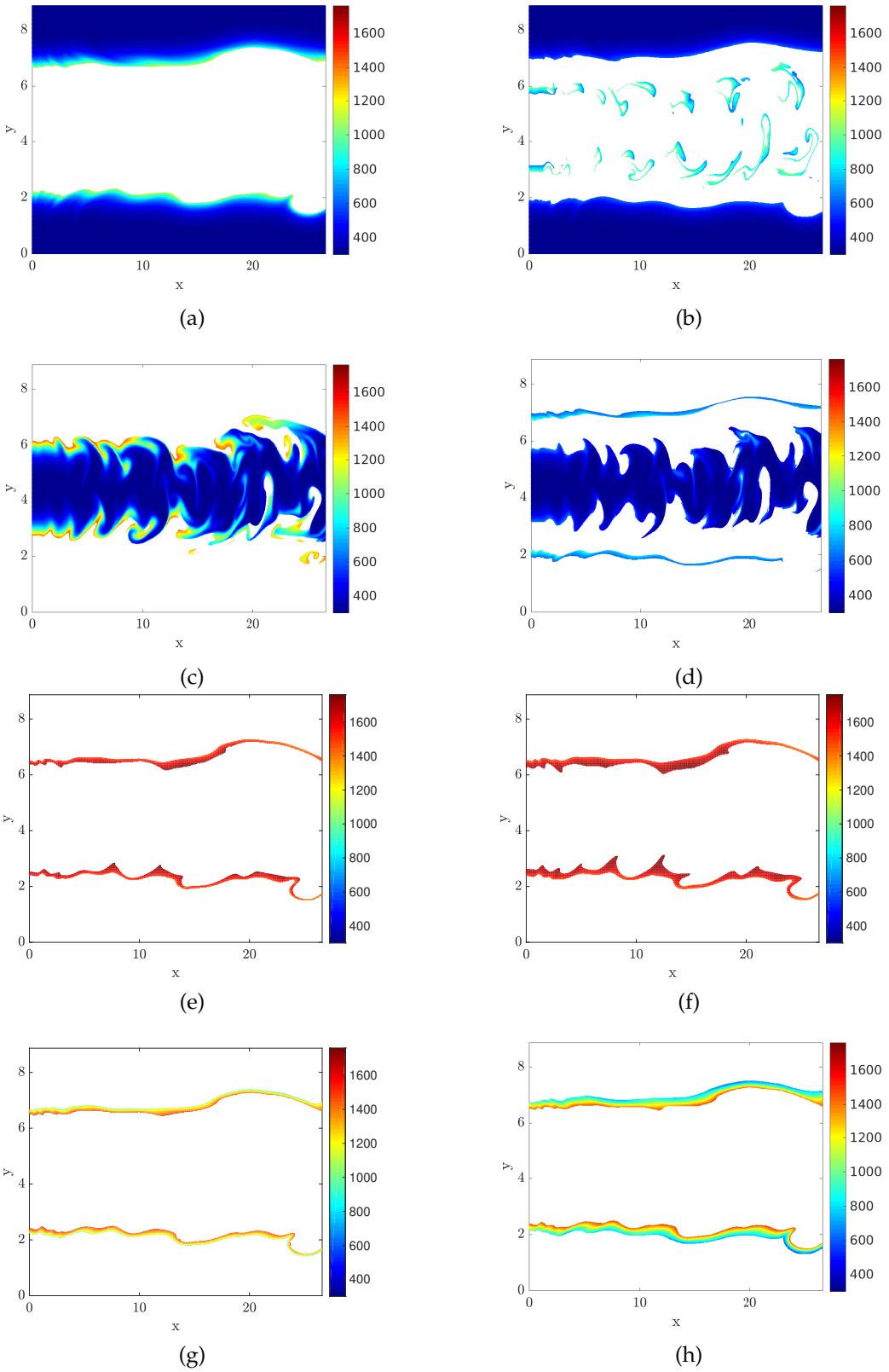


Figure 33: (a), (c), (e), (g): Clusters found with VQPCA with $k = 7$; (b), (d), (f), (h): Clusters found with feature-assisted clustering approach. Time step $2.5 \cdot 10^{-3}$ s of DNS1.

5. Conclusions and discussion

The present work investigated the application of data-driven techniques for the detection of features from turbulent combustion data sets (direct numerical simulation). Two H_2/CO flames were analyzed: a spatially-evolving (DNS1) and a temporally-evolving jet (DNS2). Methods such as Principal Component Analysis (PCA), Local Principal Component Analysis (LPCA), Non-negative Matrix Factorization (NMF) and Autoencoders were explored for this purpose. It was shown that various factors could affect the performance of these methods, such as the criteria employed for the centering and the scaling of the original data or the choice of the number of dimensions in the low-rank approximations. A set of guidelines was presented in this paper that can aid the process of identifying meaningful physical features from turbulent reactive flows data.

PCA-based techniques were firstly explored, such as Global PCA, Local PCA and Kernel PCA. The effect of the various scaling methods on the feature extraction capability of PCA was assessed. Local PCA was shown to be able to automatically detect the most important features of the data-set. As the mixture fraction was known to be an important feature for DNS1, it was remarkable to see that Local PCA could partition the data into zones of low and high values of mixture fraction even if this variable was not present in the data analyzed by Local PCA. Besides, the locally-linear nature of Local PCA , in contrast with the globally linear nature of PCA, meant that the method could better approximate the original data manifold, confirmed by the lower reconstruction errors obtained for a given number of retained principal components, and thus it could detect features that only had a local importance in the original data manifold, that were not detected by global PCA. The interpretability of the factors extracted with the Non-negative Matrix Factorization (NMF) technique showed to be a more challenging task. Due to the NMF positive constraint, some features, such as mixture fraction, could not be represented by a single mode, while this was possible with PCA. Non-linear techniques such as Kernel PCA and Autoencoders were also tested for feature extraction purposes. Kernel PCA could detect features such as mixture fraction or the reactive layer. The Autoencoder was used to find a lower two-dimensional manifold in the original data space, which proved to be a reliable choice to separate well the reactants on opposing ends of the manifold and the reactive layer in-between, indicating that such manifold could be linked to a reaction progress variable. Besides, the manifold detected by a locally-linear method such Local PCA was shown to converge to the one found by a non-linear method such as the Autoencoder.

In parallel with detecting features, the potential of the mentioned techniques for data compression purposes, thus the quality of reconstruction of the original data with a fewer number of dimensions, was also assessed, as achieving both tasks well is a decisive factor for selecting a given data-driven method for accurate modeling of combustion systems. It was shown that Local PCA was capable of achieving excellent reconstruction with $q = 2$ principal components when the number of clusters was sufficiently high. Autoencoders performed comparably well when used with the ELU activation function. The reconstruction errors associated to these two methods were lower for Local PCA only when a high (128) number of clusters was used, showing the ability of the method to more easily approximate non-linear manifolds in the original data space as, to do so, the user would just need to change, and more specifically further increase, one free parameter, while decreasing the reconstruction errors of an Autoencoder

would require the investigation of a wider range of design choices. However, it is true that using a very high number of clusters had some drawbacks: data interpretability became more complex as the number of features to analyze increased (e.g., 2 features in each one of the 128 clusters); a high number of matrices, one for each cluster, containing the local PCA modes had to be stored.

A new way to cluster data-sets based on the possibility to a priori decide what features to find in it was presented and named Feature-assisted Clustering.

Future work will include the application of different data-driven methods in the feature detection approach as well as generalization of the findings presented in this paper to a broader range of turbulent combustion data sets.

Acknowledgements

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.

References

[1] J. C. Sutherland and A. Parente, "Combustion modeling using principal component analysis," *Proceedings of the Combustion Institute*, vol. 32 I, no. 1, pp. 1563–1570, 2009.

[2] A. Parente, J. C. Sutherland, L. Tognotti, and P. J. Smith, "Identification of low-dimensional manifolds in turbulent flames," *Proceedings of the Combustion Institute*, vol. 32 I, no. 1, pp. 1579–1586, 2009.

[3] G. Aversano, A. Bellemans, Z. Li, A. Coussement, O. Gicquel, and A. Parente, "Application of reduced-order models based on PCA & Kriging for the development of digital twins of reacting flow applications," *Computers and Chemical Engineering*, vol. 121, pp. 422–441, 2019.

[4] F. Richecoeur, L. Hakim, A. Renaud, L. Zimmer, F. Richecoeur, L. Hakim, A. Renaud, L. Zimmer, B. F. Richecoeur, L. Hakim, and A. Renaud, "DMD algorithms for experimental data processing in combustion To cite this version : HAL Id : hal-00825509 DMD algorithms for experimental data processing in combustion," 2013.

[5] T. Grenga, J. F. Macart, M. E. Mueller, T. Grenga, J. F. Macart, and M. E. Mueller, "Dynamic mode decomposition of a direct numerical simulation of a turbulent premixed planar jet flame : convergence of the modes," vol. 7830, no. May, 2018.

[6] H. Mirgolbabaei, T. Echekki, and N. Smaoui, "A nonlinear principal component analysis approach for turbulent combustion composition space," *International Journal of Hydrogen Energy*, vol. 39, no. 9, pp. 4622–4633, 2014.

[7] A. G. Landge, V. Pascucci, A. Gyulassy, J. C. Bennett, H. Kolla, J. Chen, and P.-T. Bremer, "In-situ feature extraction of large scale combustion simulations using segmented merge trees," in *SC'14:*

Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, pp. 1020–1031, IEEE, 2014.

[8] N. Kambhatla and T. Leen, "Dimension reduction by local principal component analysis," *Neural Computation*, vol. 9, no. 7, pp. 1493–1516, 1997.

[9] H. R. R. A. Yetter, F. L. Dryer, "A Comprehensive Reaction Mechanism For Carbon Monoxide/Hydrogen/Oxygen Kinetics," *Combustion Science and Technology*, no. January 2011, pp. 37–41, 1991.

[10] J. C. Sutherland, P. J. Smith, and J. H. Chen, "A quantitative method for a priori evaluation of combustion reaction models," *Combustion Theory and Modelling*, vol. 11, no. 2, pp. 287–303, 2007.

[11] I. T. Jolliffe, "Principal Component Analysis, Second Edition," 2002.

[12] A. Bellemans, G. Aversano, A. Coussement, and A. Parente, "Feature extraction from reduced order models based on principal component analysis," *Computers & Chemical Engineering*, no. March, 2018.

[13] H. F. Kaiser, "The varimax criterion for analytic rotation in factor analysis," *Psychometrika*, vol. 23, no. 3, pp. 187–200, 1958.

[14] M. B. Richman, "Rotation of Principle Components," *Journal of Climatology*, vol. 6, pp. 293–335, 1986.

[15] Q. Wang, "Kernel principal component analysis and its applications in face recognition and active shape models," *arXiv preprint arXiv:1207.3538*, 2012.

[16] B. Schölkopf, A. Smola, and K.-R. Müller, "Kernel principal component analysis," in *International conference on artificial neural networks*, pp. 583–588, Springer, 1997.

[17] P. D. E. P.-g. Em and S. E. Cultura, "Socorro De Fátima Moraes Nina Trabalho , Ambiente E Saúde : Cotidiano Dos Socorro De Fátima Moraes Nina Trabalho , Ambiente E Saúde : Cotidiano Dos," vol. 401, no. October 1999, pp. 788–791, 2014.

[18] D. Cai, X. He, and J. Han, "Locally consistent concept factorization for document clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 6, pp. 902–913, 2011.

[19] P. TrendIC; Tecnologia, "Establishing a Quality Framework for ITIL An Overview of Six Sigma and its Role in ITIL," no. 1, 2002.

[20] C. Ding, X. He, and H. D. Simon, "On the Equivalence of Nonnegative Matrix Factorization and Spectral Clustering," *Proc. SIAM International Conference on Data Mining*, pp. pp. 606–610, 2005.

[21] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, "Self-normalizing neural networks," in *Advances in neural information processing systems*, pp. 971–980, 2017.

[22] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," *arXiv preprint arXiv:1511.07289*, 2015.

BIBLIOGRAPHY

[1] Gianmarco Aversano, Aurélie Bellemans, Zhiyi Li, Axel Coussement, Olivier Gicquel, and Alessandro Parente. « Application of reduced-order models based on PCA & Kriging for the development of digital twins of reacting flow applications. » In: *Computers and Chemical Engineering* 121 (2019), pp. 422–441. ISSN: 0098-1354. DOI: 10.1016/j.compchemeng.2018.09.022.

[2] Aurélie Bellemans, Gianmarco Aversano, Axel Coussement, and Alessandro Parente. « Feature extraction and reduced-order modelling of nitrogen plasma models using principal component analysis. » In: 115 (2018), pp. 504–514. DOI: 10.1016/j.compchemeng.2018.05.012.

[3] Chris M Bishop. « Training with Noise is Equivalent to Tikhonov Regularization. » In: *Neural Computation* 116 (1995), pp. 108–116.

[4] Christopher M Bishop. *Pattern Recognition and Machine Learning*. Vol. 53. 9. 2013, pp. 1689–1699. ISBN: 978-0-387-31073-2. DOI: 10.1117/1.2819119. arXiv: arXiv:1011.1669v3.

[5] K. Bizon, G. Continillo, E. Mancaruso, S. S. Merola, and B. M. Vaglieco. « POD-based analysis of combustion images in optically accessible engines. » In: *Combustion and Flame* 157.4 (2010), pp. 632–640. ISSN: 00102180. DOI: 10.1016/j.combustflame.2009.12.013. URL: http://dx.doi.org/10.1016/j.combustflame.2009.12.013.

[6] Katarzyna Bizon and Gaetano Continillo. « Reduced order modelling of chemical reactors with recycle by means of POD-penalty method. » In: *Computers and Chemical Engineering* 39 (2012), pp. 22–32. ISSN: 0098-1354. DOI: 10.1016/j.compchemeng.2011.10.001. URL: http://dx.doi.org/10.1016/j.compchemeng.2011.10.001.

[7] Katarzyna Bizon, Gaetano Continillo, Marek Berezowski, and Joanna Smuła-Ostaszewska. « Optimal model reduction by empirical spectral methods via sampling of chaotic orbits. » In: *Physica D: Nonlinear Phenomena* 241.17 (2012), pp. 1441–1449. ISSN: 01672789. DOI: 10.1016/j.physd.2012.05.004. URL: http://dx.doi.org/10.1016/j.physd.2012.05.004.

[8] Paul G. Constantine, Eric Dow, and Qiqi Wang. « Active Subspace Methods in Theory and Practice. » In: *SIAM Journal of Scientific Computation* 36.4 (2014), pp. 1500–1524.

[9] Axel Coussement, Benjamin J. Isaac, Olivier Gicquel, and Alessandro Parente. « Assessment of different chemistry reduction methods based on principal component analysis: Comparison of the MG-PCA and score-PCA approaches. » In: *Combustion and Flame* 168 (2016), pp. 83–97. ISSN: 00102180. DOI: 10.1016/j.combustflame.2016.03.021.

[10] Thierry Crestaux, Olivier Le Maître, and Jean Marc Martinez. « Polynomial chaos expansion for sensitivity analysis. » In: *Reliability Engineering and System Safety* 94.7 (2009), pp. 1161–1172. ISSN: 09518320. DOI: 10.1016/j.ress.2008.10.008.

[11] María Teresa Daza, Juan José Orteils, and Carmen Noguera. « Negative semantic priming from consciously vs. unconsciously perceived single words. » In: *Psicologica* 28.2 (2007), pp. 105–127. ISSN: 02112159. DOI: 10.1016/j.csda.2006.11.006.

[12] Tarek Echekki and Hessam Mirgolbabaei. « Principal component transport in turbulent combustion: A posteriori analysis. » In: *Combustion and Flame* 162.5 (2015), pp. 1919–1933. ISSN: 00102180. DOI: 10.1016/j.combustflame.2014.12.011.

[13] Magnus Fürst, Pino Sabia, Marco Lubrano Lavadera, Gianmarco Aversano, Mara de Joannon, Alessio Frassoldati, and Alessandro Parente. « Optimization of Chemical Kinetics for Methane and Biomass Pyrolysis Products in Moderate or Intense Low-Oxygen Dilution Combustion. » In: *Energy & Fuels* (2018), acs.energyfuels.8b01022. ISSN: 0887-0624. DOI: 10.1021/acs.energyfuels.8b01022.

[14] E. H. Glaessgen, Damage Tolerance Branch, D. S. Stargel, and Material Sciences. « The Digital Twin Paradigm for Future NASA and U . S . Air Force Vehicles. » In: (2019), pp. 1–14.

[15] Temistocle Grenga, Jonathan F Macart, Michael E Mueller, Temistocle Grenga, Jonathan F Macart, and Michael E Mueller. « Dynamic mode decomposition of a direct numerical simulation of a turbulent premixed planar jet flame : convergence of the modes. » In: 7830.May (2018). DOI: 10.1080/13647830.2018.1457799.

[16] M Guenot, I Lepot, C Sainvitu, J Goblet, and R F Coelho. « Adaptive sampling strategies for non-intrusive POD-based surrogates. » In: *Engineering Computations* 30.4 (2013), pp. 521–547. ISSN: 02644401. DOI: Doi10.1108/02644401311329352.

[17] Sebastian Haag and Reiner Anderl. « Digital twin - Proof of concept. » In: *Manufacturing Letters* (2018), pp. 10–12. ISSN: 2213-8463. DOI: 10.1016/j.mfglet.2018.02.006. URL: https://doi.org/10.1016/j.mfglet.2018.02.006.

[18] Benjamin J. Isaac, Axel Coussement, Olivier Gicquel, Philip J. Smith, and Alessandro Parente. « Reduced-order PCA models for chemical reacting flows. » In: *Combustion and Flame* 161.11 (2014), pp. 2785–2800. ISSN: 15562921. DOI: 10.1016/j.combustflame.2014.05.011.

[19] Benjamin J. Isaac, Jeremy N. Thornock, James Sutherland, Philip J. Smith, and Alessandro Parente. « Advanced regression methods for combustion modelling using principal components. » In: *Combustion and Flame* 162.6 (2015), pp. 2592–2601. ISSN: 15562921. DOI: 10.1016/j.combustflame.2015.03.008.

[20] I T. Jolliffe. « Principal Component Analysis, Second Edition. » In: (2002).

[21] Nandakishore Kambhatla and TK Leen. « Dimension reduction by local principal component analysis. » In: *Neural Computation* 9.7 (1997), pp. 1493–1516. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.7.1493.

[22] Maninder Jeet Kaur, Ved P Mishra, and Piyush Maheshwari. *The Convergence of Digital Twin, IoT, and Machine Learning: Transforming Data into Action*. Springer International Publishing, pp. 3–17. ISBN: 9783030187323. DOI: 10.1007/978-3-030-18732-3.

[23] Théa Lancien, Nicolas Dumont, Kevin Prieur, Daniel Durox, Sébastien Candel, Olivier Gicquel, and Ronan Vicquelin. « Uncertainty quantification of injected droplet size in mono-dispersed Eulerian simulations. » In: (2016).

[24] Régis Lebrun and Anne Dutfoy. « An innovating analysis of the Nataf transformation from the copula viewpoint. » In: *Probabilistic Engineering Mechanics* 24.3 (2009), pp. 312–320. ISSN: 0266-8920. DOI: 10.1016/j.probengmech.2008.08.001. URL: http://dx.doi.org/10.1016/j.probengmech.2008.08.001.

[25] Zhiyi Li, Alberto Cuoci, and Alessandro Parente. « Large Eddy Simulation of MILD combustion using finite rate chemistry : Effect of combustion sub-grid closure. » In: 37.x (2019), pp. 4519–4529. DOI: 10.1016/j.proci.2018.09.033.

[26] Guang Lin. « On the Bayesian calibration of expensive computer models with input dependent parameters. » In: *Spatial Statistics* (2017). ISSN: 22116753. DOI: 10.1016/j.spasta.2017.08.002. URL: http://dx.doi.org/10.1016/j.spasta.2017.08.002.

[27] Cheng-yuan Liou, Wei-chen Cheng, Jiun-wei Liou, and Daw-ran Liou. « Neurocomputing Autoencoder for words. » In: *Neurocomputing* 139 (2014), pp. 84–96. ISSN: 0925-2312. DOI: 10.1016/j.neucom.2013.09.055.

[28] Zheng Liu, Norbert Meyendorf, and Nezih Mrad. « The role of data fusion in predictive maintenance using digital twin The Role of Data Fusion in Predictive Maintenance Using Digital Twin. » In: 020023.April 2018 (2019).

[29] Søren N Lophaven, Jacob Søndergaard, and Hans Bruun Nielsen. « Kriging Toolbox. » In: (2002), pp. 1–28.

[30] Hessam Mirgolbabaei, Tarek Echekki, and Nejib Smaoui. « A nonlinear principal component analysis approach for turbulent combustion composition space. » In: *International Journal of Hydrogen Energy* 39.9 (2014), pp. 4622–4633. ISSN: 03603199. DOI: 10.1016/j.ijhydene.2013.12.195. URL: http://dx.doi.org/10.1016/j.ijhydene.2013.12.195.

[31] Juliane Müller, Christine A. Shoemaker, and Robert Piché. « SO-MI: A surrogate model algorithm for computationally expensive nonlinear mixed-integer black-box global optimization problems. » In: *Computers and Operations Research* 40.5 (2013), pp. 1383–1400. ISSN: 03050548. DOI: 10.1016/j.cor.2012.08.022.

[32] Pentti Paatero and Unto Tapper. « Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. » In: *Environmetrics* 5.2 (1994), pp. 111–126. ISSN: 1099095X. DOI: 10.1002/env.3170050203.

[33] A. Parente, J. C. Sutherland, L. Tognotti, and P. J. Smith. « Identification of low-dimensional manifolds in turbulent flames. » In: *Proceedings of the Combustion Institute* 32 I.1 (2009), pp. 1579–1586. ISSN: 15407489. DOI: 10.1016/j.proci.2008.06.177. URL: http://dx.doi.org/10.1016/j.proci.2008.06.177.

[34] A. Parente, J. C. Sutherland, B. B. Dally, L. Tognotti, and P. J. Smith. « Investigation of the MILD combustion regime via Principal Component Analysis. » In: *Proceedings of the Combustion Institute* 33.2 (2011), pp. 3333–3341. ISSN: 15407489. DOI: 10.1016/j.proci.2010.05.108. URL: http://dx.doi.org/10.1016/j.proci.2010.05.108.

[35] Alessandro Parente and James C. Sutherland. « Principal component analysis of turbulent combustion data: Data pre-processing and manifold sensitivity. » In: *Combustion and Flame* 160.2 (2013), pp. 340–350. ISSN: 00102180. DOI: 10.1016/j.combustflame.2012.09.016.

[36] Julien Pedel, Jeremy N Thornock, and Philip J Smith. « Ignition of co-axial turbulent diffusion oxy-coal jet flames : Experiments and simulations collaboration. » In: *Combustion and Flame* 160.6 (2013), pp. 1112–1128. ISSN: 0010-2180. DOI: 10.1016/j.combustflame.2013.01.022. URL: http://dx.doi.org/10.1016/j.combustflame.2013.01.022.

[37] Julien Pedel, Jeremy N Thornock, Sean T Smith, and Philip J Smith. « International Journal of Multiphase Flow Large eddy simulation of polydisperse particles in turbulent coaxial jets using the direct quadrature method of moments. » In: *International Journal of Multiphase Flow* 63 (2014), pp. 23–38. ISSN: 0301-9322. DOI: 10.1016/j.ijmultiphaseflow.2014.03.002. URL: http://dx.doi.org/10.1016/j.ijmultiphaseflow.2014.03.002.

[38] Mohammad Rafi, Benjamin J Isaac, Axel Coussement, Philip J Smith, and Alessandro Parente. « Principal component analysis coupled with nonlinear regression for chemistry reduction. » In: *Combustion and Flame* 187 (2018), pp. 30–41. ISSN: 0010-2180. DOI: 10.1016/j.combustflame.2017.08.012. URL: https://doi.org/10.1016/j.combustflame.2017.08.012.

[39] Rommel G. Regis and Christine A. Shoemaker. « Constrained global optimization of expensive black box functions using radial basis functions. » In: *Journal of Global Optimization* 31.1 (2005), pp. 153–171. ISSN: 09255001. DOI: 10.1007/s10898-004-0570-0.

[40] Franck Richecoeur et al. « DMD algorithms for experimental data processing in combustion To cite this version : HAL Id : hal-00825509 DMD algorithms for experimental data processing in combustion. » In: (2013).

[41] Gerasimos G Rigatos. « Feedback control and Kalman Filtering of nonlinear wave dynamics. » In: *IFAC-PapersOnLine* 48.3 (2015), pp. 1362–1367. ISSN: 2405-8963. DOI: 10.1016/j.ifacol.2015.06.276.

[42] S Sahyoun and S Djouadi. « Local Proper Orthogonal Decomposition based on space vectors clustering. » In: *Systems and Control (ICSC), 2013 3rd International Conference on* (2013), pp. 665–670. DOI: 10.1109/ICoSC.2013.6750930.

[43] Benjamin Schleich, Nabil Anwer, Luc Mathieu, and Sandro Wartzack. « Shaping the digital twin for design and production engineering. » In: *CIRP Annals - Manufacturing Technology* 66.1 (2017), pp. 141–144. ISSN: 17260604. DOI: 10.1016/j.cirp.2017.04.040. URL: http://dx.doi.org/10.1016/j.cirp.2017.04.040.

[44] R. Schöbi, B. Sudret, and J. Wiart. « Polynomial-Chaos-based Kriging. » In: (2015), pp. 1–33. ISSN: 2152-5080. DOI: 10.1615/Int.J.UncertaintyQuantification.2015012467.

[45] R. Schöbi, P. Kersaudy, B. Sudret, and J. Wiart. *Combining Polynomial Chaos Expansions and Kriging <hal-01432550>*. Tech. rep. ETH Zurich, Switzerland: Orange Labs research, 2014.

[46] Matthias Seeger. *Gaussian processes for machine learning*. Vol. 14. 2. 2004, pp. 69–106. ISBN: 026218253X. DOI: 10.1142/S0129065704001899. arXiv: 026218253X.

[47] James C. Sutherland and Alessandro Parente. « Combustion modeling using principal component analysis. » In: *Proceedings of the Combustion Institute* 32 I.1 (2009), pp. 1563–1570. ISSN: 15407489. DOI: 10.1016/j.proci.2008.06.147. URL: http://dx.doi.org/10.1016/j.proci.2008.06.147.

[48] Thomas H.J. Uhlemann, Christoph Schock, Christian Lehmann, Stefan Freiberger, and Rolf Steinhilper. « The Digital Twin: Demonstrating the Potential of Real Time Data Acquisition in Production Systems. » In: *Procedia Manufacturing* 9 (2017), pp. 113–120. ISSN: 23519789. DOI: 10.1016/j.promfg.2017.04.043.

[49] Shaomin Wu. « A review on coarse warranty data and analysis. » In: *Reliability Engineering and System Safety* 114 (2013), pp. 1–11. ISSN: 0951-8320. DOI: 10.1016/j.ress.2012.12.021.

[50] Manyu Xiao, Piotr Breitkopf, Rajan Filomeno Coelho, Catherine Knopf-Lenoir, Maryan Sidorkiewicz, and Pierre Villon. « Model reduction by CPOD and Kriging. » In: *Structural and Multidisciplinary Optimization* 41.4 (2010), pp. 555–574. ISSN: 1615-147X. DOI: 10.1007/s00158-009-0434-9. URL: http://link.springer.com/10.1007/s00158-009-0434-9.

[51] Manyu Xiao, Piotr Breitkopf, Rajan Filomeno Coelho, Catherine Knopf-Lenoir, and Pierre Villon. « Enhanced POD projection basis with application to shape optimization of car engine intake port. » In: *Structural and Multidisciplinary Optimization* 46.1 (2012), pp. 129–136. ISSN: 1615147X. DOI: 10.1007/s00158-011-0757-1.

[52] Manyu Xiao, Piotr Breitkopf, Rajan Filomeno Coelho, Catherine Knopf-Lenoir, Pierre Villon, and Weihong Zhang. « Constrained Proper Orthogonal Decomposition based on QR-factorization for aerodynamical shape optimization. » In: *Applied Mathematics and Computation* 223 (2013), pp. 254–263. ISSN: 00963003. DOI: 10.1016/j.amc.2013.07.086.

[53] Manyu Xiao, Piotr Breitkopf, Rajan Filomeno Coelho, Pierre Villon, and Weihong Zhang. « Proper orthogonal decomposition with high number of linear constraints for aerodynamical shape optimization. » In: *Applied Mathematics and Computation* 247 (2014), pp. 1096–1112. ISSN: 00963003. DOI: 10.1016/j.amc.2014.09.068. URL: http://dx.doi.org/10.1016/j.amc.2014.09.068.

[54] Jianbo Yu. « Local and global principal component analysis for process monitoring. » In: *Journal of Process Control* 22.7 (2012), pp. 1358–1373. ISSN: 09591524. DOI: 10.1016/j.jprocont.2012.06.008. URL: http://dx.doi.org/10.1016/j.jprocont.2012.06.008.

[55] Arthur Zimek and Erich Schubert. *Outlier Detection. Encyclopedia of Database Systems.* Ed. by Liu L. and Özsu M. Springer, New York, NY, 2017, pp. 1–5. ISBN: 9781489979933. DOI: 10.1007/978-1-4899-7993-3.