Principal Component Analysis coupled with nonlinear regression for chemistry reduction

Mohammad Rafi Malik[a,*], Benjamin J. Isaac[b], Axel Coussement[a], Philip J. Smith[b], Alessandro Parente[a]

[a]*Service d'Aéro-Thermo-Mécanique, Université Libre de Bruxelles, Bruxelles, Belgium*
[b]*Department of Chemical Engineering, University of Utah, Salt Lake City, UT, 84112, USA*

Abstract

Large kinetic mechanisms are required in order to accurately model combustion systems. If no parameterization of the thermo-chemical state-space is used, solution of the species transport equations can become computationally prohibitive as the resulting system contains a wide range of time and length scales. Parameterization of the thermo-chemical state-space with an a priori prescription of the dimension of the underlying manifold would lead to a reduced yet accurate description. To this end, the potential offered by Principal Component Analysis (PCA) in identifying low-dimensional manifolds is very appealing. The present work seeks to advance the understanding and application of the PC-transport approach by analyzing the ability to parameterize the thermo-chemical state with the PCA basis using nonlinear regression. In order to demonstrate the accuracy of the method within a numerical solver, unsteady perfectly stirred reactor (PSR) calculations are shown using the PC-transport approach. The PSR analysis extends previous investigations by the authors to more complex fuels (methane and propane), showing the ability of the approach to deal with relatively large kinetic mechanisms. The ability to achieve highly accurate mapping through Gaussian Process based nonlinear regression is also shown. In addition, a novel method based on local regression of the PC source terms is also investigated which leads to improved results.

Keywords: Combustion; Nonlinear Regression; Local Regression; Low-dimensional manifolds; Principal Component Analysis.

1. Introduction

The numerical modeling of turbulent combustion is a very challenging task as it combines the complex phenomena of turbulence and chemical reactions. This study becomes even more challenging when large detailed kinetic mechanisms are used in order to understand some special features such as pollutant formation. A detailed combustion mechanism for a simple fuel such as methane involves 53 species and 325 chemical reactions [1]. Moreover, the number of species and reactions increases with increasing fuel complexity. The coupling of the kinetic equations with the set of Navier-Stokes equations results in a problem that is too complex to be solved by the current computational means. In a CFD calculation, the number of species tracked impacts the memory usage and CPU time. It is thus important to minimize this number by the use of a simpler but representative set of variables. Therefore, there is a need for methods allowing to parameterize efficiently the thermo-chemical state of a reacting system with a reduced number of optimal reaction variables. Among those, Principal Component Analysis (PCA) appears as an

[*] Corresponding author. Phone + 32 2 650 26 80 Fax +32 2 650 27 10 Address: Avenue F. D. Roosevelt 50, 1050 Bruxelles, Belgium.
Email address: Rafi.Malik@ulb.ac.be (Mohammad Rafi Malik)

ideal candidate to fulfill the purpose [2–8]. PCA offers the possibility of automatically reducing the dimensionality of data sets consisting of a large number of correlated variables, while retaining most of the variation present in the original data. After reduction, the new set of variables, called principal components (PCs), are othogonal, uncorrelated and linear combinations of the original variables. By retaining the PCs containing most of the variance and transporting them in a numerical simulation, the dimensionality of the system can be higly reduced. Another advantage of PCA resides in the fact that the PCs can be obtained through data sets based on simple systems (such as canonical reactors) and then applied to a similar, more complex system [9]. A methodology based on PCA was proposed [5] for the identification of the controlling dynamics in reacting systems and for the consistent reduction of very large kinetic mechanisms. Sutherland and Parente [8] proposed a combustion model based on the concepts from PCA (PC-score approach). They derived transport equations for the principal components (PCs), and proposed a model where the state-space variables are constructed directly from the PCs. The PCA-based modeling approach was enhanced [3, 10, 11] by combining PCA with nonlinear regression techniques, allowing a nonlinear mapping of the thermo-chemical state and the corresponding source terms onto the basis identified by the principal components. As a result, the nonlinear nature of chemical manifolds is better captured, thus, maximizing the potential size reduction provided by the method. Isaac et al. [4] and Echekki and Mirgolbabaei [2] provided the first a posteriori studies on the use of the PC-score approach. In particular, Isaac et al. showed in [4] the potential of PC-transport based combustion models coupled with nonlinear regression techniques. The model was tested on an unsteady calculation of a perfectly stirred reactor (PSR) burning syngas. The authors showed that Gaussian Process Regression (GPR) technique produced the most accurate reconstruction, showing remarkable accuracy for the prediction of temperature and major and minor species with 2 transported variables instead of 11. The approach was also tested for the first time within a CFD solver.

The present work seeks to advance the understanding and application of the PC-transport approach by applying this method to more complex fuels such as methane and propane. First, 0-D simulation of a PSR is used to generate the database for model training. Then, the solution of a steady and unsteady PSR calculation using the PC-transport approach for large kinetic mechanisms is compared with the full solution. Next, the PC-transport approach is coupled with nonlinear regression (PC-GPR) in order to increase the size reduction potential of PCA. Finally, the first study on an enhancement of the classical PC-transport approach by the use of local nonlinear regression (PC-L-GPR) is also shown. It should be pointed out that the objective of the present work was to demonstrate the applicability of GPR regression for accurate source term regression. To this purpose, the choice of a PSR is quite obvious as it allows to focus on such an aspect without the influence of transport processes.

2. Principal Component Analysis

Principal Component Analysis [12] is a useful statistical technique that has found application in combustion for its ability of identifying low-dimensional manifolds. In high dimension data sets, where graphical representation is not possible, PCA can be a powerful tool as it identifies correlations and patterns in a data set. Once these patterns have been identified, the data set can be compressed by reducing the number of dimensions without much loss of information. PCA analyzes the covariance between variables in a data set and identifies a linear representation of the system through orthogonal vectors, each one having a significance proportional to its eigenvalue.

In order to perform principal component analysis, a data-set $\mathbf{X}\,(n \times Q)$ consisting of n observations of Q independent variables is needed. Then, the data must be centered (by subtracting its mean) and scaled (using an appropriate scaling method): centering is used to convert observations into fluctuations over the mean, while scaling is done in order to compare the data evenly (if they have different units or order of magnitudes):

$$\mathbf{X_{SC}} = (\mathbf{X} - \overline{\mathbf{X}})\mathbf{D^{-1}} \tag{1}$$

where $\overline{\mathbf{X}}$ is $(n \times Q)$ matrix containing the mean of each variable and $\mathbf{D}$ is a $(n \times Q)$ matrix containing the scaling factor of each variable. Several scaling methods can be found in the litterature: *auto scaling*, *range scaling, pareto scaling, variable stability scaling* and *level scaling* [6].

Then, one can compute the covariance matrix $\mathbf{S}$ defined as (the notation $\mathbf{X}$ will be used in the following instead of $\mathbf{X_{SC}}$ for the sake of simplicity):

$$\mathbf{S} = \frac{\mathbf{1}}{\mathbf{n-1}}\mathbf{X^T X}$$

The diagonal elements of $\mathbf{S}$ represent the variance of each variable, while the off-diagonal values show the covariance between two variables. Since $\mathbf{S}$ is a square matrix (of size $(Q \times Q)$), an eigenvalue decomposition can be performed yielding the eigenvectors and eigenvalues of the system:

$$\mathbf{S} = \mathbf{ALA^T}$$

where $\mathbf{A}$ $(Q \times Q)$ and $\mathbf{L}$ $(Q \times Q)$ are respectively the eigenvectors of $\mathbf{S}$ (also called principal components, PCs) and the eigenvalues of $\mathbf{S}$, in decreasing order. The eigenvectors matrix $\mathbf{A}$, also called the basis matrix, is used to obtain the principal component scores, $\mathbf{Z}$ $(n \times Q)$, by projecting the original data set $\mathbf{X}$ on that basis:

$$\mathbf{Z} = \mathbf{XA} \tag{2}$$

Eq. 2 indicates that the original data set can be uniquely recovered using the PCs and their scores:

$$\mathbf{X} = \mathbf{ZA^{-1}}$$

where $\mathbf{A^{-1}} = \mathbf{A^T}$. Then, using a subset of $\mathbf{A}$ by retaining only q PCs (with $q < Q$), noted $\mathbf{A_q}$, an approximation of $\mathbf{X}$ based on the first q eigenvectors ($\mathbf{X_q}$) is obtained:

$$\mathbf{X} \approxeq \mathbf{X_q} = \mathbf{Z_q A_q^T}$$

where $\mathbf{X_q}$ is the approximation of $\mathbf{X}$ based on the first q eigenvectors of Q, and $\mathbf{Z_q}$ is the $(n \times q)$ matrix of the principal component scores. In the PC analysis, the largest eigenvalues correspond to the first columns of $\mathbf{A}$. This means the largest amount of variance in the original variables is described by the first PCs. Thus, the truncation is made on the last eigenvectors (corresponding to the smallest eigenvalues). By removing the last PCs, the dimension of the system is reduced while retaining most of the variation in the system.

2.1. PC-score Approach

In the work of Sutherland and Parente [8], a model based on transport equations for the PCs is proposed derived from the general species transport equation:

$$\frac{\partial}{\partial \mathbf{t}}\left(\rho \mathbf{Y_k}\right) + \nabla\left(\rho \bar{\mathbf{u}} \mathbf{Y_k}\right) = \nabla\left(\rho \mathbf{D_k} \nabla \mathbf{Y_k}\right) + \mathbf{R_k} \qquad k = 1, ..., n_s \tag{3}$$

where Y_k is the mass fraction of species k and R_k is its corresponding source term (with n_s the total number of species in the system), D_k the diffusion coefficent for species k , ρ the density and $\bar{u}$ the velocity vector. Transport equations for the PC scores ($\mathbf{Z}$) can be formulated from Eq. 3 given the basis matrix $\mathbf{A}$ and the scaling factors d_k:

$$\frac{\partial}{\partial \mathbf{t}} \left(\rho \mathbf{z}\right) + \nabla \left(\rho \bar{\mathbf{u}} \mathbf{z}\right) = \nabla \left(\rho \mathbf{D_z} \nabla \mathbf{z}\right) + \mathbf{s_z} \tag{4}$$

$$\mathbf{s_z} = \sum_{\mathbf{k=1}}^{\mathbf{Q}} \frac{\mathbf{R_k}}{\mathbf{d_k}} \mathbf{A_{kq}} \tag{5}$$

where $\mathbf{z} = \mathbf{Z_i^t}$ represents an individual score realization. One of the major weaknesses of classic PCA is that a multi-linear model is used to approximate a highly nonlinear manifold. The nonlinearity of chemical manifolds can be attributed to the high nonlinearity of chemical source terms (Arrhenius). This can be visualized in Fig. 1, showing the first principal component source term s_{z1}, as a function of the first two principal components for the propane case.

In the present work, PCA is used to identify the most appropriate basis to parameterize the empirical low-dimensional manifolds and define transport equations in the new space (see Eq. 4 and 5). Then, both the state space and the source terms are non-linearly regressed onto the new basis using several approaches, described in Section 3. The non-linear regression of the chemical state space and of the corresponding source terms is intended to overcome the shortcomings associated to the multi-linear nature of PCA, and to reduce the number of components required for an accurate description of the state-space. The method belongs to the family of Empirical Low-Dimensional Manifolds (eLDMs) [7], and it is based on the idea that compositions occurring in combustion systems lie close to a low-dimensional manifold. eLDMs require samples for the construction of reduced models, which might be seen as a limitation of the approach, as all system states are required before model reduction. However, although initial studies on PCA models involved DNS data of turbulent combustion [7, 8], recent studies have demonstrated [2, 13, 14] that PCA-based models can be trained on simple and inexpensive systems, such as 0D reactors and 1D flames, and then applied to model complex systems, such as flame-vortex interaction [15], flame-turbulence interactions [4] as well as turbulent premixed flames [13].

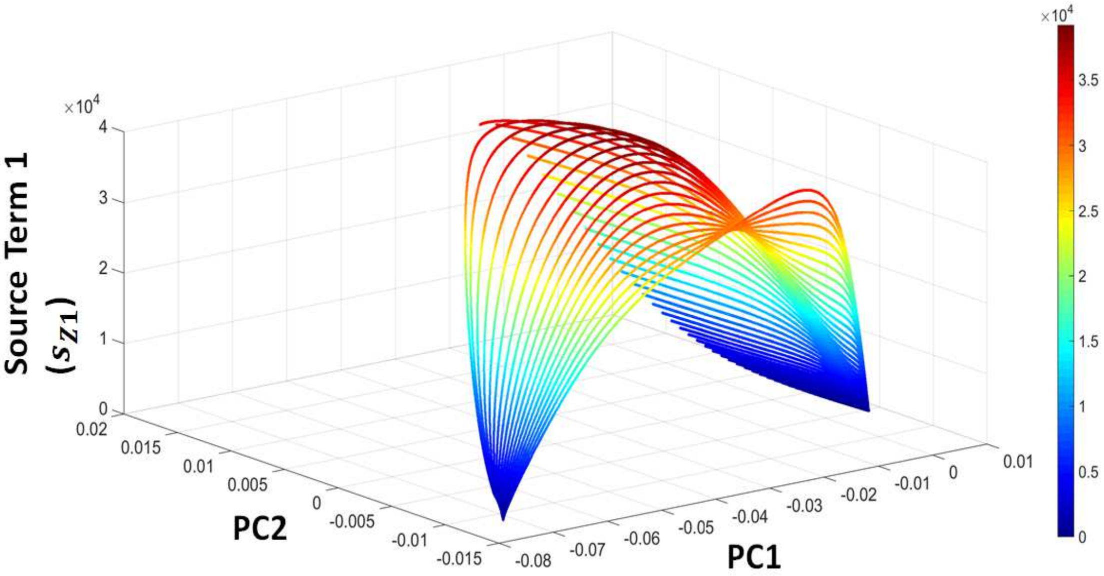


Figure 1: Manifold of Source Term 1 (s_{Z1}) in function of $PC1$ and $PC2$

3. Regression Models

In this study, the state-space variables (Y_k, T, ρ, ...) and the PC source terms (s_{Z_q}) are mapped to the PC basis using nonlinear regression:

$$\phi \approx f_\phi \left(Z_q \right)$$

where f_ϕ is the nonlinear regression function and ϕ represent the dependant variables (i.e. Y_k, T, ρ and s_{Z_q}). In a previous study [4], the authors compared different regression models in their ability to accurately map the highly nonlinear functions (such as the chemical source terms) on the plane PCA manifold. These models include:

- *Linear Regression Model* (LIN) in which the state-space is mapped to the PC using a linear function [16]
- *Mutivariate Adaptive Regression Splines* (MARS) where the model is build from product spline basis functions [17]
- *Artificial Neural Networks* (ANN) that uses the concept of networking various layers of estimation resulting in a highly accurate output layer [18]
- *Support Vector Regression* (SVR) which is a subset of support vector machines (SVM) and in which the idea is again to create a model which predicts s_Z given Z using learning machines which implement the structural risk minimization inductive principle [19]
- *Gaussian Process Regression* (GPR), which is based on the idea that dependent variables can be described by a gaussian distribution [20, 21]. In particular, it was shown that GPR produced the most accurate reconstruction of the state-space variable, using only 2 transport equations instead of 11 in the full system without regression.

In the present work, we will focus on the use of GPR for state space and source term parameterization. The choice of GPR is due to its semi-parametric nature, that increases the generality of the approach. GPR employs

Gaussian mixtures to capture information about the relation between data and input parameters, making predictions of non-observed system states more reliable than in fully parametric approaches [20].

3.1. Gaussian Process Regression

Gausian Processes (GPs) does not assume a specific model for the regression function. Rather, GPs generate data in the domain of interest such that any finite subset of the range follows a multivariate Gaussian distribution. The dependant variables can thus be described by a gaussian distribution:

$$\phi \approx GP\left(m(x), K(x, x^{'}\right)$$

where m is a mean function and K is a covariance function (or kernel). The mean function is often assumed to be zero. The covariance function used here is the Squared Exponential:

$$K(x, x^{'}) = \sigma_f^2 \ exp\left[\frac{-(x - x^{'})^2}{2l^2}\right]$$

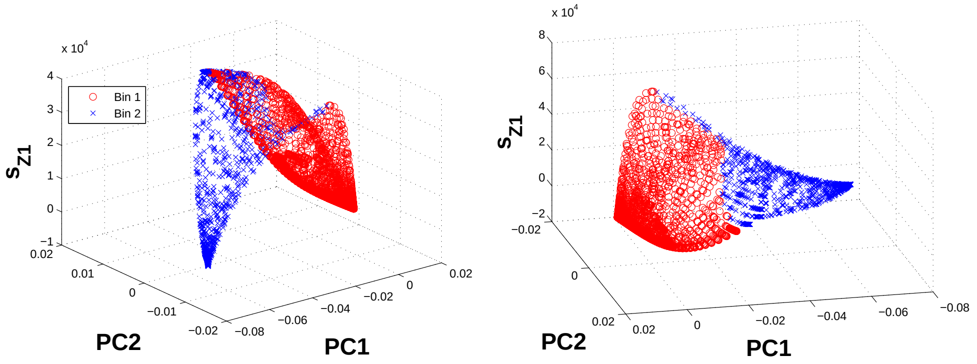
with σ_f^2 being the signal variance and l the characterictic lenght scale. These two parameters of the covariance function are called *hyper-parameters*. After an initial guess, those hyper-parameters are optimized using a Gaussian likelihood function.

4. Local regression

In order to improve further the accuracy of the regression and increase PCA's potential for size reduction, a novel approach is proposed where the PC-score approach is coupled with locally regressed state-space (PC-L-GPR). The idea is to divide the PC state-space into bins or clusters, and to perform a GP regression seperately in each of these bins. As a consequence, a better regression would be obtained (if each bin is chosen appropriately) and the computational time required for GPR will also be reduced. In order to define such bins, a conditioning variable has to be chosen. This variable should be able to capture the general characteristics of the state-space. Possible candidates are the PCs source terms, as the latter are highly nonlinear over the PC space. Clustering the source terms manifolds such as they can be approximated by quasi-linear functions in each bin would simplify and accelerate the regression algorithm. As to the author's knowledge, this approach has not yet been tested previously in the context of PC-transport approach.

4.1. Single Conditioned

In order to divide the state-space into bins, a conditioning variable has to be chosen. As stated above, the PC source terms are appropriate candidates as they are highly nonlinear over the PC space and need to be accurately mapped in order to obtain high accuracy in a simulation. Indeed, if the PC source terms are well captured by the regression, the PC's will be accurately calculated, thus all the other variables. A good candidate would be the first PC source term s_{Z_1}, as the latter is highly correlated with the major species and also contains most of the variance in the system. Figure 2 shows the first source term's manifold in a 2D PC space. The bin borders are chosen to the extrema of s_{Z_1}. This results in two bins as shown on Figure 2 , the border being represented by the red line. It can be seen that in each bin, s_{Z_1} is a rather smoothly increasing (or decreasing) function of Z_1 and Z_2. Regressing each of these two bins seperately is easier, more accurate and faster than regressing the whole manifold at once (i.e. global



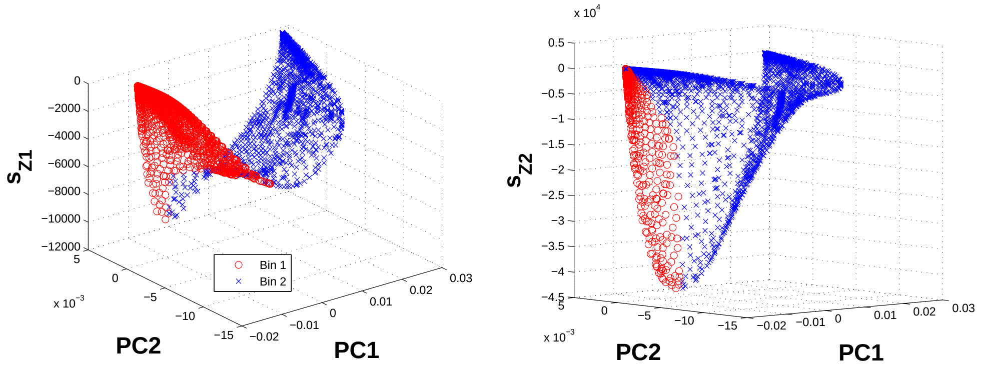
(a) (b)

Figure 2: Clustering based on the extrema of s_{Z_1} for both s_{Z_1} (a) and s_{Z_2} (b) (propane case, Polimi mechanism)

regression). It must be noted that the manifold of s_{Z_2} was also clustered based on the extrema of s_{Z_1} (Figure 2 b)), but those extrema do not necessarily fall within the ones of s_{Z_2}. Although this approach leads to improved results compared to global regression (cfr. Section 6.2.2), they can be further improved even using the double conditioning method (cfr. Section 4.2). Local regression provides better results when the bins and conditioning variable are chosen correctly (cfr. Section 6.2.2). In order to handle the discontinuities that could occur at the boundaries of the bins, the clusters were artificially extended across the bin border, by providing an overlap of 2% at the boundaries of the cluster region, to ensure smoothness of the soluton and avoid discontinuities.

4.2. Double Conditioned

In some cases, local regression with a single conditioning variable can still provide unsatisfactory results. In such cases, the accuracy of the results can be further improved by using a second conditioning variable. In the case of PC-transport where the first conditioning variable is s_{Z_1}, a natural choice for the second conditionig variable would be the second PC source term s_{Z_2}. Thus, s_{Z_1} is regressed locally based on clusters defined by its extrema, while s_{Z_2} is regressed locally in clusters defined by its own extrema (i.e. not based on s_{Z_1} extrema). Figure 3 shows the first and second source terms' manifolds for the propane case, together with the clusters borders.



(a) (b)

Figure 3: Clustering based on the extrema of s_{Z_1}(a) and s_{Z_2}(b) (propane case, San Diego mechanism)

5. Perfectly Stirred Reactor and Test Cases

The objective of the present work was to extend previous investigation on syngas [4] to more complex fuels, with a significantly large number of species and reactions. In [4], the proposed PCA approach was demonstrated on the unsteady solution of a perfectly stirred reactor (PSR). The solution from the full set of equations was compared to the standard PC-transport approach, and the PC-transport approach using nonlinear regression.

In this work, the analysis of the proposed PC model in its ability to handle complex fuels and large kinetic mechanisms was done in a similar way. The data sets for PCA were generated by performing unsteady simulations by varying the residence time in the vessel from extinction to equilibrium. For each residence time, the temporal solution was saved until steady-state was reached. The vessel was initialized at equilibrium conditions (constant pressure and enthalpy) and the inlet conditions for the reactor were set at an equivalence ratio of 1. The initial conditions for the reactor are set at the equilibrium conditions of the inlet and the system is run until a steady-state solution is reached. The PSR is modeled assuming constant volume, residence time and pressure. The ideal gas law was used to model the behaviour of the mixture. Thermodynamic properties were obtained through the Cantera software package [22]. Two different fuels were investigated:

- methane (CH_4), burned with pure oxygen. The mechanism used was the GRI 3.0 [1], without species containing nitrogen (resulting in 34 species). The inlet temperature was set to 300K. One hundred cases were run between residence time of 1$e^{-4}\,s$ to 1$e^{-6}\,s$. The PCA database generated in this way contained $\sim$ 100,000 points.
- propane (C_3H_8), burned with air. Two different kinetic schemes were used: the San Diego Mechanism [23] (subsequently referred as San Diego), without nitrogen species (50 species, 230 reactions) and the Primary Reference Fuels Polimi_PRF_PAH_HT_1412 kinetic mechanism [24] (subsequently referred as Polimi), without nitrogen species (162 species, ~6,000 reactions). The inlet temperatures were set to 1300K for the San Diego scheme and to 1500K for the Polimi mechanisms. One hundred cases were run between residence time of 1$e^{-1}\,s$ to 1$e^{-7}\,s$. The PCA database consisted of $\sim$ 110,000 points for the San Diego scheme and of $\sim$ 420,000 points for the Polimi one.

The PCA process described in the previous section is then applied to the database to create the basis matrix A_q, and the regression functions f_ϕ for the state-space variables, ϕ. Gaussian Process Regression was done using 1, and 2 PC's as independant variables. The implementation of the PSR equations was done using MATLAB together with the Cvode toolbox and Cantera. The temporal solution to the equations is obtained using the Newton nonlinear solver, and the BDF multistep method. Governing equations for species transport and energy were implemented and solved:

$$\frac{\partial m_i}{\partial t} = \dot{m}_{i,in} - \dot{m}_i + \omega_i \cdot MW_i \cdot V \tag{6}$$

where m_i (kg) and ω_i $(kmol/m^3/s)$ are the mass and the net molar production rate of the i^{th} species, MW_i is the molecular weight of the i^{th} species and V (m^3) the volume of the reactor. or the mass flow rates (kg/s), $\dot{m}_{i,in}$ is the mass flow of the i^{th} species entering the reactor and $\dot{m}_{i,out}$ is the mass flow exiting the reactor. The residence time τ $(/s)$ in the reactor is defined as:

$$\tau = \frac{\rho V}{\dot{m}}$$

where ρ is the density of the mixture inside the reactor. For the energy equation:

$$\frac{\partial H}{\partial t} = \dot{m}_{in} h_{in} - \dot{m} h + V \frac{dP}{dt} \quad (7)$$

where H is the enthalpy of the system and h is the specific enthalpy (J/kg), $\dot{m}_{in}$ is the total mass flow entering the reactor and $\dot{m}$ the total mass flow rate leaving the system. The last term of Eq. 7 being zero as the PSR operates in constant pressure conditions. In this study, no accumulation of mass inside the reactor has been assumed, thus $\dot{m}_{in} = \dot{m}_{out} = \dot{m}$, but $\dot{m}$ can change due to a change in density.

6. Results and Discussion

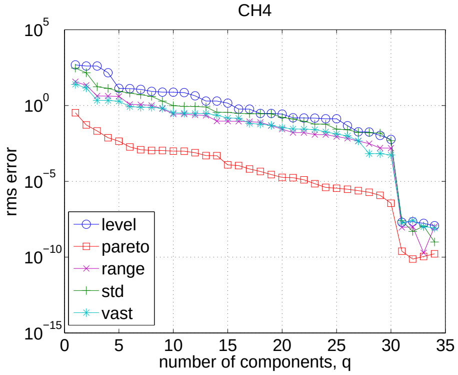
In this section, the proposed method is demonstrated in a PSR, comparing the calculations using the full set of equations to the standard PC-transport approach and to the PC-transport approach using nonlinear regression. This demonstration is done for two different fuels: a simple one, methane (CH_4), and a more complex, propane (C_3H_8). But first, an analysis is performed on the effects of several scaling methods used in PCA (Eq. 1).

6.1. Scaling

As mentioned in [6], scaling has an important effect on the accuracy of the method. It can change the PCA structure by altering the relative importance of various species, and the choice of a particular scaling method is motivated by the goal of the resulting PCA to reconstruct specific variables. In [4], the authors showed that pareto scaling method is able to achieve the greatest reduction, and produces a highly regressible surface for a syngas mechanism. This was also consistently showed in other previous investigations [6, 14, 25–27]. In order to assess the accuracy of the various scaling methods presented in Section 2, a similar study was performed for the methane and propane cases on the species and PCs source terms. The *rms error* was used as a mean of quantifying the error in the reconstruction of species mass fractions and PCs source terms. The definition of the *rms error* used here is:

$$rms\, error = \sqrt{\frac{\sum_{i=1}^{n} \left(x_{predicted,i} - x_i\right)^2}{n}}$$

Figure 4 shows the *rms error* for the mass fraction of CH_4 for the methane case, and for the various scaling methods while varying the number of principal components, q. It is clear from Fig. 4 that pareto scaling provides the lowest



(a)

Figure 4: *Rms error* values for CH_4 mass fraction while varying q, the number of PCs, and the scaling method (methane case).

error in the reconstruction for CH_4, and this for all the range of q, while all other methods show similar behaviour. It can also be seen that a significant decrease in the *rms error* is not achieved until $q = 31$, and this is observed with all the scaling methods. With $q = 31$, only a minor reduction is achieved. This is due to the linear nature of PCA based models, which try to model highly nonlinear reaction rates on a linear basis. An alternative approach to overcome this issue can be the use of nonlinear regression functions, which can be used to map the nonlinear reaction rates or nonlinear species concentrations to the lower dimensional representation given by the PCs. A similar analysis of the influence of scaling methods was also done for the propane cases, which led to the same conclusion, i.e. that pareto scaling provides the lowest error in the reconstruction of all species (major and minor).

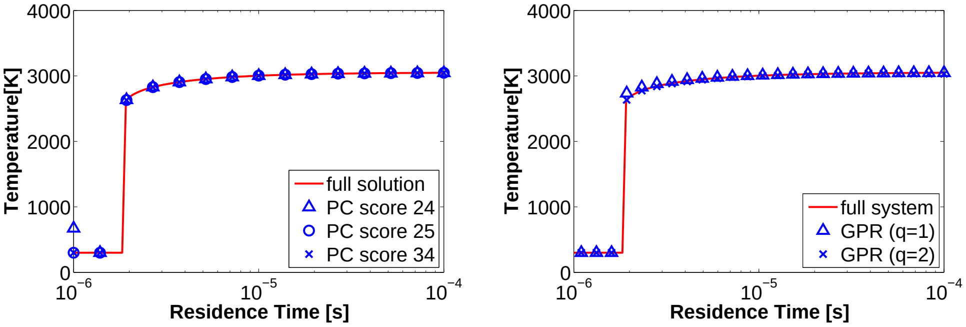
6.2. Standard PC-score Approach vs PC-score with Gaussian Process Regression

The standard PC-score approach based on Eq. 4 and 5 was tested for both methane and propane, and compared to the full solution, i.e. the solution based on the transport of all species (Eq. 6 and 7). Then, the non linear state-space variables were mapped to the linear PC basis using Gaussian Process Regression (GPR). GPR was performed on all variables (temperature, species and score source terms) using 5,000 sample points evenly distributed over the PC space. Error quantification is done through the coefficient of determination R^2:

$$R^2 = \frac{\sum_{i=1}^{n} \left(x_{predicted,i} - \bar{x}\right)^2}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}$$

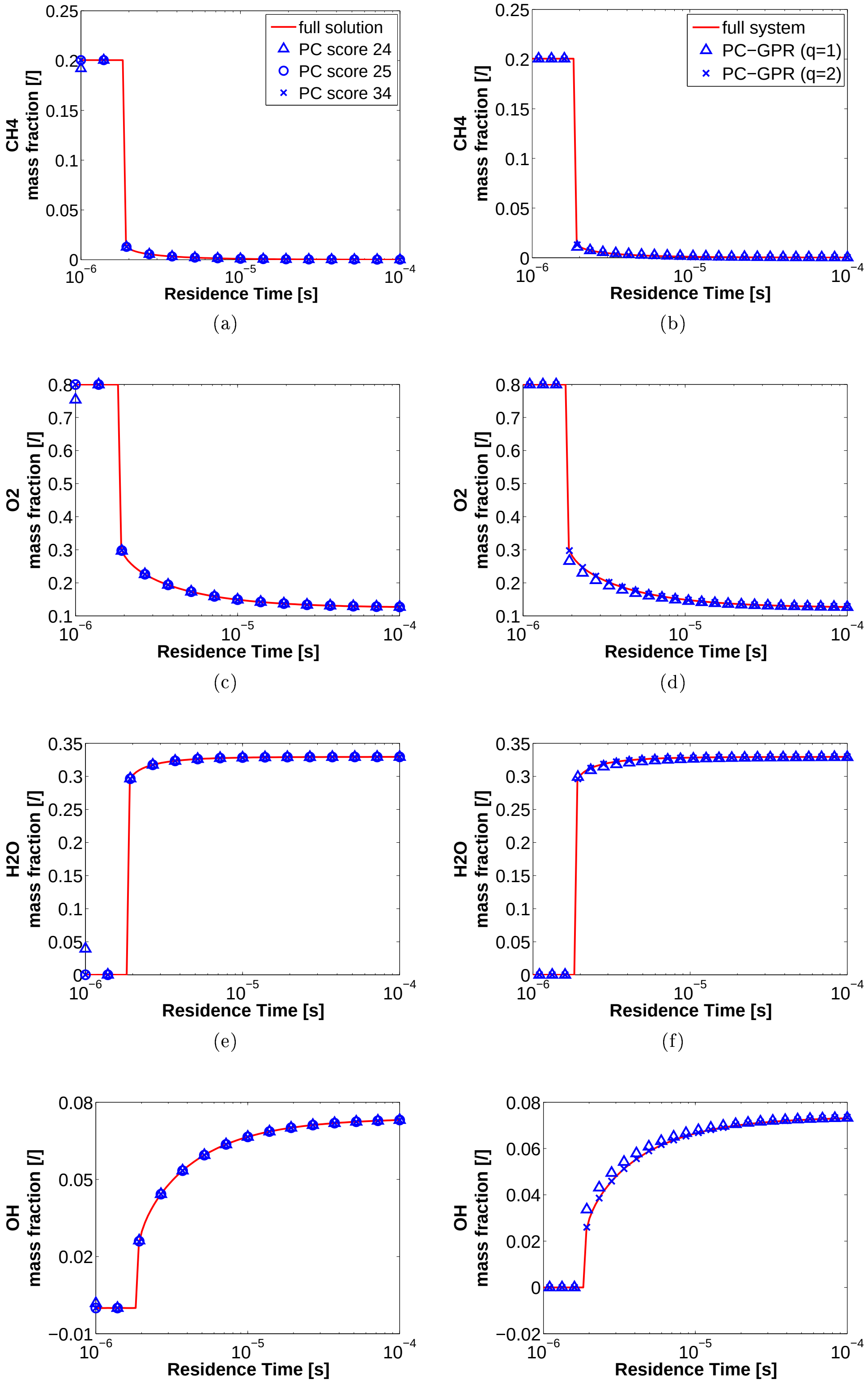
where $\bar{x}$ is the mean value of an observed variable.

6.2.1. Methane case



(a) (b)

Figure 5: PSR temperature as a function of the residence time, with the solid line representing the full solution. The markers represent the results for the standard PC-score model while varying q (a), and the PC-score with GPR regression (b) using q = 1 and 2 PCs



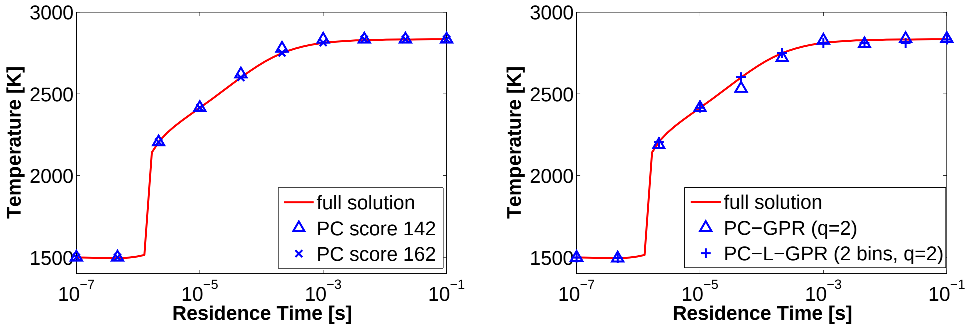
(a) (b) (c) (d) (e) (f) (g) (h)

Figure 6: Species mass fraction as a function of the residence time, with the solid line representing the full solution. The markers represent the results for the standard PC-score model while varying q (left plots), and the PC-score with GPR regression (right plots) using q = 1 and 2 PCs

Figure 5 and 6 show the solution using the standard PC-score model (i.e. without regression) and the solution using the PC-score model together with GPR (PC-GPR) for the methane case. It can be seen that using the standard PC-score approach, at least 25 components out of 34 are required in order to obtain an accurate solution, which correspond to a model reduction of 26%. However, when using GPR, the reduction potential is highly increased: using only 2 PCs, the results show remarkable accuracy for the model with regression over the range of residence times for the predicted temperatures, and both major and minor species. A similar degree of accuracy is not observed in the model without regression until $q = 25$. Also, using PC-GPR with $q = 1$ does not provide sufficient accuracy in the ignition region, where the ignition delay is under-estimated. Moving to $q = 2$ allows to capture the ignition adequately. The regression of ϕ using Gaussian Process and pareto scaling yielded an R^2 of 0.999 for all variables using $q = 2$, and an R^2 of 0.986 or higher with $q = 1$.

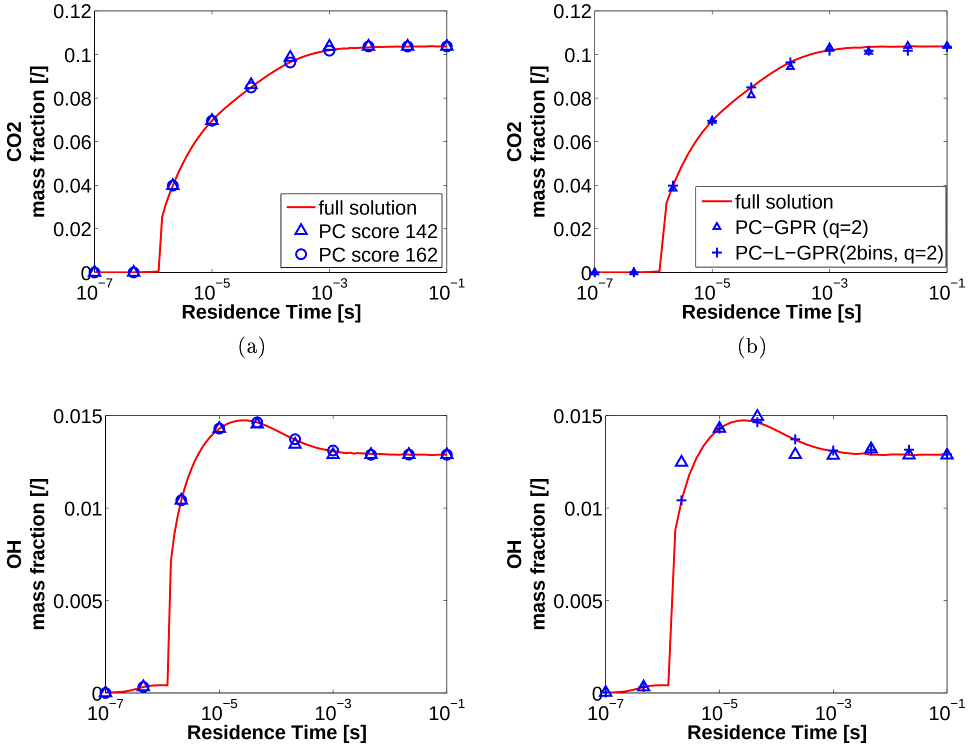
6.2.2. Propane case - Polimi mechanism

Figure 7 shows the temperature profile for the combustion of propane and air using the Polimi mechanism. As far as the standard PC-score approach is concerned, it can be seen that at least 142 components out of 162 are required in order to get an accurate description using a reduced model, which represent a model reduction of 12%. When adding the potential of GPR (PC-GPR), this number can be reduced to 2, leading to fair solution, but not yet satisfying. Indeed, a significant deviation from the full solution can be observed in the ignition/extinction region. In order to improve the model even further, the potential of using GPR locally, together with the PC-score approach (PC-L-GPR), is assessed. In this study, the first principal component's source term, s_{Z1}, was chosen as the variable on which the clustering should be conditioned. The data-set was thus single conditioned on s_{Z1}. This choice can be justified knowing that the first PC's source term is highly correlated with the major species, thus containing most of the variance in the system, and also very nonlinear. The clustering algorithm used in this work searches for the extrema of the conditioning variable, and defines the borders of the bins at those extrema. This allows to have a monotonic increasing (or decreasing) variable in each bin, thus making the job easier for the regression algorithm. In the present analysis, 2 bins were identified (cfr. Figure 2). It can be observed on Figure 7b that using local regression with only 2 components instead of 162 (reduction of 98%) improves significantly the accuracy of the model, especially in the ignition/extinction region, leading to an almost perfect match. Figure 8 shows some of the species mass fraction. Again, it can be seen that using local regression allows to increase the accuracy in the predictions, both for major and minor species.



(a) (b)

Figure 7: PSR temperature as a function of the residence time (Polimi), with the solid line representing the full solution. The markers represent the results for the standard PC-score model while varying q (a), and the PC-score with global and local GPR regression (b) using q=2 PCs and single conditioning



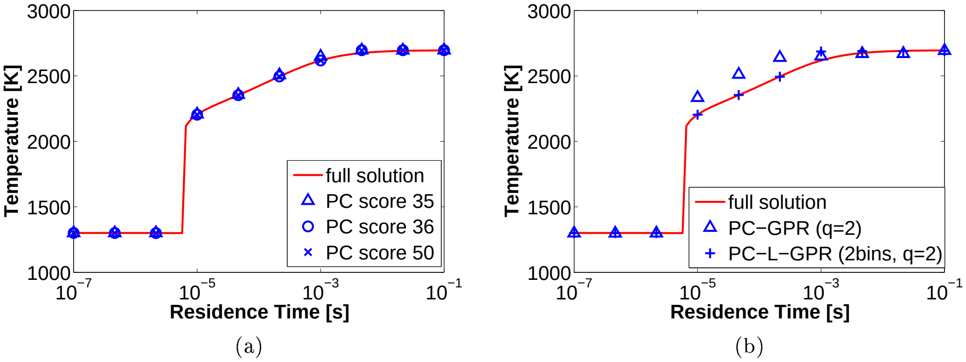
(a) (b)

(c) (d)

Figure 8: Species mass fraction as a function of the residence time (Polimi), with the solid line representing the full solution. The markers represent the results for the standard PC-score model while varying q (left plots), and the PC-score with GPR regression (right plots) using q = 2 PCs

6.2.3. Propane case - San Diego mechanism

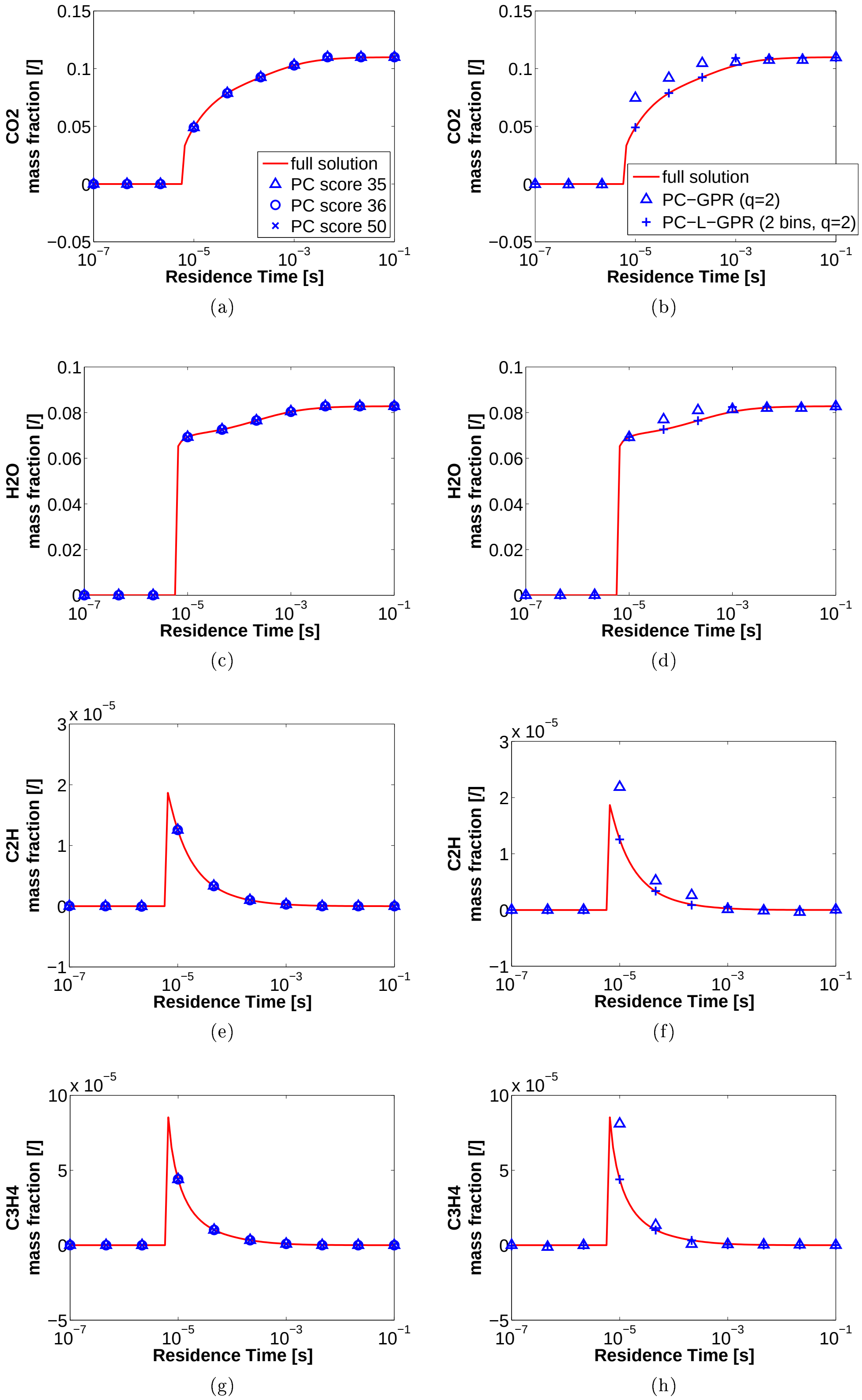
Figure 9 shows the temperature profile for the combustion of propane and air using the San Diego mechanism. It can be seen on Figure 9a that using the standard PC-score approach at least 36 components out of 50 are required in order to get an accurate description using a reduced model, which represent a model reduction of 28%. When coupling GPR with PC-score (PC-GPR), the solution obtained using only 2 components is accurate enough, except in the ignition/extinction region. In order to increase the accuracy in that region as well, the potential of PC-score with local GRP (PC-L-GPR) was assessed. Here again, the data-set was single conditioned based on s_{Z1}. Again, 2 bins were identified for the San Diego mechanism (cfr. Figure 3a). Figure 9b shows a significant improvement in the accuracy of the model in the ignition/extinction region while using only 2 components instead of 50 (reduction of 96%). Figure 10 shows some of the major and minor species mass fraction profiles, where it can be seen that using local regression allows to increase the accuracy of the predictions.



(a)

(b)

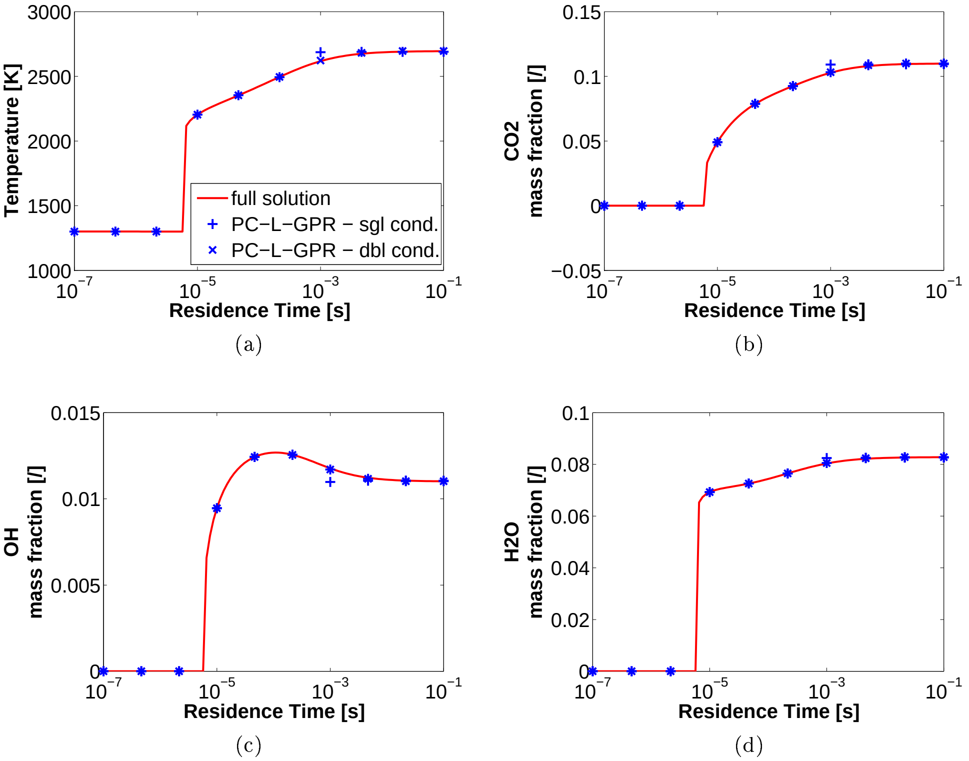
Figure 9: PSR temperature as a function of the residence time (San Diego), with the solid line representing the full solution. The markers represent the results for the standard PC-score model while varying q (a), and the PC-score with global and local GPR regression (b) using q=2 PCs and single conditioning



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 10: Species mass fraction as a function of the residence time (San Diego), with the solid line representing the full solution. The markers represent the results for the standard PC-score model while varying q (left plots), and the PC-score with GPR regression (right plots) using q = 2 PCs

The single conditioned PC-L-GPR model gives quite satisfactory results, but these could be further improved by double conditioning the data set before using GPR. Indeed, clustering the first source term based on its own extrema increased the accuracy of the regression of s_{Z1}, but that clustering does not necessarily fall on the extrema of the second source term s_{Z2}(cfr. Figure 3). By clustering the s_{Z2} based on its own extrema, its subsequent regression can be strongly improved. Figure 11a shows the temperature profile with a comparison between single conditioned and double conditioned PC-L-GPR model. It can be seen that double conditioning the data set prior to applying the regression improves the accuracy of the result even further, leading to a perfect match between the reduced model and the full solution. The same conclusion can be drawn when looking at major and minor species profiles as shown on Figure 11b-d.



(a) (b)

(c) (d)

Figure 11: Temperature (a) and species mass fraction (b-d) as a function of the residence time (San Diego), with the solid line representing the full solution. The $'+'$ markers represent the results for the PC-L-GPR model with single conditionig and the $'*'$ markers show the solution using PC-L-GPR with double conditioning.

6.3. Transient behaviour

The reduced model generation using the PC-GPR approach is now validated in a transient system. An accurate representation of the transient solution is also essential in order to guarantee reliable results. Figure 12 shows the temporal evolution of temperature and some species mass fraction for the methane case, with a residence time inside the reactor of $2 \cdot 10^{-5}s$. As previously, the reactor was initialized at the chemical equilibrium solution at constant enthalpy and pressure. It can be observed that temperature and species mass fractions are accurately predicted in time by the PC-GPR model, using only 2 PCs out of 35. Figure 13 shows the transient solution for the propane case, using the Polimi mechanism, with a residence time inside the reactor of $1 \cdot 10^{-5}s$. The temperature and species mass fraction profiles are shown for the full model and the PC-score with local GPR model, respectively. The reduced model is able to provide a very accurate representation of the transient evolution within the reactor, as for the methane case, using only 2 PCs out of 162. The ability of the reduced approach to reproduce the unsteday evolution of the chemical state using complex chemistry is very important towards its application in realistic turbulent combustion simulations.

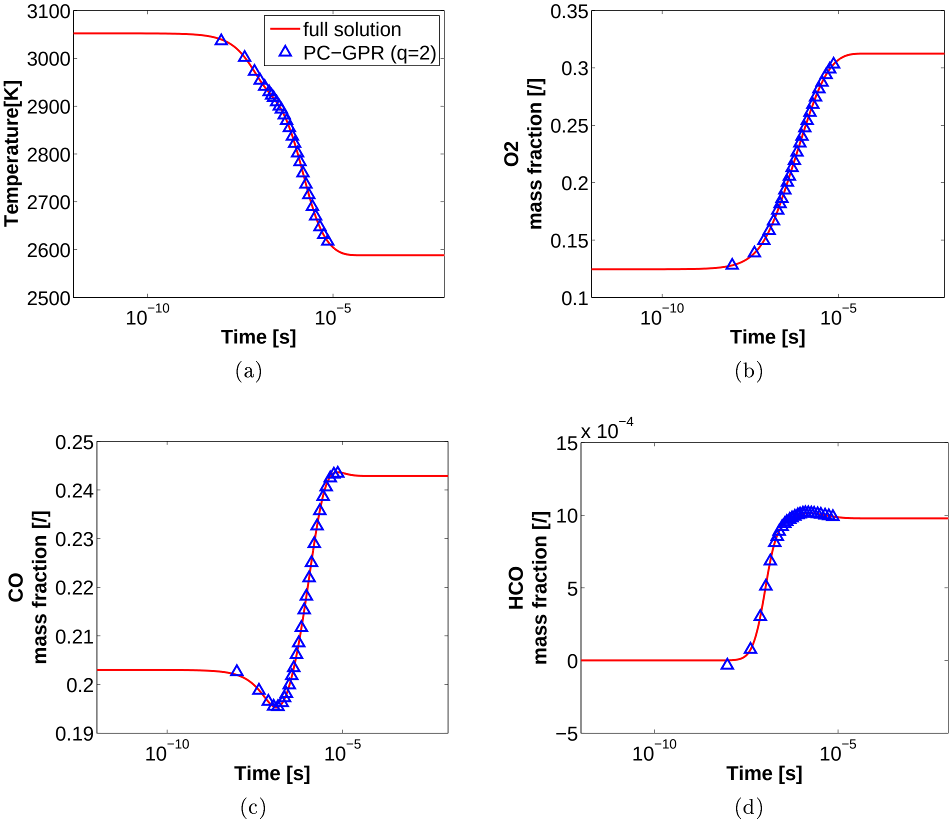
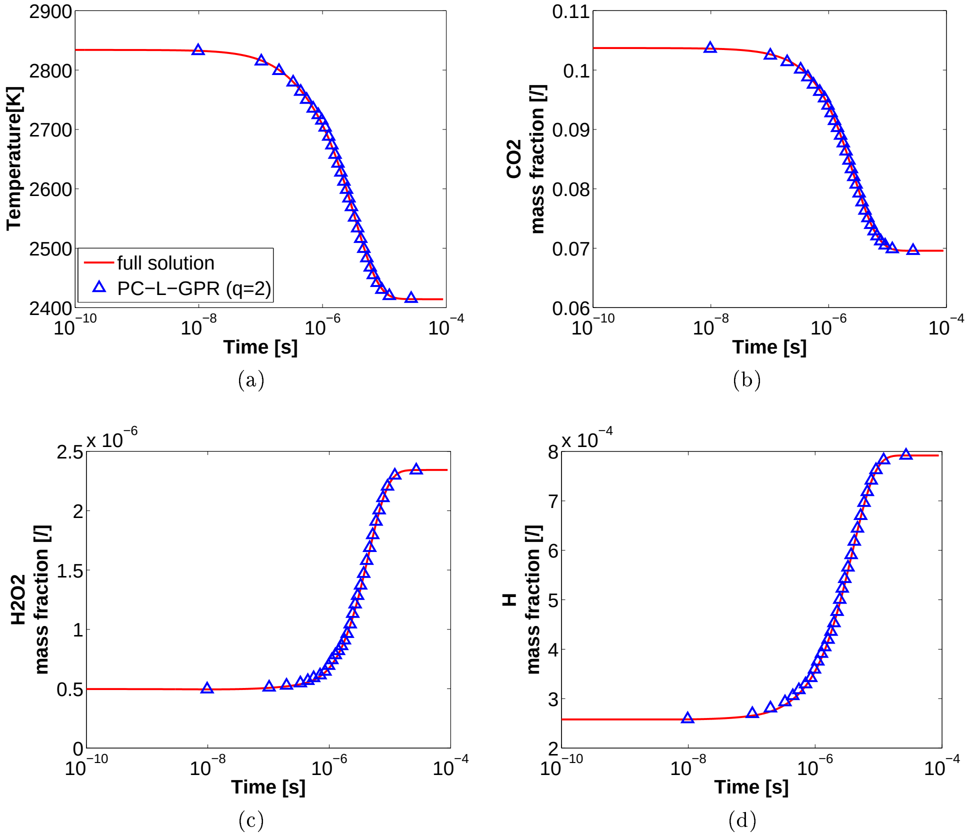


Figure 12: PSR temperature (a) and major and minor species (b-d) as a function of time (methane case), for a residence time of $2 \cdot 10^{-5} s$, with the solid line representing the full solution. The markers represent the results for PC-score with GPR regression using q = 2 PCs

7. Conclusion

The present work investigates the applicability of the PC-transport approach, focusing on the application of nonlinear regression to provide an accurate and compact parameterization of the thermo-chemical state. Steady and unsteady perfectly stirred reactor (PSR) calculations were carried out using the PC-transport approach, coupled to Gaussian Process Regression (GPR), for two different fuels (methane and propane) and three different kinetic mechanisms of increasing complexity.

The PC-GPR model showed its ability to produce very accurate representation of all state space variables, including temperature, major and minor species and source terms, using only a reduced number of principal components. In particular, for methane, the use of GPR allows to model accurately the system with only $q = 2$ principal components instead of the 34 variables in the original GRI-3.0 kinetic mechanism. For propane, the same approach lead to a very significant reduction, from 50 species, when using the San Diego mechanism, and 162 species, when using the Polimi mechanism, to only 2 PCs.



(a)

(b)

(c)

(d)

Figure 13: PSR temperature (a) and major and minor species (b-d) as a function of time (propane case, Polimi mechanism), for a residence time of $1 \cdot 10^{-5} s$, with the solid line representing the full solution. The markers represent the results for PC-score with local GPR regression using q = 2 PCs

Moreover, the application of the PC-transport model using local nonlinear regression (PC-L-GPR) was demonstrated. The use of local regressions within bins improved the accuracy of the PC-GPR approach while decreasing the computational cost associated to the generation of the reduced model. In particular, the use of PC-L-GPR provided an optimized mapping of the thermo-chemical state and the corresponding source terms.

Acknowledgements

The last author would like to acknowledge the support of Fédération Wallonie-Bruxelles, via “Les Actions de Recherche Concertée (ARC)” call for 2014–2019, to support fundamental research.

References

[1] G. P. Smith, D. M. Golden, M. Frenklach, N. W. Moriarty, B. Eiteneer, M. Goldenberg, C. T. Bowman, R. K. Hanson, S. Song, W. C. G. Jr., V. V. Lissianski, Z. Qin, 1999.

[2] T. Echekki, H. Mirgolbabaei, Combust. Flame 162 (2015) 1919–1933.

[3] J. Einbeck, B. Isaac, L. Evers, A. Parente, in: Proceedings of the 27th International Workshop on Statistical Modelling.

[4] B. Isaac, J. Thornock, J. Sutherland, P. Smith, A. Parente, Combust. Flame 162 (2015) 2592–2601.

[5] A. Parente, J. C. Sutherland, L. Tognotti, P. J. Smith, Proceedings of the Combustion Institute 32 (2009) 1579–1586.

[6] A. Parente, J. C. Sutherland, Combust. Flame 160 (2013) 340–350.

[7] S. B. Pope, Proc. Combust. Inst. 34 (2013) 1–31.

[8] J. C. Sutherland, A. Parente, Proceedings of the Combustion Institute 32 (2009) 1563–1570.

[9] A. Biglari, J. C. Sutherland, Combustion and Flame 159 (2012) 1960–1970.

[10] Y. Yang, S. Pope, J. Chen, Combust. Flame 160 (2013) 1967–1980.

[11] H. Mirgolbabaei, T. Echekki, Combust. Flame 160 (2013) 898–908.

[12] I. Jolliffe, Principal Component Analysis, Springer-Verlag New York, 2002.

[13] A. C. ans B. Isaac, O. Gicquel, A. Parente, Combust. Flame 168 (2016).

[14] A. Biglari, J. C. Sutherland, Combust. Flame 162 (2015).

[15] A. Coussement, O. Gicquel, A. Parente, Proc. Combust. Inst. 34 (2013) 1117–1123.

[16] W. Cleveland, E. Grosse, W. Shyu, Statistical models 79 (1992) 531–554.

[17] J. Friedman, The Annals of Statictics 19 (1991) 1–67.

[18] H.-T. Pao, Expert Systems with Applications 35 (2008) 720–727.

[19] A. Smola, B. Scholkopf, Statistics and Computing 14 (2004) 199–222.

[20] C. Rasmussen, Gaussian processes for machine learning, 2006.

[21] D. Nguyen-Tuong, M. Seeger, J. Peters, Advanced Robotics 23 (2009) 2015–2034.

[22] D. Goodwin, Cantera: An object-oriented software toolkit for chemical kinetics, thermodynamics, and transport processes, 2009.

[23] F. A. Williams, Chemical-kinetic mechanisms for combustion applications, 2010.

[24] S. Humer, A. Frassoldati, S. Granata, T. Faravelli, E. Ranzi, R. Seiser, K. Seshadri, Proceedings of the Combustion Institute 31 (2007) 393–400.

[25] B. Isaac, A. Coussement, O. Gicquel, P. Smith, A. Parente, Combust. Flame 161 (2014).

[26] K. Peerenboom, A. Parente, T. Kozák, A. Bogaerts, G. Degrez, Plasma Sources Sci. T. (2015).

[27] A. Bellemans, T. Magin, A. Munafò, G. Degrez, A. Parente, Physics Plasmas (2015).