Learning physics-based reduced-order models for a single-injector combustion process

by

Renee Swischuk, Boris Kramer, Cheng Huang, Karen E. Willcox
Learning physics-based reduced-order models for a single-injector combustion process

Renee Swischuk∗1, Boris Kramer†1, Cheng Huang‡2, and Karen Willcox§3

1Massachusetts Institute of Technology, Cambridge, MA, 02139
2University of Michigan, Ann Arbor, MI, 48109
3University of Texas at Austin, Austin, TX, 78712

This paper presents a physics-based data-driven method to learn predictive reduced-order models (ROMs) from high-fidelity simulations, and illustrates it in the challenging context of a single-injector combustion process. The method combines the perspectives of model reduction and machine learning. Model reduction brings in the physics of the problem, constraining the ROM predictions to lie on a subspace defined by the governing equations. This is achieved by defining the ROM in proper orthogonal decomposition (POD) coordinates, which embed the rich physics information contained in solution snapshots of a high-fidelity computational fluid dynamics (CFD) model. The machine learning perspective brings the flexibility to use transformed physical variables to define the POD basis. This is in contrast to traditional model reduction approaches that are constrained to use the physical variables of the high-fidelity code. Combining the two perspectives, the approach identifies a set of transformed physical variables that expose quadratic structure in the combustion governing equations and learns a quadratic ROM from transformed snapshot data. This learning does not require access to the high-fidelity model implementation. Numerical experiments show that the ROM accurately predicts temperature, pressure, velocity, species concentrations, and the limit-cycle amplitude, with speedups of more than five orders of magnitude over high-fidelity models. Moreover, ROM-predicted pressure traces accurately match the phase of the pressure signal and yield good approximations of the limit-cycle amplitude.

Nomenclature

A = System matrix for linear part
B = Input matrix
H = Matricized quadratic tensor
q(t) = State vector in finite dimensions
u(t) = External input vector
Q = Snapshot matrix
V = Matrix of POD basis vectors
x, y = Spatial coordinates
d = Number of physical variables
nx = Spatial discretization dimension
r = Reduced model dimension
t = Time
\tilde{\rho}(t, x, y), \tilde{\rho}_c(t, x, y), \tilde{\rho}_L(t, x, y) = State vector in primitive, conservative, and learning variables
p(t, x, y), p(t) = Pressure, continuous and discretized
T(t, x, y), T(t) = Temperature, continuous and discretized
\rho(t, x, y), \rho(t) = Density, continuous and discretized
\xi(t, x, y), \xi(t) = Specific volume, continuous and discretized
v_x(t, x, y), v_x(t) = Velocity in x direction, continuous and discretized

∗Graduate Student, MIT Center for Computational Engineering, swischuk@mit.edu, Student Member AIAA.
†Postdoctoral Associate, Department of Aeronautics and Astronautics, bokramer@mit.edu.
‡Assistant Research Scientist, Department of Aerospace Engineering, huangche@umich.edu. Member AIAA
§Director, Oden Institute for Computational Engineering and Sciences, kwillcox@oden.utexas.edu, AIAA Fellow.
\[ Y_l(t, x, y) = \text{Species mass fraction, } l = 1, 2, \ldots, n_{\text{sp}} \]
\[ c_l(t, x, y) = \text{Species molar concentrations, } l = 1, 2, \ldots, n_{\text{sp}} \text{ (also denoted as } [S]) \]
\[ \otimes = \text{Kronecker product} \]
\[ \hat{\cdot} = \text{Notation for ROM quantities} \]

**Abbreviations**

- **CFD** = Computational Fluid Dynamics
- **GEMS** = General equation and mesh solver; a CFD code
- **PDE** = Partial differential equation
- **POD** = Proper orthogonal decomposition
- **ROM** = Reduced-order model

**I. Introduction**

This paper presents an approach to learning low-dimensional surrogate models for a complex, nonlinear, multi-physics, multi-scale problem in form of a multi-species combustion process. The need for repeated model evaluations in optimization, design, uncertainty quantification and control of aerospace systems has driven the development of reduced-order models (ROMs) for applications in aerodynamics [1–7], reacting flows [8–10] and combustion [11–13]. ROMs combine the rich information embedded in high-fidelity simulations with the efficiency of low-dimensional surrogate models; yet, effective and robust ROM methods for nonlinear, multi-scale applications such as combustion have remained an open challenge.

Most existing nonlinear model reduction methods are intrusive—that is, they derive the ROM by projecting the high-fidelity model operators onto a low-dimensional subspace. In doing so, the physics of the problem is embedded in the reduced-order representation. The proper orthogonal decomposition (POD) [14, 15] is the most common way to define the low-dimensional subspace, using the singular value decomposition to identify low-dimensional structure based on training data. For some problems, the projection approach is amenable to rigorous error analysis and structure-preservation guarantees [16–19], but these rigorous guarantees do not apply to nonlinear, multi-physics, multi-scale models, for which projection-based ROMs remain challenging to implement (due to the need for access to the high-dimensional operators). Furthermore, ROMs for these problems typically require relatively high dimensionality (and thus high cost) to avoid problems with robustness and stability [12, 20].

There is increasing attention to non-intrusive model reduction methods (sometimes called black-box or data-driven methods) that learn a model based on training data, without requiring explicit access to the high-fidelity model operators. The non-intrusive philosophy aligns directly with the field of machine learning, where representations such as neural networks have been shown to induce nonlinear model forms that can approximate many physical processes [21]. However, neural networks require a large amount of training data, limiting their utility when the data comes from expensive large-scale partial differential equation (PDE) simulations [22]. Moreover, the parametrization of the learned low-dimensional model is critical to the predictive accuracy and success of the learned model—in particular, it is critical to determining whether the ROM can issue reliable predictions in regimes outside of the training data.

For large-scale PDE models, an important class of non-intrusive learning approaches tackle this challenge of model parametrization by embedding the structure of the problem into the learning formulation. Some approaches use sparse learning techniques to identify PDE model terms that explain the data [23–25]. When the model can be expressed in the form of a dynamical system with polynomial terms, then the learning problem can be formulated as a parameter estimation problem, as in the operator inference approach of [26]. An important advantage of non-intrusive learning approaches is that the user has the flexibility to choose the variables that drive the learning. This opens the way for variable transformations that expose system structure and, in doing so, transform the ROM learning task into a structured form. In some cases, the governing PDEs naturally admit variable transformations that reveal polynomial form, such as the specific volume representation of the Euler equations [27]. More generally, one can introduce new auxiliary variables to the problem—known as lifting—to produce a system that is polynomial in its expanded set of state variables [28–30]. This allows for a much broader class of nonlinear systems to be learned using the operator inference framework.

In this work, we build on the operator inference framework to learn structured, polynomial ROMs from simulated snapshot data of a single-injector combustion model. While in our case the polynomial model parametrization is a model approximation, we show that the predictive capabilities of the learned ROM are excellent beyond the training data. Our proposed approach follows the steps below, which we develop in detail in the ensuing sections:

1. We obtain high-dimensional simulation snapshot data for a spatially two-dimensional combustion process from the General Equation and Mesh Solver (GEMS) CFD code [31] developed at Purdue University. The governing
equations and combustion problem setup are described in Section II.

2) We identify a set of state variables in which many of the terms in the governing equations have quadratic form. We transform the snapshot data to these new state variables, as described in Section III.C.

3) We use operator inference to learn a ROM that evolves the combustion dynamics in a low-dimensional subspace. Details of the model learning are given in Section III and Section IV.B.

We present numerical results comparing our learned ROMs with GEMS test data in Section IV and conclude the paper in Section V.

II. Combustion model

Section II.A defines the computational domain under consideration, Section II.B presents the governing equations for the combustion model, and Section II.C briefly summarizes the numerical implementation. The combustion model follows the implementation of the General Equation and Mesh Solver (GEMS) CFD code [31] and more details can be found in Huang et al. [11].

A. Computational domain

A single-injector combustor as in [32] is shown in Figure 1a with the computational domain outlined in red dashed lines. Our domain is a simplified two-dimensional version of the computational domain, shown in Figure 1b which also shows the four locations where we monitor the state variables.

![Combustor assembly](image)

**Fig. 1** Setup and geometry of single-injector combustor.

B. Governing equations

The dynamics of the combustor are governed by the conservation equations for mass, momentum, energy and species mass fractions. For this two-dimensional problem, the conservation equations are

\[
\frac{\partial \tilde{\mathbf{q}}}{\partial t} + \nabla \cdot (\tilde{\mathbf{K}} - \tilde{\mathbf{K}}_v) = \tilde{S}
\]

and they describe the evolution of the conservative variables

\[
\tilde{\mathbf{q}} = [\rho \quad \rho v_x \quad \rho v_y \quad \rho e \quad \rho Y_1 \quad \ldots \quad \rho Y_{n_{sp}}]^{T},
\]

where \(\rho\) is the density \((\text{kg} / \text{m}^3)\), \(v_x\) and \(v_y\) are the \(x\) and \(y\) velocity \((\text{m} / \text{s})\), \(e\) is the total energy \((\text{J} / \text{m}^3)\), and \(Y_l\) is the \(l\)th species mass fraction with \(l = 1, 2, \ldots, n_{sp}\) and \(n_{sp}\) is defined as the number of chemical species that are included in the model.

The total energy is defined as

\[
e = \sum_{l=1}^{n_{sp}} h_l Y_l + \frac{1}{2} (v_x^2 + v_y^2) - \frac{P}{\rho} = h^0 - \frac{P}{\rho},
\]

where \(h_l\) is the specific enthalpy of the \(l\)th species \((\text{J} / \text{kg})\), \(P\) is the pressure \((\text{Pa})\), and \(\rho\) is the density \((\text{kg} / \text{m}^3)\).
where pressure $p$ is given in (Pa), $h_l = h_l(T)$ is the enthalpy corresponding to the $l$th species and is a highly nonlinear function of temperature, $T$, and $h^h$ is the stagnation enthalpy. The inviscid flux $\vec{K}$ and viscous flux $\vec{K}_v$ in Eq. (1) are:

$$\vec{K} = \begin{bmatrix} \rho v_x \\ \rho v_x^2 + p \\ \rho v_x v_y \\ \rho v_x Y_1 \\ \vdots \\ \rho v_x Y_{nsp} \end{bmatrix} \tilde{t} + \begin{bmatrix} \rho v_y \\ \rho v_x v_y \\ \rho v_y v_x + p \\ \rho v_y Y_1 \\ \vdots \\ \rho v_y Y_{nsp} \end{bmatrix} \tilde{j}, \quad \vec{K}_v = \begin{bmatrix} 0 \\ \tau_{xx} \\ \tau_{xy} \\ \tau_{yy} \\ 0 \end{bmatrix} \tilde{t} + \begin{bmatrix} \tau_{xx}v_x + \tau_{xy}v_y - j_x^q \\ -j_{1,x}^m \\ \vdots \\ -j_{nsp,x}^m \end{bmatrix} \tilde{t} + \begin{bmatrix} \tau_{xy}v_x + \tau_{yy}v_y - j_y^q \\ -j_{1,y}^m \\ \vdots \\ -j_{nsp,y}^m \end{bmatrix} \tilde{j},$$

The two-dimensional viscous shear tensor is defined as

$$\tau = \begin{bmatrix} \tau_{xx} & \tau_{xy} \\ \tau_{xy} & \tau_{yy} \end{bmatrix} = \tilde{\mu} \begin{bmatrix} \frac{1}{3} \frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} & \frac{\partial v_y}{\partial x} + \frac{1}{3} \frac{\partial v_x}{\partial y} \\ \frac{\partial v_y}{\partial x} & \frac{\partial v_y}{\partial y} \end{bmatrix},$$

where $\tilde{\mu}$ is the mixture viscosity coefficient. The diffusive heat flux vector is defined as

$$\tilde{j}^q = \begin{bmatrix} j_x^q & j_y^q \end{bmatrix}^T = -\kappa \nabla T + \rho \sum_{l=1}^{nsp} D_l h_l \nabla Y_l,$$

where $\kappa$ quantifies thermal conductivity and $D_l$ is the diffusion coefficient for the $l$th species into the mixture, which is an approximation used to model the multi-component diffusion as the binary diffusion of each species into a mixture. The two terms in the definition of the heat flux (Eq. (4)) represent heat transfer due to conductivity and species diffusion. The diffusive mass flux vector of species $l$ is modeled as

$$\tilde{j}_l^m = \begin{bmatrix} j_{l,x}^m & j_{l,y}^m \end{bmatrix}^T = \begin{bmatrix} \rho D_l \frac{\partial Y_l}{\partial x} & \rho D_l \frac{\partial Y_l}{\partial y} \end{bmatrix}^T.$$

The source term, $\vec{S}$ in Eq. (1) is

$$\vec{S} = \begin{bmatrix} 0 & 0 & 0 & 0 & \omega_1 & \ldots & \omega_{nsp} \end{bmatrix}^T \begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix},$$

and is defined by considering a 1-step combustion reaction governed by

$$\text{CH}_4 + 2\text{O}_2 \rightarrow \text{CO}_2 + 2\text{H}_2\text{O},$$

as presented in [33], with $n_{sp} = 4$. The corresponding general stoichiometric equation is defined as $0 = \sum_{l=1}^{n_{sp}} \nu_l \chi_l$, where $\chi_1 = \text{CH}_4$, $\chi_2 = \text{O}_2$, $\chi_3 = \text{CO}_2$, $\chi_4 = \text{H}_2\text{O}$ and $\nu_l$ is the net stoichiometric coefficients of each species with $\nu_1 = -1$, $\nu_2 = -2$, $\nu_3 = 1$ and $\nu_4 = 2$. The molar concentration of the $l$th species is denoted by $c_l$. In our case, $l \in \{1, 2, 3, 4\}$, so $c_1 = [\text{CH}_4]$, $c_2 = [\text{O}_2]$, $c_3 = [\text{CO}_2]$, and $c_4 = [\text{H}_2\text{O}]$ are the molar concentrations. Here, we use the standard bracket notation [ ] to indicate molar concentration of a species. The general relationship between a species molar concentration, $c_l$, and a species mass fraction, $Y_l$, is

$$Y_l = \frac{c_l M_l}{\rho},$$

where $Y_l$ is the mass fraction of CH$_4$, $Y_2$ is the mass fraction of O$_2$, $Y_3$ is the mass fraction of CO$_2$ and $Y_4$ is the mass fraction of H$_2$O. The production rate of the $l$th species in the source term $\vec{S}$ in Eq. (6) is modeled as

$$\dot{\omega}_l = M_l \frac{\partial c_l^{\text{reaction}}}{\partial t} = \nu_l \Gamma_r,$$

where $c_l^{\text{reaction}}$ are chemical reaction source terms whose dynamics are described below and $\Gamma_r$ is the reaction rate. The molar mass of CH$_4$ is $M_1 = 16.04$ mol, the molar mass of O$_2$ is $M_2 = 32.0$ mol, the molar mass of CO$_2$ is $M_3 = 44.01$ mol, and the molar mass of H$_2$O is $M_4 = 18.0$ mol. The reaction rate is approximated by

$$\Gamma_r = k \prod_{l=1}^{n_{reaction}} c_l^{\omega_l},$$
where \( n_{\text{reactant}} = 2 \) is the number of reactants, \( k \) is the rate coefficient and \( o_l \) is the reaction order of the \( l \)th reactant. In our case \( o_1 = 0.2 \) and \( o_2 = 1.3 \). The rate coefficient, \( k \), is described by the Arrhenius equation as

\[
k = A \exp \left( -\frac{E_a}{R_u T} \right),
\]

(9)

where \( R_u = 8.314 \frac{J}{mol \cdot K} \) is the universal gas constant, \( A = 2 \times 10^{10} \) is the pre-exponential constant and \( E_a = 2.025 \times 10^5 \) is the energy required to reach a chemical reaction, measured in Joules and referred to as the activation energy. In this work, we use the ideal gas state equation that relates density and pressure to temperature

\[
\rho = \frac{p}{RT},
\]

(10)

where \( R = \frac{R_u}{M} \) and \( M = \left( \sum_{j=1}^{n_{sp}} \left( \frac{Y_j}{\rho} \right) \right)^{-1} \) is the average molar mass of the mixtures.

At the downstream end of the combustor, we impose a non-reflecting boundary condition while maintaining the chamber pressure via

\[
p_{\text{back}}(t) = p_{\text{back,ref}}[1 + A \sin(2\pi ft)],
\]

(11)

where \( p_{\text{back,ref}} = 1.0 \times 10^6 \) Pa, \( A = 0.1 \) and \( f = 5000 \)Hz. The top and bottom wall boundary conditions are no-slip conditions, and for the upstream boundary we impose constant mass flow at the inlets.

C. Numerical model

GEMS uses the finite volume method to discretize the governing equations, which are solved in primitive variables. For a spatial discretization with \( n_x \) cells, this results in a \( dn_x \)-dimensional system of nonlinear ordinary differential equations (ODEs)

\[
\frac{dq}{dt} = G(q, u(t)), \quad q(0) = q_0.
\]

(12)

for \( 0 < t \leq T \), where \( d \) is the number of unknowns in the PDE governing equations and here \( d = 8 \) (four flow variables and four species concentrations). In Eq. (12), \( q(t) \in \mathbb{R}^{dn_x} \) is the discretized state vector at time \( t \) (for GEMS, it is the discretization of the primitive variables \( q_p = [p \ v_x \ v_y \ T \ Y_1 \ \ldots \ Y_{n_{sp}}]^\top \), \( q_0 \) are the specified initial conditions, and \( \frac{dq}{dt} \) is the time derivative of the state vector at time \( t \). The \( m \) inputs \( u(t) \in \mathbb{R}^m \) arise from the time-dependent boundary condition, defined in Eq. (11), applied at the combustor downstream end. The nonlinear function \( G : \mathbb{R}^{dn_x} \times \mathbb{R}^m \rightarrow \mathbb{R}^{dn_x} \) maps the discretized states \( q \) and the input \( u \) to the state time derivatives, representing the spatial discretization of the governing equations described in Section II.B.

Solving these high-dimensional nonlinear ODEs is expensive, motivating the derivation of a ROM that can yield approximate solutions at reduced cost. The nonlinear multi-scale dynamics represented by these equations makes this a challenging task. To maintain computational efficiency in the ROM, state-of-the-art nonlinear model reduction methods combine POD with a sparse interpolation method (often called hyperreduction) by evaluating the nonlinear functions only at a select number of points. For instance, POD together with the discrete empirical interpolation method (DEIM) has had some success, but also encountered problems in combustion applications \[12\] . Of particular challenge is the need to include a large number of interpolation points in the POD-DEIM approximation, which means that the ROM loses its computational efficiency. Robustness and stability of the POD-DEIM models is also a challenge \[20\]. In the next section, we present a different approach that uses non-intrusive ROM learning to enable variable transformations that expose system structure. This structure is then exploited in the derivation of the ROM and removes the need for the DEIM approximation.

III. Non-intrusive learning of a combustion reduced model

This section presents our approach to learn ROMs for the unsteady combustion dynamics simulation from GEMS. Section III.A writes a general nonlinear system in a form that exposes the underlying structure of the governing equations and shows how projection preserves that structure. Section III.B presents the operator inference approach from \[26\], which learns structured ROM operators from simulation data. Section III.C describes variable transformations that lead to the desired polynomial structure for the combustion governing equations presented in Section II. These transformations yield the structure needed to apply the operator inference approach.
A. Projection preserves polynomial structure in the governing equations

Consider a large-scale system of nonlinear ODEs written in polynomial form
\[
\frac{dq}{dt} = Aq + H(q \otimes q) + C(q \otimes \dot{q} \otimes q) + Bu + c + \text{HOT}.
\] (13)

Relating this equation to the general nonlinear system in Eq. (12), we see that \(Aq\) are the terms in \(G(\cdot)\) that are linear in the state \(q\), with \(A \in \mathbb{R}^{dn \times dn}\); \(H(q \otimes q)\) are the terms in \(G(\cdot)\) that are quadratic in \(q\), with \(H \in \mathbb{R}^{dn \times (dn)^2}\); \(C(q \otimes q \otimes q)\) are the terms in \(G(\cdot)\) that are cubic in \(q\), with \(C \in \mathbb{R}^{dn \times (dn)^3}\); \(Bu\) are the terms in \(G(\cdot)\) that are linear in the input \(u\), with \(B \in \mathbb{R}^{dn \times m}\); and \(c \in \mathbb{R}^{dn}\) are constant terms in \(G(\cdot)\) that do not depend on state or input. The abbreviation “HOT” in Eq. (13) denotes higher-order terms, and represents terms that are quartic and higher order, as well as any other nonlinear terms that cannot be represented in polynomial form.

We emphasize that we are not (yet) introducing approximations—rather, we are explicitly writing out the discretized projections of the state variables via the process of lifting [28, 29] to convert these terms to polynomial form. For terms in the governing equations that are not in polynomial form (such as terms involving \(1/T\) and the Arrhenius reaction terms) we discuss in Section III.C the introduction of variable transformations and auxiliary variables via the process of lifting [28, 29] to convert these terms to polynomial form.

B. Operator inference for learning reduced models

Here we summarize the steps of the operator inference approach from [20]. First, we collect \(K\) snapshots of the state by solving the high-fidelity model. We store the snapshots and the inputs used to generate them in the matrices:
\[
Q = [q_0 \ldots q_K] \in \mathbb{R}^{dn \times K}, \quad U = [u_0, \ldots, u_K] \in \mathbb{R}^{m \times K},
\] (15)

where \(u_i \equiv u(t_i)\) and \(q_i \equiv q(t_i)\) with \(0 = t_0 < t_1 < \cdots < t_K = T\). In general, \(dn \gg K\), so the matrix \(Q\) is tall and skinny. Second, we identify the low-dimensional subspace in which we will learn the ROM. In this work, we use the POD to define the low-dimensional subspace, by computing the singular value decomposition of the snapshot matrix
\[
Q = V \Sigma W^T.
\]

where \(V \in \mathbb{R}^{dn \times K}, \Sigma \in \mathbb{R}^{K \times K}\) and \(W \in \mathbb{R}^{K \times K}\). The \(r \ll dn\) dimensional POD basis, \(V_r = [v_1, \ldots, v_r]\), is given by the first \(r\) columns of \(V\). Third, we project the state snapshot data onto the POD subspace spanned by the columns of \(V_r\) and obtain the reduced snapshot matrices
\[
\hat{Q} = V_r^T Q = [\hat{q}_0 \ldots \hat{q}_K] \in \mathbb{R}^{r \times K}, \quad \hat{Q} = [\hat{q}_0, \hat{q}_1, \ldots, \hat{q}_K] \in \mathbb{R}^{r \times K},
\] (16)

where the columns of \(\hat{Q}\) are computed from \(Q\) using any time derivative approximation (see, e.g., [34–36]), or can be obtained—if available—by collecting and projecting snapshots of \(G(q, u)\).

Operator inference solves a least squares problem to find the reduced operators that yield the ROM that best matches the projected snapshot data in a minimum residual sense. For the quadratic ROM
\[
\hat{q} = \hat{A}q + \hat{H}(q \otimes \hat{q}) + \hat{B}u + \hat{c},
\] (17)
where $\mathbf{1}_K \in \mathbb{R}^K$ is the length $K$ column vector with all entries set to unity. Note that this least squares problem is linear in the coefficients of the unknown ROM operators $\hat{\mathbf{A}}, \hat{\mathbf{H}}, \hat{\mathbf{B}}$ and $\hat{\mathbf{c}}$. Also note that the operator inference approach permits us to compute the ROM operators $\hat{\mathbf{A}}, \hat{\mathbf{H}}, \hat{\mathbf{B}}$ and $\hat{\mathbf{c}}$ without needing explicit access to the original high-dimensional operators $\mathbf{A}, \mathbf{H}, \mathbf{B}$ and $\mathbf{c}$.

We combine the unknown operators of Eq. (17) in the matrix

$$
\mathbf{O} = [\hat{\mathbf{A}} \quad \hat{\mathbf{H}} \quad \hat{\mathbf{B}} \quad \hat{\mathbf{c}}] \in \mathbb{R}^{r \times (r + r^2 + m + 1)},
$$

and the known low-dimensional data in the data matrix

$$
\mathbf{D} = \begin{bmatrix}
\hat{\mathbf{Q}}^\top & (\hat{\mathbf{Q}} \otimes \hat{\mathbf{Q}})^\top & \mathbf{U}^\top & \mathbf{1}_K
\end{bmatrix} \in \mathbb{R}^{K \times (r + r^2 + m + 1)},
$$

and then solve the minimization problem

$$
\min_{\mathbf{O} \in \mathbb{R}^{r \times (r + r^2 + m + 1)}} \left\| \mathbf{D} \mathbf{O}^\top - \hat{\mathbf{Q}} \right\|_2^2.
$$

It was proven in [26] that Eq. (20) can be written as $r$ independent least squares problems of the form $\min_{\mathbf{o}_i \in \mathbb{R}^{r + r^2 + m + 1}} \| \mathbf{D} \mathbf{o}_i - \mathbf{r}_i \|_2^2$, for $i = 1, \ldots, r$, where $\mathbf{o}_i$ is a column of $\mathbf{O}^\top$ (row of $\mathbf{O}$) and $\mathbf{r}_i$ is a column of $\hat{\mathbf{Q}}^\top$. This makes the operator inference approach efficient and scalable.

Regularization becomes necessary to avoid overfitting and to infer operators that produce a stable ROM. In this work, we use an $L_2$ regularization penalty on the off-diagonal elements of the operator $\hat{\mathbf{A}}$ and on all elements of the remaining operators. With this regularization, our least squares problem becomes

$$
\min_{\mathbf{o}_i \in \mathbb{R}^{r + r^2 + m + 1}} \| \mathbf{D} \mathbf{o}_i - \mathbf{r}_i \|_2^2 + \lambda \| \mathbf{P}_i \mathbf{o}_i \|_2^2 \quad \text{for } i = 1, \ldots, r,
$$

where $\lambda$ is the regularization parameter and $\mathbf{P}_i$ is the $r + r^2 + m + 1$ identity matrix with the $i$th diagonal set to zero so that we avoid regularizing the diagonal elements of $\mathbf{A}$. It should be noted that the regularization parameter, $\lambda$, is problem specific and should be chosen accordingly. In Section IV.B, we discuss details of the operator inference implementation, a method for selecting $\lambda$, and the removal of redundant terms in the least squares problem in Eq. (21).

C. A structure-exploiting ROM learning formulation for GEMS

A key contribution of this work is to recognize that the non-intrusive operator inference approach gives us complete flexibility in the set of physical variables we work with to define the ROM. We can identify choices of physical variables that expose the desired polynomial structure in the governing equations, and then extract snapshots for those variables by applying transformations to the snapshot data— we do not need to make any modifications to the high-fidelity CFD simulation model itself. In theory, a classical intrusive ROM approach could work with transformed variables (e.g., in the work of [37]); however, this would involve rewriting the high-fidelity simulator, a task that would be not only time-consuming but also fraught with mathematical pitfalls, especially for unusual choices of variables. This is where the data-driven perspective of machine learning becomes extremely valuable.

The Euler equations admit a quadratic representation in the specific volume variables; in that case, a transformation of the snapshots from conservative (or primitive) variables to specific volume variables can be exploited to create quadratic ROMs [27]. Other PDEs may not admit polynomial structure via such straightforward transformations, but the process of lifting the equations via the introduction of new auxiliary variables can produce a set of coordinates in which the governing equations become polynomial in the lifted state [28–30]. For example, the tubular reactor example of [29] includes Arrhenius-type reaction terms similar to those in Eq. (9). The introduction of auxiliary variables permits the governing equations to be written equivalently with quartic nonlinearity in the lifted variables [3].
Lifting to polynomial form for the GEMS equations described in Section II is made difficult by several of the terms, in particular through some of the gas thermal properties such as the nonlinear dependence of enthalpy on temperature. A complete lifting that converts all equations to a polynomial form is possible, but would require the introduction of a large number of auxiliary variables and would also result in the introduction of some algebraic equations. However, as the analysis below shows, the GEMS governing equations admit a transformation for which many terms in the governing equations take polynomial form when we use the variables

\[ \vec{q}_L = \left[ p \ v_x \ v_y \ \xi \ c_1 \ c_2 \ c_3 \ c_4 \right]^T. \]

Here \( \xi = \frac{1}{\rho} \) is the specific volume, and recall that \( c_1 = [\text{CH}_4], c_2 = [\text{O}_2], c_3 = [\text{CO}_2], \) and \( c_4 = [\text{H}_2\text{O}] \) are the molar concentrations with \( c_1 = \frac{\nu_{\text{H}_4}}{M_{\text{H}_4}}. \)

Below, we derive the governing PDEs for specific volume \( \xi \) and velocities \( v_x, v_y \). These three governing PDEs all turn out to be quadratic in the learning variables \( \vec{q}_L \). In Appendix A we present the lifting transformations for the source term dynamics \( \vec{c}_S^\text{action} \) in the vector \( \vec{S} \) in Eq. (6). In Appendix B we derive the equations governing the pressure \( p \) and the species molar concentrations \( c_i \). These equations have some terms that are not polynomial in the chosen learning variables \( \vec{q}_L \).

To keep notation clean in application of the chain rule, let the conservative variables be denoted as \( g_1 = \rho, g_2 = \rho v_x, g_3 = \rho v_y, g_4 = \rho e, g_5 = \rho Y_1, g_6 = \rho Y_2, g_7 = \rho Y_3, g_8 = \rho Y_4. \) Throughout, we frequently use the relationship

\[ \frac{\partial \xi}{\partial \rho} = -\frac{1}{\rho^2} \frac{\partial \rho}{\partial x} = -\xi^2 \frac{\partial \rho}{\partial x}, \]

and similarly for \( \frac{\partial \xi}{\partial \vec{x}}. \) Note also that we are assuming the existence of these partial derivatives, that is, we do not consider the case of problems with discontinuities.

**Specific Volume** \( \xi = 1/\rho \). We use the constitutive relationship for the density \( \rho \) in Eq. (1) in the derivation:

\[ \frac{\partial \xi}{\partial t} = \frac{1}{\rho} \frac{\partial g_2}{\partial t} = \frac{1}{\rho} \frac{\partial g_2}{\partial \xi} \frac{\partial \xi}{\partial t} \rho = \xi^2 \nabla \cdot (\rho v_x \vec{i} + \rho v_y \vec{j}) = \xi^2 \left[ \frac{\partial \rho}{\partial x} v_x + \rho \frac{\partial v_x}{\partial x} + \xi \frac{\partial \rho}{\partial y} v_y + \rho \frac{\partial v_y}{\partial y} \right]. \]

Inserting Eq. (23) into the above, we obtain

\[ \frac{\partial \xi}{\partial t} = -\xi \frac{\partial v_x}{\partial x} - v_y \frac{\partial \xi}{\partial y} + \xi \frac{\partial v_y}{\partial y}, \]

which is quadratic in the learning variables \( \xi, v_x, v_y \).

**Velocities** \( v_x, v_y \): We have

\[ \frac{\partial v_x}{\partial t} = \frac{\partial g_2}{\partial t} = \frac{1}{g_1} \frac{\partial \xi}{\partial t} \rho = \frac{1}{g_1} \frac{\partial \xi}{\partial \rho} \frac{\partial \rho}{\partial t} = \frac{1}{g_1} \frac{\partial \xi}{\partial t} \rho = \frac{1}{\rho} \frac{\partial \rho}{\partial t} v_x \]

and from Eq. (1) we have \( \frac{\partial \rho}{\partial t} = -\nabla \cdot (\rho v_x \vec{i} + \rho v_y \vec{j}), \) and \( \frac{\partial \xi}{\partial \rho} = \frac{1}{\rho} \frac{\partial \xi}{\partial \rho} \frac{\partial \rho}{\partial t} = \frac{1}{g_1} \frac{\partial \xi}{\partial t} \rho = \frac{1}{g_1} \frac{\partial \xi}{\partial \rho} \frac{\partial \rho}{\partial t} = \frac{1}{\rho} \frac{\partial \rho}{\partial t} v_x \)

Thus, we obtain

\[ \frac{\partial v_x}{\partial t} = \frac{1}{\rho} \nabla \cdot (\rho v_x \vec{i} + \rho v_y \vec{j}) = \frac{v_x}{\rho} \nabla \cdot (\rho v_x \vec{i} + \rho v_y \vec{j}) - \xi \frac{\partial v_x}{\partial x} - \xi \frac{\partial v_x v_y}{\partial y} + \xi \frac{\partial v_x v_y}{\partial y} + \frac{v_x}{\rho} \nabla \cdot (\rho v_x \vec{i} + \rho v_y \vec{j}) \]

\[ = \frac{1}{\rho} \nabla \cdot (\rho v_x \vec{i} + \rho v_y \vec{j}) - \xi \frac{\partial v_x}{\partial x} - \xi \frac{\partial v_x v_y}{\partial y} + \xi \frac{\partial v_x v_y}{\partial y} + \frac{v_x}{\rho} \nabla \cdot (\rho v_x \vec{i} + \rho v_y \vec{j}) \]

\[ = \frac{v_x}{\rho} \nabla \cdot (\rho v_x \vec{i} + \rho v_y \vec{j}) - \xi \frac{\partial v_x}{\partial x} - \xi \frac{\partial v_x v_y}{\partial y} + \xi \frac{\partial v_x v_y}{\partial y} + \frac{v_x}{\rho} \nabla \cdot (\rho v_x \vec{i} + \rho v_y \vec{j}) \]

\[ = -\xi \frac{\partial v_x}{\partial x} - \xi \frac{\partial v_x v_y}{\partial y} + \xi \frac{\partial v_x v_y}{\partial y} + \frac{v_x}{\rho} \nabla \cdot (\rho v_x \vec{i} + \rho v_y \vec{j}) \]

\[ = -\xi \frac{\partial v_x}{\partial x} - \xi \frac{\partial v_x v_y}{\partial y} + \xi \frac{\partial v_x v_y}{\partial y} + \frac{v_x}{\rho} \nabla \cdot (\rho v_x \vec{i} + \rho v_y \vec{j}) \]

and we get a similar expression for \( \frac{\partial v_y}{\partial t} \). Both dynamics are quadratic in the learning variables \( p, v_x, v_y, \xi \).

As noted above, Appendix A and Appendix B present the derivations for the chemical source terms, pressure, and chemical species.
IV. Numerical Results

We now apply the variable transformations and operator inference framework to learn a predictive ROM from GEMS high-fidelity combustion simulation data. Section IV.A describes the problem setup and GEMS dataset. Section IV.B discusses implementation details and Section IV.C presents our numerical results. Additional numerical results can be found in [38].

A. GEMS Dataset

The computational domain shown in Figure 1 is discretized with \( n_x = 38523 \) spatial discretization points. Each CFD state solution thus has dimension \( d_n = 308184 \). The problem considered here has fuel and oxidizer input streams with constant mass flow rates of \( 5.0 \times 10^{-3} \) kg/s and \( 0.37 \) kg/s, respectively. The fuel is composed of gaseous methane and the oxidizer is 42% gaseous \( \text{O}_2 \) and 58% gaseous \( \text{H}_2\text{O} \), as described in [11]. The forcing input Eq. (11) is applied at the right side of the domain.

To generate training data, GEMS is simulated for a time duration of 1ms with a time step size of \( \Delta t = 1 \times 10^{-7} \) s resulting in \( K = 10000 \) snapshots. The GEMS output is transformed to the variables given in Eq. (22). The recorded snapshot matrix is thus

\[
Q = [q_0 \quad q_1 \quad \ldots \quad q_K] \in \mathbb{R}^{d_n \times K} = \mathbb{R}^{308184 \times 10000}
\]

Our numerical experiments were parallelized on a cluster with two computing nodes. Each node has two 10-core Intel Xeon-E5 processors (20 cores per node) and 128 GB RAM. The training data generation took approximately 200h in CPU time for the 1ms, 10000 snapshots of high-fidelity CFD data.

The range of variable values for the training data is shown in Table 1. Note that the data covers a wide range of scales. Pressure is of the order \( 10^6 \) while species concentrations can be as low as \( 10^{-12} \). This large scaling difference presents a challenge when learning models from data. To deal with the numerical issues related to large differences in scaling and small species concentrations and velocities, we scale each variable to the interval \([-1, 1]\). Variables are scaled before computing the POD basis and projecting the data.

<table>
<thead>
<tr>
<th>State variable</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pressure ( p ) in Pa</td>
<td>( 9.226 \times 10^5 )</td>
<td>( 1.142 \times 10^6 )</td>
<td>( 1.433 \times 10^6 )</td>
</tr>
<tr>
<td>Velocity ( v_x ) in ( m/s )</td>
<td>-222.930</td>
<td>69.637</td>
<td>307.147</td>
</tr>
<tr>
<td>Velocity ( v_y ) in ( m/s )</td>
<td>-206.990</td>
<td>1.304</td>
<td>186.548</td>
</tr>
<tr>
<td>Specific volume ( \xi = \rho^{-1} ) in ( m^3/kg )</td>
<td>0.0533</td>
<td>0.220</td>
<td>0.674</td>
</tr>
<tr>
<td>Molar concentration ([\text{CH}_4])</td>
<td>0.0</td>
<td>0.063</td>
<td>1.169</td>
</tr>
<tr>
<td>Molar concentration ([\text{O}_2])</td>
<td>0.0</td>
<td>0.056</td>
<td>0.097</td>
</tr>
<tr>
<td>Molar concentration ([\text{CO}_2])</td>
<td>0.0</td>
<td>0.002</td>
<td>0.012</td>
</tr>
<tr>
<td>Molar concentration ([\text{H}_2\text{O}])</td>
<td>0.0</td>
<td>0.154</td>
<td>0.232</td>
</tr>
</tbody>
</table>

To obtain snapshots of the projected state time derivative, we approximate the derivative with a five-point approximation \( \dot{\hat{\mathbf{q}}}_i = (\mathbf{\hat{q}}_{i+2} - 8\mathbf{\hat{q}}_{i+1} + 8\mathbf{\hat{q}}_{i-1} - \mathbf{\hat{q}}_{i-2})/(12\Delta t) \). This approximation is fourth order accurate. The first two and last two time derivatives are computed using first-order forward and backward Euler approximations, respectively.

B. Learning a quadratic reduced-order model

To learn the operators of the quadratic ROM, we solve the regularized least squares problem shown in Eq. (21). In what follows next, we describe several important implementation details.

1. Singular value decomposition implementation and POD basis selection

Due to the large size of this dataset, we implement the randomized singular value decomposition algorithm, introduced in [39], to compute the leading 500 singular values and vectors of the snapshot matrix. The randomized

singular value decomposition algorithm can be implemented in a scalable way for very large datasets. The POD basis is chosen as the \( r \) leading left singular vectors. The dimension \( r \) is typically chosen so that the cumulative energy contained in the subspace is greater than a user specified tolerance \( \epsilon \), i.e.,

\[
\sum_{k=1}^{r} \sigma_k^2 > \epsilon,
\]

where \( \sigma_k^2 \) are the squared singular values of the data matrix \( \mathbf{Q} \). To guide the choice of \( r \) we also use the relative projection error

\[
\mathcal{E}_{\text{proj}} = \frac{\| \mathbf{Q} - \mathbf{V}_r \mathbf{V}_r^T \mathbf{Q} \|^2_F}{\| \mathbf{Q} \|^2_F} = 1 - \frac{\sum_{k=1}^{r} \sigma_k^2}{\sum_{k=1}^{\infty} \sigma_k^2}
\]

(31)

2. Removing redundant terms in least squares problem

Due to the redundant terms computed in \( \mathbf{Q} \odot \hat{\mathbf{Q}} \) in equation (19), the least squares problem may become ill-posed. Thus, the Kronecker product is replaced with the term

\[
\hat{\mathbf{Q}}^2 = \begin{bmatrix} \hat{q}_0^2 & \hat{q}_1^2 & \ldots & \hat{q}_K^2 \end{bmatrix} \in \mathbb{R}^{s \times K},
\]

where \( s = \frac{r(r+1)}{2} \). Each vector \( \hat{q}_j^2 \) is defined, according to [26], as

\[
\hat{q}_j^2 = \begin{bmatrix} q_{j}^{(1)} \\ \vdots \\ q_{j}^{(r)} \end{bmatrix} \in \mathbb{R}^r, \quad \text{where,} \quad q_{j}^{(i)} = \hat{q}_{i,j} \begin{bmatrix} \hat{q}_{i,j} \\ \vdots \\ \hat{q}_{r,j} \end{bmatrix} \in \mathbb{R}^i,
\]

and \( \hat{q}_{i,j} \) is the \( i \)th element of the vector \( \hat{q}_j \). Now, instead of learning the operator \( \hat{\mathbf{H}} \in \mathbb{R}^{r \times r^2} \), the operator \( \hat{\mathbf{F}} \in \mathbb{R}^{r \times s} \) is learned, which satisfies the equivalent least squares problem

\[
\min_{\hat{\mathbf{A}} \in \mathbb{R}^{r \times r}, \hat{\mathbf{F}} \in \mathbb{R}^{r \times s}, \hat{\mathbf{B}} \in \mathbb{R}^{s \times m}, \hat{\mathbf{e}} \in \mathbb{R}^r} \left\| \hat{\mathbf{Q}}^T \hat{\mathbf{A}}^T + (\hat{\mathbf{Q}}^2)^T \hat{\mathbf{F}}^T + \mathbf{U}^T \hat{\mathbf{B}}^T + \mathbf{1}_K \hat{\mathbf{e}}^T - \frac{\hat{\mathbf{q}}}{\| \hat{\mathbf{q}} \|^2} \right\|_2^2
\]

(32)

The least squares problem is again of the form (20) but we use the data matrix \( \mathbf{D} = [\hat{\mathbf{Q}}^T \ (\hat{\mathbf{Q}}^2)^T \ \mathbf{U}^T \ \mathbf{1}_K] \in \mathbb{R}^{K \times (r+s+m+1)} \) and solve for the operators \( \mathbf{O} = [\hat{\mathbf{A}} \ \hat{\mathbf{F}} \ \hat{\mathbf{B}} \ \hat{\mathbf{e}}] \in \mathbb{R}^{r \times (r+s+m+1)} \). Once we have solved for the operator \( \hat{\mathbf{F}} \) we can easily transform it to obtain \( \hat{\mathbf{H}} \).

3. Regularization

We use an L_2-regularization (also known as Tikhonov regularization or ridge regression) to solve the operator inference problem, as shown in Eq. (32). The regularization term introduces a trade-off between operators that fit the data well and operators with small values. This regularization is used to avoid overfitting to the data, which in this setting causes our learned ROMs to be unstable. The regularization parameter \( \lambda \) affects the performance of this algorithm—we require enough regularization to avoid overfitting, but if \( \lambda \) is too large, the data will be poorly fit.

To help determine appropriate values of \( \lambda \), we consider the “L-curve” criterion discussed in [40]. The L-curve is a way of visualizing the effects of different values of \( \lambda \) on the norm of the residual (data fit) against the norm of the solution. The L-curve criterion recommends choosing a value for \( \lambda \) that lies in the corner of the curve, nearest the origin. In our numerical experiments, we compute the L-curve to help determine appropriate values for \( \lambda \).

Regularization also helps to reduce the condition number of the regularized least squares data matrix, \( [\mathbf{D} \ \lambda \mathbf{P}]^T \). In Figure 2 we show the condition number of the data matrix \( \mathbf{D} \) from Eq. (32) (with the data already scaled to [-1,1]) when we include 2500, 5000, 7500 and 10000 snapshots in the training set. The condition number is quite large for this application, yet it decreases as we add more training data. This effect is due to the fact that as we add more training data, the first \( r \) POD basis vectors become richer. This in turn means that the projected data in \( \mathbf{D} \) are richer for a given dimension of the POD basis. Our numerical experiments reinforced this finding by confirming that 10000 training snapshots were required to achieve a sufficiently rich POD basis to obtain accurate ROMs, an indication of the complexity of the combustor dynamics.
4. Error measures

To evaluate the ROM performance, we define appropriate error metrics for each variable. The error is computed after the ROM solutions have been reconstructed back into the full CFD model dimension and scaled back to the original variable ranges. Recall Table 1 which showed the range of values for each variable. Below we provide details on how error is computed for each variable:

- For pressure and temperature, the values are always positive and well above zero, so we use a standard relative error, defined as
  \[
  E_{\text{relative},i} = \frac{|\zeta_{\text{CFD},i} - \zeta_{\text{ROM},i}|}{|\zeta_{\text{CFD},i}|},
  \]
  (33)
  where \( \zeta \in \{p, T\} \) and \( \zeta_{-,i} \) denotes the \( i \)th entry of a vector, and \( \zeta_{\text{CFD}} \) is the CFD solution variable and \( \zeta_{\text{ROM}} \) is the solution variable obtained from the ROM simulation.

- Due to the small values of species concentrations (on the order of \( 10^{-12} \)), dividing by the true value can skew a small error. Similarly, velocities range from positive to negative, including zero. Thus, for species concentrations and velocities, we use a normalized absolute error, defined as
  \[
  E_{\text{nabs},i} = \frac{|\zeta_{\text{CFD},i} - \zeta_{\text{ROM},i}|}{\max_l(|\zeta_{\text{CFD},l}|)},
  \]
  (34)
  where \( \zeta \in \{v_x, v_y, [\text{CH}_4], [\text{O}_2], [\text{CO}_2], [\text{H}_2\text{O}]\} \) and \( \max_l(|\zeta_{\text{CFD},l}|) \) denotes the maximum entry of \( |\zeta_{\text{CFD}}| \), i.e., the maximum absolute value over the discretized spatial domain.

C. Learned reduced model performance

The given \( K = 10000 \) snapshots representing 1ms of GEMS simulation data are used to learn the ROM. We are also given another 1ms of testing data at the monitor locations shown in Figure 1b, which we use to assess the predictive capabilities of our learned ROMs beyond the range of training data.

The cumulative energy of the singular values of the snapshot data matrix \( Q \) is shown in Figure 3. The singular values that correspond to a cumulative energy of 97.5\% and 99\% are indicated with the red triangle and green square, respectively. We use basis sizes of \( r = 17 \), capturing 97.5\% of the total energy, and \( r = 29 \), capturing 99\% of the total energy.

In Figures 4a and 4b we show the L-curve for each basis size. The L-curve for a basis of \( r = 17 \) is somewhat uninformative in this case. The regularization parameters chosen were \( \lambda = 3.0\text{E+04} \) and \( 1.0\text{E+05} \), which lie in the middle of the L-curve and are therefore a recommended choice by the L-curve criterion discussed above. For a basis of size of \( r = 29 \), the L-curve indicates a regularization parameter around \( \lambda = 3.0\text{E+04} \). Stable systems are produced for \( \lambda = 3.0\text{E+04} \) and \( 5.0\text{E+04} \).

We simulate the learned ROM for the two model sizes \( r = 17 \) and \( r = 29 \) with the same initial value and time step size \( (\Delta t = 1 \times 10^{-7} \text{s}) \) as those used to generate the training set. Figures 5 and 6 compare the time trace of pressure computed by GEMS (our “truth” data) with the ROM predictions for all 20000 time steps at the cell located at \((0.0, 0.0225)\) in the...
The cumulative energy computed with Eq. (31). The leading $r = 17$ singular values capture 97.5% of the energy and the leading $r = 29$ capture 99%.

Fig. 3 The cumulative energy computed with Eq. (31). The leading $r = 17$ singular values capture 97.5% of the energy and the leading $r = 29$ capture 99%.

Fig. 4 The L-curve for different basis sizes and regularization parameters $\lambda$. The horizontal axis shows the squared norm of learned operators, $\|O\|_2^2$, and the vertical axis shows the least squares residual, $\|DO^T - Q^T\|_2^2$.

The L-curve for different basis sizes and regularization parameters $\lambda$. The horizontal axis shows the squared norm of learned operators, $\|O\|_2^2$, and the vertical axis shows the least squares residual, $\|DO^T - Q^T\|_2^2$.

domain (denoted as monitor location 1 in Figure 1(b). The performance of the ROM on the training data (first 1ms of data) is good in both cases, although the $r = 29$ ROM (Figure 6) is more accurate. For test data predictions beyond the training data, Figure 6 shows that a basis of size of $r = 29$ yields accurate phase predictions and pressure oscillation amplitudes that are good approximations of the truth. We notice an improved amplitude prediction when using a smaller regularization value.

We also compute the average error of each variable over the entire domain at the last time step of the training set (the 10000th time step). The normalized absolute error, defined in Eq. (34), is shown for each species and for $x$ and $y$ velocity in Figure 7. This figure also shows the relative error, defined in Eq. (33), for pressure and temperature. These plots show that overall, the error is decreasing with an increasing basis size. At a basis size of 18, 22, and 24 the systems were unstable and so these basis sizes are excluded from the figure. The cause of this may be due to the fact that the same regularization parameter was used for each of these, $\lambda = 3.0E+04$, and ideally one would pick a parameter specific for the basis size.

In Figure 8 we show the integrated species concentrations over time. To compute these, at each time step in our simulation, we sum all elements of a species vector. This measure monitors whether our ROM conserves species mass, an important feature of a physically meaningful simulation. As the discretization of the high-fidelity model becomes finer, point-wise error may become large and misleading if the mass is shifted slightly into the neighboring cells. The integrated species concentration complements the evaluation of point-wise errors and provides a global view of the error in the domain.

We also compare the state variables over the entire domain predicted by the learned ROM with the given GEMS data
Fig. 5  Pressure time traces for basis size $r = 17$. Training with 10000 snapshots. Black vertical line denotes the end of the training data and the beginning of the test data.

Fig. 6  Pressure time traces for basis size of $r = 29$. Training with 10000 snapshots. Black vertical line denotes the end of the training data and the beginning of the test data.

Fig. 7  Error measures vs. basis size, averaged over the spatial domain at the last time step of training data. Normalized absolute error (34) of species CH$_4$, O$_2$, CO$_2$, H$_2$O and $v_x$ and $v_y$ velocities; Relative error (33) given for pressure and temperature.
Fig. 8 Integrated species at each time step for different basis sizes. Training with 10000 snapshots.

at the last time step $K = 10000$ (which corresponds to $t = 0.0159999$s). We provide the true field, the ROM-predicted field, and an error field for each variable in Figures 9–16. Again, for pressure and temperature, we use a relative error from Eq. (33). For x and y velocity and for species molar concentrations we use a normalized absolute error from Eq. (34). The plots show that the ROM predictions are, as expected, not perfect, but indeed they capture well the overall structure and many details of the solution fields.

Fig. 9 Predictive results for pressure at the last time step of training data. Training with 10000 snapshots, a basis size of $r = 29$, and regularization set to $\lambda = 3.0E + 04$.

Table 2 shows timing results for the ROM generation and simulation, as performed using python 3.6.4 on a dual-core Intel i5 processor with 2.3 GHz and 16GB RAM. We report the following CPU times: solving the operator inference least squares problem (32); ROM runtime for the two different basis sizes for 2ms of real time prediction; and reconstruction of the high-dimensional, unscaled combustion variables. The latter is required since the ROM is evolving the dynamics in the $[-1, 1]$ scaled variables, and thus a post-processing step is required to obtain the true magnitudes of the variables. In comparison with the approximately 200h of CPU time needed on a 40-core architecture (more details in Section IV.A) to compute the first 10000 snapshots of data, the ROMs provide five to six orders of magnitude in computational speedup.
Fig. 10  Predictive results for velocity $v_x$ at the last time step of training data. Training with 10000 snapshots, a basis size of $r = 29$, and regularization set to $\lambda = 3.0E + 04$.

Fig. 11  Predictive results for velocity $v_y$ at the last time step of training data. Training with 10000 snapshots, a basis size of $r = 29$, and regularization set to $\lambda = 3.0E + 04$.

Table 2  CPU times for two ROMs with time step size $\Delta t = 1 \times 10^{-7}s$ and 20000 time steps.

<table>
<thead>
<tr>
<th>ROM order</th>
<th>LS solve of (21)</th>
<th>ROM simulation</th>
<th>Reconstructing high-dim. field</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r = 17$</td>
<td>0.97s</td>
<td>2.88s</td>
<td>0.03s</td>
</tr>
<tr>
<td>$r = 29$</td>
<td>6.05s</td>
<td>3.73s</td>
<td>0.05s</td>
</tr>
</tbody>
</table>
Fig. 12  Predictive results for temperature at the last time step of training data. Training with 10000 snapshots, a basis size of $r = 29$, and regularization set to $\lambda = 3.0E + 04$.

Fig. 13  Predictive results for CH$_4$ molar concentration at the last time step of training data. Training with 10000 snapshots, a basis size of $r = 29$, and regularization set to $\lambda = 3.0E + 04$.

Fig. 14  Predictive results for O$_2$ molar concentration at the last time step of training data. Training with 10000 snapshots, a basis size of $r = 29$, and regularization set to $\lambda = 3.0E + 04$. 
Fig. 15  Predictive results for CO$_2$ molar concentration at the last time step of training data. Training with 10000 snapshots, a basis size of $r = 29$, and regularization set to $\lambda = 3.0E + 04$.

Fig. 16  Predictive results for H$_2$O molar concentration at the last time step of training data. Training with 10000 snapshots, a basis size of $r = 29$, and regularization set to $\lambda = 3.0E + 04$. 
V. Conclusion

Operator inference is a data-driven method for learning reduced-order models (ROMs) of dynamical systems with polynomial structure. This paper demonstrates how variable transformations can expose quadratic structure in the nonlinear system of partial differential equations describing a complex combustion model. This quadratic structure is preserved under projection, providing the mathematical justification for learning a quadratic ROM using operator inference. An important feature of the approach is that the learning of the ROM is non-intrusive—it requires state solutions generated by running the high-fidelity combustion model, but it does not require access to the discretized operators of the governing equations. This is important because it means that the variable transformations can be applied as a post-processing step to the simulation data set, but no intrusive modifications are needed to the high-fidelity code. While the quadratic model form is an approximation for this particular application problem, the numerical results show that the learned quadratic ROM can predict relevant quantities of interest and can also conserve species accurately. Many nonlinear equations in scientific and engineering applications admit variable transformations that expose polynomial structure. This combined with the non-intrusive nature of the approach make it a viable option for deriving ROMs for complex nonlinear applications where traditional intrusive model reduction is impractical and/or unreliable.

Appendix A: Lifting chemical source terms

The chemical source terms in $\dot{S}$ in Eq. (6) are $\dot{\omega}_l = M_i \dot{c}_l^{\text{reaction}}$, see Eq. (8). The dynamics for the source terms $\dot{c}_l^{\text{reaction}}$ are given by

$$\dot{c}_1^{\text{reaction}} = A \exp \left( -\frac{E_a}{R_a T} \right) c_1^{0.2} c_2^{1.3}$$

(35a)

$$\dot{c}_2^{\text{reaction}} = 2c_1$$

(35b)

$$\dot{c}_3^{\text{reaction}} = -c_1$$

(35c)

$$\dot{c}_4^{\text{reaction}} = -2c_1$$

(35d)

To lift this system to polynomial form, we introduce the auxiliary variables

$$w_1 = c_1^{0.2}, \quad w_2 = c_1^{-1}, \quad w_3 = c_2^{1.3}, \quad w_4 = c_2^{-1}, \quad w_5 = \exp \left( -\frac{E_a}{R_a T} \right), \quad w_6 = \frac{1}{T^2}. \tag{36}$$

The source term dynamics (35) are then cubic in $w_1, w_3, w_5$:

$$\dot{c}_1^{\text{reaction}} = A w_1 w_3 w_5,$$

(37a)

$$\dot{c}_2^{\text{reaction}} = 2A w_1 w_3 w_5,$$

(37b)

$$\dot{c}_3^{\text{reaction}} = -A w_1 w_3 w_5,$$

$$\dot{c}_4^{\text{reaction}} = -2A w_1 w_3 w_5.$$

We next derive the dynamics for the auxiliary variables $w_1, \ldots, w_6$. For instance, we have $\dot{w}_1 = 0.2 w_1^{0.2-1} c_1 = 0.2 w_1 w_2 c_1$. Similarly, we obtain the system of auxiliary dynamics:

$$\dot{w}_1 = 0.2 w_1 w_2 c_1$$

(38)

$$\dot{w}_2 = -w_2^2 c_1,$$

$$\dot{w}_3 = 1.3 w_3 w_4 c_2$$

(39)

$$\dot{w}_4 = -w_4^2 c_2,$$

$$\dot{w}_5 = \frac{E_a}{R_a} \frac{1}{T^2} w_5 \dot{T} = \frac{E_a}{R_a} w_5 w_6 \dot{T}$$

(40)

$$\dot{w}_6 = -2 \frac{1}{T^3} = -2 T w_6^2$$

The dynamics of the lifted variables are quintic in the variables $w_1, \ldots, w_6$. If we further include an additional auxiliary variable $w_7 = w_1 w_3 w_5$, then we obtain the system of equations

$$\dot{w}_1 = 0.2 A w_1 w_2 w_7,$$

(41)

$$\dot{w}_2 = -A w_1^2 w_7,$$

$$\dot{w}_3 = 2.6 A w_3 w_4 w_7,$$

(42)

$$\dot{w}_4 = -2 A w_2^2 w_7,$$

$$\dot{w}_5 = \frac{E_a}{R_a} w_5 w_6 \dot{T},$$

(43)

$$\dot{w}_6 = -2 T w_6^2,$$

$$0 = w_7 - w_1 w_3 w_5$$

(44)

The temperature $T = \frac{p \xi}{R(Y_f)}$ can be obtained from the states $\xi, p, Y_f$ via the ideal gas relationship in Eq. (10).
Appendix B: Equations for pressure and chemical species

Here, we give the complete governing equation for the species molar concentrations $c_l$ and pressure $p$. Recall from Section III.C the notation used to denote the conservative variables $g_l$: $g_1 = \rho$, $g_2 = \rho v_x$, $g_3 = \rho v_y$, $g_4 = \rho e$, $g_5 = \rho Y_1$, $g_6 = \rho Y_2$, $g_7 = \rho Y_3$, $g_8 = \rho Y_4$.

**Species molar concentrations $c_l$:** For the species molar concentration dynamics, we use the relationship $c_l = \frac{\rho Y_l}{M_l}$ from Eq. (4), where the constants $M_1, \ldots, M_4$ are molar masses. We obtain for $l = 1, 2, \ldots, n_{sp}$:

$$\frac{\partial c_l}{\partial t} = \frac{1}{M_l} \frac{\partial \rho Y_l}{\partial t}$$

and since $Y_l = M_l c_l \xi$, we have that

$$\rho \frac{\partial Y_l}{\partial t} = \rho M_l \frac{\partial c_l \xi}{\partial t} = M_l \left( \frac{\partial c_l}{\partial x} + \rho \frac{\partial \xi}{\partial x} \right)$$

Overall, we have that

$$\frac{\partial c_l}{\partial t} = \dot{c}_l^{\text{reaction}} - \left( \frac{\partial v_x c_l}{\partial x} + \frac{\partial v_y c_l}{\partial y} + \frac{\partial c_l}{\partial x} + \rho c_l \frac{\partial \xi}{\partial x} \right) = \dot{c}_l^{\text{reaction}} - \left( M_l c_l \xi \right)$$

which is quadratic in the learning variables $v_x, v_y, c_l$ with the exception of the terms $\rho c_l \frac{\partial \xi}{\partial x}$ and $\rho c_l \frac{\partial \xi}{\partial y}$. We note that if $\rho$ were included as a lifted variable (in addition to $\xi$), these terms would become quadratic in the lifted state.

**Pressure $p$:** We start with the energy equation (2). By multiplying with density $\rho$, we have $\rho e = \rho h^0 - p$, so from the conservation equation (1) for $\rho e$ we obtain

$$\frac{\partial (\rho h^0 - p)}{\partial t} + \frac{\partial \rho v_x h^0}{\partial x} + \frac{\partial \rho v_y h^0}{\partial y} + \frac{\partial}{\partial x} (v_x \tau_{xx} + v_y \tau_{yx} - j_x^q) + \frac{\partial}{\partial y} (v_x \tau_{xy} + v_y \tau_{yy} - j_y^q) = 0.$$  \hspace{1cm} (52)

This directly gives an equation for the time evolution of pressure:

$$\frac{\partial p}{\partial t} = \frac{\partial \rho h^0}{\partial t} + \frac{\partial \rho v_x h^0}{\partial x} + \frac{\partial \rho v_y h^0}{\partial y} + \frac{\partial}{\partial x} (v_x \tau_{xx} + v_y \tau_{yx} - j_x^q) + \frac{\partial}{\partial y} (v_x \tau_{xy} + v_y \tau_{yy} - j_y^q) = 0.$$  \hspace{1cm} (53)

Moreover, per definition of $h^0$ in Eq. (2) and with $c_l = \frac{\rho Y_l}{M_l}$ we have that

$$\frac{\partial c_l}{\partial t} = \frac{\partial \rho h^0}{\partial t} = \rho \frac{\partial Y_l}{\partial t}$$

Overall, we have that

$$\frac{\partial p}{\partial t} = \sum_{i=1}^{n_p} \frac{\partial p h_l}{\partial t} + \frac{1}{2} \left( v_x^2 + v_y^2 \right) \frac{\partial \rho v_x}{\partial t} + \rho \frac{\partial (v_x + v_y)}{\partial t} - \frac{\partial \rho Y_l}{\partial x} + \frac{\partial \rho Y_l}{\partial y}$$

$$= \sum_{i=1}^{n_p} M_l \frac{\partial h_l c_l}{\partial t} + \frac{1}{2} \left( v_x^2 + v_y^2 \right) \frac{\partial \rho v_x}{\partial t} + \rho \frac{\partial (v_x + v_y)}{\partial t} - \frac{\partial \rho Y_l}{\partial x} + \frac{\partial \rho Y_l}{\partial y}$$

$$+ \frac{\partial}{\partial x} (v_x \tau_{xx} + v_y \tau_{yx} - j_x^q) + \frac{\partial}{\partial y} (v_x \tau_{xy} + v_y \tau_{yy} - j_y^q).$$  \hspace{1cm} (54)
This equation remains nonlinear in our chosen learning variables $\vec{q}_L$. In particular, the enthalpies $h_l = h_l(T)$ and their time derivatives are nonlinear functions of temperature. The other terms show some polynomial structure; for example, in Section III.C we showed that $\frac{\partial (v_x + v_y)}{\partial t}$ is quadratic in the learning state variables $p, v_x, v_y, \xi$. However, to write this equation exactly in a polynomial form would require introducing a large number of auxiliary variables along with their corresponding dynamics. We have instead chosen to introduce an approximation by learning a ROM in the variables $\vec{q}_L$ with quadratic form.

**Funding Sources**

This work has been supported in part by the Air Force Center of Excellence on Multi-Fidelity Modeling of Rocket Combustor Dynamics award FA9550-17-1-0195, and the Air Force Office of Scientific Research (AFOSR) MURI on managing multiple information sources of multi-physics systems award numbers FA9550-15-1-0038 and FA9550-18-1-0023.

**References**


