A Filter-Independent Model Identification Technique for Turbulent Combustion Modeling

Amir Biglari, James C. Sutherland*

University of Utah, 155 South 1452 East, Room 350, The Institute for Clean and Secure Energy, Salt Lake City, Utah, USA 84112

Abstract

In this paper, we address a method to reduce the number of species equations that must be solved via application of Principal Component Analysis (PCA). This technique provides a robust methodology to reduce the number of species equations by identifying correlations in state-space and defining new variables that are linear combinations of the original variables. We show that applying this technique in the context of Large Eddy Simulation allows for a mapping between the reduced variables and the full set of variables that is insensitive to the size of filter used. This is notable since it provides a model to map state variables to progress variables that is a closed model.

As a linear transformation, PCA allows us to derive transport equations for the principal components, which have source terms. These source terms must be parameterized by the reduced set of principal components themselves. We present results from *a priori* studies to show the strengths and weaknesses of such a modeling approach. Results suggest that the PCA-based model can identify manifolds that exist in state space which are insensitive to filtering, suggesting that the model is directly applicable for use in Large Eddy Simulation. However, the resulting source terms are not parameterized with an accuracy as high as the state variables.

Keywords: Manifold, Data analysis, Dimensionality reduction, Principal component analysis, PCA, Turbulent combustion modeling

1. Introduction and Background

Modeling turbulent combustion processes requires solution of a large number of equations due to the large number of reacting species present. Furthermore, the computational cost of resolving turbulent flows scales as Re^3. Reducing the range of scales that must be resolved as well as the number of equations to be solved is, therefore, of utmost importance to achieve simulations of practical combustion systems.

Classical turbulence theory indicates that resolution requirements scale with the Reynolds number as Re^3 for isotropic, homogeneous turbulent flow [1]. Species with large Schmidt numbers further increase the range of scales. In addition to the separation of length scales due to turbulence, the large number of species involved in combustion and the stiff chemistry associated with the reactions further increase the cost of direct simulation so that it is prohibitively expensive for all but the simplest of systems.

*Corresponding author, tel: +1-801-585-1246, fax: 801-585-1456
Email addresses: amir.biglari@utah.edu (Amir Biglari), james.sutherland@utah.edu (James C. Sutherland)

Typically, time averaging (RANS) or spatial filtering (LES) is used to reduce the resolution requirements. To reduce the number of thermochemical degrees of freedom, there are two broad approaches:

- mechanism reduction, where the chemical mechanism is modified to reduce the number of species and the stiffness, and
- state-space parameterization, where the state of the system is assumed to be parameterized by a small number of variables which are evolved in the CFD.

The techniques proposed in this paper fall into the second category: they seek to obtain a set of variables that parameterizes the thermochemical state, and these variables are then evolved in the CFD calculation.

There have been numerous efforts to reduce the dimensionality of a combustion process (see, *e.g.* [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12] for a few). Flamelet models such as Steady Laminar Flamelet Method (SLFM) [2, 3, 4], flamelet-generated manifold (FGM) [5, 6, 7, 9] or flamelet-prolongation of ILDM model (FPI) [11, 12, 13] are examples of state-space parameterization model.

The remainder of this paper is organized as follows: we first identify the datasets that will be used to evaluate the proposed model in section 2. We then review Principal Component Analysis (PCA) as a technique to obtain a reduced parameter set (section 3), discuss how PCA can be formulated as a predictive model (section 3.1.2), and introduce adaptive regression to enable parameterization of nonlinear functions of the principal components (section 3.2). Section 4 then examines the model in the context of turbulent closure and shows that the model is closed, *i.e.* it requires no explicit closure model for the thermochemistry. Finally, conclusions are presented in section 6.

2. Datasets

In the discussions below we will consider two datasets:

1. A dataset from a One-Dimensional Turbulence (ODT) simulation which has been done on a temporally evolving nonpremixed CO/H_2-air jet with extinction and reignition [14, 15]. This was shown to be a statistically accurate representation of a corresponding high-fidelity DNS dataset [15]. The calculations include detailed chemical kinetics, thermodynamics, and transport and exhibit significant local extinction and reignition and the dataset is, therefore, a modeling challenge. The state variables are: temperature and species mass fractions for H_2, O_2, O, OH, H_2O, H, HO_2, CO, CO_2, HCO and N_2.
2. Sandia TNF CH_4/air Flame D [16]. This flame does not exhibit significant amounts of extinction or reignition, and is a standard modeling target flame. The state variables are temperature and mass fractions for O_2, N_2, H_2, H_2O, CH_4, CO, CO_2, OH and NO.

The Flame D dataset is "incomplete" in that it does not contain species reaction rates or a complete set of species. The ODT/DNS dataset, on the other hand, is "complete" in that it has the full set of species, reaction rates, etc. resolved in space and time, but relies on simulation to obtain the data, and is only as accurate as the thermodynamic, kinetic and transport properties that were used in the simulation.

When comparing against the datasets, we report R^2 values to measure the accuracy with which the model represents the original data,

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(\phi_i - \phi_i^*)^2}{\sum_{i=1}^{N}(\phi_i - \langle\phi\rangle)^2}, \tag{1}$$

where ϕ_i is the observed value, ϕ_i^* is the predicted value, and $\langle\phi\rangle$ is the mean value of ϕ. For the PCA analysis, we consider data sampled from all space and time in the ODT dataset, and the full dataset for the TNF data. In other words, the PCA does not vary in $\vec{x}$ or t since we sample all $\vec{x}$ and t to obtain the PCA.

3. Parameterization using Principal Component Analysis

3.1. Principal Component Analysis

Principal Component Analysis (PCA) provides a robust methodology to reduce the number of species equations by identifying correlations in state-space and defining new variables (principal components) that are linear combinations of the original variables (state variables) [17, 18, 19, 20]. Details of the formulation have been published elsewhere [19, 20, 21, 22], and here we only review the concepts behind the PCA analysis. The basic process of a PCA reduction is

1. Identify a new basis in the multidimensional dataset that is a rotation of the original basis. We call this new basis $\boldsymbol{\eta}$ and the original data $\boldsymbol{\phi}$. The new basis is obtained via an eigenvalue decomposition of the correlation matrix for many observations of $\boldsymbol{\phi}$ for a system. At this point, we have only performed a rotation, and no information loss has occurred.
2. Truncate the new basis and project the data onto the new basis to obtain an approximation (compression) of the data on the new basis.
3. Given an observation in the truncated basis, we can approximate the value of the original data. This "reconstruction" is a linear reconstruction and is thus very efficient.

Steps 1 and 2 are illustrated conceptually in Figure 1.

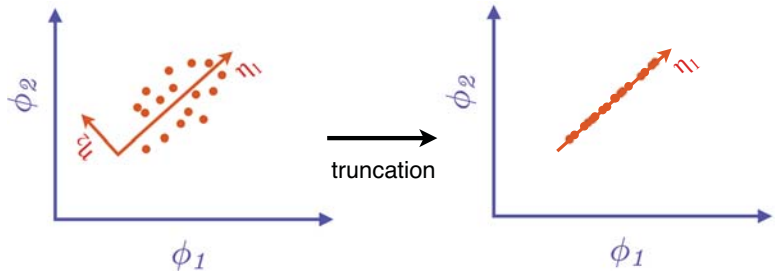


Figure 1: Illustration of the principal components of a hypothetical 2D data set where we retain a single principal component.

The PCA modeling approach thus requires "training" data which should (ideally) be observations of the a system at conditions close to where we wish to apply the model. Once η_i is known, the original state variables (*e.g.* T, y_j) can be easily obtained. Furthermore, the accuracy of the parameterization is obtained *a priori*, and can be adjusted to obtain arbitrary accuracy by increasing the number of retained PCs. This is illustrated in Figure 2, where the eigenvalues (relative importance of a given PC in representing the data) as well as the percent of the variance in the original variables are shown as a function of the number of retained eigenvalues. The eigenvalues can assist in determining how many PCs should be retained to maintain a desired level of accuracy in the resulting model.

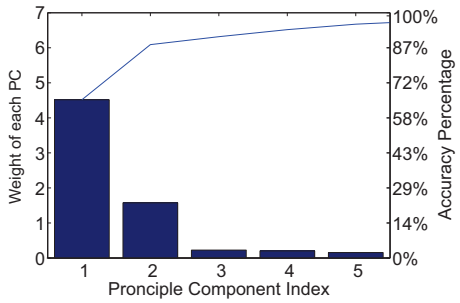


Figure 2: Eigenvalue magnitude (left axis, bars) and percent variance captured (right axis, line) by retaining the given number of components.

Table 1: Brief descriptions of the scaling options considered here.

Method	Factor used to scale each state variable
STD	Standard deviation
VAST	Ratio of the standard deviation to the mean.
Range	Maximum-minimum
Level	Mean
Max	Maximum
Pareto	Square-root of the standard deviation

3.1.1. Effects of Scaling

Prior to applying PCA, the original data should be centered and scaled [23, 24, 19, 25, 20]. There are many different scaling options, some of which are enumerated in Table 1. Further details regarding scaling may be found in the aforementioned sources. For the purposes of this paper, it is sufficient to recognize that the choice of scaling affects the accuracy of the resulting PCA parameterization. To illustrate this point, consider the results shown in Table 2, where R^2 values are shown for parameterization of the original state variables by three PCs for various choices of scaling. The effects of scaling will become even more pronounced

Table 2: R^2 values for PCA projection of state variables with different scaling methods using 2 PCs on the temporal CO/H_2 dataset.

Scaling	T	H_2	O_2	O	OH	H_2O	H	HO_2	CO	CO_2	HCO	**Average**
VAST	0.999	0.952	1.000	0.777	0.853	0.954	0.822	0.075	0.996	0.985	0.893	**0.846**
STD	0.978	0.939	0.995	0.862	0.920	0.927	0.841	0.056	0.988	0.984	0.911	**0.855**
Level	0.944	0.947	0.978	0.906	0.951	0.877	0.824	0.037	0.976	0.965	0.933	**0.849**
Range	0.987	0.946	0.999	0.817	0.881	0.952	0.862	0.059	0.996	0.985	0.876	**0.851**
Max	0.984	0.945	0.999	0.820	0.884	0.950	0.866	0.057	0.996	0.984	0.875	**0.851**
Pareto	1.000	0.948	0.998	0.746	0.832	0.956	0.809	0.085	1.000	0.981	0.864	**0.838**

when source terms are considered in section 3.2.1. However, from Table 2, it is evident that scaling can have an appreciable impact in the accuracy of a PCA reconstruction.

Unless explicitly stated otherwise, the results presented in this paper were obtained with VAST scaling.

3.1.2. Transport Equations for PCs

The governing equations can be written as

$$\frac{\partial \rho\phi}{\partial t} = -\nabla \cdot \rho\phi\vec{u} - \nabla \cdot \vec{J}_\phi + S_\phi, \tag{2}$$

where $\phi = \{1, \vec{u}, y_i, T\}$ (or any suitable energy variable in place of T), $\vec{u}$ is the mass-averaged velocity, $\vec{J}_\phi$ is the diffusive flux of $\rho\phi$, and S_ϕ is the volumetric rate of production of $\rho\phi$. Due to the large number of species present in combustion, Eq. (2) represents a large number of strongly coupled partial differential equations that must be solved. The thermochemical state variables (T, p and $n_s - 1$ species mass fractions Y_i) define an $(n_s + 1)$-dimensional state space which is widely recognized to have lower-dimensional attractive manifolds [26].

Since PCA is a linear transformation, we may apply it directly to the subset of Eq. (2) associated with T and y_i to derive the transport equations for the PCs. The full derivation has been presented elsewhere [21], and results in

$$\frac{\partial \rho\eta_i}{\partial t} = -\nabla \cdot \rho\eta_i\vec{u} - \nabla \cdot \vec{J}_{\eta_i} + S_{\eta_i}. \tag{3}$$

The source term for the PCs, S_{η_i}, is a linear combination of the original (scaled) species and temperature source terms, and must be parameterized in terms of $\boldsymbol{\eta}$ to close the model. It is important to note that (3) requires that the PCA definition is independent of space and time so that commutativity with differential operators is maintained. This can be achieved by using data from all space and time in constructing the PCA reduction, and all analyses presented herein adhere to this principle.

In previous work where this approach was originally proposed [21], preliminary results were shown where PCA was performed locally in mixture fraction space (*i.e.* conditioned on mixture fraction). Here we consider unconditional PCA, and extend the analysis to examine: 1) the effects of scaling (see section 3.1.1) on the source term parameterization, 2) the effects of filtering on the accuracy of the source term parameterization, and 3) multivariate regression, which will be discussed in section 3.2.

Just as the species source terms are highly nonlinear functions of y_i and T, the PC source terms (S_{η_i}) are highly nonlinear functions of the PCs. The original state variables are well-parameterized by the PCs (given a sufficient number of retained PCs) because this is the objective of PCA: to identify correlations in the original variables. However, the PCA transformation does not necessarily identify the ideal basis for representing source terms. Furthermore, although the original state variables can be well-characterized by linear functions of $\boldsymbol{\eta}$, the same is not necessarily true for S_{η_i}. Thus, several questions remain to be addressed relative to a model based on PCA:

1. Can the truncated basis (see section 3.1) adequately represent the PC source terms?
2. Given that the relationship between η_i and S_{η_j} is highly nonlinear, can an adaptive regression technique be employed to obtain the functions

$$S_{\eta_j} = \mathcal{F}_j(\eta_1, \eta_2, \ldots \eta_{n_\eta}) \tag{4}$$

   for the $j = 1 \ldots n_\eta$ retained PCs?
3. Are the functions represented by Eq. (4) sensitive to filtering? In other words, is $\mathcal{F}_j$ a function of the filter width, Δ?

This paper aims to address these questions using *a priori* analysis of high-fidelity combustion data. We next

Table 3: R^2 values for MARS regression of state variables on principal components with different scaling methods in PCA using 2 PCs on the temporal CO/H_2 dataset [15].

Scaling	T	H_2	O_2	O	OH	H_2O	H	HO_2	CO	CO_2	HCO	**Average**
VAST	1.000	0.996	1.000	0.996	0.994	0.997	0.989	0.937	0.999	0.998	0.998	**0.991**
STD	0.995	0.996	0.999	0.986	0.995	0.994	0.997	0.938	0.999	0.991	0.997	**0.990**
Level	0.990	0.996	0.997	0.982	0.991	0.991	0.994	0.938	0.997	0.982	0.997	**0.987**
Range	0.997	0.996	1.000	0.991	0.996	0.995	0.993	0.925	0.999	0.993	0.996	**0.989**
Max	0.996	0.996	1.000	0.989	0.995	0.995	0.994	0.924	0.999	0.992	0.996	**0.989**
Pareto	1.000	0.993	1.000	0.987	0.984	0.997	0.985	0.921	1.000	0.997	0.953	**0.983**

turn our attention to question 2 and outline a methodology to obtain $\mathcal{F}_j$.

3.2. Multivariate Adaptive Nonlinear Regression

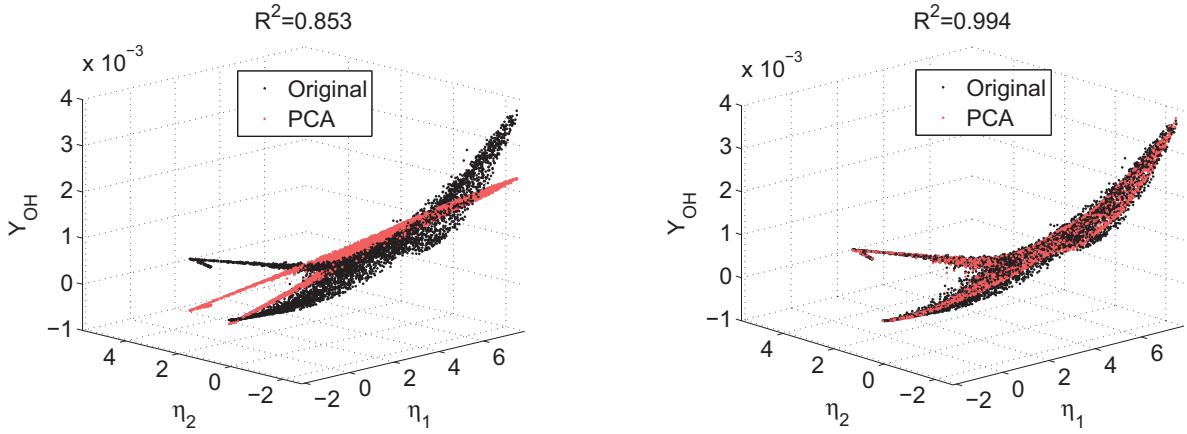
Because we have no physical insight into the appropriate basis functions to form $\mathcal{F}_j$ in Eq. (4), we need an adaptive method. Multivariate Adaptive Regression Splines (MARS) [27, 28, 29] is a technique that allows adaptive selection of basis functions to obtain nonlinear functions such as $\mathcal{F}_j$. At each iteration of the MARS algorithm, a basis function is selected that results in the largest reduction in the regression error. The iterative procedure is repeated until convergence is achieved. To avoid over-fitting the data, we choose lower-order basis functions (typically quadratic or cubic at most) and subdivide the high-dimensional space into only a few sub-spaces to fit the data (5 sub-spaces were used for the results presented here).

Table 3 shows the results of applying MARS to map the state variables onto the PCs. Comparing Table 3 to Table 2, where the state variables were mapped onto the PCs directly via the (linear) PCA transformation, we note an increase in the accuracy of all state variables (but particularly minor species and most notably HO_2), indicating a nonlinear relationship between the state variables and PCs. This nonlinear relationship has also been observed elsewhere [19, 20, 22], but the MARS approach allows us to capture the nonlinearity between $\boldsymbol{\eta}$ and $\boldsymbol{\phi}$ quite well. Figure 3 shows the OH mass fraction, Y_{OH} projected into the two-dimensional space defined by the first two principal components, (η_1, η_2) . Also shown is a reconstruction of Y_{OH} using the (linear) PCA reconstruction (Figure 3a) and the nonlinear MARS reconstruction (Figure 3b). This clearly illustrates the advantages of the nonlinear reconstruction.

3.2.1. MARS for Parameterizing PC Source Terms

In contrast to the state variables themselves, where the PCA defines a linear relationship with the PCs, the PC source terms have no linear relationship to the PCs, and adaptive regression is the only plausible method to obtain $\mathcal{F}_j$ in Eq. (4). Table 4 shows the R^2 values for the regression of the source terms for various scaling approaches. Notably, that there is a much more significant influence of the choice of scaling on the accuracy with which the PC source terms can be represented than for the state variables (shown in Tables 2 and 3).

With regard to question 1 posed in section 3.1.2 (can the S_{η_i} be parameterized by $\boldsymbol{\eta}$?) we observe that the source terms have more error in their representation than the original state variables. This suggests that the basis selected by the PCA, which seeks to identify correlations among the state variables, may not be optimal for the representation of the PC source terms. Therefore, other methods that identify a basis that simultaneously optimizes parameterization of both the state variables and the PC source terms should be explored. Nevertheless, as the number of retained PCs increases, the accuracy of the S_{η_i} parameterization



(a) PCA (linear) reconstruction of Y_{OH} in (η_1, η_2)-space.

(b) MARS (nonlinear) reconstruction of Y_{OH} in (η_1, η_2)-space.

Figure 3: Comparison of PCA and MARS reconstructions for OH mass fraction for a two-dimensional model based on principal components η_1 and η_2. VAST scaling was used.

Table 4: R^2 values for MARS regression of source terms on principal components with different scaling methods in PCA using 3 PCs on the temporal CO/H_2 dataset [15].

Number of PCs	1	2			3			
Scaling method	S_{η_1}	S_{η_1}	S_{η_2}	**Average**	S_{η_1}	S_{η_2}	S_{η_3}	**Average**
VAST	0.838	0.949	0.929	**0.939**	0.968	0.938	0.223	**0.710**
STD	0.041	0.276	0.491	**0.383**	0.349	0.535	0.183	**0.356**
Level	0.073	0.331	0.509	**0.420**	0.437	0.600	0.178	**0.405**
Range	0.369	0.603	0.551	**0.577**	0.698	0.661	0.291	**0.550**
Max	0.407	0.669	0.619	**0.644**	0.751	0.735	0.253	**0.580**
Pareto	0.877	0.960	0.956	**0.958**	0.966	0.963	0.973	**0.967**

also increases. We should note that the definition for S_{η_i} remains unchanged as n_η increases, *i.e.* S_{η_1} for $n_\eta = 1$ is defined in the same manner as S_{η_1} for $n_\eta = 3$. However, their definitions are different for different scaling methods.

4. Filtering & Turbulent Closure

The techniques and results presented in section 3 were discussed in the context of fully-resolved quantities. For filtered/averaged quantities, several additional issues arise:

1. How sensitive is the PCA mapping to filtering? In other words, is the PCA mapping itself affected by filtering?
2. Are the source term functions valid for filtered quantities, *i.e.*, is $\mathcal{F}_i(\eta_1, \eta_2, \ldots \eta_n) = \mathcal{F}_i(\bar{\eta}_1, \bar{\eta}_2, \ldots \bar{\eta}_n)$?

We consider each of these issues in the following sections.

4.1. PCA Sensitivity to Filtering

To determine the sensitivity of PCA to filtering, we examine computational data from a fully resolved CO/H_2 jet flame (see section 2). The data is filtered and a PCA is applied to the filtered variables. This

is performed using a top-hat filter for several filter widths to determine if/how the PCA structure itself is affected by filtering. Figure 4 shows the temperature field extracted along a line-of-sight and shows the effect of the filter on the temperature profile. Δx is the grid spacing of the original data set, whereas Δ refers to the filter width so that $\Delta/\Delta x = 1$ implies no filtering. Figure 4 indicates that the largest filter width employed here ($\Delta/\Delta x = 32$) has a substantial effect on the temperature field.

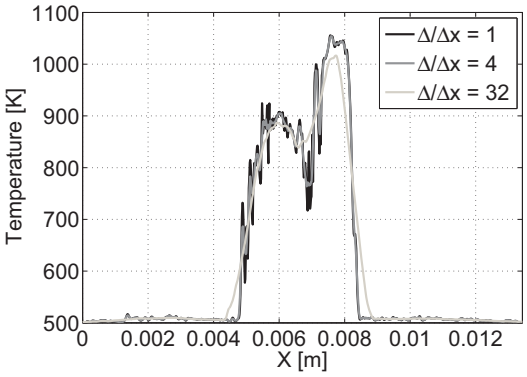


Figure 4: Effect of filtering on temperature profile for a specific time and realization from the temporal CO/H_2 dataset [15].

Figure 5 shows the relative size of the kinetic energy fluctuations, $\frac{K-\bar{K}}{\bar{K}} = \frac{K'}{\bar{K}}$, at filter widths of $\Delta/\Delta x = 4$ and 16. Note that $\Delta/\Delta x = 16$ results in a significant fraction of the kinetic being unresolved, and substantiates the observation from Figure 4 that $\Delta/\Delta x = 16$ is an appreciable filter width. Figure 6 shows

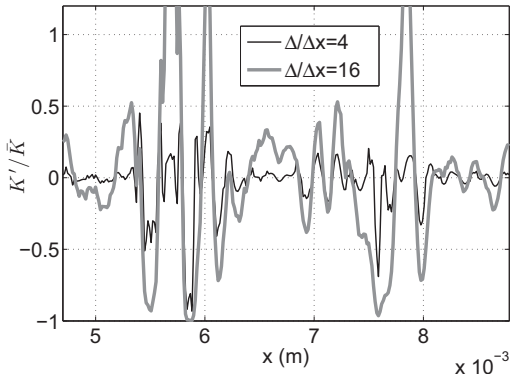


Figure 5: Instantaneous profile of $\frac{K-\bar{K}}{\bar{K}} = \frac{K'}{\bar{K}}$ indicating the magnitude of the unresolved kinetic energy at $\Delta/\Delta x = 4$ and 16.

the largest five contributions to the first three eigenvectors, which define the rotated basis or the principal components. Consider the first eigenvector. The results indicate that the definition of this eigenvector/PC is almost entirely unaffected by filtering. The same results are observed for the second and third eigenvectors. This shows that the PCA reduction itself is insensitive to filtering. The remaining eigenvectors, which are associated with exponentially diminishing eigenvalues (see Figure 2), exhibit the same behavior and are not shown for brevity. These results are of significant importance, since the PCA reduction plays a key role in the proposed modeling strategy outlined in section 3.

The results in Figure 6 suggest that, over a substantial range of filter widths, the structure of a PCA remains unchanged. This is an important result since it implies that the definition of a (linear) manifold for

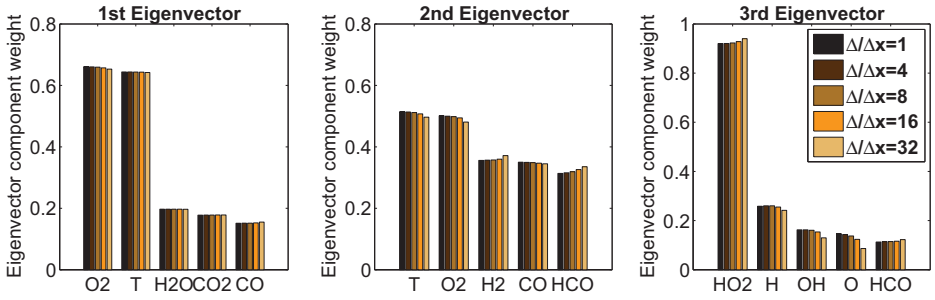


Figure 6: Changes in largest (most important) components of the first three eigenvectors with respect to changes in normalized filter width in temporal CO/H_2 dataset [15].

the state variables is insensitive to filtering.

4.2. Turbulent Closure

We now turn our attention to the question of whether a mapping $\phi = \mathcal{G}(\boldsymbol{\eta})$ is valid for the averaged/filtered quantities, *i.e.* $\bar{\phi} \stackrel{?}{=} \mathcal{G}(\bar{\boldsymbol{\eta}})$. This is particularly important for the source terms that appear in the averaged/filtered PC transport equation,

$$\frac{\partial \bar{\rho}\tilde{\eta}}{\partial t} = -\nabla \cdot \bar{\rho}\tilde{\eta}\vec{\tilde{u}} - \nabla \cdot \vec{J}_{\eta}^{T} + \bar{S}_{\eta}, \tag{5}$$

where $\tilde{\eta}$ is the Favre-averaged/filtered value of η and $\vec{J}_{\phi}^{T}$ is the turbulent diffusive flux.

In traditional state-space parameterization approaches, one defines the parameterization variables and then the mapping between the state variables $\boldsymbol{\phi}$ and reaction variables $\boldsymbol{\eta}$, *e.g.* $\phi = \mathcal{G}(\boldsymbol{\eta})$. Then the joint probability density function (PDF) of all $\boldsymbol{\eta}$, $p(\boldsymbol{\eta})$, is used to obtain mean/filtered values of ϕ,

$$\bar{\phi} = \int \phi(\boldsymbol{\eta})\, p(\boldsymbol{\eta})\, \mathrm{d}\boldsymbol{\eta}.$$

The problem then becomes how to approximate $p(\boldsymbol{\eta})$. If a function

$$\phi = \mathcal{G}(\boldsymbol{\eta}) \tag{6}$$

exists so that

$$\bar{\phi} = \mathcal{G}(\bar{\boldsymbol{\eta}}), \tag{7}$$

then there is no turbulent closure problem and the joint PDF of all $\boldsymbol{\eta}$ is not required.

4.2.1. Ensemble Averaging

We first consider ensemble-averaged data from Flame D (see section 2). Ensemble averages are formed by number-averaging all samples from a given spatial location. Table 5 shows results for all of the species available for flame D. The results show a PCA reconstruction (linear) for two and three retained PCs as well as a (nonlinear) MARS reconstruction based on the same two and three PCs. The "original data" refer to the data processed directly from the flame D dataset where PCA was applied to the entire dataset. PCA and MARS regressions were performed to obtain $\phi = \mathcal{G}(\boldsymbol{\eta})$ and the resulting R^2 values reported. The "ensemble data" used the PCA and MARS regression obtained from the original data and applied it to the ensemble-averaged values for the PCs. Specifically, 1) PCA was applied to the original dataset, 2) MARS was performed to obtain $\mathcal{G}(\boldsymbol{\eta})$, 3) using the PCA obtained in step 1, the PCs were computed from the original

data and then ensemble-averaged to obtain $\bar{\boldsymbol{\eta}}$, 4) $\bar{\phi}^* = \mathcal{G}(\bar{\boldsymbol{\eta}})$ was calculated and compared with the directly averaged values of $\bar{\phi}$ to obtain an R^2 value.

Table 5: R^2 values for different variables in flame D showing that no closure is required to reconstruct the original variables, $\boldsymbol{\phi}$ from the principal components, $\boldsymbol{\eta}$.

Approach	Data Type	T	O_2	CO_2	NO	H_2O	N_2	H_2	CH_4	CO	OH
PCA, $n_\eta = 2$	Original	0.990	0.981	0.966	0.860	0.979	1.000	0.385	0.914	0.439	0.495
	Ensemble	0.995	0.995	0.987	0.899	0.990	1.000	0.539	0.987	0.611	0.687
PCA, $n_\eta = 3$	Original	0.991	0.991	0.988	0.911	0.989	1.000	0.944	0.916	0.890	0.599
	Ensemble	0.996	0.998	0.998	0.939	0.998	1.000	0.982	0.990	0.972	0.650
MARS, $n_\eta = 2$	Original	0.997	0.991	0.976	0.926	0.989	1.000	0.625	0.941	0.666	0.650
	Ensemble	0.998	0.999	0.995	0.950	0.999	1.000	0.917	0.992	0.933	0.744
MARS, $n_\eta = 3$	Original	0.997	0.997	0.991	0.974	0.993	1.000	0.938	0.941	0.904	0.745
	Ensemble	0.998	0.999	0.999	0.899	0.999	1.000	0.971	0.993	0.979	0.755

There are several noteworthy points relative to Table 5:

1. As n_η increases from 2 to 3, the R^2 value uniformly increases, indicating the increase of accuracy of a PCA-based model as the number of retained components increases. This has been discussed in detail elsewhere [19, 20, 21, 22].
2. The MARS representation of the data is more accurate than the corresponding direct PCA reconstruction, indicating that there is an underlying nonlinear relationship between the $\boldsymbol{\phi}$ and $\boldsymbol{\eta}$ that the linear PCA-based reconstruction cannot accurately capture.
3. The ensemble-averaged data shows R^2 values that are nearly always higher than their corresponding original data values. This suggests that the PCA based models $\phi = \mathcal{G}(\boldsymbol{\eta})$ do not incur any additional error when evaluated using mean values, $\bar{\phi}^* = \mathcal{G}(\bar{\boldsymbol{\eta}})$. This is true for the linear reconstruction as well as the nonlinear (MARS) reconstruction.

4.2.2. Spatial Filtering

We next consider spatial filtering with the CO/H_2 dataset discussed in section 2. Figure 4 illustrates the effect of different filter lengths on an extracted line-of-sight represented by the ODT data for the temporal CO/H_2 dataset. For this particular dataset, a filter width of $\Delta/\Delta x = 16$ induces substantial filtering on the data.

First, a PCA was performed on the fully resolved data and either $n_\eta = 2$ or $n_\eta = 3$ PCs were retained. This provides a linear mapping between the PCs ($\boldsymbol{\eta}$) and the state variables ($\boldsymbol{\phi}$). Using this mapping, we then compute $\bar{\boldsymbol{\eta}}$ directly from the dataset and then use the mapping to approximate $\bar{\boldsymbol{\phi}}^*$, which is then compared to $\bar{\boldsymbol{\phi}}$ calculated directly from the dataset. These results are shown in Figure 7. Finally, a MARS regression was performed to map the original variables onto the PCs at the fully resolved scale, providing $\phi_i = \mathcal{G}_i(\boldsymbol{\eta})$. Then, $\bar{\boldsymbol{\eta}}$ was calculated directly from the data and $\bar{\phi}_i^*$ was approximated as $\bar{\phi}_i^* = \mathcal{G}_i(\bar{\boldsymbol{\eta}})$, and this was compared to the value of $\bar{\phi}_i$ calculated directly from the dataset. The profiles in Figure 7 show these results. From Figure 7, it is apparent that the error is well-controlled as the filter width is increased, indicating that the PCA-based models require no explicit closure. This is consistent with the results for the ensemble-averaged analysis performed in section 4.2.1.

Figure 8 shows extracted spatial profiles (over a small portion of the domain corresponding to an active flame region) for CO_2 and OH mass fractions for two different filter widths ($\Delta/\Delta x = 1$ and 16) and $n_\eta = 2$.

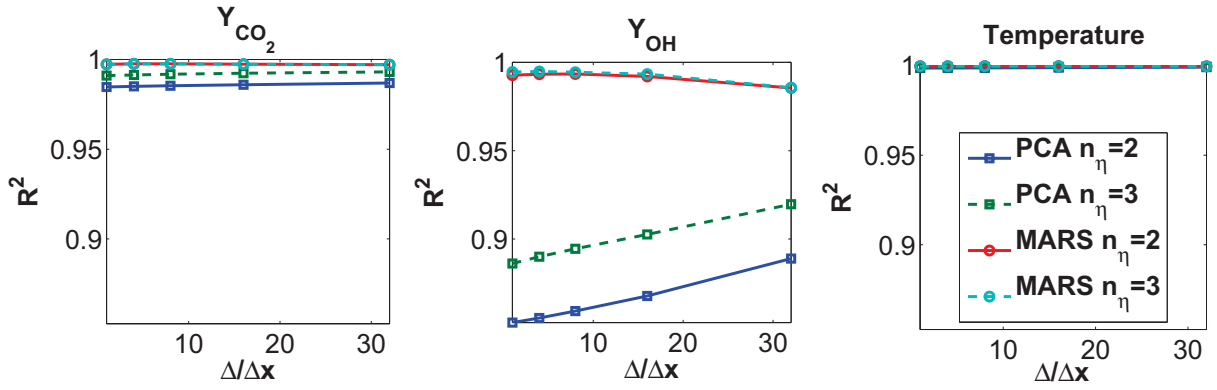


Figure 7: R^2 value changes with respect to the changes in normalized filter width (normalized with grid spacing length) for several variables in temporal CO/H_2 dataset [15].

The solid lines represent the profiles extracted directly from the data, whereas the dashed lines are the reconstructed profiles using the PCA/MARS model. These results demonstrate the ability of the PCA/MARS modeling approach to reconstruct the unfiltered and filtered quantities, and also indicate the strong filtering that is occurring at $\Delta/\Delta x = 16$. It is particularly remarkable that the OH profiles are reconstructed so well by a two-parameter model, and that the filtered profiles are also reconstructed with reasonable accuracy.

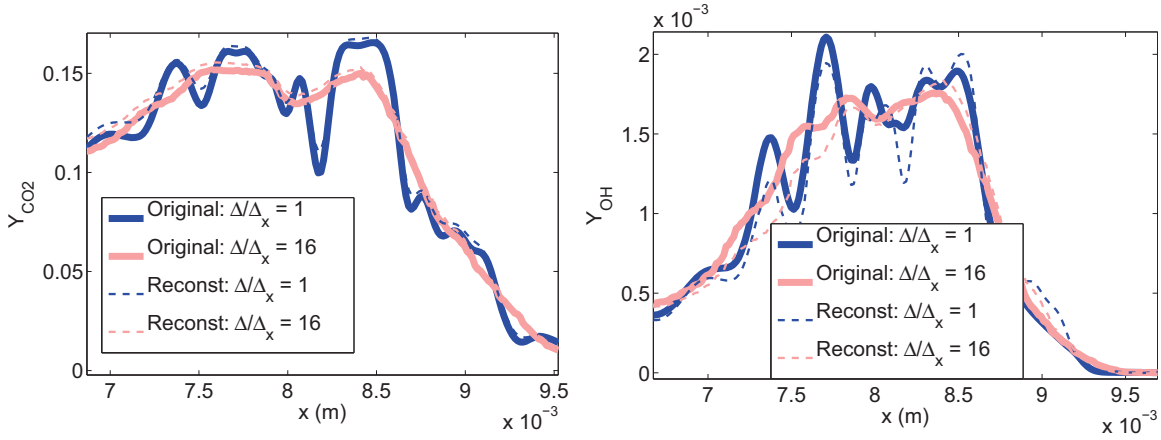


Figure 8: Original (solid) and reconstructed (dashed) profiles for CO_2 and OH for no filtering and a filter width of $\Delta/\Delta x = 16$ and $n_\eta = 2$.

4.2.3. Source Term Parameterization

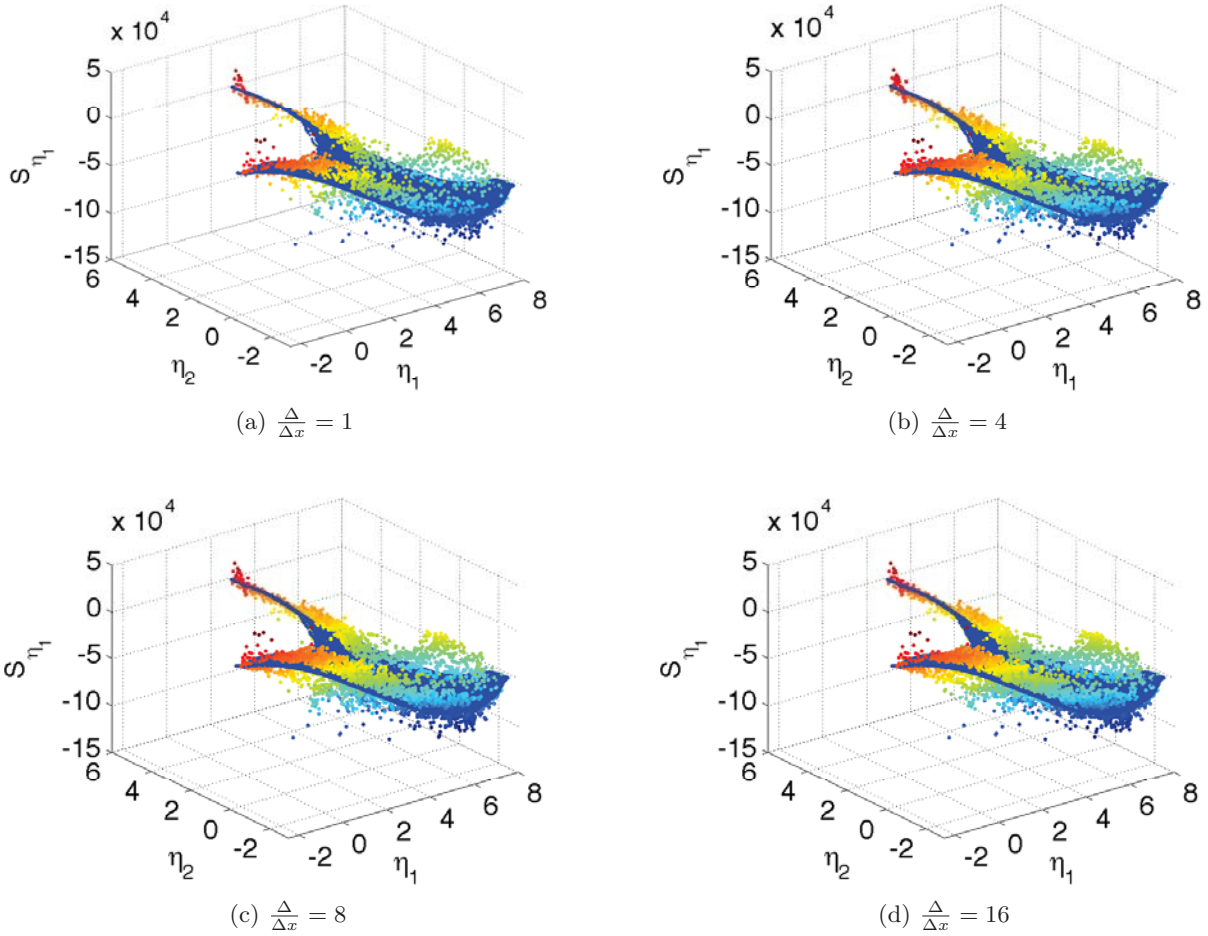
We now turn our attention to the parameterization of the PCA source terms in Eq. (3) and Eq. (5), and seek to answer question 2 posed in section 3.1.2 and question 2 in section 4: can a function $S_{\eta_i} = \mathcal{F}_i(\boldsymbol{\eta})$ be found, and is $\bar{S}_{\eta_i} = \mathcal{F}_i(\bar{\boldsymbol{\eta}})$?

To ascertain the performance of the PCA-based model in representing $\bar{S}_{\eta_i} = \mathcal{F}_i(\bar{\boldsymbol{\eta}})$, we first calculate S_{η_i} and then obtain the regressing function $S_{\eta_i} = \mathcal{F}_i(\boldsymbol{\eta})$ via MARS. Next, $\bar{\boldsymbol{\eta}}$ and $\bar{S}_{\eta_i}$ are calculated directly from the data, and compared against $\mathcal{F}_i(\bar{\boldsymbol{\eta}})$. Figure 9 illustrates the results of this in state space while Figure 10 shows the associated R^2 values. From these results, as well as those previously presented in Table 4, several conclusions may be drawn:

1. S_{η_i} is parameterized with less accuracy than ϕ_i. This is not surprising given that the PCA was designed

to parameterize ϕ well, and it is well-known that S_{ϕ_i} is a highly nonlinear function of $\boldsymbol{\phi}$ so that S_{η_i} will also be a highly nonlinear function of $\boldsymbol{\eta}$.

2. The error in the approximation $\bar{S}^*_{\eta_i} = \mathcal{F}_i(\bar{\boldsymbol{\eta}})$ is bounded and well behaved with the moderate range of filter widths considered in this study. Indeed, the structure of $\bar{S}^*_{\eta_i}(\boldsymbol{\eta})$ is largely unaffected by filtering as shown in Figure 9 for $\bar{S}^*_{\eta_1}(\bar{\eta}_1, \bar{\eta}_2)$, and more quantitatively in Figure 10.



(a) $\frac{\Delta}{\Delta x} = 1$ (b) $\frac{\Delta}{\Delta x} = 4$

(c) $\frac{\Delta}{\Delta x} = 8$ (d) $\frac{\Delta}{\Delta x} = 16$

Figure 9: $\bar{S}_{\eta_1}(\bar{\eta}_1, \bar{\eta}_2)$ obtained directly from the data (points) as well as the prediction based on PCA/MARS (surface) for various filter widths using the temporal CO/H_2 dataset. Figure 10 shows the corresponding R^2 values.

Figure 11 shows extracted spatial profiles for S_{η_1} and S_{η_2} for a model with $n_\eta = 2$ and filter widths of ($\Delta/\Delta x = 1$ and 16). These results were obtained using Pareto scaling, as this provided the best reconstruction of the source terms as shown in Table 4. The solid lines represent the profiles extracted directly from the data, whereas the dashed lines are the reconstructed profiles using the PCA/MARS model. These results correspond to the R^2 values reported in Figure 10 at $\Delta/\Delta x = 16$. While the general trend for S_η is captured in both cases, it is clear that the detailed profiles for S_η are not captured fully, and this is also reflected in the relatively low R^2 values shown in Table 4.

Another interesting feature of Figure 11 is the structure of S_{η_1} and S_{η_2} are very similar, although their magnitudes are different. Figure 12 shows the weights that define the first three PCs using Pareto scaling. The first PC (η_1) is defined almost exclusively by temperature (this is commonly the case when Pareto scaling

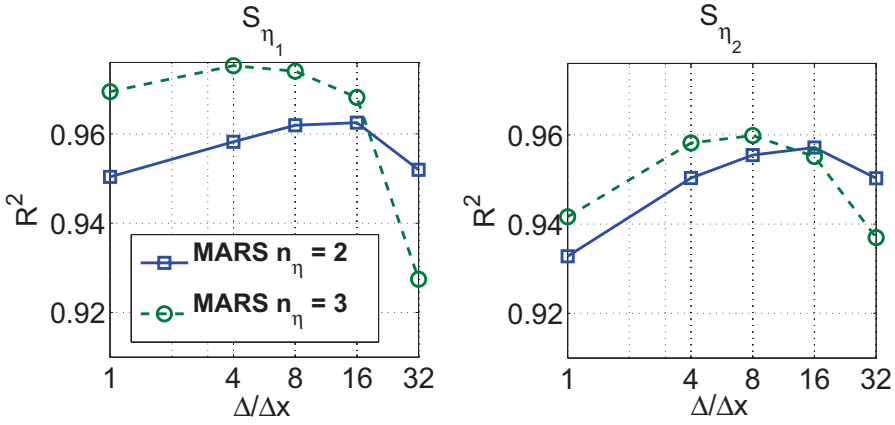


Figure 10: R^2 value changes with respect to the changes in normalized filter width, $\Delta/\Delta x$ for the source term of the first and the second PCs, in temporal CO/H_2 dataset [15].

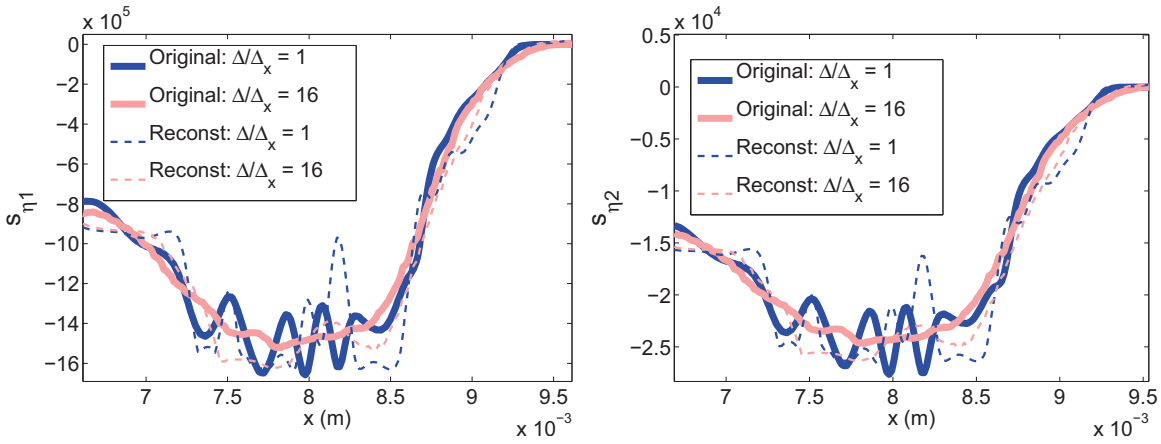


Figure 11: Original (solid) and reconstructed (dashed) profiles for S_{η_1} and S_{η_2} for no filtering and a filter width of $\Delta/\Delta x = 16$ and $n_\eta = 2$.

is used). The second PC is defined primarily by the reactants CO, H_2 and O_2. Therefore, $S_{\eta_1} \approx S_T$ while S_{η_2} is primarily comprised of S_{CO}, S_{H_2} and S_{O_2}. Since these reaction rates are all spatially correlated, it is not surprising that S_{η_1} and S_{η_2} are also highly correlated spatially. As expected, the profiles for η_1 and η_2 are quite different from one another since η_1 follows the temperature profile (which peaks in the reaction zone) whereas η_2 follows the reactants (which are strongly depleted in the reaction zone). It is important to note that a different choice of scaling (resulting in a different PC structure) will have a major influence on the resulting source term profiles. This needs to be investigated further and will be the subject of future research.

4.3. Comparison with SLFM

To provide a reference point with a very common combustion modeling approach, we compare the SLFM model with the PCA model in their respective abilities to reproduce the CO/H_2 dataset described in section 2. At the outset, we note that the SLFM model (for the purposes of this paper) is parameterized by three parameters: the averaged/filtered mixture fraction ($\bar{Z}$), its variance ($\sigma_Z'^2$), and the scalar dissipation rate (χ). The mapping $\phi = \mathcal{G}(Z, \sigma_Z'^2, \chi)$ is obtained through solving the flamelet equations [2] and convoluting them with a β-PDF for the mixture fraction. Formally, this implies that Z and χ are statistically independent and

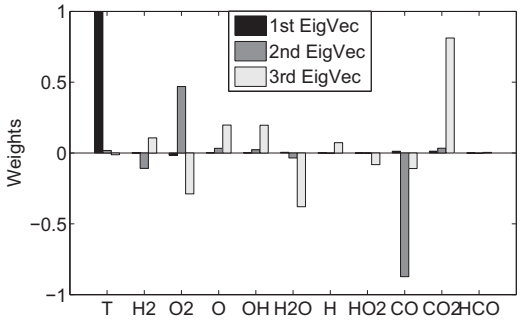


Figure 12: Weights for the first three eigenvectors (which define the first three PCs) for Pareto scaling.

that the PDF of the scalar dissipation rate is approximated as $\delta(\bar{\chi})$. This is a common modeling assumption, and may be justified in part by observations in previous DNS studies that suggest errors in $\phi(Z,\chi)$ overshadow errors in approximating $p(\chi) = \delta(\bar{\chi})$ [8].

Figure 13 shows parity plots and R^2 values comparing the observed and reconstructed values of T and OH for the SLFM and PCA/MARS models at the fully resolved scale (*i.e.* no filtering). In this case, the PCA/MARS model employed two PCs, consistent with the number of parameters in the SLFM model (Z,χ). While one would not expect the SLFM model to perform well in this case since extinction and reignition are present, this does demonstrate the ability of the PCA/MARS model to identify and parameterize the state space effectively.

Figure 14 shows realizations of the original and reconstructed data for the PCA/MARS and SLFM models in state space. Here it is much more clear that the PCA/MARS identifies a manifold in state space whereas the SLFM model does not. Again, while the SLFM model is not expected to perform well in this situation, the comparison is illustrative of the differences between the models with same number of parameters.

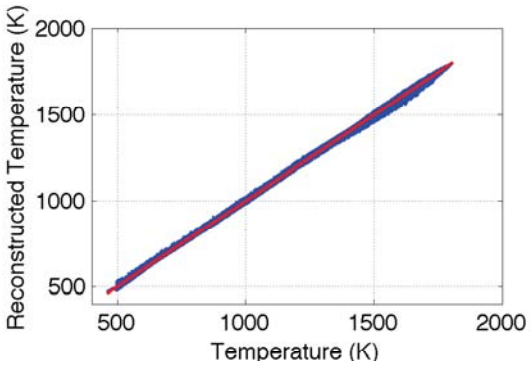
When averaging/filtering is applied, the mixture fraction variance is typically introduced as an additional parameter, with a presumed PDF for the mixture fraction parameterized in terms of $\bar{Z}$ and $\sigma_Z'^2$ so that the state variables are obtained via

$$\phi = \int \phi(Z,\bar{\chi})\, p(Z;\bar{Z},\sigma_Z'^2)\, \mathrm{d}Z.$$

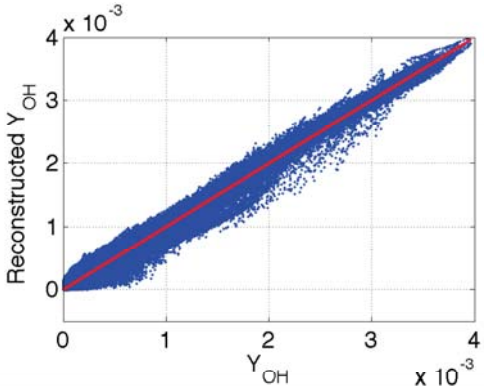
For comparison purposes, we explore the performance of several models, summarized in Table 6. Figure 15 shows the performance of these models for several state variables as a function of the filter width. There are several important observations to be made:

- The SLFM model error is bounded if the β-PDF model is also used, but the error increases (indicated by the decrease of R^2 with increasing Δ) if no closure is made.
- The PCA based models demonstrate no sensitivity to filtering, requiring no explicit closure.
- The two-parameter PCA model is more accurate than the three parameter SLFM/PDF model.

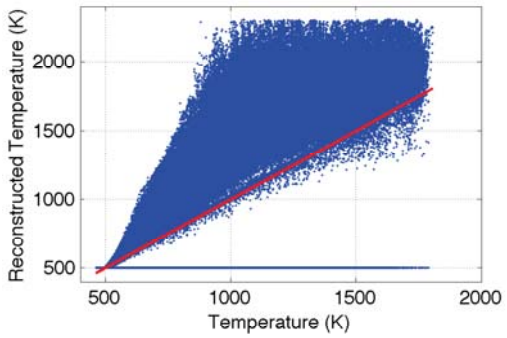
While SLFM is not expected to accurately capture the thermochemical state of this system that involves extinction and reignition, these results illustrate the degree of accuracy that can be obtained using PCA to identify parameterizing variables for use in defining models, and illustrates that proper selection of parameterizing variables can lead to significant improvements in model accuracy even for a relatively small number parameters, n_η.



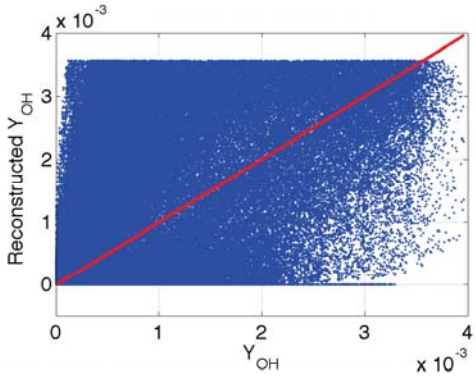
(a) MARS with n_η=2 (R^2=1.000)



(b) MARS with n_η=2 (R^2=0.994)



(c) SLFM (R^2=0.456)



(d) SLFM (R^2=0.055)

Figure 13: Parity plots for temperature (left) and OH (right) reconstructions using PCA/MARS (top row) and SLFM (bottom row).

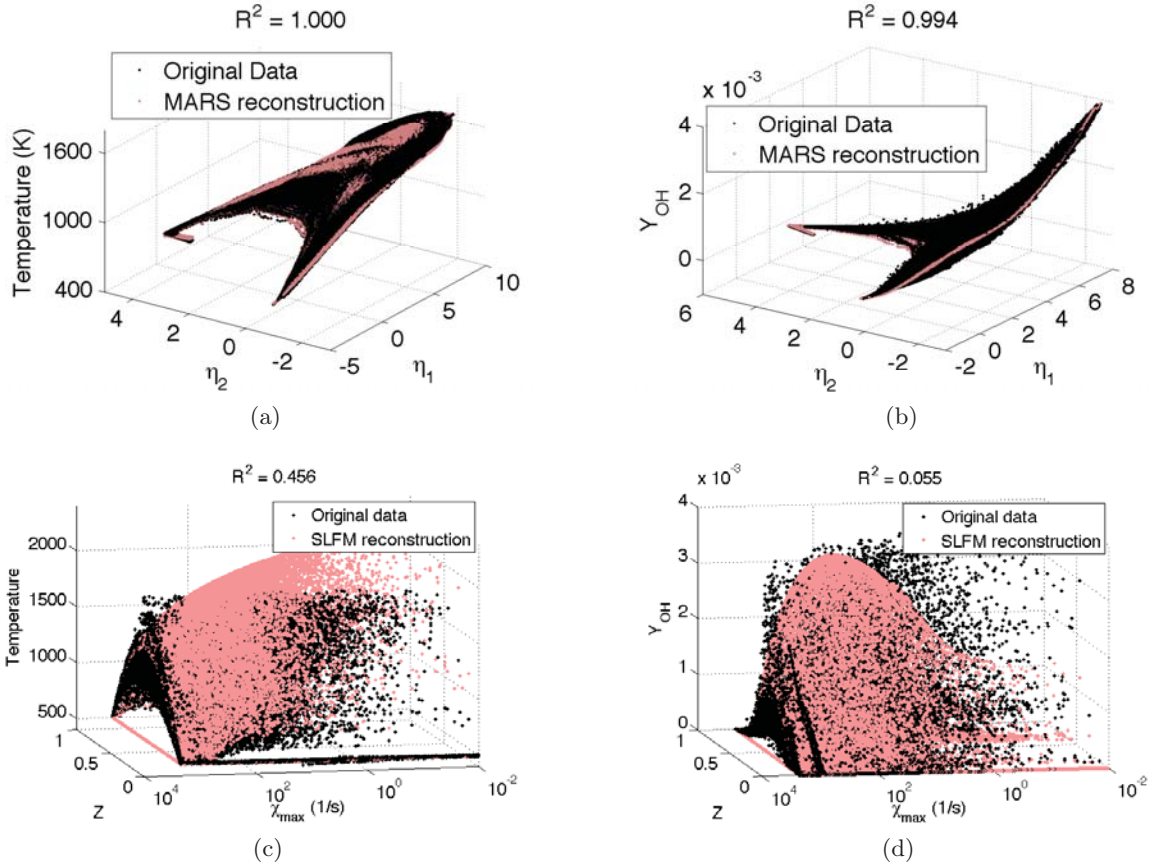


Figure 14: Reconstruction of temperature (left) and OH mass fraction (right) for a 2D PCA/MARS model (top) and the SLFM model (bottom) for the temporal ODT data set [15]. The R^2 value of these reconstructions are reported on each plot as well.

Table 6: Summary of models compared in Figure 15.

Model	Parameters	Comments
PCA, $n_\eta = 2$	(η_1, η_2)	Linear reconstruction using two PCs
PCA, $n_\eta = 3$	(η_1, η_2, η_3)	Linear reconstruction using three PCs
PCA/MARS, $n_\eta = 2$	(η_1, η_2)	Nonlinear reconstruction using two PCs
PCA/MARS, $n_\eta = 3$	(η_1, η_2, η_3)	Nonlinear reconstruction using three PCs
SLFM	$(\bar{Z}, \bar{\chi})$	Reconstruction using SLFM without closure
SLFM/PDF	$(\bar{Z}, \bar{\chi}, \sigma_Z'^2)$	Reconstruction using SLFM with a presumed PDF on Z

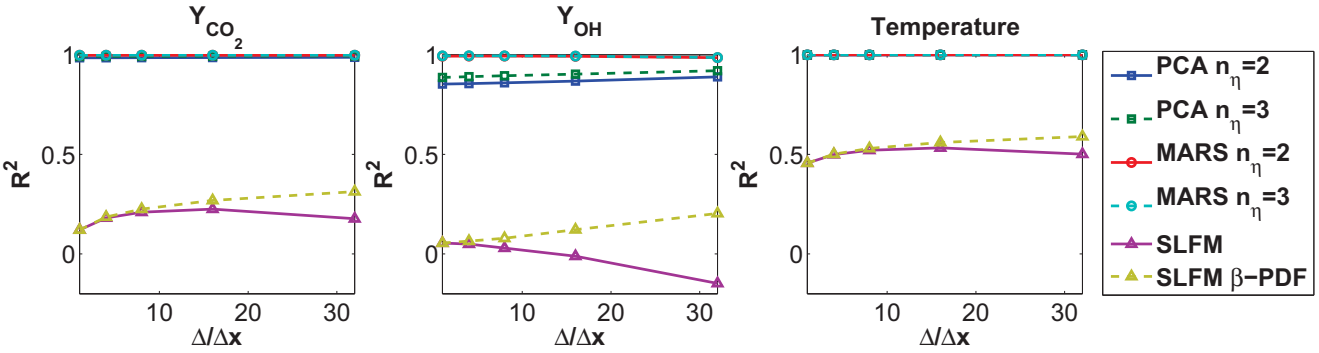


Figure 15: R^2 value changes with respect to the changes in normalized filter width ($\Delta/\Delta x$) for several state variables comparing PCA and MARS results with SLFM and SLFM-β-PDF results for the temporal CO/H_2 dataset. See Table 6 for more information.

5. Considerations for Model Generation

Using PCA as the basis for combustion modeling is still a new concept, and there are several outstanding issues regarding its application.

In traditional combustion modeling approaches, a canonical reactor configuration is adopted and solved parametrically to obtain a mapping between the state variables $\boldsymbol{\phi}$ and the parameterizing variables $\boldsymbol{\eta}$. Such models are, therefore, inherently limited by the assumptions inherent in the canonical reactor. Although PCA can be applied to canonical systems such as the flamelet equations [2], we favor using models such as ODT that allow for a wide range of coupling in length and time scales and provide statistical sampling of the state over a wider range of applicable conditions than traditional canonical reactors. PCA is particularly well-suited for application to such systems because it allows "adaptive" selection of the optimal parameterizing variables. However, it remains to be seen how sensitive the PCA is to the chosen canonical reactor. This may not be critically important in the case of using ODT as the model since it can provide reasonable results for combustion systems [15].

Additionally, the sampling density for the data used to obtain the PCA may be an important consideration. For example, the PCA may be influenced by the over-representation of fuel and oxidizer and relatively small number of observations from the flame regions. Identifying biasing from over/under-sampling is not a trivial task and future work will seek to address this issue.

6. Conclusions & Future Work

This paper discusses a novel state-space parameterization method. It belongs to the same family as other parameterization methods such as equilibrium, steady laminar flamelet (SLFM) [2], flamelet-prolongation of ILDM [12, 11], *etc.*, but extracts the parameterization directly from data rather than presuming a functional form for the parameterization. We use PCA to identify the model parameters and then MARS (a multidimensional adaptive regression technique) to obtain the functional form between the progress variables (principal components), $\boldsymbol{\eta}$ and the state variables, $\boldsymbol{\phi}$. For the jet flame considered in this paper, we observe that the structure (definition) of progress variables thus identified is independent of spatial filtering, and that the functional dependency between $\boldsymbol{\eta}$ and $\boldsymbol{\phi}$ is likewise independent of filter width. The same observations hold for Flame D [16] in the context of Reynolds-averaging rather than filtering. To the extent that this is a universal feature of PCA-based models, these results imply that no explicit closure model is required for the thermochemistry. Further investigation into this observation, particularly using data at higher Reynolds numbers, is certainly warranted to corroborate these observations.

The "principal components" that form the independent variables to which the state variables are mapped, are not conserved scalars and their source terms, S_η, must be parameterized as functions of $\boldsymbol{\eta}$. We have explored using MARS to achieve this parameterization, and found reasonably accurate mappings. However, further work is required here to achieve mappings that are sufficiently accurate for predictive modeling. A significant finding presented here is that the functional form is independent of filter width so that given $S_{\eta_i} = \mathcal{F}_i(\boldsymbol{\eta})$, $\bar{S}_{\eta_i} = \mathcal{F}_i(\bar{\boldsymbol{\eta}})$.

We have also considered the effects of scaling (preprocessing the data) on the accuracy of the resulting PCA-based models and have shown that there can be significant influence, particularly on the accuracy with which PC source terms, S_{η_i}, can be represented.

For a point of reference, we have compared the PCA-based models with SLFM and demonstrated that the PCA model is able to, with the same number of parameters, achieve significantly higher accuracy than SLFM.

Future work will focus on improving the accuracy with which the PCA transformation can parameterize source terms and consider *a posteriori* analysis of the modeling approach. Also, it remains to be seen how universal a given PCA definition is. There are several factors that influence this, including sampling density in state space, the dataset from which the PCA is obtained, etc. These important issues will be considered in future work.

7. Acknowledgements

The authors gratefully acknowledge support from the Department of Energy under award number DE-NT0005015 and the National Science Foundation PetaApps project 0904631.

References

[1] H. Tennekes, J. L. Lumley, A First Course in Turbulence, MIT Press, 1972.

[2] N. Peters, Prog. Energy Combust. Sci. 10 (1984) 319–339.

[3] N. Peters, Proc. Combust. Inst. 24 (1986) 1231–1250.

[4] H. Pitsch, N. Peters, Combust. Flame 114 (1998) 26–40.

[5] J. A. van Oijen, L. P. H. de Goey, Combust. Sci. and Tech. 161 (2000) 113–137.

[6] J. A. van Oijen, Flamelet-Generated Manifolds: Development and Application to Premixed Flames, Ph.D. thesis, Eindhoven University of Technology, 2002.

[7] J. A. van Oijen, L. P. H. de Goey, Combust. Theory Modelling 6 (2002) 463–478.

[8] J. C. Sutherland, Evaluation Of Mixing And Reaction Models For Large-Eddy Simulation Of Non-premixed Combustion Using Direct Numerical Simulation, Ph.D. thesis, University of Utah, 2004.

[9] H. Bongers, J. A. van Oijen, L. M. T. Sommers, L. P. H. de Goey, Combust. Sci. and Tech. 177 (2005) 2373–2393.

[10] J. C. Sutherland, P. J. Smith, J. H. Chen, Combust. Theory Modelling 11 (2007) 287–303.

[11] O. Gicquel, N. Darabiha, D. Thévenin, Proc. Combust. Inst. 28 (2000) 1901–1908.

[12] B. Fiorina, R. Baron, O. Gicquel, D. Thevenin, S. Carpentier, N. Darabiha, Combust. Theory Modelling 7 (2003) 449–470.

[13] B. Fiorina, O. Gicquel, S. Carpentier, N. Darabiha, Combust. Sci. Technol. 176 (2004) 785–797.

[14] E. R. Hawkes, R. Sankaran, J. C. Sutherland, J. H. Chen, in: Proc. Combust. Inst., volume 31, pp. 1633–1640.

[15] N. Punati, J. C. Sutherland, A. R. Kerstein, E. R. Hawkes, J. H. Chen, Proc. Combust. Inst. 33 (2011) 1515–1522.

[16] International workshop on measurement and computation of turbulent nonpremixed flames, ????

[17] I. T. Jolliffe, Principal Component Analysis, Springer Series in Statistics, New York: Springer, 2nd edition, 2002.

[18] J. Shlens, A Tutorial on Principal Component Analysis, Technical Report, University of California, San Diego, La Jolla, 2009.

[19] A. Parente, Experimental and Numerical Investigation of Advanced Systems for Hydrogen-based Fuel Combustion, Ph.D. thesis, Università di Pisa, 2008.

[20] A. Parente, J. C. Sutherland, L. Tognotti, P. J. Smith, Proc. Combust. Inst. 32 (2009) 1579–1586.

[21] J. C. Sutherland, A. Parente, Proc. Combust. Inst. 32 (2009) 1563–1570.

[22] A. Parente, J. C. Sutherland, B. B. Dally, L. Tognotti, P. J. Smith, Proc. Combust. Inst. 33 (2011) 3333–3341.

[23] H. C. Keun, T. M. D. Ebbels, H. Antti, M. E. Bollard, O. Beckonert, E. Holmes, J. C. Lindon, J. K. Nicholson, Analytica Chimica Acta 490 (2003) 265–276.

[24] R. A. van den Berg, H. C. Hoefsloot, J. A. Westerhuis, A. K. Smilde, M. J. van der Werf, BMC Genomics 7 (2006) 142.

[25] I. Noda, Journal of Molecular Structure 883-884 (2008) 216–227.

[26] U. Maas, S. B. Pope, Proc. Combust. Inst. 24 (1992) 103–112.

[27] J. H. Friedman, Annals of Statistics 19 (1991) 1–67.

[28] J. H. Friedman, Fast MARS, Technical Report 110, Stanford University Department of Statistics, 1993.

[29] J. H. Friedman, C. B. Roosen, Statistical Methods in Medical Research 4 (1995) 197–217.

UU IR Author Manuscript