Combustion and Flame 162 (2015) 2592–2601

Contents lists available at ScienceDirect

Combustion and Flame

journal homepage: www.elsevier.com/locate/combustflame

Advanced regression methods for combustion modelling using principal components

Benjamin J. Isaac [a,b,*], Jeremy N. Thornock [a], James Sutherland [a], Philip J. Smith [a], Alessandro Parente [b]

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

ARTICLE INFO

Article history:
Received 4 August 2014
Received in revised form 11 March 2015
Accepted 12 March 2015
Available online 8 April 2015

Keywords:
Combustion
Nonlinear regression
Low-dimensional manifolds
Principal Component Analysis
Reacting flows
Reduced-order modelling

ABSTRACT

Modelling the physics of combustion remains a challenge due to a large range of temporal and physical scales which are important in these systems. Detailed chemical kinetic mechanisms are used to describe the chemistry involved in the combustion process yielding highly coupled partial differential equations for each of the chemical species used in the mechanism. Recently, Principal Components Analysis (PCA) has shown promise in its ability to identify a low-dimensional manifold describing the reacting system. Several PCA-based models have been developed which may be well-suited for combustion problems; however, several challenging aspects of the model must be addressed. In this paper, the parameterization of state-space variables and PC-transport equation source terms are investigated. The ability to achieve highly accurate mapping through various nonlinear regression methods is shown. In addition, the effect of PCA-scaling on the ability to regress the surface is investigated. Finally, the present work demonstrates the capabilities of the model by solving a reduced system represented by several PC-transport equations for a perfectly stirred reactor (PSR) configuration.

1. Introduction

The ability to accurately model a turbulent combustion system remains challenging due to the complex nature of combustion systems. A simple fuel such as CH_4 requires 53 species and 325 chemical reactions [1] to be accurately described. More complex fuels require increasingly complex chemical mechanisms. Each resolved chemical species requires a conservation equation which is a coupled, nonlinear partial differential equation. Such systems are only possible to solve under very limited situations at this time due to computational costs. Current computational expenses result in a need for reduced models which can adequately describe the chemical reactions. Many methods attempt to reduce the complexity of the mechanism by splitting the system into slow and fast variables, using equilibrium assumptions for fast chemical processes, and occupying the computational resources on the more pertinent evolution of species within the reacting system [2,3]. Indeed, in these complex combustion reaction mechanisms many of the species evolve at time-scales much larger than the time-scales of interest, allowing for decoupling of fast and slow processes while maintaining accuracy. Low-dimensional manifolds exist in these systems which can describe the governing characteristics of the flames. Several models take advantage of this, including the steady laminar flamelet model (SLFM) [4–6], flamelet-generated manifolds (FGM) [7,8], or the flame prolongation of ILDM (FPI) [9–11] to name a few. As a fundamental example, the steady laminar flamelet model uses the mixture fraction and mixture fraction variance to describe the flame as an ensemble of steady laminar diffusion flames undergoing various strain rates. In some cases, this provides a good representation of the entire system with a reduced number of variables.

Recently, principal component analysis (PCA) has been investigated for its use in combustion modelling. Several advantages of PCA include: its ability to identify orthogonal variables which are the best linear representation of the system; its ability to reduce in dimensionality requiring fewer coordinates; and the ability to do the analysis on canonical systems, such as the counter diffusion flames or empirical data-sets containing highly complex turbulent chemistry interaction. Parente et al. [12,13] used PCA to identify the low-dimensional manifold in one-dimensional turbulence and experimental data. Biglari and Sutherland [14] and Yang and Pope [15,16] enhanced the capability of the PCA concept by combining the analysis with nonlinear regression, allowing a nonlinear mapping between state-space variables and the linear PCA basis.

* Corresponding author at: 155 South 1452 East, Room 350, Salt Lake City, UT 84112, USA.
E-mail addresses: Benjamin.J.Isaac@utah.edu (B.J. Isaac), J.Thornock@utah.edu (J.N. Thornock), James.Sutherland@utah.edu (J. Sutherland), Philip.Smith@utah.edu (P.J. Smith), Alessandro.Parente@ulb.ac.be (A. Parente).

http://dx.doi.org/10.1016/j.combustflame.2015.03.008
0010-2180/

The work of Biglari and Sutherland showed that the PC parameterization is superior to the standard flamelet parameterization, for the ODT data-set investigated in the study. Mirgolbabaei and Echekki [17] extended the nonlinear mapping concept using artificial neural networks and investigated the potential of kernel PCA [18,19], showing the high compression potential derived by transforming the initial problem into a non-linear featured space where linear PCA is carried out. In addition, several combustion models have been proposed based on the concepts from PCA. Sutherland and Parente [20] derived transport equations for the principal components (PCs), and discussed the feasibility of a model where the PCs are used directly to construct state-space variables. Biglari and Sutherland [14] extended the concept of transporting the PCs by suggesting the nonlinear regression in order to increase the accuracy and reducibility of the model. Coussement et al. [21], Isaac et al. [22] and other groups [23] proposed transporting a reduced set of state-space variables and used the PC basis for reconstructing the variables which are not represented. Najafi-Yazdi et al. [24] used PCA to identify optimal progress variables to use the flamelet-generated manifold framework.

The present work seeks to advance the understanding and application of the PC-transport approach of Sutherland and Parente [20,14] by first analyzing the effect of several scaling methods on the PC basis, and the resultant ability to regress the nonlinear state-space variables to the PC basis. Various regression methods used in previous studies [14,17], as well as several alternative methods are analyzed in their ability to approximate the reacting state-space from the PCs. In order to demonstrate the accuracy of the method within a numerical solver, an unsteady perfectly stirred reactor (PSR) calculation is shown using the PC-transport approach. The PSR provides a validation of the approach by comparing the reduced model to the detailed simulation results. To the authors knowledge all published analysis on the PC-transport concept using nonlinear regression has been carried out on various data-sets using *a priori* analysis [14,17,19,18]. Only recently, *a posteriori* work has begun in this area. Specifically, the work of Mirgolbabaei [25], who provides an *a posteriori* demonstration of the nonlinear PC-transport approach using one-dimensional turbulence (ODT) simulations.

2. Theory

A principal component analysis is done by taking a data-set consisting of n observations and Q independent variables and organizing it as an $n \times Q$ matrix ($\mathbf{X}$). The data $\mathbf{X}$ is centered to zero by its corresponding means $\bar{\mathbf{X}}$, and scaled by the diagonal matrix, $\mathbf{D}$, containing a scaling value for each of the k variables:

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

For sake of simplicity, $\mathbf{X^s}$ will be simply indicated as $\mathbf{X}$ in the following. In a PC analysis, the principal components ($\mathbf{Z}$) are identified by performing an eigenvalue decomposition of the covariance matrix of $\mathbf{X}$:

$$\frac{1}{Q-1}\mathbf{X}^T\mathbf{X} = \mathbf{A}^{-1}\mathbf{L}\mathbf{A} \tag{2}$$

The eigenvector matrix $\mathbf{A}$ (referred to here as a 'basis matrix') is then used to project the original state-space into PC space:

$$\mathbf{Z} = \mathbf{X}\mathbf{A} \tag{3}$$

Now given a subset of the basis matrix $\mathbf{A}$, denoted as $\mathbf{A}_q$ and applying the previous equation, an approximation of the original centered and scaled state-space can be made using the following:

$$\mathbf{X} \approx \mathbf{Z_q}\mathbf{A}_q^T. \tag{4}$$

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. Accordingly, when one truncates the basis matrix ($\mathbf{A}_q$), the resultant approximation from Eq. (4) may be very accurate, while representing the system with fewer variables.

In the work of Sutherland and Parente [20], a combustion model is proposed where conservation equations for the PCs are derived from the general species transport equation [26]:

$$\frac{\partial}{\partial t}(\rho Y_k) + \frac{\partial}{\partial x_i}(\rho u_i Y_k) = \frac{\partial}{\partial x_i}\left(\rho D_k \frac{\partial Y_k}{\partial x_i}\right) + R_k \tag{5}$$

where R_k is the net production rate of species k. One can easily derive the transport equations for the PCs ($\mathbf{Z}_q$) given the basis matrix $\mathbf{A}_q$, the scaling vector d_k, being the diagonal components of $\mathbf{D}$, and the centering vector $\bar{Y}_k$:

$$\frac{\partial}{\partial t}(\rho Z_q) + \frac{\partial}{\partial x_i}(\rho u_i Z_q) = \frac{\partial}{\partial x_i}\left(\rho D_{Z_q}\frac{\partial}{\partial x_i}(Z_q)\right) + s_{Z_q} \tag{6}$$

$$s_{Z_q} = \sum_{k=1}^{Q}\frac{R_k}{d_k}A_{kq} \tag{7}$$

where s_{Z_q} is simply the net production rate of the principal component. The term $D_{Z_q}\frac{\partial}{\partial x_i}(Z_q)$ is the diffusion flux for the principal component. For a more detailed discussion on the treatment of the PCs diffusive flux, where molecular diffusion is important refer to [27]. According to the proposed formulation, one can theoretically use PCA with its inherent advantages. These advantages include: the ability to represent the system with a reduced number of variables; the option to include a predetermined amount of reconstruction error (dependent on q, the number of retained PCs), and possibly a reduction in stiffness if the selected PCs are highly weighted with reacting species that change more slowly, such as the major species.

In order to use PCA to its fullest potential, several aspects of PCA must be studied. One of these aspects, is how the data is scaled (Eq. (1)). The various effects of scaling have been studied previously in [14,28,22]. The same approach has been followed in the present paper to find the best scaling option for the present application of PCA, using a data-set which exhibits physics of interest. A one-dimensional turbulence (ODT) data-set of a non-premixed synthesis/air jet has been considered here [29,30]. The simulation includes 11 chemical species [31] (H_2, O_2, O, OH, H_2O, H, HO_2, CO, CO_2, HCO, N_2), and 21 chemical reactions and it is initialized with a temperature of 500 K, with air as the oxidizer (0.7241 N_2 and 0.2759 O_2 by mass) and a fuel stream containing 0.0078 H_2, 0.5511 CO, and 0.4411 N_2 by mass. The ODT realizations are saved on a uniform grid of 672 grid points evenly spaced over a 0.01 m domain. The velocity field is initialized with a Reynolds numbers of 2500. The ODT data-set is particularly interesting because of the turbulence/chemistry interaction observed in the data, including physical effects such as extinction and re-ignition. Similarly to previous investigations [14,28,22], the *a priori* analysis showed that pareto scaling has a distinct advantage for major species and source terms reconstruction.

The *a priori* analyses showed, however, that at least 8 PCs were required to accurately reconstruct the ODT data-set and the corresponding source terms, due to the linear nature of the PC-based model. Considering the original 11 degrees of freedom of the system (with differential diffusion, enthalpy and elemental mass fractions are not constant), $q = 8$ implies only a minor problem reduction. An alternative to the direct reconstruction of $\mathbf{X}$ is to use 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. Biglari and Sutherland [14] suggest applying a nonlinear mapping to the linear

underlying surface by using nonlinear regression. It has been shown [14,17,15,18,19] that nonlinear regression allows to fully exploit the underlying manifold identified by the PCs. It is important to note that the linear basis derived from the PCs is critical as it allows for the derivation of simple transport equations; however, by using nonlinear functions on top of the basis, the model can capture the nonlinearities which are present in combustion systems.

2.1. Regression models

In this study, nonlinear regression is used to model the highly nonlinear state-space variables as a function of the principal components ($\mathbf{Z}$). In place of Eq. (4), now the various state-space variables and PC source terms ($\mathbf{s_Z}$) are mapped to the PC basis using the nonlinear regression function f_{Φ}:

$$\mathbf{\Phi} \approx f_{\Phi}(\mathbf{Z_q}) \tag{8}$$

where $\mathbf{\Phi}$ represents the state-space variables, or in terms of regression, the dependent variables (i.e. Y_i, T, ρ, and, $\mathbf{s_Z}$).

Until now, two nonlinear regression methods have been applied to mapping $\mathbf{\Phi}$ to $\mathbf{Z}$. In the work of Biglari and Sutherland [14] and Pope [15], multivariate adaptive regression splines are used. In the work of Mirgolbabaei and Echekki [17,18], artificial neural networks are investigated. Here, in addition to previously used regression techniques, several other methods are investigated, including support vector regression [32], and gaussian process regression [33,34]. In summary, the following regression techniques are investigated:

- *Linear Regression Model* (LIN).
  The linear model applied in multiple dimensions is of the form:
  $$\mathbf{\Phi} = \mathbf{Z}a + v \tag{9}$$
  where a is the regression coefficient vector and v is the intercept vector [35]. The implementation for the linear model found in the statistical computing software R [36] was used for the regression analysis.
- *Multivariate Adaptive Regression Splines* (MARS).
  Multivariate adaptive regression splines use the concept of building up the model from product spline basis functions. This model creates a number of basis functions, and automatically determines knot location and implements splines at knot boundaries. The model is of the form:
  $$\mathbf{\Phi} = \sum_{m=1}^{M} a_m B_m(\mathbf{Z}) \tag{10}$$
  where B_m are the basis functions and a_m are the expansion coefficients [37]. The implementation of MARS, found in the mda package of the statistical computing software R [36], was used for the regression analysis. The default options for MARS were used. The mda package determines the degree of the polynomials as well as the number of knot boundaries, given user settings such as: degree (default is 1, specifying the interaction degree), threshold (default is 0.001), and penalty (default is 2, specifying the cost per degree of freedom charge).
- *Artificial Neural Networks* (ANN).
  Artificial neural networks uses the concept of networking various layers of estimation resulting in a highly accurate output layer. Following the theory of Pao [38], the model works as follows: first, t hidden networks (NET_t) are calculated as a weighted (w_t) sum of the training data inputs ($k_i = [\mathbf{Z}, \mathbf{\Phi}]$):
  $$\mathrm{NET}_t = \sum_{i=1}^{N} w_{ti} k_i + b_i \tag{11}$$
  A sigmoid transfer function is then used to generate an output for the network:
  $$Z_t = [1 + exp(-\mathrm{NET}_t)]^{-1} \tag{12}$$
  Next, the output networks are calculated:
  $$\mathrm{NET} = \sum_{t=1}^{h} v_t Z_t + b_o \tag{13}$$
  Again, the network is scaled and a prediction of $\mathbf{\Phi}$ is then given:
  $$\mathbf{\Phi} = [1 + exp(-\mathrm{NET})]^{-1} \tag{14}$$
  In the present study, the implementation of ANN (ANNGA) in R [36] was used. One hidden layer with 20 neurons and one additional neuron in the output layer were used for the design, 1000 chromosomes for the population of each generation, a mutation rate of 0.2 was used, and crossover rate of 0.6.
- *Support Vector Regression* (SVR).
  Support vector regression is a subset of support vector machines (SVM). The idea behind SVR is again to create a model which predicts $\mathbf{s_Z}$ given $\mathbf{Z}$ using learning machines which implement the structural risk minimization inductive principle. The basic model form is
  $$\mathbf{\Phi} = \sum_{i=1}^{N} (\alpha_i^* - \alpha_i) K(\mathbf{Z}_0, \mathbf{Z}_i) \tag{15}$$
  where α_i^* and α_i are Lagrange multipliers, and $K(\mathbf{Z}_0, \mathbf{Z}_i)$ is the kernel operator. In the current study, a radial-based kernel was used and the optimum kernel hyper-parameter as well as the insensitive-loss function were determined by doing various calculations over a range of input parameters. The implementation of SVM within the e1071 package for R was used for the regression analysis of SVR. The kernel hyper parameter gamma was optimized by running a series of SVM fits over a range of values ($exp(-3)$ to $exp(3)$), a value of 1e-3 was used for epsilon, and the cost was optimized by running over a range of values ($exp(-3)$ to $exp(3)$).
- *Gaussian Process Regression* (GPR).
  Gaussian process regression is founded on the idea that dependent variables can be described by a gaussian distribution [33,34]:
  $$\mathbf{\Phi} \sim N(0, \mathbf{K}(\mathbf{Z}, \mathbf{Z}) + \sigma_n^2 \mathbf{I}) \tag{16}$$
  Here $\mathbf{Z}$ is the data matrix containing all sample points in PC space; $\mathbf{K}(\mathbf{Z}, \mathbf{Z})$ is the kernel function for $\mathbf{Z}$; in the current study, the gaussian kernel is used:
  $$\mathbf{K}(\mathbf{Z_p}, \mathbf{Z_q}) = \sigma_s^2 exp\left(-\frac{1}{2}(\mathbf{Z_p}, \mathbf{Z_q})^T \mathbf{W}(\mathbf{Z_p}, \mathbf{Z_q})\right) \tag{17}$$
  Given query points $\mathbf{Z}_*$ it can be shown that a prediction $\mathbf{\Phi}_*$ can be made using the following formula:
  $$\mathbf{\Phi}_* = \mathbf{K}_*^{\mathbf{T}} (K + \sigma_{\mathbf{n}}^{\mathbf{2}} \mathbf{I})^{-1} \mathbf{\Phi} \tag{18}$$
  where $\mathbf{K}_* = \mathbf{K}(\mathbf{Z}, \mathbf{Z}_*)$ and $\mathbf{K} = \mathbf{K}(\mathbf{Z}, \mathbf{Z})$. A value of 1 was used as the initial guess for the kernel's hyper-parameters: the characteristic length scale, and signal variance. A gradient-based marginal likelihood optimization was used find the optimal values. The GPR implementation from the MATLAB toolbox gpml [34] was used for the regression analysis of GPR. The hyper parameters were found using the gradient-based marginal likelihood functions in the toolbox.

In order to map the highly nonlinear reaction rate surface (dependent variables) to PC space (independent variables) it is useful to understand how nonlinear the reaction rates and other state-

space variables are with respect to the underlying manifold represented by the principal components. A simple way to do this in multiple dimensions is to divide the independent variable space onto a coarse grid, and assess locally the variation of dependent variables within a local section of the independent variable space. Locally, if the dependent variable has a large variation, then the ability to regress the dependent variable locally will be more difficult because of the nonlinear nature or even local scatter in the data. The following equation is used to calculate the locally normalized variance for the ith coarse grid cell (χ^i_{Φ}):

$$\chi^i_{\Phi} = \frac{v(\Phi(\mathbf{Z}^i_{\mathbf{q}}))}{v(\Phi(\mathbf{Z}_{\mathbf{q}}))} \tag{19}$$

where $v(x) = \left\langle (x - \langle x \rangle)^2 \right\rangle$ is the variance function which is calculated on the observations within the ith coarse grid cell ($\Phi(\mathbf{Z}^i_{\mathbf{q}})$) or for all observations ($\Phi(\mathbf{Z}_{\mathbf{q}})$). Now, summing over all course grid cells in PC space, we obtain the overall manifold nonlinearity for dependent variable Φ:

$$\chi_{\Phi} = \sum_{i=1}^{c} \chi^i_{\Phi} \tag{20}$$

Table 1 shows the manifold nonlinearity calculation for the various dependent variables in the ODT data-set mentioned previously. It is clear from the analysis that some scaling methods have distinct advantages for several of the dependent variables. In particular, pareto scaling has an advantage when comparing several major species (O_2, CO, CO_2, and N_2), temperature, and density, with a weaker performance for some of the radical species (OH, H). All methods show the regression for s_{Z_1} is challenging; however, the regression for s_{Z_2} appears promising with pareto scaling.

Given the results for both the state-space reconstruction and the manifold nonlinearity, it is clear that the pareto scaling method has some unique advantages for this particular data-set dealing with syngas combustion. Several other studies have reached similar conclusions with the pareto scaling method as shown in [21,22,28]. With this observation in mind, the various regression models are now tested with the pareto scaling method. The nonlinear regression analysis is done using a combination of computing software packages including the statistical computing software R [36], and MATLAB [39], as described previously. The R code implementations for LIN, MARS, ANN, and SVR were used. For GPR, the MATLAB toolbox gpml [34] was employed. The models are trained on $n = 5000$ sample points evenly distributed over Z space, with $q = 2$ or 3. The models are then tested on another subset of points of the same size, ensuring that training points are not used again as testing points. This is done to ensure that over-fitting is avoided.

Table 1
Manifold nonlinearity (χ_{Φ}) for state-space variables, Φ while using different scaling methods.

χ_{Φ}	Std	Range	Pareto	Vast	Level
H_2	5.7	11.4	10.8	12.3	3.5
O_2	4.0	1.9	0.3	0.7	4.9
O	12.6	11.8	17.2	28.8	7.5
OH	16.6	17.3	21.5	41.5	6.8
H_2O	6.1	5.1	4.9	5.3	7.0
H	14.6	22.3	30.0	46.1	5.2
HO_2	7.1	9.6	6.2	3.3	7.2
CO	2.4	1.3	0.1	1.8	1.7
CO_2	5.0	5.0	0.8	3.0	6.2
HCO	6.9	14.6	18.1	29.4	2.5
N_2	1.7	0.7	0.1	0.7	1.4
T	7.0	6.5	2.0	4.0	9.2
ρ	7.8	6.9	2.5	5.0	9.6
s_{Z_1}	256.5	292.2	300.5	404.0	210.1
s_{Z_2}	150.0	172.7	25.8	143.7	95.9

Table 2
Nrms error and R^2 statistics for the prediction of $\mathbf{s}_{Z_1}$ while using pareto scaling and $q = 2$ or $q = 3$.

Method	*nrms error* ($q = 2$)	R^2 ($q = 2$)	*nrms error* ($q = 3$)	R^2 ($q = 3$)
LIN	0.99	0.02	0.67	0.55
MARS	0.30	0.91	0.26	0.93
ANN	0.22	0.95	0.20	0.96
SVR	0.23	0.95	0.19	0.97
GPR	0.22	0.95	0.18	0.97

Table 2 shows the regression results for s_{Z_1} as a function of $\mathbf{Z}$, with $q = 2$ and $q = 3$, using normalized root mean squared error ($nrms(x_p, x)/max(\sigma(x_p), \sigma(x))$) or R^2 error ($\sum_{i=1}^{N}(x_{p,i} - \bar{x})^2 / \sum_{i=1}^{N}(x_i - \bar{x})^2$) metrics. As expected, the linear regression method has difficulty mapping the highly nonlinear dependent variables. Complex methods also struggle with the mapping while $q = 2$. Table 1 shows that s_{Z_1} is highly non-linear. One can easily conclude that methods such as linear regression will fail, polynomial methods such as MARS may also struggle given the degree of non-linearity. Methods which use local tuning (ANN, SVR, GPR) may be able to better approximate the problematic regions of the manifold. When moving to $q = 3$, the later 3 methods are beginning to show higher accuracy. In this particular case, GPR produces the most accurate reconstruction. The approximation shows a vast improvement especially if compared with the results of the direct computation (Eq. (7)), with the same level of accuracy being achieved with $q = 8$.

It is important to note that the results given in Table 2 are related to the specific implementation of the regression methods, as well as to any tuning or optimization that was performed for each method. Indeed, the results for the GPR regression may be optimal because of the robust optimization of the hyper-parameters that the implementation utilizes. The various regression methods may indeed improve given more tuning, or using different implementations. However, tuning the different regression methods is not the purpose of the present study. The focus of the present investigation is the benchmark of various non-linear approaches, based on state-of-the art implementations found in the literature.

Ultimately, the PC-Transport approach will be utilized within a CFD solver. Several factors are important in deciding which regression method to use. In addition to the methods accuracy, the methods ease of use, its applicability to different problems, its ability to optimize tuning parameters, and its expense within a CFD algorithm are important factors. Because of the numerous variations and implementations of the regression methods, general conclusions about the methods cannot be made. However, these factors can be addressed for the implementations used in the current study. Table 3 summarizes these factors for the various regression methods.

Table 3
Summary of the relative accuracy, ease of use, applicability to problems of a certain size, optimization, and relative cost for the various regression methods. A scale, ranging from 1 to 3 is used to rank the regression methods, 1 representing poor performance, and 3 excellent performance.

Method	Accuracy	Ease	Problem size	Optimization	Cost
LIN	1	3	3	–	3
MARS	2	2	3	2	3
ANN	3	1	3	2	2
SVR	3	1	1	2	1
GPR	3	1	1	3	1

While MARS and LIN are easier to use, the authors found the implementation of ANN, SVR and GPR the most difficult to use, due to the complexity of the methods and the various inputs required to use them. Both SVR and GPR methods employ *qxq* matrix inversions (q being the number of observations), which make the method slow with larger data sets. GPR often took the longest to run, but required the smallest amount of optimization work from the user due to the minimization functions, which optimize the methods hyper-parameters. As far as run-time costs, all methods except for SVR and GPR may be suitable. It is however possible to tabulate the regression results and use a simple table look-up to reduce the run-time costs associated with the expensive methods.

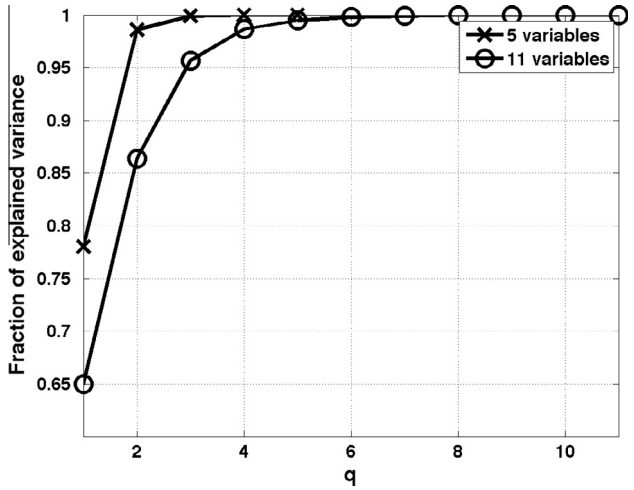

Fig. 1. Scree plot from the eigenvalue matrix, showing the fraction of explained variance (y-axis) as a function of the number of PCs (q) for the system containing a subset of the original species ('**x**' markers), and the full system ('**o**' markers).

Table 4
nrms error and R^2 statistics for the prediction of $\mathbf{\Phi}$ while using pareto scaling and $q = 2$.

$\mathbf{\Phi}$	*nrms error*	R^2
H_2	0.05	0.997
O_2	0.04	0.999
O	0.06	0.996
OH	0.07	0.995
H_2O	0.06	0.997
H	0.05	0.997
HO_2	0.17	0.969
CO	0.05	0.998
CO_2	0.05	0.997
HCO	0.03	0.999
N_2	0.04	0.998
T	0.04	0.998
ρ	0.04	0.999
s_{Z_1}	0.22	0.949
s_{Z_2}	0.16	0.974

Table 5
Eigenvector matrix, A, from the PC analysis.

Species weight	Z_1	Z_2	Z_3	Z_4	Z_5
H_2	0.047	0.117	−0.302	0.900	0.288
O_2	−0.627	0.119	−0.034	−0.230	0.734
H_2O	0.176	−0.186	0.895	0.222	0.292
CO	0.624	0.656	−0.040	−0.243	0.348
CO_2	0.431	−0.713	−0.325	−1.124	0.414

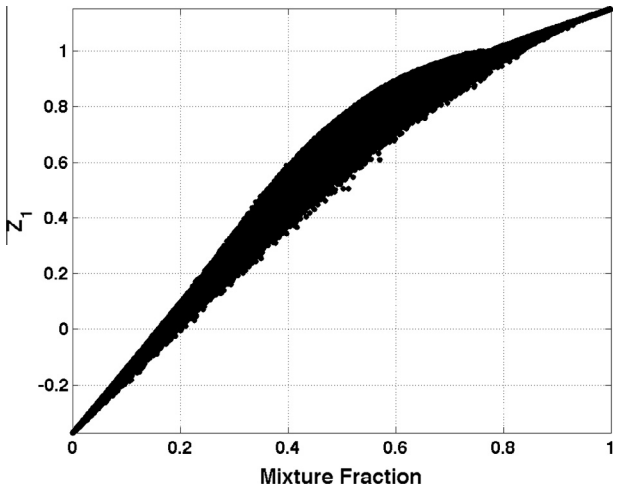

Fig. 2. A scatter plot of mixture fraction (x-axis) versus Z_1 (y-axis), illustrating the correlation between the variables.

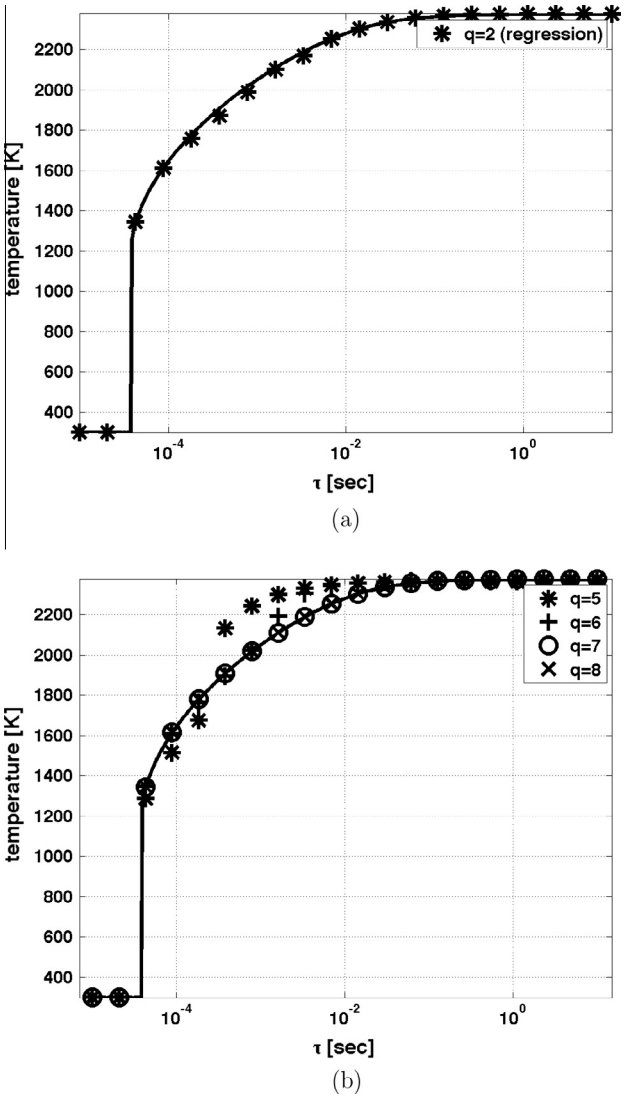

Fig. 3. PSR temperature as a function of the residence time, with the solid-line representing the full solution. The markers represent the results for the model with GPR regression (a) using $q = 2$ PCs, and the standard model without regression (b) while varying q.

2.2. Subset PCA

In the work of Mirgolbabaei and Echekki [17], the PCA analysis is done on a subset of species in order to recover sufficiently accurate source terms. This has the benefit of removing certain species which may be contributing highly nonlinear source terms to $\mathbf{s}_{Z_q}$. The drawback to doing this is that there is no guarantee that the underlying manifold computed from the subset will be able to adequately predict the species removed from the analysis. In the current study, the retained species are selected by choosing variables which tend to pertain to the slower chemical time-scales of the system, such as the major species. The following subset of species were selected for the present analysis: H_2, O_2, H_2O, CO and CO_2. With the selected subset of species, the PCA analysis is repeated, again with pareto scaling. Figure 1 shows the scree plot [40], which gives the percentage of variance accounted for while selecting q PCs. The figure compares the full PCA version using 11 variables and the subset PCA using 5. It is clear that the PCA based on the subset of variables represents the variation in the system with fewer variables.

Table 4 shows the error statistics for the entire set of state variables while using GPR and pareto scaling. It is interesting to note that even though several of these variables were not included in the analysis, the PCA basis computed from the major species in combination with the nonlinear regression is sufficient for mapping these highly nonlinear minor species.

The subset PCA also allows to more easily associate a physical interpretation to the PC structure. Table 5 shows the basis matrix weights from the PCA analysis on the major species. The weights from the first PC have large positive values for carbon containing variables (CO, CO_2), and a large negative value on the oxidizer (O_2). This appears to be very similar in nature to Bilger's mixture fraction [41], ξ. Figure 2 shows a plot of Z_1 against ξ; the plot shows that Z_1 is clearly correlated with ξ. The weights for Z_2 show positive correlations for H_2, O_2 and CO, with negative correlations for H_2O and CO_2. These weights appear to be related to the extent of reaction, where reactants have negative stoichiometric coefficients, and products have positive reaction coefficients. With a larger initial mass-based concentration of CO (compared with H_2), a large amount of CO_2 is produced, and a

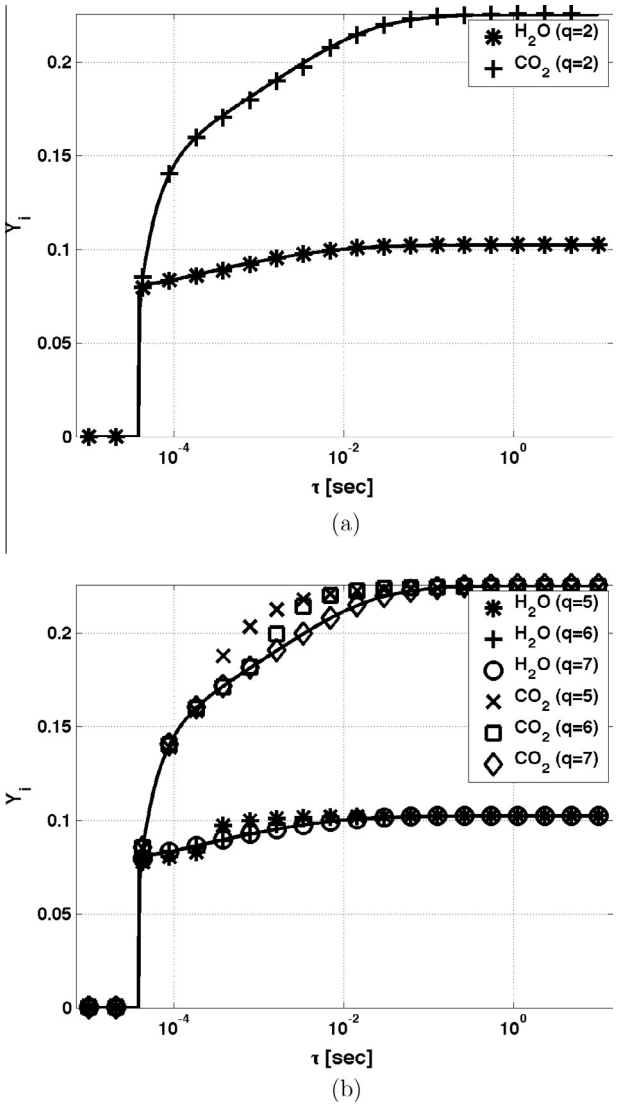

Fig. 4. Major species products as a function of the residence time, with the solid-line representing the full solution. The markers represent the results for the model with GPR regression (a) using $q = 2$ PCs, and the standard model without regression (b) while varying q.

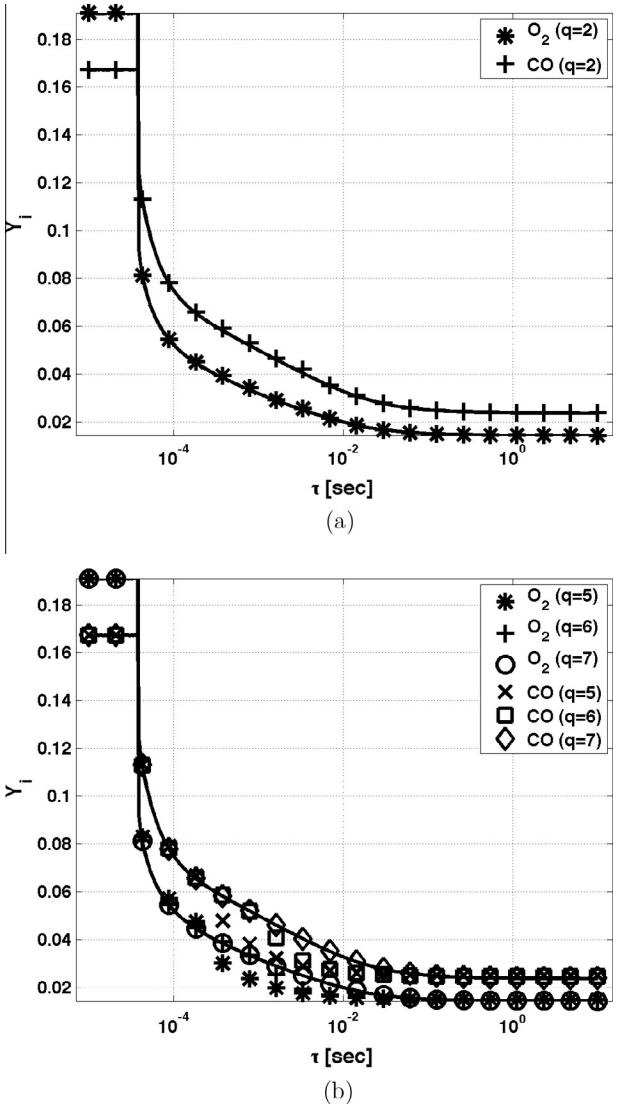

Fig. 5. Major species reactants as a function of the residence time, with the solid-line representing the full solution. The markers represent the results for the model with GPR regression (a) using $q = 2$ PCs, and the standard model without regression (b) while varying q.

much smaller amount of H_2 is present leading to a smaller positive weight on H_2 and smaller negative weight for the product H_2O. It is interesting to point out that without any prior understanding or assumptions of the combustion systems, the PC analysis is able to identify two important variables which are often used to characterize combustion systems.

It is evident that the linear PC model in conjunction with a nonlinear regression has the potential of delivering accurate state-space variables as well as relatively accurate reaction rates for the ODT data-set that has been studied in the current section.

3. Results and discussion

As a first step in advancing the PCA based models, a perfectly stirred reactor is used, which contains complexity in reaction space, without complexity from mixing. This system is ideal for demonstrating the approach as it is simple to implement, compute, and validate.

3.1. Perfectly stirred reactor

An implementation for the perfectly stirred reactor was made using MATLAB. The following governing equations were implemented and solved using the CVODE toolbox in MATLAB [42]:

$$\frac{d\rho H}{dt} = \frac{\rho}{\tau}H^0 - \frac{\rho}{\tau}H \tag{21}$$

$$\frac{d\rho Y_i}{dt} = \frac{\rho}{\tau}Y_i^0 - \frac{\rho}{\tau}Y_i + R_i W_{s,i} \tag{22}$$

where H is the mixture enthalpy, Y_i and R_i are the ith species mass fraction and molar reaction rate (kmole/m^3/s), τ (seconds) is a constant representing the residence time through the reactor, $W_{s,i}$ is the ith species molecular mass, and ρ is the density (kg/m^3). The temporal solution to the equations are solved using the Newton nonlinear solver, and the BDF multi-step method. The problem is initially solved using a stoichiometric mixture of syngas-air using the same mechanism which was used for the ODT data-set [31], where the mechanism includes 11 chemical species and 21 reactions. The inlet conditions for the reactor (Y_i^0) are set at an

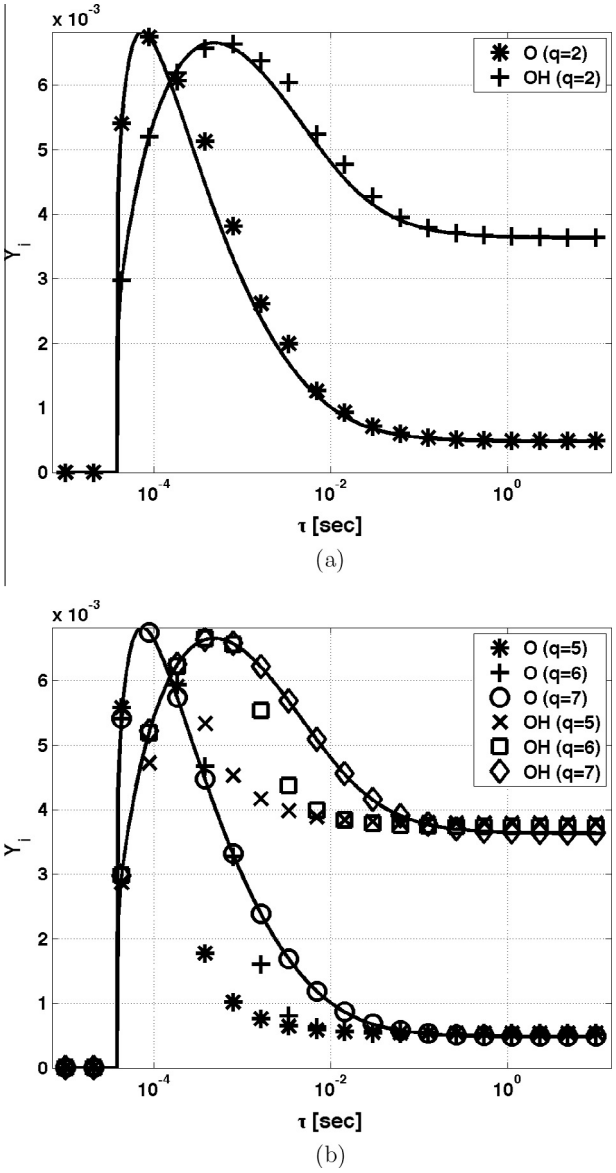

Fig. 6. Minor species as a function of the residence time, with the solid-line representing the full solution. The markers represent the results for the model with GPR regression (a) using $q = 2$ PCs, and the standard model without regression (b) while varying q.

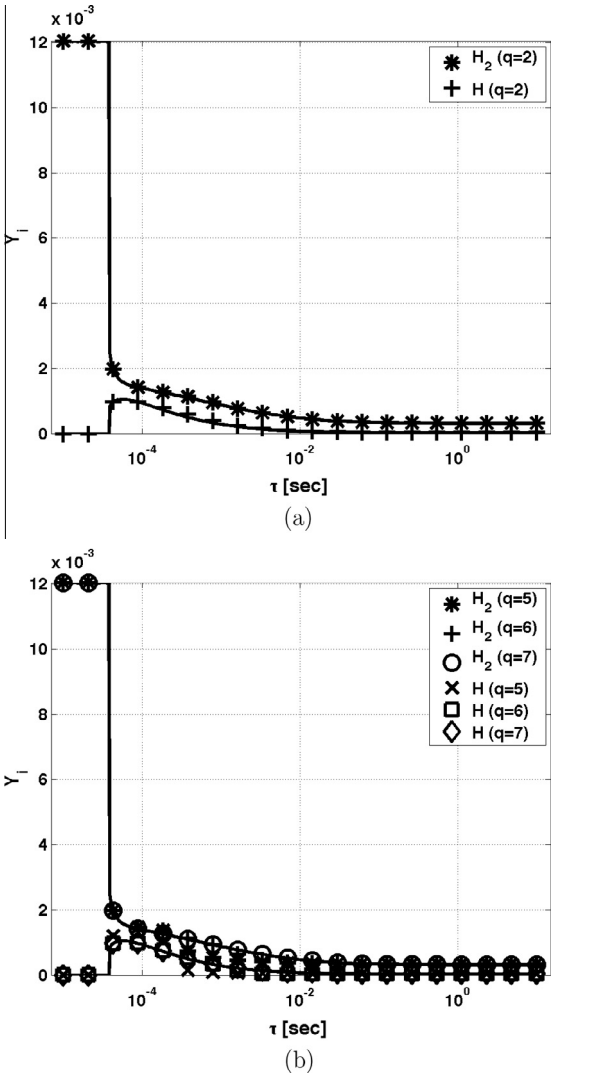

Fig. 7. Minor species as a function of the residence time, with the solid-line representing the full solution. The markers represent the results for the model with GPR regression (a) using $q = 2$ PCs, and the standard model without regression (b) while varying q.

equivalence ratio of 1 with a temperature of 300 K. The initial conditions for the reactor (Y_i) are set at chemical equilibrium using a Gibbs free energy minimization method (constant enthalpy and pressure). The elemental composition and enthalpy of the inlet mixture yield an equilibrium solution which is set as the initial condition for all of the PSR cases. The temporal solution of the system is then solved until a steady-state solution is reached. This process is repeated for various residence times between 10^{-5} and 10 s. All PSR simulations (including the transient solution) are then assembled into one data-set. The PCA process described in Section 2 is then applied to the data to create the basis matrix $\mathbf{A_q}$, and the regression functions f_Φ for the state-space variables, $\mathbf{\Phi}$. The approach is then tested with various values of τ, which were not used when creating the data-set.

The regression of $\mathbf{\Phi}$ is carried out using $q = 2$ resulting in R^2 of 0.9995 or higher for all variables including s_{Z_q}. The simulations are then performed with 2 transport equations instead of 11, yielding a significant reduction. Figures 3a, 4–8b show the temperature and species mass fractions of the system. The markers show the steady-state solution for a given τ using the PC-transport model. The underlying solid-lines in the figures show the full solution calculated over a range of residence times. The top plot (a) shows the results of the model using GPR for the nonlinear mapping with $q = 2$, and the results on the bottom (b) show the standard model without the regression step while varying q. 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 $q = 7$. In the current system, constant enthalpy and elemental mass is observed yielding 7 degrees of freedom, which would imply virtually no reduction due to the degrees of freedom.

Although the previous figures have shown the accuracy of the models for the steady-state solution, accurate representation of the transient solution is also essential. Figure 9 shows the transient solution for a reactor with a residence time of 10^{-4} s. Figure 9a shows the evolution of temperature and Fig. 9b the evolution of the OH radical mass fraction. The 'o' markers in the figures show the results for the regression method using only $q = 2$ PCs. As observed, an accurate transient solution is achieved given the significant reduction provided by the method.

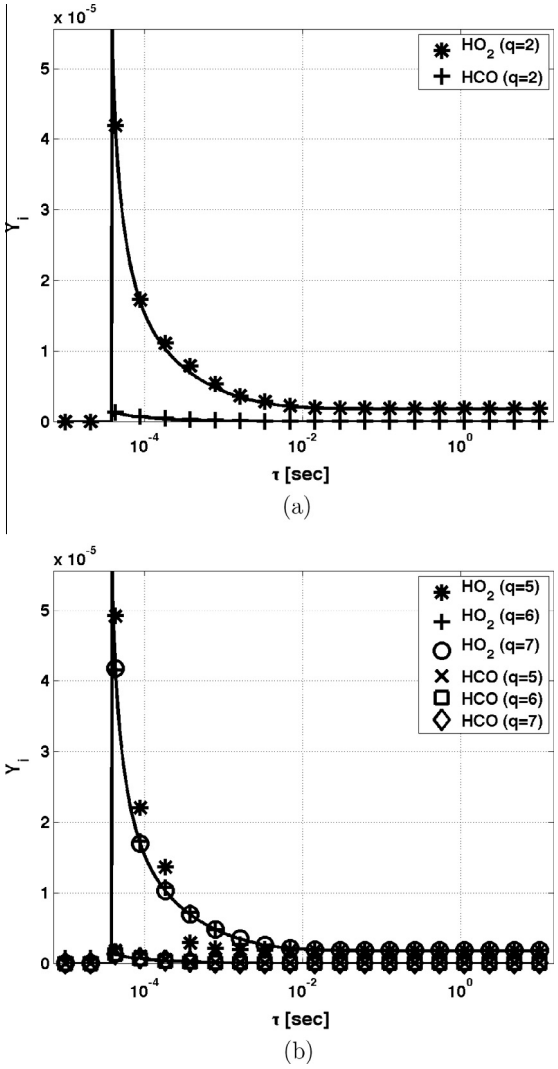

Fig. 8. Minor species as a function of the residence time, with the solid-line representing the full solution. The markers represent the results for the model with GPR regression (a) using $q = 2$ PCs, and the standard model without regression (b) while varying q.

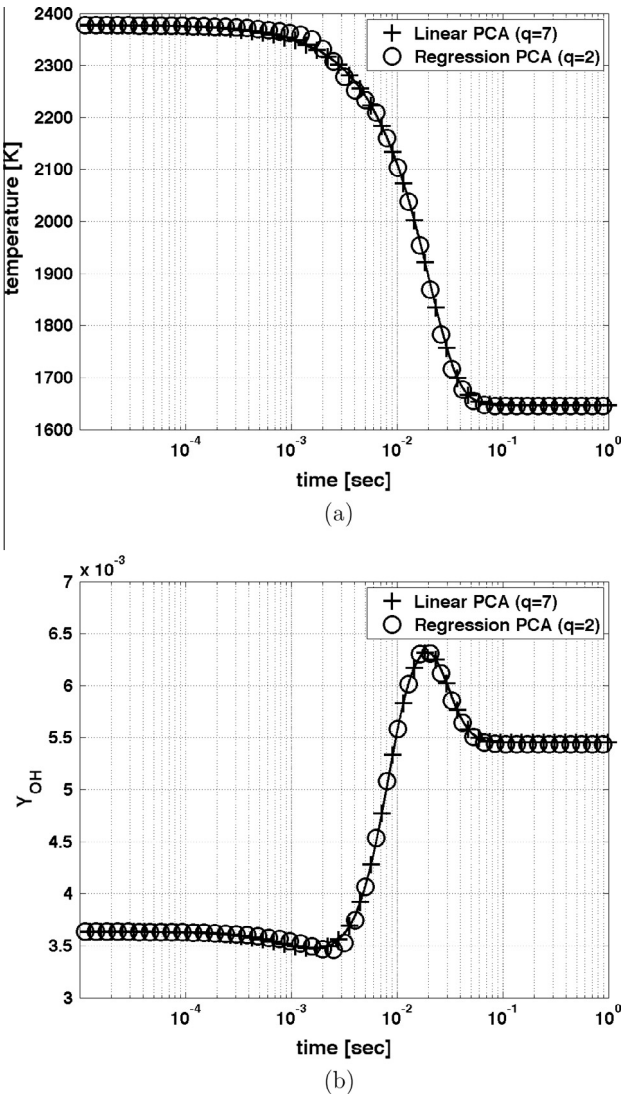

Fig. 9. Temperature [K] (a), and OH radical mass fraction (b) as a function of time. Given a residence time of 10^{-4} [s], and the chemical equilibrium solution (constant enthalpy and pressure) as the initial condition, the temporal evolution is shown. The solid-line represents the solution given the full system of equations. The markers represent the results for the either model, with 'o' markers for the solution using regression ($q = 2$ PCs), or '+' markers for the solution using the standard model ($q = 7$ PCs).

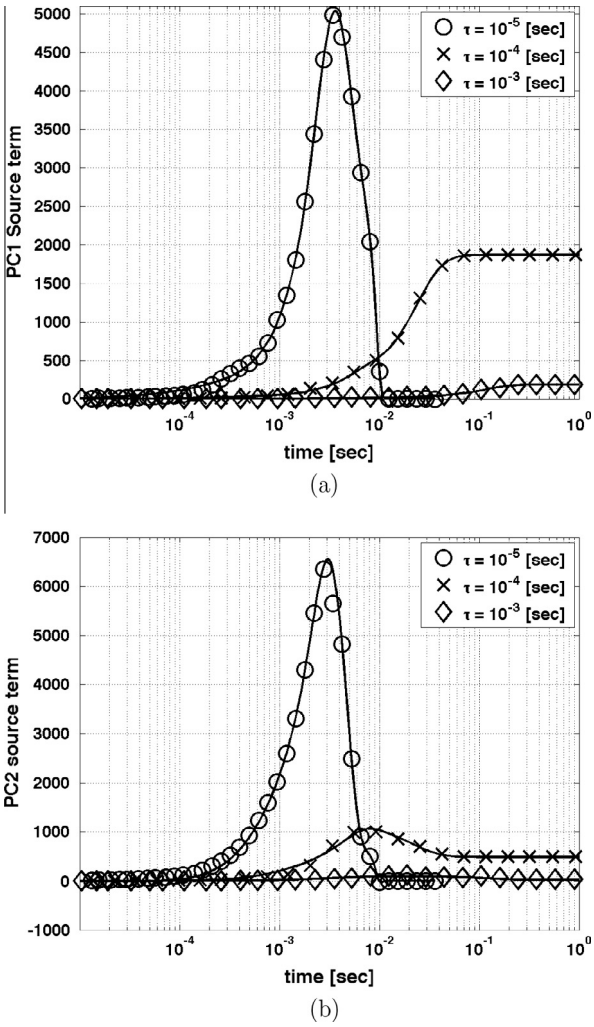

Fig. 10. Comparison of regressed PC source terms as a function of time, with (a) and (b) showing the results for the first and seconds PC source terms. Several cases are shown, with the following residence times: 10^{-5} s ('o' markers), 10^{-4} s ('x' markers), and 10^{-3} s ('◇' markers). The solid-line is the actual PC source term for the various residence times.

Accurate prediction of the PC transport source terms is essential to the PCA based model. In order to illustrate this, three cases with residence times of 10^{-5} s,10^{-4} s and 10^{-3} s were selected. The PC source terms with no approximation from the training data set are computed using: $s_Z = \frac{R}{d}A$. These source terms are then compared with the source-terms computed from the regression analysis at run-time. Figure 10 shows the transient results of the first and second PC source terms for the three different cases. It is evident that the regression method gives a good approximation of the actual source terms (indicated with the solid black line). As observed, both the first and second source terms are accurately predicted, temporally, by the regression method. One non-linear regression is able to accurately predict the source terms for three different residence times. These results indicate that the PCs yielded an optimal basis for regression, being able to parameterize the non-linear source terms.

4. Conclusion

The current work has addressed the ability to use nonlinear regression methods to estimate source-terms for the PC-transport combustion model. Various nonlinear regression methods have been analyzed showing the ability to produce accurate estimation, even when using a lower number of **Z**. In particular, the SVM and GPR methods have shown improved accuracy in estimating **Φ**. A method for defining the regressibility of a manifold has been presented. In addition, the effect of the various PCA-scaling methods on the regressibility of the system has been assessed. The pareto scaling method appears to achieve the greatest reduction with fewer components, and produces the most regressible surface. The current work outlines an example of an *a priori* analysis which provides the best regression and scaling method for a given turbulent combustion data-set.

The work includes the first demonstration of the PC-transport model using nonlinear regression within a numerical solver. In the case of the PSR, the model provided a computational reduction factor of 0.71, resulting in an accurate representation of the original system with $q = 2$ variables of the 7 degrees of freedom in the system. Future work will include a validation study, looking into how the approach compares with experimental values, and with other combustion models.

Acknowledgments

We are grateful to our sponsor for which part of the present research was funded: The National Nuclear Security Administration under the Accelerating Development of Retrofittable CO_2 Capture Technologies through Predictivity program through DOE Cooperative Agreement DE NA 00 00 740.

References

[1] G. Smith, D. Golden, M. Frenklach, N. Moriarty, B. Eiteneer, M. Goldenberg, C. Bowman, R. Hanson, S. Song, W. Gardiner Jr., et al., Gri-mechanism 3.0.
[2] R. Fox, Computational Models for Turbulent Reacting Flows, Cambridge University Press, 2003.
[3] W. Jones, R. Stelios, Combust. Flame 142 (2005) 223–234.
[4] N. Peters, Prog. Energy Combust. Sci. 10 (1984) 319–339.
[5] N. Peters, Proc. Combust. Inst. 24 (1986) 1231–1250.
[6] H. Pitsch, N. Peters, Combust. Flame 114 (1998) 26–40.
[7] J.v. Oijen, L.d. Goey, Combust. Sci. Technol. 161 (1) (2000) 113–137.
[8] J. Van Oijen, L. De Goey, Combust. Theor. Model. 6 (3) (2002) 463–478.
[9] O. Gicquel, N. Darabiha, D. Thévenin, Proc. Combust. Inst. 28 (2) (2000) 1901–1908.
[10] B. Fiorina, R. Baron, O. Gicquel, D. Thevenin, S. Carpentier, N. Darabiha, et al., Combust. Theor. Model. 7 (3) (2003) 449–470.
[11] B. Fiorina, O. Gicquel, S. Carpentier, N. Darabiha, Combust. Sci. Technol. 176 (5–6) (2004) 785–797.
[12] A. Parente, J.C. Sutherland, P.J. Smith, L. Tognotti, Proc. Combust. Inst. 32 (2009) 1579–1586.
[13] A. Parente, J.C. Sutherland, B.B. Dally, L. Tognotti, P.J. Smith, Proc. Combust. Inst. 33 (2) (2011) 3333–3341.
[14] A. Biglari, J.C. Sutherland, Combust. Flame. 159 (5) (2012) 1960–1970.
[15] S.B. Pope, Proc. Combust. Inst. 34 (2013) 1–31.
[16] Y. Yang, S.B. Pope, J.H. Chen, Combust. Flame 160 (10) (2013) 1967–1980.
[17] H. Mirgolbabaei, T. Echekki, Combust. Flame 160 (2013) 898–908.
[18] H. Mirgolbabaei, T. Echekki, Combust. Flame 161 (1) (2014) 118–126, http://dx.doi.org/10.1016/j.combustflame.2013.08.016. <http://www.sciencedirect.com/science/article/pii/S0010218013003209>.
[19] H. Mirgolbabaei, T. Echekki, N. Smaoui, Int. J. Hydrogen Energy 39 (9) (2014) 4622–4633, http://dx.doi.org/10.1016/j.ijhydene.2013.12.195. <http://www.sciencedirect.com/science/article/pii/S036031991303187X>.
[20] J. Sutherland, A. Parente, Proc. Combust. Inst. 32 (2009) 1563–1570.
[21] A. Coussement, O. Gicquel, A. Parente, Proc. Combust. Inst. 34 (2013) 1117–1123.
[22] B. Isaac, A. Coussement, O. Gicquel, P. Smith, A. Parente, Combust. Flame 161 (2014) 2785–2800.
[23] Y. Yang, S.B. Pope, J.H. Chen, Combust. Flame 160 (2013) 1967–1980.
[24] A. Najafi-Yazdi, B. Cuenot, L. Mongeau, Combust. Flame 159 (2012) 1197–1204.
[25] H. Mirgolbabaei, ow-dimensional manifold simulation of turbulent reacting flows using linear and nonlinear principal components analysis (Ph.D. thesis), North Carolina State University, 2014, <http://www.lib.ncsu.edu/resolver/1840.16/9479>.
[26] T. Poinsot, D. Veynante, Theoretical and Numerical Combustion, R.T. Edwards, Inc., 2001.
[27] T. Echekki, H. Mirgolbabaei, Combust. Flame 162 (2015) 1919–1933.
[28] A. Parente, J.C. Sutherland, Combust. Flame 160 (2013) 340–350.

[29] E.R. Hawkes, R. Sankaran, J.C. Sutherland, J.H. Chen, Proc. Combust. Inst. 31 (1) (2007) 1633–1640.
[30] N. Punati, J.C. Sutherland, A.R. Kerstein, E.R. Hawkes, J.H. Chen, Proc. Combust. Inst. 33 (1) (2011) 1515–1522.
[31] S.G. Davis, A.V. Joshi, H. Wang, F. Egolfopoulos, Proc. Combust. Inst. 30 (2005) 1283–1292.
[32] A.J. Smola, B. Schölkopf, Statist. Comput. 14 (3) (2004) 199–222.
[33] D. Nguyen-Tuong, M. Seeger, J. Peters, Adv. Rob. 23 (15) (2009) 2015–2034.
[34] C.E. Rasmussen, Gaussian processes for machine learning, Citeseer, 2006.
[35] W.S. Cleveland, E. Grosse, W.M. Shyu, Statist. Models S (1992) 309–376.
[36] R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2011. <http://www.R-project.org/>.
[37] J.H. Friedman, Ann. Statist. (1991) 1–67.
[38] H.-T. Pao, Expert Syst. Appl. 35 (3) (2008) 720–727.
[39] MATLAB, version 7.10.0 (R2010a), The MathWorks Inc., Natick, Massachusetts, 2010.
[40] I.T. Jolliffe, Principal Component Analysis, Springer, New York, NY, 1986.
[41] R. Bilger, Symp. (Int.) Combust., vol. 22, Elsevier, 1989, pp. 475–488.
[42] S.D. Cohen, A.C. Hindmarsh, Comput. Phys. 10 (2) (1996) 138–143.