Flow, Turbulence and Combustion
https://doi.org/10.1007/s10494-020-00176-4

Data-driven Classification and Modeling of Combustion Regimes in Detonation Waves

Shivam Barwey[1] · Supraj Prakash[1] · Malik Hassanaly[1,2] · Venkat Raman[1]

Received: 4 December 2019 / Accepted: 27 May 2020

Abstract
A data-driven approach to classify combustion regimes in detonation waves is implemented, and a procedure for domain-localized source term modeling based on these classifications is demonstrated. The models were generated from numerical datasets of canonical detonation simulations. In the first phase, delineations of combustion regimes within the detonation wave structure were analyzed through a clustering procedure. The clustering output usefully illuminated distinctions between detonation, deflagration, and intermediary regimes within the wave structure. In the second phase, the resulting delineated fields from the clustering step were used to guide localized source term modeling via artificial neural networks (ANNs), enabling a type of classification-based regression approach for source term estimation. A comparison of the estimations obtained from the local ANNs (trained for a subset of the domain given by a particular cluster) with the global ANN counterparts (trained agnostic to the clustering) showed general improvement of estimations provided by the domain-localized modeling in most cases. Ultimately, this work illuminates the useful role of data-driven classification and regression techniques for both physical analysis of the complex wave structure and for the development of new models which may serve as suitable pathways for long-time simulations of complex combustion systems (such as rotating detonation combustors).

Keywords Rotating detonation combustors · Data-driven modeling · Clustering · Neural networks · Turbulent combustion modeling

1 Introduction

The field of combustion modeling has experienced significant shifts in recent years with the emergence of data-driven techniques (Raman and Hassanaly 2019). These techniques, many of which are rooted in machine learning (ML)-based algorithms, have naturally risen in popularity as a result of (a) their inherent versatility in dealing with abundance of data, and (b) their compatibility with new high-performance computing (HPC) architectures

✉ Shivam Barwey
sbarwey@umich.edu

[1] Department of Aerospace Eng, University of Michigan, Ann Arbor, MI, USA

[2] Present Address: National Renewable Energy Laboratory (NREL), Golden, CO, USA

(i.e. GPU-centric systems). The datasets used to develop such models, which contain the underlying effects of nonlinear combustion processes of interest, can be obtained from either high-fidelity numerical simulations (Kapoor et al. 2001; Franke et al. 2017), high-resolution experimental laser diagnostics (Barwey et al. 2019), or operational real-world sensor data streams from well-instrumented propulsion devices (Giorgetti et al. 2019). As a result, a variety of data-driven models have been developed for modeling and feature extraction purposes in combustion applications (Barwey et al. 2019; Ranade et al. 2019; Malik et al. 2018; Akintayo et al. 2016; Tammisola and Juniper 2016; Landge et al. 2014). A particularly useful application of these techniques specific to the combustion field lies in the modeling of chemical source terms for computational fluid dynamics (CFD) simulations of complex combustion processes.

In turbulence-driven combustion processes, computational modeling of complex chemical reactions pose a perpetual challenge (Raman and Hassanaly 2019). This is primarily due to the computational burden associated with solving highly nonlinear stiff equations for the advancement of chemical state, which becomes especially prohibitive in combustion environments that require detailed chemical mechanisms to properly describe the effect of turbulence-chemistry interactions at small length and time scales. Thus, in a general sense, the primary goal of source-term modeling is to reduce the cost of these computations through the construction of physics-inspired reduced representations using data-driven techniques or other means. This objective has ultimately led to the development of a wide variety of modeling approaches related to the underlying goal of computationally-efficient source term estimation (Pitsch 2006; Raman and Hassanaly 2019).

There are essentially three different approaches. First, the chemical source term computation can be accelerated through adaptive algorithms (Pope 1997; Bell et al. 2000). For instance, the in-situ adaptive tabulation (ISAT) approach replaces computationally expensive direct integration of stiff ODEs with linear interpolation in "trust" regions. Second, reduced-order models that exploit the low dimensionality of the state space spanned by the thermochemical composition could be used. The very popular flamelet (Peters 2000) or flamelet-generated manifold (FGM) (Van Oijen and de Goey 2004; Jha and Groth 2012; Fiorina et al. 2005) approaches leverage this computational reduction. Third, the chemistry mechanism could be reduced by removing species/reactions that are not relevant for particular operating conditions (Lu and Law 2005; Warnatz et al. 1996; Malik et al. 2018). This last approach includes the so-called virtual chemistry formulations, whereby entirely new chemistry mechanisms satisfying certain physical constraints are developed from simulations of equivalent, secondary systems (Fiorina 2019; Cailler et al. 2017). These virtual chemistry approaches do not require the development and storage of the source term table, which are some of the critical issues of flame-generated methods. However, the validity of the model is closely linked to the selection of canonical systems.

An alternate route for modeling that must also be mentioned is enabled by artificial neural networks (ANNs). ANNs are well-suited for source term modeling and build upon the above techniques for many reasons: (1) they utilize nonlinear representations of linear combinations of the thermochemical state to generate source term predictions, (2) they can be more memory-efficient than traditional tabulation-based approaches, and (3) they are well-primed for the newer GPU-accelerated HPC architectures. ANNs have been used both recently and in previous decades for combustion modeling for these above reasons (Christo et al. 1996; Kempf et al. 2005; Sen and Menon 2010) and they have also been extended to in-situ settings to good effect (Kapoor et al. 2001; Chen et al. 2000).

While each of the above techniques have been used for different combustion systems, the focus here is on an emerging propulsion concept, termed the rotating detonation

combustor (RDC), which is increasingly being considered for different applications (Bykovskii et al. 2006; Frolov et al. 2013; Zhou et al. 2016; Lu and Braun 2014; Sato et al. 2018). A unique need in the modeling of such combustors is the simulation of long-time behavior. Currently, the computational cost associated with chemistry limits the total simulation time to tens of milliseconds, which is not sufficient to ensure that the system has reached a statistically stationary state (Sato and Raman 2019). The focus of this work is to develop a machine learning-based chemistry representation that vastly accelerates the integration of chemical source terms to enable long-time simulations of RDCs.

In a typical RDC as shown in Fig. 1, a detonation wave propagates azimuthally in an annular chamber, processing a mixture of fuel and air injected axially (or radially, depending on the configuration) from the bottom of the device. However, the combustion processes within an RDC can be highly unsteady and chaotic, with both detonative and deflagrative combustion due to mixture non-uniformity and adverse wave behavior. The detonation process is highly non-ideal, with a reaction zone significantly broader than the theoretical expectation, flanked by parasitic deflagration ahead of the wave and slow, delayed heat release that extends far behind the wave and homogenizes the mixture (Chacon and Gamba 2019; Prakash and Raman 2019). Since detonations occur over short length and time scales, the computational grids used to simulate the geometries need to fully resolve the chemical reactions and flow gradients. Further, the stability of the detonation wave is highly influenced by the fuel-air mixing process which is driven by the high-shear generated from the injection scheme. Hence, capturing the turbulent mixing mechanisms and their interplay within the unsteady detonation process is crucial, but expensive. From the modeling side, challenges emerge from the fact that the detonation domain is segmented into distinct detonation, deflagration, and transitional (or intermediary) regions (Prakash et al. 2019). Thus, it is crucial from both perspectives to systematically differentiate between the different forms of combustion. These combustion regimes vary spatially, resulting in highly stiff chemistry contained within small portions of the domain that greatly increase the computational cost of these studies. Therefore, to enable the long-time simulations of RDCs, associated models should both (a) take into account differences

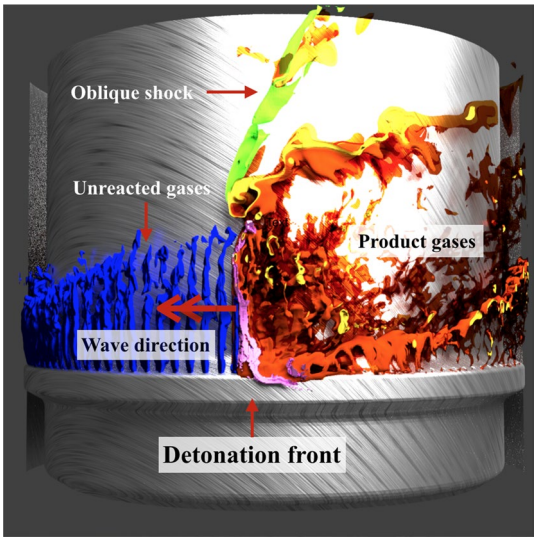

Fig. 1 Full-scale simulation snapshot with the detailed flow features of a typical RDC combustor (Sato et al. 2019)

between these complex spatially-varying regimes and (b) properly leverage the rich datasets generated by the well-resolved numerical simulations.

The contribution of this work lies in the demonstration of a different data-driven combustion modeling approach which addresses the above mentioned numerical and modeling challenges. The approach consists of two steps. In the first step, a clustering algorithm is used to extract the varying regimes of combustion in the detonation wave structure in an unsupervised fashion. In the second step, the clustering is used to drive the training of several ANNs, each of which produces source term estimations that are localized to the regimes identified in the first step. More specifically, using datasets generated from direct numerical simulations of canonical detonating flows relevant to the study of RDC physics, the primary focus of this work is to show that (a) the clustering output recovers physically relevant delineations of combustion regimes in the detonation wave structure in an unsupervised manner, and (b) the source term estimations obtained from ANNs tailored to these different combustion regimes (i.e. local ANNs) are more accurate than the estimations obtained when *not* considering the clustering output during ANN training (i.e. global ANN).

The paper proceeds as follows. In Sect. 2, the sample detonation datasets derived from past direct numerical simulations are described, and the reasoning for their selection is explained. In Sect. 3, the clustering results are discussed. In Sect. 4, the localized training of ANNs derived from the clustering in Sect. 3 is carried out. Both improvements and advantages of the classification-based learning approach are discussed. Lastly, in Sect. 5, concluding remarks and directions for future work are presented. Other additions to the clustering approach are discussed in the Appendix.

2 Description of Data

A direct numerical simulation (DNS) approach is used to generate the models for the detonation wave structure. In this section, the numerical details of the associated solver are described and the details of the training/testing datasets used for the data-driven modeling results throughout the paper are presented.

2.1 Numerical Solver and Chemical Mechanism

The numerical simulation must capture the small-scale turbulence structures and the induction region, which is very thin compared to other length scales that dictate detonation wave behavior. For most chemical compositions and flow conditions, the detonation wave—a primary shock wave coupled with a reaction zone—is on the order of a few micrometers in width. The flow gradients must therefore be accurately resolved to capture the interactions between the fuel-air mixture and the detonation wave. Furthermore, detailed chemical mechanisms are required to model the species transient behavior across the wave accurately. To this end, the governing equations for fluid flow consist of the mass, momentum, and energy conservation equations augmented by the species conservation equations that incorporate chemical reactions. The system of equations is closed using the ideal gas equation of state.

The Navier-Stokes equations in compressible form are incorporated in the in-house compressible flow solver UTCOMP, which has been validated for a range of shock-containing flows (Fiévet and Raman 2018; Fiévet et al. 2015, 2017). The detailed chemical

kinetics for hydrogen-oxygen combustion with a nitrogen diluter is derived from the 9-species 19-reaction chemical mechanism of (Mueller et al. 1999) using CHEMKIN-based subroutines. The solver has been validated for hydrogen-air and hydrogen-oxygen detonation using a series of one-, two-, and three-dimensional canonical cases (Prakash et al. 2019). A structured grid is utilized with a cell-centered, collocated variable arrangement. A 5th order weighted essentially non-oscillatory (WENO) scheme (Jiang and Peng 2000) is used for computing the non-linear convective fluxes and the non-linear scalar terms are calculated using a quadratic upstream interpolation for convective kinematics (QUICK) scheme (Herrmann et al. 2006). A 4th order central scheme is used to calculate the diffusive terms and a 4th order Runge-Kutta scheme is used for temporal discretization. Additional details on solver parameters used in this work are provided in Ref. Prakash et al. (2019).

2.2 Training and Testing Data

In the investigation of RDC flow physics, the primary flow features stem from the injector dynamics and the mixture inhomogeneity within the annulus. The fuel-air mixture is highly stratified due to the turbulent mixing characterized by recirculation zones and free-shear/jet interactions within the combustor. These interactions can be simulated through canonical channel flow geometries. The training dataset is derived from case 1b of the linear injector array simulation of Ref. Prakash et al. (2019). The testing dataset is sourced from a channel detonation simulation with a stratified fuel-air mixture, as implemented in Ref. Prakash et al. (2019). The schematic of the numerical configuration and fuel-air distribution in the training and testing datasets are given in Fig. 2. The two cases are at similar operating conditions and the same combustion regimes are present, allowing for a convenient way to assess confidence in the data-driven approach.

The full *training dataset* is a set of snapshots from a linear injector array, known as the linearized model detonation engine (LMDE), as depicted in Fig. 2a. In the LMDE configuration, a fully-developed detonation wave processes a partially-premixed stoichiometric hydrogen-air mixture injected from an array of 15 injectors. Here, the jet turbulent mixing controls the local gas composition before the detonation wave processes the mixture. For numerical cost considerations, the operating pressure was lowered to 0.5 atm with an ambient temperature of 297 K when simulating the LMDE. Upon running the respective

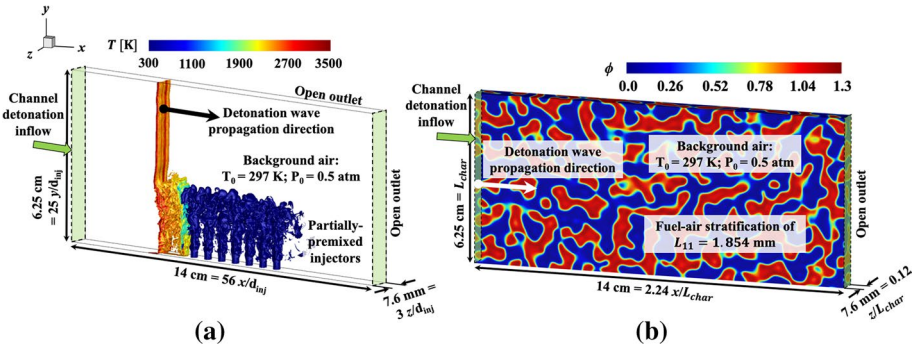

Fig. 2 Configurations of the **a** linear injector array with isocontour of $H_2 = 0.016$ and OH = 0.0053 colored by temperature and **b** channel with stratified fuel-air mixture of integral length scale $L_{11} = 1.854$ mm, which serve as the training and testing datasets, respectively

simulations at these prescribed operating conditions, a set of thirteen consecutive two-dimensional snapshots, separated by $dt = 0.25\ \mu$s, was extracted along the depthwise mid-plane of the domain. The thirteen snapshots capture the detonation wave structure as it passes through the 12th injector.

The full *testing dataset* is composed of ten snapshots ($dt = 0.25\ \mu$s) sourced from a canonical simulation of channel detonation with a quiescent stratified fuel-air mixture. The range of equivalence ratios is constrained from 0 to 1.3. This is shown in Fig. 2b. In a confined channel of dimensions identical to the LMDE, fuel-air stratification is introduced through a method used for turbulent mixing studies in homogeneous isotropic turbulence; further details are provided in Ref. Prakash et al. (2019). Similar to the training dataset, the ambient pressure and temperature are 0.5 atm and 297 K. The snapshots are recorded as the detonation wave reaches near the end of the channel. Though the testing set data may seem less physically complex, the gas mixture spans a greater range of thermochemical state space in a more controlled manner in comparison to the training dataset, allowing for ideal model testing conditions. Further, this testing set particularly useful as it contains similar detonation-induced combustion regimes as the LMDE training case. However, the differences in spatial fuel-air distribution and influence of each regime on wave propagation due to the presence of fuel stratification make it a good candidate for assessing model extrapolation.

As a visual example, numerical Schlieren images of the depthwise mid-plane of a single snapshot (one of thirteen) of the training (left) and testing (right) datasets are shown in Fig. 3a. The datasets are cropped to the outlined regions which capture all of the combustion regimes around a detonation wave: (a) an ambient fuel-air mixture, (b) a shock-separated region in the absence of fuel, (c) a primary detonation region, and d) a

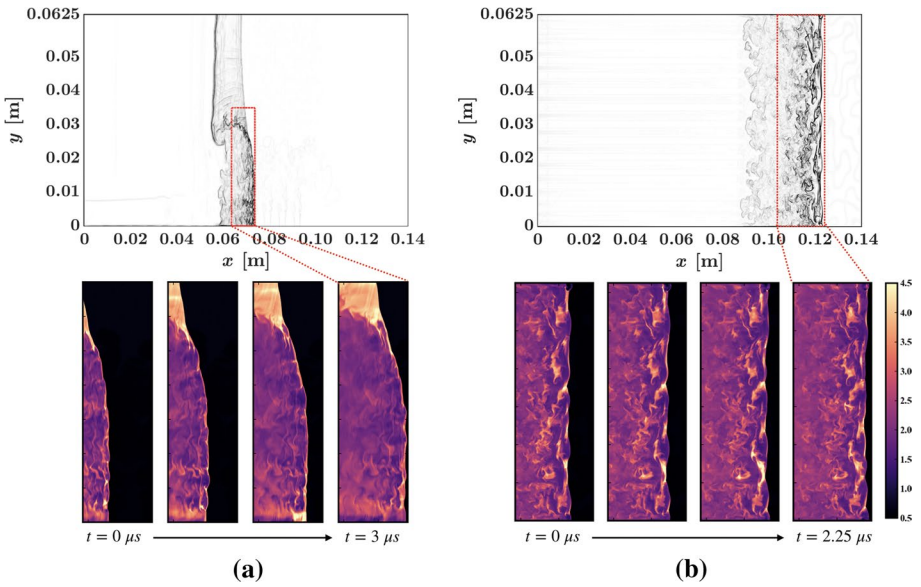

Fig. 3 The top row shows numerical Schlieren images of the detonation wave through **a** the linear injector array and **b** a stratified fuel-air mixture in a confined channel. The bottom row shows a few of the training and testing set snapshots displaying the time evolution of density (units of kg/m^3). For convenience, the first snapshot in the dataset sequence is denoted as the initial state ($t = 0$)

post-detonation deflagrative combustion region trailing the shock front. After cropping, each training snapshot is composed of 1011 pixels in the x direction and 363 in the y direction, resulting in 3.7×10^5 grid points per snapshot. Each testing snapshot consists of 411 pixels in the x direction and 858 pixels in the y direction, for a total of 3.5×10^5 grid points per snapshot.

As a final note, in the regime classification analysis (Sect. 3), all 13 of the LMDE snapshots are used in training and all 10 of the stratified mixture snapshots are used in testing. To demonstrate the source-term estimation concept in Sect. 4, only one of 13 available LMDE snapshots is used in training (i.e. data from one snapshot is used to train the neural networks), and one of 10 available stratified mixture snapshots in testing. Additionally, for better assessment of the model dependence on configuration, the source term estimation ANN models will also be evaluated on an unseen snapshot from the same LMDE training configuration at a future time.

3 Regime Classification

The goal in this section is to (a) classify grid points within the domain to macroscopic regions of interest, and (b) assess the physical significance of these classifications. In particular, the demarcations of regions associated with detonation and deflagration are desired, as they will guide a localized modeling procedure in Sect. 4. Although there are many potential routes available to the user for successful labeling, the one used here is an unsupervised machine learning approach known as K-means clustering. Algorithmic details are given in Appendix A.1. The finer mathematical details of K-means can be found in Refs. Arthur and Vassilvitskii (2007), Steinley (2006).

3.1 Clustering Methodology

Consider a set of grid points $\mathcal{S} = \{S_1, S_2, \ldots, S_N\}$ extracted from one or more numerical or experimental snapshots. For example, in the training dataset described in Sect. 2, there are 13 snapshots each composed of 3.7×10^5 grid points, yielding a total number of grid points $N = (3.7 \times 10^5) \times 13$. Further, consider another set $\mathcal{F} = \{F_1, F_2, \ldots, F_{N_f}\}$ of features associated with each grid point in $\mathcal{S}$ (i.e. each grid point is of dimension N_f). The elements in $\mathcal{F}$ can include velocities, temperature, density, mass fractions, progress variable, or any other quantities of interest. In this work, the clustering is performed over the set $\mathcal{S}$. In non-hierarchical hard-clustering algorithms such as K-means, the three facets that govern the clustering output are (1) how the grid point is represented via $\mathcal{F}$, (2) how the similarity between two grid points is defined (through a distance function), and (3) the number of clusters, N_k.

The K-means algorithm upon convergence produces a set of *centroids* $\mathcal{C} = \{C_1, C_2, \ldots C_{N_k}\}$ and *labels* for each grid point in $\mathcal{S}$. These centroids are statistical quantities that are similar to the grid points in that they are also made up of N_f features. A subset of grid points in $\mathcal{S}$ which have similar feature values is represented by the same centroid. The centroid that represents a grid point is the one that is the closest to the grid point based on a distance measure. Here, the L_2-norm in the $\mathbb{R}^{N_f}$ space is used to compute this distance. As such, the grid point is labeled to cluster k if it is closest to centroid k in the Euclidean sense. This restricts each grid point in $\mathcal{S}$ to be labeled with only one centroid. The N_k *clusters* can then be visualized in physical space through the grid point labels, as

all members in $\mathcal{S}$ which share the same centroid label occupy the same cluster. Further, the same set $\mathcal{C}$ produced from one dataset can be used to classify any other dataset so long as each grid point in this new dataset contains the same number of features N_f as the original data used to generate the centroids. In the results below, the clustering generated by the training set (LMDE described in Sect. 2) will also be extended to the testing set (stratified fuel-air mixture) to validate the physical significance of the clusters.

It should be noted that the selection of features $\mathcal{F}$ requires some level of physical intuition by the user. In this application, the features should properly characterize differences in regions of differing chemical mechanisms in the detonation and deflagration regimes. With this in mind, $\mathcal{F}$ should contain at minimum temperature, pressure, and intermediary species as these are the primary drivers of the combustion chemistry in the computation of reaction source terms and in the determination of local flow enthalpies. Thus, $\mathcal{F} = \{\rho, T, P, Y_H, Y_O, Y_{H_2O}, Y_{OH}, Y_{HO_2}, Y_{H_2O_2}\}$ is selected as a starting point, producing $N_f = 9$. This set should be treated as an upper bound on the total "required" number of elements in $\mathcal{F}$—it is not guaranteed that the contribution of each feature to the resulting classifications is equal. As such, to reduce the level of supervision required in feature selection, a downselection procedure for $\mathcal{F}$ using an importance metric is provided in Appendix A.3.

The selection of the number of clusters, N_k, is non-trivial and can vary based on the clustering application (Barwey et al. 2019, 2020; Kaiser et al. 2014; Steinley 2006). Broadly speaking, the number of clusters should be high enough such that there is sufficient resolution of the macroscopic regions in the domain and low enough such that there is a sufficient number of grid points per cluster. Note that although there are two overarching regions of interest in the domain (detonation and deflagration regimes), this does not necessarily translate to the selection of $N_k = 2$. Since the level of chemical complexity in these regimes is not the same, a higher level of refinement is expected to establish finer distinctions between detonation and deflagration. Through a user-guided inspection, it was found that cluster numbers beyond 6 for the dataset used here did not result in significant differences in the delineation structure. As such, all of the clustering results below are shown for $N_k = 6$. Further detail regarding the subtlety of N_k selection can be found in Refs. Barwey et al. (2020), Steinley (2006).

3.2 Clustering Results

The clustering procedure was carried out on the training set for $N_k = 6$. Before proceeding with the physical interpretation of the clusters themselves, an illustration of the way in which the resulting delineations can be visualized is shown in Fig. 4. The grid point assignments of one training snapshot for each of the 6 clusters is shown on the left. Each delineated region represents a single cluster whose grid points are colored in white. Figure 4 shows that 1) some clusters represent more grid points than others, and 2) the cluster structures preserve a large amount of coherence in space. These two points will be assessed further below. Additionally, since these grid point labels are non-overlapping (a grid point is only assigned to one cluster), all of the cluster labels can be consolidated into a single image called the *segmented field*, shown on the right of Fig. 4. Such segmented flowfields provide a concise juxtaposition of the various cluster patterns and will be used to facilitate the discussion of the clustering output below.

A summary of the labeling results as produced by the clustering is provided in Fig. 5 for both training (left) and testing (sets). In each of the training and testing sets, three snapshots are used to concisely illustrate the effect of the time-evolution of the segmented

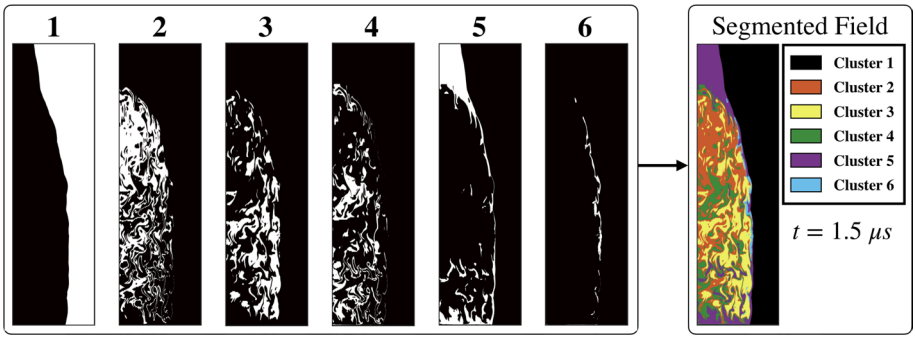

Fig. 4 (Left) Grid point labels for the 6 clusters for one snapshot, where white regions correspond to cluster assignment. (Right) Combination of the labels into segmented field, where different colors indicate clusters

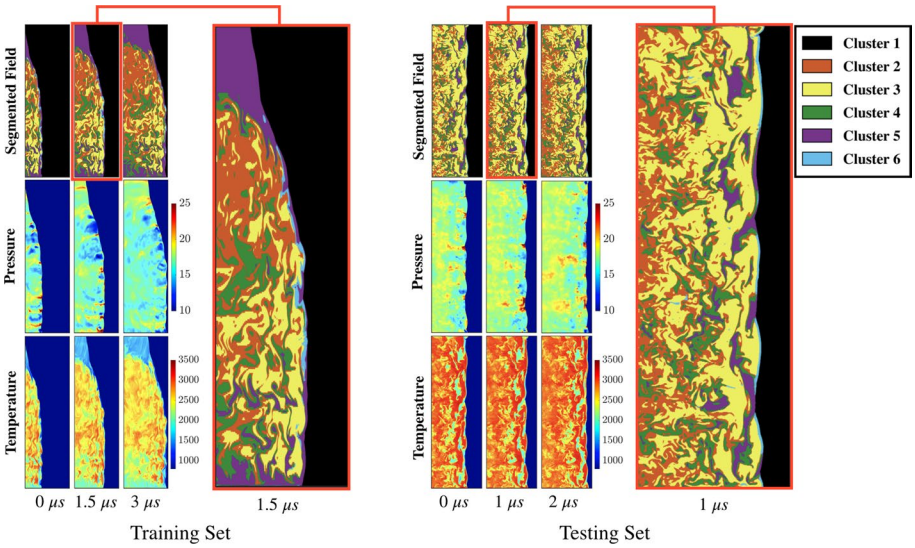

Fig. 5 Spatial classification of the detonation wave in (left) the training and (right) testing datasets. Shown is the time progression for each configuration given by a smaller subset of the available snapshots. The evolution of the segmented field is compared to the pressure (atm) and temperature (K) distributions. A blow-up of an intermediate snapshot is provided in each case for clearer visualization of the segmented field

fields. Pressure and temperature fields at the same time instances are also shown. It should be noted that although only three training snapshots are shown in Fig. 5, all 13 were used to produce the K-means output.

From the training data results in the left of Fig. 5, it can be seen that the selection of $N_k = 6$ produces distinct clusters corresponding to the different flow states within the training data domain. Cluster 1 corresponds to the ambient mixture region of varying equivalence ratio at 0.5 atm and 297 K. Cluster 5 signifies the shock-separated region where the lack of fuel causes the detonation wave to transition to a leading pressure wave followed by a lagging region of high-temperature products. Furthermore, these regions appear along the

entire detonation wavefront, highlighting the finite distance behind the shock front before the gas mixture begins to react. Thus, cluster 5 represents a region where the gas has been compressed but is largely non-reacting. The shock wave acts as a near-sonic throat, where past analysis has shown that the Mach number is approximately 1.3 (Prakash et al. 2019). Cluster 6, the smallest of all clusters (in terms of proportion of represented grid points) is classified in regions around and including the triple points at the detonation wavefront.

The triple points are concentrations of high-pressure and temperature, resulting in high heat release rates. Along the shock front, sections of the wavefront where cluster 6 is observed signify strong detonation, where the leading shock wave and the reaction zone are closely coupled. Here, the profile of the thermodynamic parameters across the wave closely resembles the ideal Zeldovich-von Neumann-Döring (ZND) profile. It is important to note that cluster 6 exists only in parts of the wavefront. Surrounding cluster 6, regions identified by cluster 5 signify shocked gas, either with entrained non-reacting gas or gas undergoing the induction process. Interestingly, the triple points highlighted by cluster 6 appear in the transition to the shock-separated region, becoming progressively smaller as the pressure concentrations are dissipated within the upper portion of the frame. Thus, the three-dimensional detonation wave structure varies spatially along the front, leading to a corrugated temperature and pressure profile. As the detonation wave moves through the partially-premixed distribution of fuel and air, triple points propagate along the wave front. As they near regions of non-detonable mixture, the pressure concentrations diminish, such as near the shock-separated region and base of the channel. The propagation of these triple points is crucial to maintaining wave strength and the trailing transverse waves form vortical structures that support post-detonation homogenization.

Following the detonation mode, combustion transitions to deflagration represented by clusters 2, 3, and 4. Here, the post-detonation gases burn at high temperatures, rendering temperature a main driver for the distinction between different combustion regimes. Due to post-detonation expansion, temperature and pressure decrease. The separation of the detonation front above the fill height of the injectors, highlighted by cluster 5, creates a shear layer that bisects the wavefront. Post-detonation gases travel upwards behind the wavefront and are turned by the shear layer, increasing the mixing of residual gases. It is interesting to note that this results in more complete combustion of gases. The post-detonation region is identified by cluster 3, where high temperature deflagration and gas expansion supports detonation wave propagation. However, regions of slower heat release within clusters 2 and 4 extend farther behind the wave front. The post-detonation region within the primary fill height of the injected mixture is highlighted by cluster 3, signifying a more ideal detonation process within this height range. Cluster 2 exists within regions of lower temperature deflagration and increased mixing enforced by the shear layer at the fill height. The base of the channel where regions of air in between the injectors are entrained is highlighted by cluster 5 (shocked gas) and cluster 4 (weaker, low temperature deflagration). Relating the temperature and pressure contours to the classified regions, deflagration is captured primarily by clusters 2 and 3 whereas cluster 4 represents the transition to cluster 5.

The application of the learned classifications to the testing data (right of Fig. 5) essentially shows where the same features/regimes recovered from the training LMDE dataset (the 6 clusters) are located in the stratified testing case. Interestingly, the representation of the features along the detonation wavefront are quite similar in physical significance to that of the training set, which validates the above analysis. For example, strong detonation where the shock and reaction fronts are closely attached occur only in areas where fuel exists, denoted by cluster 6 along the wave front. Cluster 5 is entrained within the post-shock region due to regions of non-detonable gas processed by the wave front that is

characterized by high pressures due to shock compression. In the training dataset, cluster 5 was visible near the fringes of the detonable mixture. Here, the compressed gas exists throughout the post-detonation region in regions roughly corresponding to the size of the stratification. Cluster 3 represents high-temperature deflagration and gas expansion in the post-detonation region, where a slower heat release process consumes the residual reactant mixture. The eddy structures stemming from the collision of triple points and transverse waves ensure that the partially burnt gases are consumed through the deflagration process. Similar to the training dataset, clusters 2 and 4 exist further from the wave front. However, as the wave front is largely perpendicular to the propagation direction, the transition between clusters 2, 3, and 4 is more evident. Cluster 4 exists closer to the wave front and may be characterized by lower temperature deflagration. The appearance of cluster 2 at distances far from the front signify that this region represents a more homogeneous mixture where the reacting gas has reached some equilibrium condition. This further qualifies that the regions represented by cluster 2 in the training dataset may represent a more homogeneous mixture due to the mixing enforced by the shear interactions at the injector fill height.

The above discussion provided a physical interpretation of the delineations produced by the clustering. With this context, the localized source term modeling can now be performed. Note that the clustering output can be extended in other useful ways to reveal further insight into physical processes; details into the time evolution of cluster sizes, for example, are shown in Appendix A.2. Further, an approach to remove redundant features from $\mathcal{F}$ without affecting clustering results is provided in Appendix A.3.

4 Source Term Regression

In this section, the delineated regions for $N_k = 6$ obtained in Sect. 3 are used to guide the modeling for thermochemical source terms of interest. The ANN-based source term modeling approach is of particular interest here since the relations between the thermochemical state and the source terms are highly nonlinear and render computationally intensive chemistry routines intractible in reacting detonation solvers. The goal is to show that proper utilization of the segmented fields produced in Sect. 3 can lead to more robust modeling procedures that better enable long-time simulations. Specific details into ANN methodology are omitted here, but can be found in Refs. Murphy (2012), Goodfellow et al. (2016).

4.1 ANN Architecture

Two types of ANNs are considered: global and local. The global ANN is trained agnostic to the cluster labels and uses the full set grid points as the training data. In contrast, the local ANN refers to the model trained only for data belonging to cluster k. This means that there are a total of 1 global ANN and 6 local ANNs (1 for each cluster). This is similar to the method used in Ref. Barwey et al. (2019) to display the advantage of domain-localized modeling, but the difference here is that the domain localization is learned via the clustering step. Throughout the results below, local versus global ANN prediction accuracy for the source terms (conditioned on cluster number) will be discussed.

All ANNs contain two hidden layers, each with 50 neurons. The hyperbolic tangent function is used for hidden layer activations. As mentioned in the end of Sect. 2, in the ANN implementation, the training set is restricted to a single snapshot from the LMDE

simulation. The testing set contains a snapshot from the stratified fuel-air mixture case. Additionally, to assess model performance in both similar and different configurations as the training data, the testing routine also consists of unseen snapshot from the LMDE configuration at a later timestep than the training set. These training and testing snapshots are shown in Fig. 6a.

The full ANN architecture is visualized in Fig. 6b. The ANN inputs are the same as the *reduced* feature set defined in Appendix A.2; that is, all input grid points to the ANNs are described by $\{\rho, T, P, Y_H, Y_O, Y_{H_2O}, Y_{OH}\}$. The output source terms are given by $\{\dot{T}, \dot{Y}_H, \dot{Y}_O, \dot{Y}_{H_2O}, \dot{Y}_{OH}\}$. All inputs and outputs were standardized before training. Gradients of a mean-squared error (MSE) loss function are found using backpropagation, and the Adam algorithm was used for parameter optimization (Hecht-Nielsen 1992; Kingma and Ba 2017). Training samples were randomly shuffled with 10% of the dataset set aside for validation to monitor overfitting. The neural networks were implemented with the PyTorch library (Paszke et al. 2017).

4.2 ANN Results

The MSE values obtained after completion of the training procedure for both local and global ANN models are shown in Fig. 7. In particular, the MSE values for the LMDE training set (top row), LMDE testing set (middle row), and stratified mixture testing set (bottom row) are compared for each of the five source terms and are conditioned on the cluster number (the columns in Fig. 7). Note that since the primary purpose of Fig. 7 is to show the improvements provided by the local ANN models with reference to the global model, the MSE values are derived from the standardized representation of the source terms to enable cross-feature comparisons. Further, the output for cluster 1 is excluded for the sake of brevity as it represents an ambient uninteresting region with regards to source term output.

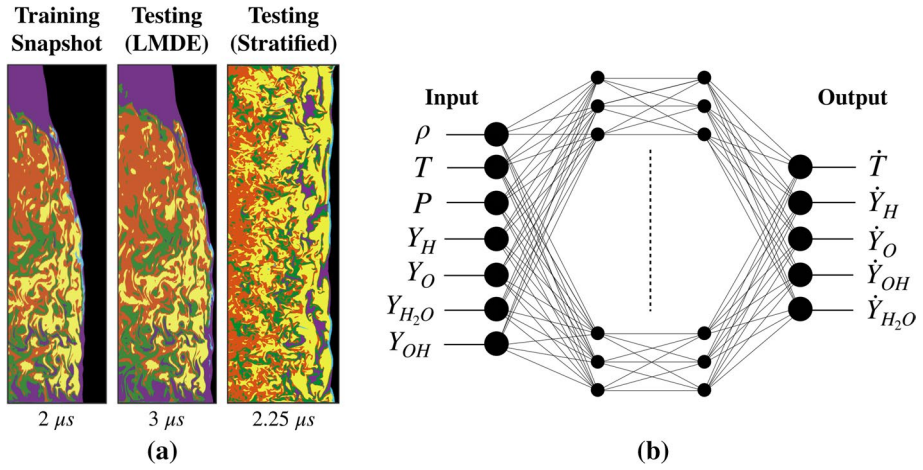

Fig. 6 **a** Representation of the training and two testing snapshots for the demonstration of ANN source term regression (corresponding time instances are given at bottom). **b** Illustration of ANN architecture, where the input corresponds to a grid point represented by the reduced feature set as described in Sect. A.3, and the output is a set of source terms for the same grid point

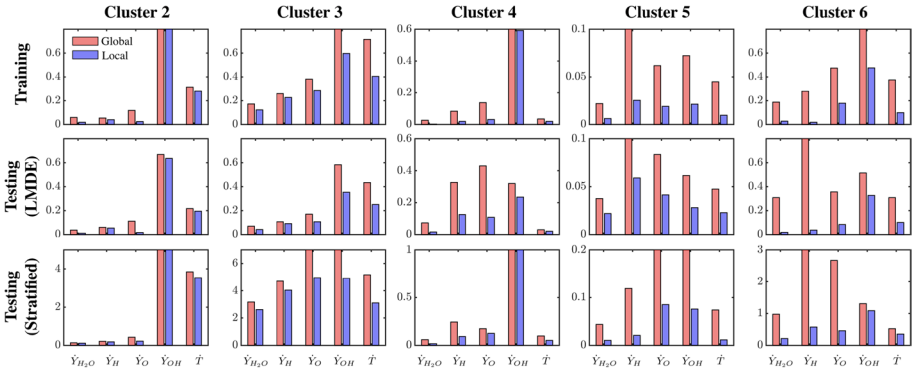

Fig. 7 Comparison of MSE values for each cluster for global (red bars) and local (blue bars) ANN source term predictions. First row corresponds to the training LMDE dataset, second row to the testing set with the same LMDE configuration but at a future timestep, and last row to the testing set in a stratified fuel-air mixture configuration. Within each plot, MSE is compared over all five features as listed at bottom. MSE is computed over standardized outputs to better facilitate comparisons across multiple features

For the training set, the MSE values indicate general improvement provided by the domain-localized modeling. For example, in clusters 5 and 6 (which represent the complex near-wavefront regions as shown in Fig. 5), the local ANNs display significantly lower errors when compared to global counterparts. This is less pronounced for the deflagration clusters 2, 3, and 4, although improvements provided by the localized modeling is still seen in these clusters for all tested source terms. Further, the MSE of the source term predictions for $\dot{Y}_{OH}$ in the training set are significantly higher than the rest for most clusters in both local and global ANN settings (especially in cluster 4).

The middle row of Fig. 7 shows the respective MSE comparisons for a future snapshot (the last of the available 13 as described in Sect. 2) in the *same* configuration as the training set (the LMDE testing set). It is convincing that the overall trends are preserved for the unseen data. Interestingly, in clusters 2, 3, and 6, the local ANN MSE values are lower than those observed in the training set for all source term features. This drop in error for the LMDE testing set is not observed in cluster 5, which indicates that the extrapolative power is not distributed uniformly throughout cluster index. However, the fact that significant improvement provided by the localized models is still seen across the board is a form of confirmation of the similar trends observed in the training set. The bottom row of Fig. 7 shows similarly structured MSE plots, but for a testing snapshot (the last of 10 snapshots as described in Sect. 2) from the stratified mixture configuration. In this testing set, MSE values are significantly higher as expected—the configuration and spatial distribution of the detonation structure are quite different from the training set, and as such, extension of the ANN models to this setting becomes more difficult. Despite this, the MSE trends again show comparatively much more accurate predictions generated by the localized ANN models, especially in clusters 5 and 6. It should be noted that even though such improvements are provided by the local models, instances of significantly high error are seen again for $\dot{Y}_{OH}$ in clusters 2 and 4 in the stratification testing set, even for the local ANNs.

The MSE plots in Fig. 7 were obtained by averaging error quantities over all grid points—to visualize performance in a more direct sense, scatter plots of predicted source term values versus the ground-truth are shown in Fig. 8 for $\dot{T}$ and $\dot{Y}_{H_2O}$ outputs (other source term outputs excluded as all relevant trends are readily identified by these two

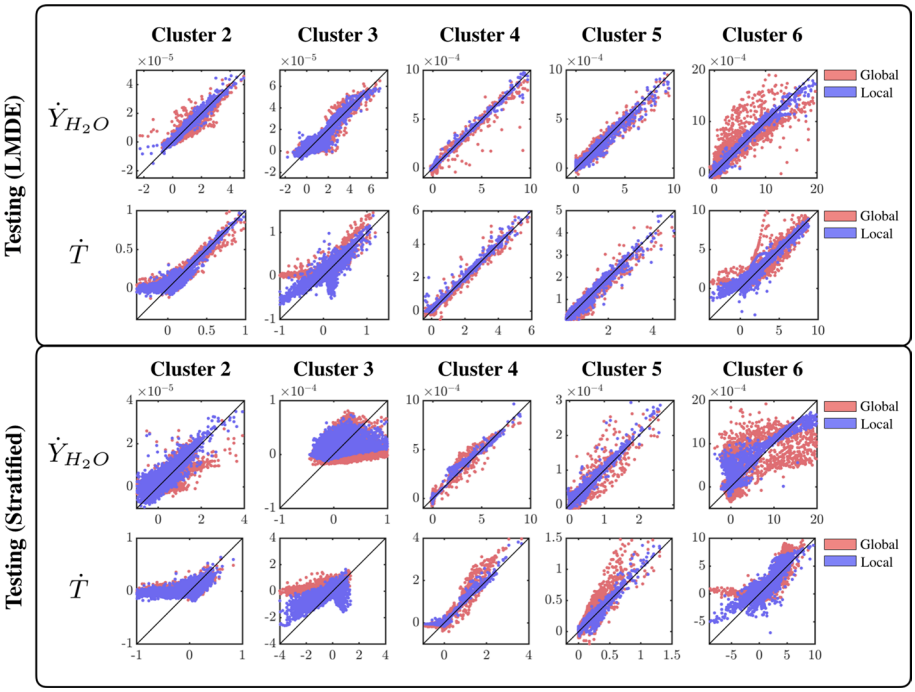

Fig. 8 Scatter plots showing global (red points) and local (blue points) predictions on the y-axis versus ground-truth on the x-axis for clusters 2 to 6. Diagonal solid black lines correspond to exact solutions. Top block corresponds to LMDE testing snapshot (same configuration as training snapshot) and the bottom block to testing snapshot for the stratified mixture configuration. Within each block, the upper row of plots correspond to $\dot{Y}_{H_2O}$ predictions and lower row to $\dot{T}$ predictions. In these plots, source terms have been scaled by the simulation timestep (same value for all grid points) such that the $\dot{Y}_{H_2O}$ is unitless and $\dot{T}$ is in units of Kelvin

outputs). Predictions for both testing sets (LMDE at future timestep and stratified mixture cases) are shown. Here, source term values in each of these plots have been unscaled, and the dashed unit-slope lines represents an ideal prediction.

The topmost set of plots in Fig. 8 show the results for the LMDE testing set (same configuration as training set, but at a future time instance). The reduction in variance of predicted quantities around the exact solution can be seen when considering the local ANN models—the largest improvement is observed for cluster 6, which represents the region near the detonation wavefront and triple points. However, the improvements provided by the localized modeling are more readily seen for $\dot{Y}_{H_2O}$ predictions than for the $\dot{T}$ counterparts. For example, in cluster 6, the local ANN is slightly better at capturing the negative temperature source term values than the global model, but is still overall inaccurate in this region. Further, consider the $\dot{T}$ predictions for cluster 2 versus cluster 3. The local model fails to address negative values in cluster 2; in cluster 3, though still not perfectly accurate due to the innate complexity of temperature source term distributions in detonating flows, the localized modeling better resolves the negative temperature source term than the global counterpart. In clusters 4 and 5, predictions for both local and global models are accurate in the LMDE testing set and follow the exact line quite well, though the local models observe less variation around the exact solution.

The bottom set of plots in Fig. 8 show the results for the stratified fuel-air testing case (different configuration from the training snapshot). As implied by Fig. 7, both global and local model performances are altogether poorer than those seen in the LMDE testing set, though this performance loss is much less severe in clusters 4 and 5. Despite the performance drop, improvement is still seen in the local model predictions, particularly for the $\dot{Y}_{H_2O}$ output. For example, in cluster 6, the $\dot{Y}_{H_2O}$ local ANN prediction curtails much of the sporadic variation from the global model. The same is true to a lesser extent in clusters 2, 4, and 5—cluster 3 predictions, on the other hand, are altogether unsatisfactory, as both local and global predictions show very little correlation with the ground truth. In the predictions for the stratified testing case for $\dot{T}$, noticeable local model improvement is only seen in clusters 4, 5, and 6. The local ANN predictions for $\dot{T}$ in cluster 2 are almost identical to the global ANN predictions, and the same issue with negative temperature source term as discussed for the LMDE testing snapshot is again apparent for the stratified mixture testing snapshot.

From the results in Fig. 8, it can essentially be concluded that (a) noticeable localized model improvement is indeed observed in most cases, with highest improvements seen for cluster 6, and (b) the extrapolation error is much more pronounced in the unseen stratified mixture configuration. A useful summary of these trends lies in the visualization of the source term fields themselves, as shown in Fig. 9. Fig. 9 (left) illustrates the increase in accuracy provided by the local ANN models in the LMDE testing snapshot, though the accuracy gain is more apparent for $\dot{Y}_{H_2O}$ than for $\dot{T}$. In particular, the region above the wavefront (represented by cluster 5 in Fig. 5) and the fluctuations of source term behind the wavefront present in the local ANN predictions better resemble the ground truth. On the other hand, the predictions generated for the stratified mixture testing snapshot are much less accurate for both global and local models. Though some regions are better captured by the local models, such as the absence of $\dot{Y}_{H_2O}$ fluctuation in the far-left domain and the general source term structure in the region immediately behind the wavefront for $\dot{T}$, it is apparent that extension of the neural network-based source term models into different configurations is less plausible.

Ultimately, the application of cluster-based localization of source term modeling is quite promising when considering configurations similar to the training set; such restrictions are expected in light of the highly nonlinear nature of chemical source terms in detonating flows. More importantly, this demonstration of localized ANN modeling highlights the potential for in-situ neural network-based source term modeling (where the configuration

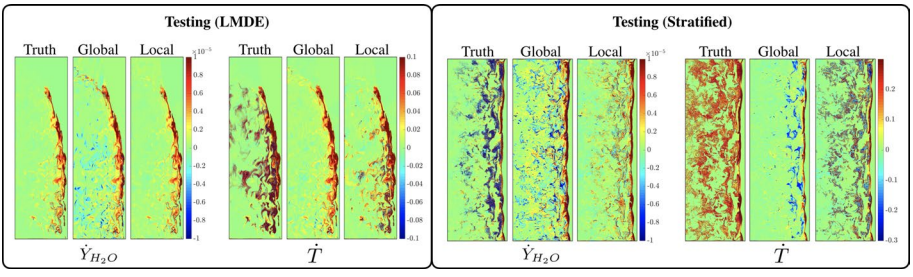

Fig. 9 Exact and predicted source term fields for the LMDE testing snapshot (left block) and stratified mixture snapshot (right block). Within each block, the left and right group of fields show $\dot{Y}_{H_2O}$ and $\dot{T}$, respectively. In these plots, source terms have been scaled by the simulation timestep (same value for all grid points) such that $\dot{Y}_{H_2O}$ is unitless and $\dot{T}$ is in Kelvin

does not change), a setting in which significant computational savings can be achieved and, as a result, long-time simulations of complex geometries exhibiting detonating reacting flow behavior (such as RDCs) be realized. It is possible that increasing the ANN parameter space, including time-history, including soft physical constraints during the parameter estimation phase, and/or tuning the input/output space for the networks may lead to better results with regards to extension into different configurations (e.g. the stratified mixture case)—all of these comments warrant a more detailed analysis into the localized modeling procedure and will be explored in future work.

5 Conclusions

Using direct numerical simulation datasets of canonical detonation configurations, a data-driven modeling procedure was developed with the goal of providing a pathway to better realize long-time simulations for complex combustor geometries such as RDCs. The modeling approach consisted of two linked phases, where the first concerned the extraction of different combustion regimes within the wave structure, and the second concerned the localized development of source term models guided by the extracted regions obtained from the first phase.

Specifically, in the first phase of the procedure (Sect. 3), a classification of the flowfield was obtained from a clustering on the progression of a detonation wave through a linear injector array. These clusters represented the ambient fuel-air mixture, a shock-separated region in the absence of fuel, a strong detonation region, and post-detonation deflagration regions within the reaction zone. This assignment of physical representations of each cluster allowed for the development of a useful coarse-grained perspective on detonation chemistry in physical space. Further, the extension of the clustering to the testing dataset (stratified fuel-air mixture) provided a manner in which detonation environments can be compared across different simulation configurations.

A comparison of the source term estimations obtained from the local ANNs (i.e. ANNs trained for each cluster in isolation) with the global ANN counterparts (i.e. a single ANN trained for the whole domain) showed general improvement provided by the domain-localized modeling in the training datasets. When predicting the source terms for unseen detonation waves in the same configuration at a future timestep (the LMDE testing snapshot), the improvements provided by the cluster-localized modeling became especially apparent and promising. When ambitiously extending the trained networks to the unseen data at a different configuration, it was found that although in many regions the localized model alleviated some of the large source term error variation seen in the global model, the source term predictions overall were much poorer. Despite this, the successful demonstration of this cluster-based localization of source term estimation on the unseen LMDE data is promising with regards to the role of domain-localized neural networks in the enabling of long-time simulations for complex combustion chemistry for situations in which the operating configuration does not change.

This work illuminates the promising role of data-driven classification and regression techniques both for concisely extracting segmented fields that delineate different combustion regimes and for guiding localized combustion modeling procedures in complex detonating environments. However, there are many promising pathways for future work. For example, for the ANN implementation, including physical constraints and tuning the input/output space for the local models may lead to better results when extending to different configurations. Further,

a detailed investigation into computational cost savings is necessary. One of the major motivations for converting the source term computation into an ANN evaluation is the compatibility with GPU architectures. As such, the computational savings stemming from both algorithm and architecture change is currently being pursued.

Acknowledgements This research is supported by the NASA Aeronautics Research Mission Directorate (ARMD) Fellowship under grant No. 80NSSC18K1735 with Dr. Tomasz Drozda of NASA Langley Research Center as technical adviser. The authors would like to thank NASA High-End Computing Capability (HECC) for the generous allocation of computational resources on the NASA Pleiades and Electra supercomputers which were used to generate the datasets.

Compliance with Ethical Standards

Conflict of interest The authors declare that they have no conflict of interest.

Appendix

K-Means Algorithm

Recall that centroid C_i represents a region of phase space that contains some proportion of the total number of snapshots. This region of phase space, denoted $\mathbb{C}_i$, is a cluster. The centroid that represents each grid point is the one that is the closest to the grid point based on a distance measure. Here, the L_2-norm in the $\mathbb{R}^{N_f}$ space is used to compute this distance, which is given by

$$d_{i,k} = \sqrt{\sum_{j=1}^{N_f} (S_i^j - C_k^j)^2}, \tag{1}$$

where $d_{i,k}$ represents the L_2-norm between the i-th grid point and k-th centroid. The superscript j in Eq. 1 indexes the number of dimensions N_f, or features, in the centroid/grid point vectors. This snapshot-centroid assignment is represented in an association matrix,

$$T_{i,k} = \begin{cases} 1 \text{ if } S_i \in \mathbb{C}_k, \\ 0 \text{ otherwise.} \end{cases} \tag{2}$$

With the above descriptions, the full K-means algorithm as detailed in Ref. Arthur and Vassilvitskii (2007) is summarized as follows:

1. Determine the initial distribution of N_k centroids $\mathcal{C}$ with k-means++.
2. Assign all N grid points to the nearest centroid as per Eq. 1, accumulating the association matrix $T_{i,k}$ as per Eq. 2.
3. Update the k-th centroid by computing the center of mass of all the grid points in the k-th cluster,

$$C_k = \frac{\sum_{i=1}^{N} S_i T_{i,k}}{\sum_{i=1}^{N} T_{i,k}}, \quad k = 1, \ldots, N_k, \quad i = 1, \ldots, N. \tag{3}$$

4. Repeat (2) and (3) until convergence, where convergence is defined as the point at which an additional iteration would cause all centroids to change by a negligible distance.

Time-Evolution of Cluster Size

The discussion in Sect.3 assigned physical significance to each cluster in the context of the training data, and then extended the findings into the testing set. To assess the effect of time progression on the segmented flowfield, Fig. 10 shows the time evolution of the *cluster size* for both training (solid lines) and testing (dashed lines) datasets. The cluster size is defined as the proportion of total number of grid points per snapshot occupied by a particular cluster—as such, for each time instance, the cluster sizes sum to unity. To see which clusters dominate at a particular time instant, the absolute cluster sizes are shown in the left plot of Fig. 10. To better represent how cluster sizes evolve, the right plot shows the absolute sizes normalized by the initial sizes at $t = 0$. Since the clusters themselves have physical meaning, the evolution of cluster size is useful metric when combined with the physical space representation in Fig. 5 in that it both 1) provides a coarse-grained picture of the time evolution of the detonation structure and 2) allows for the comparison of this structure evolution across different datasets.

From Fig. 5, the spatial structure of the delineated clusters is maintained with time progression, with introduction of clusters 2 and 4 on the left-hand side far from the wave front as the detonation wave passes to the right. Regions identified by clusters 2 and 4 are coherent between snapshots as they convect behind the wave front. However, there is transition between clusters 3 and clusters 2/4 as the mixing behavior in the post-detonation region changes with time. The triple point propagation is identified by the movement of cluster 6 with time in a direction along the wave front surface. Thus, the clustering process identifies

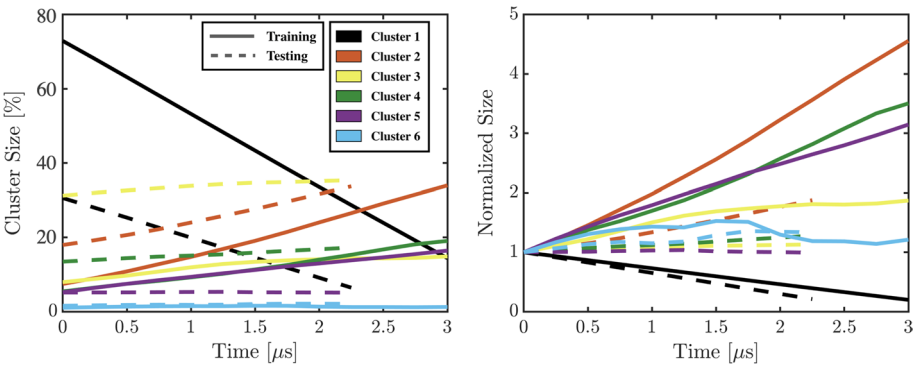

Fig. 10 Absolute cluster sizes as percentage of total grid points (left) and cluster sizes normalized by the sizes at $t = 0$ (right). Colors represent different cluster numbers. Solid lines correspond to training set and dashed lines to the testing set. Note that lines corresponding to testing set end at a lower maximum time (less testing snapshots)

and tracks the movement of the triple points with reasonable accuracy. Most importantly, the cluster classification is consistent between snapshots, allowing for each cluster to be interpreted with physical relevance to flow structures with time.

Figure 10 displays trends that are both shared different across both datasets. An expected result is the decrease in size of cluster 1 (the ambient region) in both datasets—as the wave progresses through the domain, the ambient domain proportion is reduced with time. Further, the identification of the triple point regions, cluster 6, is nearly consistent with time. This is an important result as this region occurs primarily near the shock front. As a fixed portion of the detonation front exists within each snapshot, the total number of points corresponding to this cluster remains stable. For both the training and testing dataset, cluster 6 is the lowest populated cluster as the triple points are concentrated regions due to wave interactions. In the training dataset, the size of the post-detonation cluster 3 along with deflagration clusters 2 and 4 increase in time. However, as cluster 3 was observed to be contained to a finite region behind the wave front, the cluster size increases slightly as the detonation wave fully enters the frame and reaches a stable size - this is evident in the normalized cluster size evolution (right plot of Fig. 10) of the training dataset. On the other hand, the sizes of cluster 2 and 4 increase a greater rate, with the increase in cluster 2 most dominant for both the training and testing dataset. This is expected, as the post-detonation regions represented by cluster 2 increase in size to match the reduction in cluster 1. Cluster 5 increases rapidly for the training dataset whereas this region is largely constant for the testing set; the shock-separated region increases as the wave progresses to the right in the LMDE, but in the stratified mixture case, the amount of entrained high-pressure gas is consistent prior to homogenization due to post-detonation mixing.

The rate of change in cluster size in the training dataset is more drastic in comparison to that of the testing set as the the training set features distinct regions, such as the shock-separated region, that are enforced by the fuel-air distribution due to injection. In the testing set, the stratification ahead of the wave is distributed more randomly, and does not ensure one cluster will dominate within different locations perpendicular to the shock front. Ultimately, the interpretation of time-evolution of cluster size allows for a unique and useful way to (a) assess a coarse-grained surrogate of the evolution of the detonation wave structure, and to (b) compare the evolution across different detonation configurations.

Feature Importance in Classification

The discussion below explores the finer details related to conditional distributions of grid points within each cluster. In particular, the information provided by the various features as it relates to the output delineations is analyzed, and a pathway for quantifying feature importance is provided.

Figure 11 shows typical pressure versus specific volume relations for the training set. The points corresponding to each cluster are identified by their respective colors, and the centroids are indicated by the larger white-enclosed markers. An expected trend is observed for cluster 1, which is an outlier in this space since this cluster represents ambient chemical conditions. Some separation is seen along the density axis for cluster 5, whose centroid occupies the highest density value. This is evidenced by the its spatial distribution in Fig. 5, which revealed that this cluster strongly represents a high-compression minimum-reaction region. Interestingly, the peak pressures near 50 atm are realized by points belonging to the cluster 6 (the region near the wavefront encapsulating triple point structures). However, the presence of overlapping densities in the pressure versus specific volume space (especially

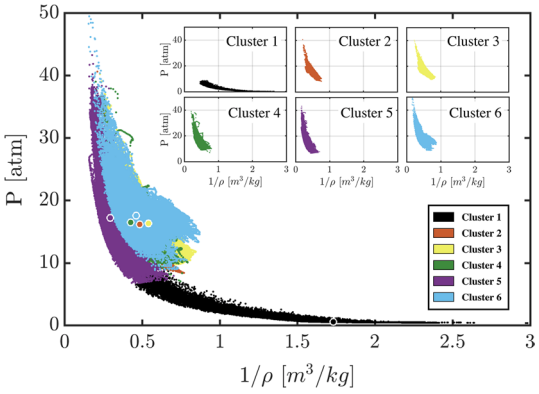

Fig. 11 Representation of pressure versus specific volume for all clusters. The larger points with white outline indicate centroids. Insets indicate cluster-specific scatter plots. Colors indicate cluster number

for clusters 2, 3, 4, and 6) is telling with regards to detonation wave structure classification: the primary role of the pressure and density features lies in 1) creating the delineation between reacting and ambient portions of the domain, and 2) identifying cluster 5 as a particularly high-density region. Ultimately, the remaining 7 features must contain the additional contribution to the variation in cluster-based probability densities of the grid points for the production of the segmented field structures. In other words, many more axes must be added to Fig. 11 to fully explain the regime delineations produced in Fig. 5.

The above discussion regarding Fig. 11 necessitates a more complete analysis of feature contribution to the overall delineation output. This relies on the interpretation and manipulation of probability distribution functions (PDFs) conditioned on both cluster number and feature. To illustrate this, Fig. 12 shows cluster-conditional PDFs for three of the nine total features: temperature (left), pressure (middle), and Y_{HO_2} (right). Each color corresponds to the grid point distribution in a single cluster. By analyzing such PDFs for each feature, one can assess how effective each feature is in determining the differences between grid points belonging to different clusters. If cluster PDFs for a feature are similar to one another (in both mean and variance), then that particular feature contributes less to the distance measure used in the clustering. Therefore, the degree of separation, or dissimilarity, between cluster distributions in Fig. 12 is related directly to the "importance" of those respective features in the segmented field representation. As expected, in both temperature and pressure distributions, cluster 1 is the outlier. Of the non-ambient clusters, there are much greater differences in distribution in for temperature than for pressure (though the pressure distributions still display differences in variation about the mean). Thus, for these

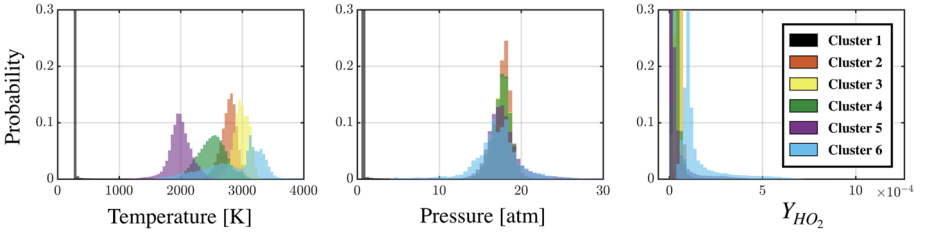

Fig. 12 Cluster PDFs conditioned on temperature (left), pressure (middle), and Y_{HO_2} (right) for the testing dataset (trends for training set were identical)

clusters, it can be surmised that temperature is a more contributing feature in determining the delineations shown in Fig. 5, in particular for regions within the deflagration regime. On the other hand, for Y_{HO_2}, the PDFs for all the clusters are much more stacked, implying lower overall delineation contribution.

The analysis of the distributions in Fig. 12 allows for a qualitative assessment of feature importance in the classifications shown in Fig. 5 based on cluster distributions. For a more quantitative analysis, a PDF-based measure known as the Earth Mover's Distance (EMD (Rubner et al. 2000)) can be used to create an importance metric for the features. The EMD is defined for two PDFs (which can be empirical) over a specified domain support, and is given by $\mathcal{D}(p_1, p_2)$, where p_1 and p_2 are PDFs, and $\mathcal{D}$ is the EMD function. The following distance properties hold: $\mathcal{D}(p_1, p_2) \geq 0$, $\mathcal{D}(p_1, p_1) = \mathcal{D}(p_2, p_2) = 0$, and $\mathcal{D}(p_1, p_2) = \mathcal{D}(p_2, p_1)$. A summary of the mathematical formulation and definition is provided in Appendix A.4. Informally, if p_1 and p_2 are visualized as piles of dirt, $\mathcal{D}(p_1, p_2)$ represents the "cost" of morphing one pile of dirt into the other. The factors which contribute to this "cost" are 1) the distance required to move the dirt, and 2) the amount of dirt present in the movement. The EMD is an appropriate measure here over other PDF-based distances such as Kullback-Leibler (KL) divergence, as it is well-defined for PDFs with nonzero densities in the feature space. The EMD usefully takes into account differences in distribution means as well as variance when determining the final dissimilarity score (the EMD of two identical distributions centered around different values will result in a positive distance, as will the EMD of two distributions with the identical centers but different standard deviations). Thus, using the EMD, a quantifiable metric for feature importance, denoted $\mathcal{I}(F_i)$, where F_i is the i-th feature in $\mathcal{F}$, is defined as

$$\mathcal{I}(F_i) = \sum_{j=1}^{N_k} \left(\sum_{k=1}^{N_k} \mathcal{D}(p_j^{F_i}, p_k^{F_i}) \right), \quad i = 1, \ldots, N_f. \tag{4}$$

In Eq. 4, $p_j^{F_i}$ represents the PDF of the j-th cluster corresponding to feature F_i. The importance metric is essentially a sum of the EMD combinations of cluster distributions for a particular feature. The inner summation in Eq. 4 represents the contribution of j-th cluster to the overall importance. By this metric, if feature 1 has lower importance than feature 2, feature 1 is less significant overall in the clustering output, i.e. it is less crucial in illuminating differences between grid points belonging to different clusters. This metric can be used to effectively downsample the feature selection used to generate the labels in Fig. 5.

Figure 13 shows importance metrics for each feature in the training and testing datasets. Note that scaled quantities were used in the computation to allow for comparison of importance across different features. The colors in each bar correspond to the cluster contribution to the importance of that particular feature, which itself can be useful in the identification of feature-regime relationships. For example, in the case of Y_{H_2O}, cluster 3 (active predominantly in the deflagration regime) contributes a large amount to the importance. On the other hand, for Y_H and Y_O, cluster 6 (active near the wavefront and triple point regions) dominates the importance value. Thus, an ability to quantitatively assign a most "relevant" cluster to a particular feature in the context of delineation power can be obtained.

Interestingly, Fig. 13 shows that the distribution of importance is similar between training and testing sets. In both cases, the metric implies that Y_{HO_2} and $Y_{H_2O_2}$ are noticeably the least significant in the clustering. To illustrate the utility of the EMD-based metric, the clustering is performed again using a reduced feature set *without* Y_{HO_2} and $Y_{H_2O_2}$. These results are shown in Fig. 14a, c for a single training and testing snapshot, respectively. Clustering using the downsampled set of features results in a segmented field that is nearly

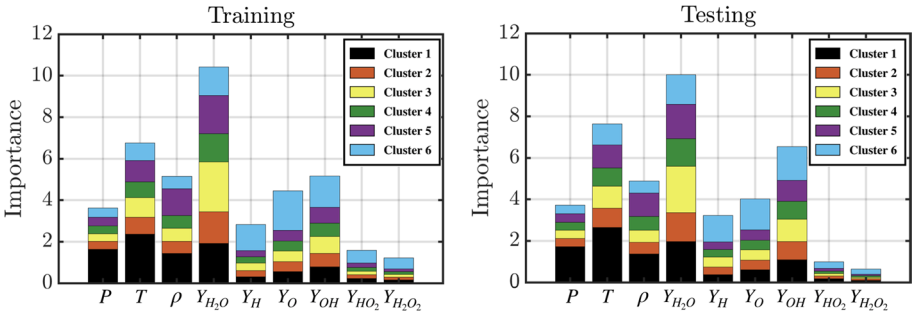

Fig. 13 Feature importance measures with cluster contributions for the training (left) and testing (right) datasets. Cluster contributions for importance value for each feature by respective colors

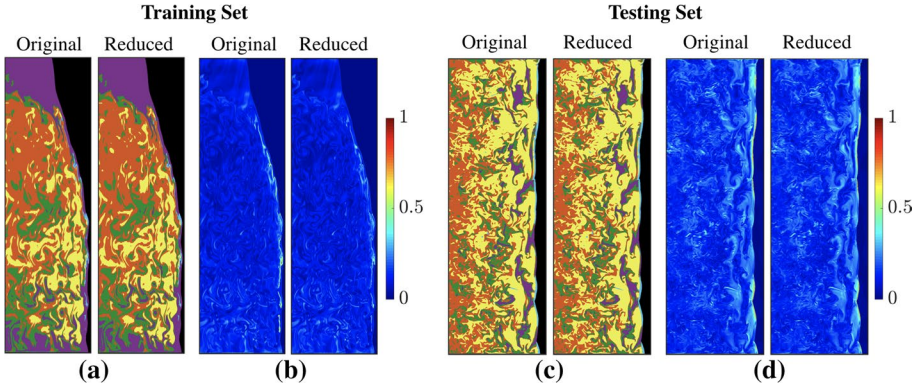

Fig. 14 Comparison of the clustering output (only one snapshot shown) between original full feature set and the reduced set. **a** Segmented fields for training dataset. **b** Distance fields for training dataset. **c** Segmented fields for testing dataset. **d** Distance fields for testing dataset

identical to that generated from the original, larger feature set. The same characteristics of the detonation wave and the regions of interest are captured by the reduced feature set. Thus, the importance metric provides useful insight into the thermodynamic properties and species necessary to demarcate the different modes of combustion.

The comparison with the original and reduced feature sets is also shown in Fig. 14b, d. These are distance plots, which is a proxy measure of uncertainty in the labels used to generate the segmented fields. They depict the distances of each grid point to its assigned centroid; high values in the distance field correlate directly with high classification uncertainty. In both training and testing sets, points of highest distance occur near the triple points. This means that despite the fact that the presence of the triple points is captured within clusters 5 and 6 to an acceptable level, the uncertainty in classification near the triple points is relatively high with the chosen parameters. This is an indicator to the chemical complexity in this portion of the detonation wave. However, alongside this, an important note is that this distance field is practically unchanged when clustering with the reduced feature set (i.e. the removal of Y_{HO_2} and $Y_{H_2O_2}$ did not cause noticeable increase in label uncertainty), meaning that the clustering output as a whole has been preserved in the process of reducing the feature set. For this reason, it is important that the initial classification is performed with a full (or very large) feature

set available from the data for general applications. Using the original clustering followed by the application of the importance metric, a lower feature set can then be obtained to identify both redundant and important regime-dependent features and additionally, reduce the computational cost for labeling in potential online applications.

Earth Mover's Distance

The formulation of the EMD is presented here in the context of Sect. 3. For a given feature f, consider two discrete densities $p(A)$ and $p(B)$—these can be thought of as distributions of data corresponding to the feature f in two clusters. These discretized densities (not necessarily normalized) can be represented as histograms in the set notation

$$\begin{aligned} \mathcal{H}_A &= \{(\alpha_1, p_1(A)),\ ..., (\alpha_M, p_M(A))\}, \\ \mathcal{H}_B &= \{(\beta_1, p_1(B)),\ ..., (\beta_N, p_N(B))\}, \end{aligned} \tag{5}$$

where α_i (resp. β_j) represents the center of bin i (resp. j) and $p_i(A)$ (resp. $p_j(B)$) represents the number of data points in bin i (resp. j). Note that the number of bins or the number of discrete points in each histogram does not have to be equal; here, $\mathcal{H}_A$ and $\mathcal{H}_B$ have a total of M and N bins respectively.

As stated in Sect. A.3, simply put, the EMD metric is a strictly positive measure that quantifies the degree of separation between two PDFs. This measure is obtained by minimizing a weighted sum of distances known as the work W, where

$$W(p(A), p(B), Q) = \sum_{i=1}^{M} \sum_{j=1}^{N} q_{i,j} d_{i,j}. \tag{6}$$

In Eq. 5, $Q = \{q_{i,j}\}$ is the set of weights (sometimes referred to as the optimal flow), and $d_{i,j}$ represents a distance measure between two points in the discrete PDFs (often taken as the Euclidean distance, not to be confused with the similar term in Eq. 1). The weights are subject to the following constraints:

$$q_{i,j} \geq 0, \tag{7}$$

$$\sum_{j=1}^{N} q_{i,j} \leq p_i(A), \tag{8}$$

$$\sum_{i=1}^{M} q_{i,j} \leq p_j(B), \tag{9}$$

$$\sum_{i=1}^{M} \sum_{j=1}^{N} q_{i,j} = \min \{ \sum_{i=1}^{M} p_i(A), \sum_{j=1}^{N} p_j(B) \}. \tag{10}$$

In other words, the upper bound of each weight $q_{i,j}$ is provided by the number of data points in bins i and j of $\mathcal{H}_A$ and $\mathcal{H}_B$ respectively. Finally, the EMD is defined as

$$\mathcal{D}(p(A), p(B)) = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N} q_{i,j}^{*} d_{i,j}}{\sum_{i=1}^{M} \sum_{j=1}^{N} q_{i,j}^{*}}, \tag{11}$$

where $Q^{*} = \{q_{i,j}^{*}\}$ is the optimal set of weights obtained by minimizing Eq. 5 (Rubner et al. 2000).

References

Akintayo, A., Lore, K.G., Sarkar, S., Sarkar, S.: Prognostics of combustion instabilities from hi-speed flame video using a deep convolutional selective autoencoder. Int. J. Prognost. Health Manage. **7**(023), 1 (2016)

Arthur, D., Vassilvitskii, S.: k-means++: The advantages of careful seeding. In: Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms, (2007)

Barwey, S., Hassanaly, M., An, Q., Raman, V., Steinberg, A.: Experimental data-based reduced-order model for analysis and prediction of flame transition in gas turbine combustors. Combust. Theory Modell. **23**(6), 994 (2019)

Barwey, S., Hassanaly, M., Raman, V., Steinberg, A.: Using machine learning to construct velocity fields from OH-PLIF images. Combust. Sci. Technol. **23**(6), 1 (2019)

Barwey, S., Ganesh, H., Hassanaly, M., Raman, V., Ceccio, S.: Data-based analysis of multimodal partial cavity shedding dynamics. Exp. Fluids **61**(4), 1 (2020)

Bell, J.B., Brown, N.J., Day, M.S., Frenklach, M., Grcar, J.F., Propp, R.M., Tonse, S.R.: Scaling and efficiency of PRISM in adaptive simulations of turbulent premixed flames. Proc. Combust. Inst. **28**(1), 107 (2000)

Bykovskii, F.A., Zhdan, S.A., Vedernikov, E.F.: Continuous spin detonations. J. Propuls. Power **22**(6), 1204 (2006)

Cailler, M., Darabiha, N., Veynante, D., Fiorina, B.: Building-up virtual optimized mechanism for flame modeling. Proc. Combust. Inst. **36**(1), 1251 (2017)

Chacon, F., Gamba, M.: Study of parasitic combustion in an optically accessible continuous wave rotating detonation engine. AIAA Paper **2019–0473**, (2019)

Chen, J.Y., Blasco, J.A., Fueyo, N., Dopazo, C.: An economical strategy for storage of chemical kinetics: fitting in situ adaptive tabulation with artificial neural networks. Proc. Combust. Inst. **28**(1), 115 (2000)

Christo, F., Masri, A., Nebot, E.: Artificial neural network implementation of chemistry with PDF simulation of H2/CO2 flames. Combust. Flame **106**(4), 406 (1996)

Fiévet, R.: Effect of vibrational nonequilibrium on isolator shock structure. J. Propuls. Power **34**(5), 1334–1344 (2018)

Fiévet, R., Koo, H., Raman, V.: Numerical simulation of a scramjet isolator with thermodynamic nonequilibrium. AIAA Paper **2015–3418**, (2015)

Fiévet, R., Voelkel, S.J., Raman, V., Varghese, P.L.: Numerical investigation of vibrational relaxation coupling with turbulent mixing. AIAA Paper **2017–0663**, (2017)

Fiorina, B.: Accounting for complex chemistry in the simulations of future turbulent combustion systems. AIAA Paper **2019–0995**, (2019)

Fiorina, B., Gicquel, O., Vervisch, L., Carpentier, S., Darabiha, N.: Approximating the chemical structure of partically premixed and diffusion counterflow flames using FPI flamelet tabulation. Combust. Flame **140**, 147 (2005)

Franke, L.L., Chatzopoulos, A.K., Rigopoulos, S.: Tabulation of combustion chemistry via artificial neural networks (ANNs): methodology and application to LES-PDF simulation of Sydney flame L. Combust. Flame **185**, 245 (2017)

Frolov, S.M., Dubrovskii, A.V., Ivanov, V.S.: Three-dimensional numerical simulation of the operation of a rotating-detonation chamber with separate supply of fuel and oxidizer. Russian J. Phys. Chem. B **7**(1), 35 (2013)

Giorgetti, S., Coppitters, D., Contino, F., Paepe, W.D., Bricteux, L., Aversano, G., Parente, A.: Surrogate-assisted modeling and robust optimization of a micro gas turbine plant with carbon capture. J. Eng. Gas Turbines Power **142**(1), 1 (2019)

Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press, Cambridge (2016)

Hecht-Nielsen, R.: Theory of the backpropagation neural network. In: Wechsler, H. (ed.) Neural Networks for Perception, pp. 65–93. Elsevier, Amsterdam (1992)

Herrmann, M., Blanquart, G., Raman, V.: A bounded QUICK scheme for preserving scalar bounds in large-eddy simulations. AIAA J. **44**(12), 2879 (2006)

Jha, P., Groth, C.: Evaluation of flame-prolongation of ildm and flamelet tabulated chemistry approaches for laminar flames. Combust. Theory Modell. **16**, 31 (2012)

Jiang, G., Peng, D.: Weighted ENO schemes for hamilton–jacobi equations. SIAM J. Sci. Comput. **21**(6), 2126 (2000)
Kaiser, E., Noack, B.R., Cordier, L., Spohn, A., Segond, M., Abel, M., Daviller, G., Östh, J., Krajnović, S., Niven, R.K.: Cluster-based reduced-order modelling of a mixing layer. J. Fluid Mech. **754**, 365 (2014)
Kapoor, R., Lentati, A., Menon, S.: Simulations of methane-air flames using ISAT and ANN. AIAA Paper **2001–3847**, (2001)
Kempf, A., Flemming, F., Janicka, J.: Investigation of lengthscales, scalar dissipation, and flame orientation in a piloted diffusion flame by LES. Proc. Combust. Inst. **30**(1), 557 (2005)
Kingma, D. P., Ba, J.: Adam: A method for stochastic optimization. (2017)
Landge, A. G., Pascucci, V., Gyulassy, A., Bennett, J. C., Kolla, H., Chen, J., Bremer, P.-T.: In-situ feature extraction of large scale combustion simulations using segmented merge trees. Supercomputing Conference Paper SC.2014.88, (2014)
Lu, F.K., Braun, E.M.: Rotating detonation wave propulsion: experimental challenges, modeling, and engine concepts. J. Propuls. Power **30**(5), 1125 (2014)
Lu, T., Law, C.K.: A directed relation graph method for mechanism reduction. Proc. Combust. Inst. **30**(1), 1333 (2005)
Malik, M.R., Isaac, B.J., Coussement, A., Smith, P.J., Parente, A.: Principal component analysis coupled with nonlinear regression for chemistry reduction. Combust. Flame **187**, 30 (2018)
Mueller, M.A., Kim, T.J., Yetter, R.A., Dryer, F.L.: Flow reactor studies and kinetic modeling of the H2/O2 reaction. Int. J. Chem. Kinet. **31**(2), 113 (1999)
Murphy, K.P.: Machine Learning: A Probabilistic Perspective. MIT Press, Cambridge (2012)
Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in PyTorch. 31st Conference on Neural Information Processing Systems (NIPS), (2017)
Peters, N.: Turbulent Combustion. Cambridge University Press, Cambridge (2000)
Pitsch, H.: Large-eddy simulation of turbulent combustion. Ann. Rev. Fluid Mech. **38**, 453 (2006)
Pope, S.B.: Computationally efficient implementation of combustion chemistry using in-situ adaptive tabulation. Combust. Theory Modell. **1**, 41 (1997)
Prakash, S., Raman, V.: Detonation propagation through inhomogeneous fuel-air mixtures. In: Proceedings of the 27th International Colloquium on the Dynamics of Explosions and Reactive Systems (ICDERS), (2019)
Prakash, S., Fiévet, R., Raman, V.: The effect of fuel stratification on the detonation wave structure. AIAA Paper **2019–1511**, (2019)
Prakash, S., Fiévet, R., Raman, V., Burr, J., Yu, K. H.: Analysis of the detonation wave structure in a linearized rotating detonation engine. AIAA J. pp. 1–15 (2019)
Raman, V., Hassanaly, M.: Emerging trends in numerical simulations of combustion systems. Proc. Combust. Inst. **37**(2), 2073 (2019)
Ranade, R., Alqahtani, S., Farooq, A., Echekki, T.: An ANN based hybrid chemistry framework for complex fuels. Fuel **241**, 625 (2019)
Rubner, Y., Tomasi, C., Guibas, L.J.: The earth mover's distance as a metric for image retrieval. Int. J. Comput. Vis. **40**(2), 99 (2000)
Sato, T., Raman, V.: Hydrocarbon Fuel Effects on Non-premixed Rotating Detonation Engine Performance (American Institute of Aeronautics and Astronautics, 2019). AIAA SciTech Forum. https://doi.org/10.2514/6.2019-2023(2019)
Sato, T., Voelkel, S., Raman, V.: Analysis of detonation structures with hydrocarbon fuels for application towards rotating detonation engines. AIAA Paper **2018–4965**, (2018)
Sato, T., Fabian, C., Duvall, J., Gamba, M., Raman, V.: Dynamics of rotating detonation engines with a pintle-type injector. 24th International Society for Air Breathing Engines (ISABE) Conference Paper, (2019)
Sen, B.A., Menon, S.: Linear eddy mixing based tabulation and artificial neural networks for large eddy simulations of turbulent flames. Combust. Flame **157**(1), 62 (2010)
Steinley, D.: K-means clustering: a half-century synthesis. Br. J. Math. Stat. Psychol. **59**(1), 1 (2006)
Tammisola, O., Juniper, M.P.: Coherent structures in a swirl injector at Re = 4800 by nonlinear simulations and linear global modes. J. Fluid Mech. **792**, 620 (2016)
Van Oijen, J.A., de Goey, L.P.H.: A numerical study of confined triple flames using a flamelet-generated manifold. Combust. Theory Modell. **8**(1), 141 (2004)
Warnatz, J., Maas, U., Dibble, R.W.: Combustion. Springer, New York (1996)
Zhou, R., Wu, D., Wang, J.: Progress of continuously rotating detonation engines. Chin. J. Aeronaut. **29**(1), 15 (2016)