Combustion and Flame 227 (2021) 172–185

Contents lists available at ScienceDirect

Combustion and Flame

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

Data-assisted combustion simulations with dynamic submodel assignment using random forests

Wai Tong Chung[a,*], Aashwin Ananda Mishra[b], Nikolaos Perakis[a,c], Matthias Ihme[a]

[a] Department of Mechanical Engineering, Stanford University, Stanford, CA 94305, USA
[b] SLAC National Accelerator Laboratory, Menlo Park, CA 94025, USA
[c] Chair of Space Propulsion, Technical University of Munich, Garching 85748, Germany

ARTICLE INFO

Article history:
Received 8 September 2020
Revised 23 December 2020
Accepted 24 December 2020

Keywords:
Combustion modeling
Model assignment
Random forests
Machine learning
Classification
Turbulent combustion

ABSTRACT

This investigation outlines a data-assisted approach that employs random forest classifiers for local and dynamic submodel assignment in turbulent-combustion simulations. This method is demonstrated in simulations of a single-element GOX/GCH4 rocket combustor; *a priori* as well as *a posteriori* assessments are conducted to (i) evaluate the accuracy and adjustability of the classifier for targeting different quantities of interest (QoIs), and (ii) assess improvements, resulting from the data-assisted combustion model assignment, in predicting target QoIs during simulation runtime. Results from the *a priori* study show that random forests, trained with local flow properties as input variables and combustion model errors as training labels, assign three different combustion models – finite-rate chemistry (FRC), flamelet progress variable (FPV) model, and inert mixing (IM) – with reasonable classification performance even when targeting multiple QoIs. Applications in *a posteriori* studies demonstrate improved predictions from data-assisted simulations, in temperature and CO mass fraction, when compared with monolithic FPV calculations. An additional *a posteriori* data-assisted simulation of a modified configuration demonstrates that the present approach can be successfully applied to different configurations, as long as thermophysical behavior can be represented by the training data. These results demonstrate that this data-driven framework holds promise for dynamic combustion submodel assignments in reacting flow simulations.

1. Introduction

High-fidelity simulations of turbulent reacting flows can incur high computational costs due to the complexity required for employing finite-rate chemical mechanisms and resolving relevant scales. Numerous strategies [1] have been employed for reducing the computational cost of detailed chemical mechanisms, such as (i) removing non-essential species and reactions [2,3], (ii) lumping similar species and reaction pathways [4,5], (iii) time-scale analysis [6,7], and (iv) stiffness reduction [8,9].

Alternatively, a significant portion of combustion research has been devoted to the development of cost-efficient models for representing the combustion chemistry and turbulent scales [10]. The most popular of these low-order manifold models are categorized under flamelet methods, which represent combustion chemistry through solutions of representative flame configurations, such as laminar counterflow diffusion flames, freely propagating premixed flames, or homogeneous reactor systems. Examples of flamelet methods include the Burke-Schumann solution [11], the flame-prolongation in intrinsic lower-dimensional manifold (FPI) [12], the flamelet-generated manifold (FGM) method [13], and the flamelet/progress variable (FPV) method [14,15]. These reduced manifold models are commonly employed to describe specific combustion regimes – a multitude of which can exist within practical combustors. However, expert knowledge and experimental data is often required to correctly assign the most appropriate combustion model.

One solution to this issue is provided by dynamic adaptive chemistry methods [16–18] that save computational cost by reducing detailed chemical mechanisms, and transitioning between smaller sets of chemical models to represent combustion regimes of different chemical fidelity. A general mathematical framework was proposed by Wu et al. [19,20] through the Pareto-efficient combustion (PEC) approach. In this approach, the compliance of a combustion submodel with the underlying flow-field representation is assessed through the construction of a so-called drift term, taking into consideration user-specific requirements about quantities of interest (QoI) and computational cost [21]. While

* Corresponding author.
E-mail address: wtchung@stanford.edu (W.T. Chung).

https://doi.org/10.1016/j.combustflame.2020.12.041
0010-2180/

mathematically rigorous, these techniques are limited by their reliance on local information regarding the chemical composition and the construction of the model-compliance indicator. In contrast, data-driven methods can potentially offer a universal solution by allowing for the consideration of a wider range of conditions and processes that cannot be easily represented in the form of mathematical expressions accessible to an indicator function.

Data-driven methods involve the extraction of knowledge from data [22]. These methods can be useful as long as a substantial corpus of data is available to infer relationships between input variables and QoIs. As such, the employment of learning algorithms are gaining popularity in the simulation of turbulent flows. These methods have shown success in quantifying uncertainty [23], and augmenting closure models [24,25] in Reynolds-averaged Navier–Stokes (RANS) simulations and large eddy simulations (LES) [26].

In simulations of turbulent reacting flows, data-driven methods have also been applied with generating additional subgrid-scale closure that arises from the inclusion of combustion chemistry. In particular, artificial neural networks have been employed for regressing thermophysical quantities in LES of turbulent flames [27–31]. *A priori* studies have been performed to demonstrate that convolutional neural-networks can provide accurate closure for turbulent combustion models [32]. Ranade and Echekki [33] conducted an *a posteriori* study to show that artificial neural networks (ANNs) can be trained with experimental data to generate closure models for chemical scalars in RANS simulations of turbulent jet flames. Henry de Frahan et al. [34] evaluated the use of ANNs, random forests, and generative learning methods for predicting the sub-filter probability density function in a turbulent combustion LES. Seltz et al. [35] employed convolutional neural networks to generate closure for unresolved terms in the filtered progress variable transport equation. Yao et al. [36] demonstrated that ANNs can be used to approximate the conditional scalar dissipation rate in spray flame LES.

To reduce computational costs that arise from complex combustion chemistry, various strategies have been employed through learning algorithms. Artificial neural networks were first successfully integrated within simulations of turbulent reacting flows as an alternative for representing chemical reactions [27,28,37]. Chatzopoulos and Rigopoulos [38], and Franke et al. [39] demonstrated that training data extracted from 100 laminar flamelets was sufficient for training ANNs for representing chemistry in simulations more complex turbulent flame configurations. With this generic training set, ANNs showed a small capacity for extrapolation, but it was noted that accurate predictions were challenging if the target predictions deviated too largely away from the training set. Sen and Menon [31], and Alqahtani and Echekki [40] also demonstrated that ANNs can be used for replacing stiff ODE solvers in turbulent flame simulations, with good accuracy and CPU performance. Ihme et al. [29], Kempf et al. [30], and Owoyele et al. [41] used optimal ANN tabulation to replace conventional tabulation methods in manifold-based simulations.

These aforementioned approaches typically involve the use of regression for estimating numerical predictions. Regression models in flow-physics problems are still in its infancy, and face challenges when extrapolating without an appropriate training set – resulting in errors that arise from numerical predictions that only match specific flow configurations represented by the training data [25]. The present study ameliorates this issue by employing a classification algorithm that assigns well-tested physics-based combustion submodels of varying fidelity and complexity within the simulation domain. Thus, the potential approximation errors made by the machine-learning algorithm are limited by the predictive capability of the lowest performing submodel.

In the approach that is proposed in this work, local thermophysical quantities in the flow field are utilized as features for a random forest algorithm that spatially and dynamically assigns combustion submodels. Random forests are an ensemble learning method commonly used in both classification and regression problems. Errors made by submodels, when predicting user-defined QoI, are used to construct the labels used for training the random forest. Overall computational fidelity and cost of the simulation is determined by a user-defined submodel error threshold during training. This approach couples the assigned combustion submodels in the *a posteriori* simulations by employing the mass-conserving approach developed by Wu et al. [20], but with a data-driven assignment approach that replaces the drift-term in the original PEC formulation.

This investigation is performed with the following objectives:

- To introduce classification algorithms for combustion submodel assignment, and assess the resulting data-assisted simulations.
- To evaluate the suitability, accuracy, and adjustability of random forests for submodel assignment.

To this end, random forests are assessed for the purpose of local and dynamic model assignment in simulations of a gaseous-oxygen/gaseous-methane (GOX/GCH4) single-element rocket combustor [42,43]. The mathematical models for simulating the turbulent combustion are presented in Section 2. The experimental configuration, computational setup and baseline simulations using monolithic combustion models are discussed in Section 3. The data-driven framework is introduced in Section 4. Results from *a priori* and *a posteriori* assessments of the random forests are presented and discussed in Section 5, before offering concluding remarks in Section 6.

2. Mathematical models

2.1. Computational method

The governing equations that are solved in the present study are the Favre-filtered conservation equations for mass, momentum, energy, and chemical species:

$$\partial_t \overline{\rho} + \nabla \cdot (\overline{\rho}\widetilde{\boldsymbol{u}}) = 0 \tag{1a}$$

$$\partial_t (\overline{\rho}\widetilde{\boldsymbol{u}}) + \nabla \cdot (\overline{\rho}\widetilde{\boldsymbol{u}}\widetilde{\boldsymbol{u}}) = -\nabla \cdot (\overline{p}\boldsymbol{I}) + \nabla \cdot (\overline{\boldsymbol{\tau}}_v + \boldsymbol{\tau}_t) \tag{1b}$$

$$\partial_t (\overline{\rho}\widetilde{e}) + \nabla \cdot [\widetilde{\boldsymbol{u}}(\overline{\rho}\widetilde{e} + \overline{p})] = -\nabla \cdot (\overline{\boldsymbol{q}}_v + \boldsymbol{q}_t) + \nabla \cdot [(\overline{\boldsymbol{\tau}}_v + \boldsymbol{\tau}_t) \cdot \widetilde{\boldsymbol{u}}] \tag{1c}$$

$$\partial_t (\overline{\rho}\widetilde{\boldsymbol{\phi}}) + \nabla \cdot (\overline{\rho}\widetilde{\boldsymbol{u}}\widetilde{\boldsymbol{\phi}}) = -\nabla \cdot (\overline{\boldsymbol{J}}_v + \boldsymbol{J}_t) + \overline{\dot{\boldsymbol{S}}} \tag{1d}$$

with density ρ, velocity vector $\boldsymbol{u}$, specific total energy e, stress tensor $\boldsymbol{\tau}$, and heat flux vector $\boldsymbol{q}$; $\overline{\cdot}$ denotes a filtered quantity and $\widetilde{\cdot}$ is a Favre-filtered quantity. Subscripts v and t denote viscous and turbulent quantities, respectively. Pressure p is computed from the ideal gas equation of state. $\boldsymbol{\phi}$, $\boldsymbol{J}$, and $\dot{\boldsymbol{S}}$ are the transported scalars, scalar diffusive flux, and scalar source term for the candidate combustion models. Molecular fluxes are modeled using the mixture-averaged diffusion model. The combustion models that are employed in the present study are described in detail in Section 2.2.

Simulations are performed by employing an unstructured compressible finite-volume solver [20,44,45]. A central scheme, which is 4th-order accurate on uniform meshes, is used along with a 2nd-order ENO scheme. The ENO scheme is activated only in regions of high local density variation using a threshold-based sensor. A Strang-splitting scheme is employed for time-advancement, combining a strong stability preserving 3rd-order Runge-Kutta (SSP-RK3) scheme for integrating the non-stiff operators with a semi-implicit Rosenbrock-Krylov scheme [46] for advancing

the chemical source terms. The dynamic Smagorinsky model [47] is used as closure for the subgrid-scale stresses. Turbulence/chemistry interaction is accounted for using the dynamic thickened-flame model [48], employing a maximum thickening factor of 3, which is estimated through 1D flame calculations *a priori*. Outside the flame region, both turbulent Prandtl and Schmidt numbers are prescribed at constant values of 0.7.

2.2. Combustion models

In this work, we perform LES calculations that employ three different combustion submodels, namely a finite-rate chemistry (FRC) model, the flamelet/progress variable (FPV) model [14,15], and an inert mixing (IM) model. The FRC model is defined by solving the species transport equation, Eq. (1d), through direct integration. This method does not rely on strong assumptions on flame structure and is suitable for representing complex flows as well as intermediate species and unsteady effects. Despite the high-fidelity offered by FRC, since the cost of evaluating the chemical source terms scale linearly with the number of species, the utilization of a large chemical mechanism can be prohibitively costly. FPV approach aims to alleviate the computational cost of combustion chemistry by representing the thermochemical state space using a low-dimensional manifold based on flamelets, a series of one-dimensional diffusion flames. FPV relies on the observation that laminar diffusion flames are weakly affected by the presence of turbulence, which allows the turbulent diffusion flame to be represented by flamelets. While FPV is computationally efficient, it assumes adiabaticity and cannot model effects of heat-flux across boundaries well. Lastly, IM models can only consider mixing without combustion chemistry.

The representation of scalar $\widetilde{\boldsymbol{\phi}}$ between FRC and the two tabulated chemistry models is dissimilar: FRC uses a chemical state-vector $\widetilde{\boldsymbol{\phi}} = [\widetilde{Y}_1, \ldots, \widetilde{Y}_{N_S}]^T$, consisting of N_S number of chemical species, while the FPV and IM state-vector is represented in terms of a low-dimensional manifold $\widetilde{\boldsymbol{\phi}} = \mathcal{M}(\widetilde{\boldsymbol{\psi}})$, where $\widetilde{\boldsymbol{\psi}}$ is the state vector that is used to parameterize the manifold. With the flame being artificially thickened as discussed in Section 2.1, FPV is parameterized by the mixture fraction and progress variable $\widetilde{\boldsymbol{\psi}} = [\widetilde{Z}, \widetilde{C}]^T$ which differs from the conventional practice of using a presumed-PDF closure [20]. The progress variable is defined as a linear combination of species mass fractions [49]: $C = Y_{CO_2} + Y_{H_2O} + Y_{CO} + Y_{H_2}$. For an inert and adiabatic mixture, the thermochemical state is fully parameterized by a single scalar, $\widetilde{\boldsymbol{\psi}} = [\widetilde{Z}]$.

The present framework resolves the discrepancy in scalar representation when coupling different combustion models with the approach developed by Wu et al. [20]. In this approach, a transport equation for mixture fraction is solved holistically in all models. Reconstruction of the chemical state-vector needed for FRC involves interpolation from the chemistry tables that stores all species, whereas the reconstruction of the progress variable needed for tabulated chemistry involves the sum of all major combustion product species: CO_2, CO, H_2O, and H_2. To ensure consistency between the submodels, the aforementioned reconstruction is applied for the inactive combustion model at the submodel interface at every timestep. Since the conservation laws for mass, momentum, and energy are universal among all combustion submodels, these properties are conserved throughout the domain. In addition, the choice of the dynamically-thickened flame model for the FRC and both manifold-based models avoids potential complications, since this closure model has been successfully applied to previous non-premixed flame simulations employing FRC and tabulated chemistry models [20,50,51].

The GRI-3.0 model [52], involving $N_S = 33$ chemical species, is used to describe the reaction chemistry in all combustion models. FRC is incorporated into the LES solver using the Cantera library interface [53]. The molecular diffusion of chemical species is modeled with constant Lewis numbers, which are calculated at equilibrium condition of a stoichiometric CH_4 and O_2 mixture. The chemistry table employed in the FPV-model is constructed from the solution of steady-state counterflow diffusion flames that are solved in composition space [54]. The Lewis numbers for the mixture fraction and progress variable are set at unity.

3. Experimental configuration, computational setup and baseline simulations

3.1. Experimental configuration

To evaluate the merit of the data-assisted classification method, we perform simulations of a single-element GOX/GCH4 rocket combustor [42,43]. The experimental configuration consists of a coaxial injector element where the oxidizer flows through a central jet with diameter $d_o = 4$ mm and the fuel is injected via an annulus with inner and outer diameters $d_{f,i} = 5$ mm and $d_{f,o} = 6$ mm. The combustion chamber with a total length of 285 mm has a cylindrical shape with diameter $d_{ch} = 12$ mm. A conical nozzle is attached at the end of the combustion chamber, having a contraction ratio of 2.5. This setup results in a Mach number of approximately 0.25 in the combustion chamber, which is similar to typical flight configurations. The combustor operates at a nominal operating pressure of 20 bar and a global oxidizer-to-fuel ratio of 2.6, with mass flow rates of oxidizer $\dot{m}_o$ and fuel $\dot{m}_f$ measured at 34.82 g/s and 13.39 g/s, respectively. The temperature of the oxidizer and the fuel supplied at the injector inlet are $T_o = 275$ K and $T_f = 269$ K. Static wall pressure and wall heat flux are measured through thermocouples and pressure transducers, installed along the chamber wall.

3.2. Computational setup

In this model-assignment problem, we consider an axisymmetrical domain that is representative of the single-element GOX/GCH4 rocket combustor, as shown in Fig. 1. The domain consists of a 3° combustor sector, with a truncation at 0.4 mm to remove the singularity at the centerline. Axisymmetric simulations of rocket combustors have been frequently employed to obtain insight in the turbulent combustion process [55,56], while offering feasible computational costs. This was found to be crucial for the exploration of a wider range of parameters in the data-assisted method, especially with the use of a detailed FRC-model consisting of 33 chemical species in the present study.

At the inlets, the fuel and oxidizer mass flow rates and temperature are prescribed following the experimental measurements [42,43]. At the chamber and nozzle walls, the temperature profile is defined as a Dirichlet boundary condition, which is obtained from the measurements by Perakis and Haidn [57]. The bottom and axisymmetric faces are prescribed with symmetry boundary conditions. All remaining boundaries are defined as adiabatic non-slip walls with the exception of the exhaust, which is modeled as a pressure outlet. The computational domain is discretized by a block-structured mesh consisting of 2×10^5 cells. The wall-normal direction is resolved down to 30 μm, and a wall model [58] is employed for the viscous sublayer. Simulations are performed using 600 Intel Xeon (E5-2680v2) processors. The solution is advanced using a typical timestep of 25 ns, corresponding to a convective CFL number of 1.0.

3.3. Baseline results from monolithic LES combustion simulations

Simulations of the rocket combustor are first performed using monolithic FRC and monolithic FPV simulations. Flow fields

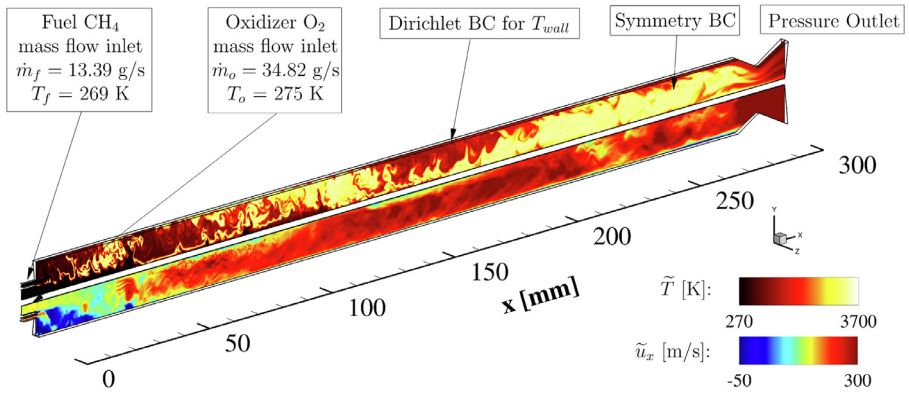


Fig. 1. Computational domain presented in conjunction with instantaneous temperature (top) and axial velocity (bottom) fields from monolithic FRC simulations.

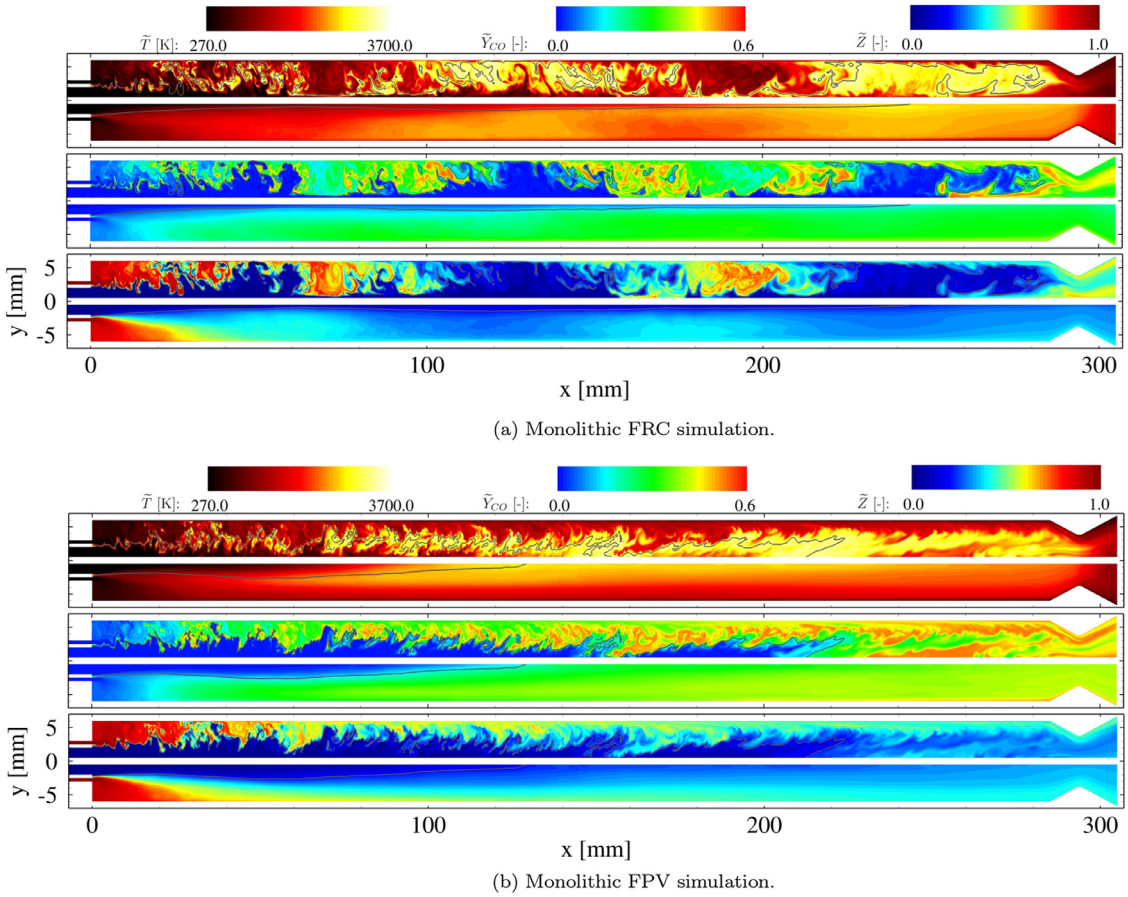


Fig. 2. Temperature, CO mass fraction, and mixture fraction fields (from top to bottom) for (a) monolithic FRC and (b) monolithic FPV simulations. Upper half: instantaneous fields, bottom half: time-averaged fields. The location of the stoichiometric mixture, $\widetilde{Z}_{st} = 0.2$, is shown by black lines.

are initialized with equilibrium products and temperature, thus allowing the monolithic FPV simulation to ignite. Instantaneous and time-averaged fields of temperature, CO mass fraction, and mixture fraction from monolithic FRC calculations and monolithic FPV simulations are shown in Fig. 2(a) and (b), respectively. Results from the FRC simulations are qualitatively similar to previous simulations [55,59], where a non-uniform mixture fraction field, a long oxygen core, and an agglomeration of cold rich gases to the chamber wall are observed. In contrast, some notable differences are observable from the FPV simulations, shown in Fig. 2b. In particular, a thicker thermal boundary layer is seen for the FPV simulation. This difference is consistent with other LES studies [60] which have shown that an adiabatic FPV model, as employed in the present study, mispredicts the wall-heat loss and exothermic CO-recombination in the boundary layer [59].

4. Data-assisted simulation framework

In this investigation, the present data-driven framework uses a supervised learning algorithm for combustion submodel assignment. During training, the supervised learning algorithm learns a function $f : \boldsymbol{x} \mapsto y$ that maps with data containing input vector $\boldsymbol{x} \in \mathcal{X}$, and the corresponding true response $y \in \mathcal{Y}$. A trained supervised learning model can then provide an approximation for any output $y \in \mathcal{Y}$, when fed with a new input set $\boldsymbol{x}$. The procedure for

incorporating a supervised learning algorithm for combustion submodel assignment is as follows:

1. Generate data either from experimental measurements or numerical simulations. In this work, we use the instantaneous flow-field solutions from the FRC simulation of the GOX/CH4 rocket combustor as learning dataset, discussed in Section 3.
2. Assign labels to the training data. Prior to training, each training datapoint is typically assigned a true response. In this work, we present a multiclass classification problem for optimal assignment of three combustion models $\mathcal{Y} = \{\text{IM, FPV, FRC}\}$. Hence, we use the local combustion submodel error of two essential local QoIs, namely T and Y_{CO}, to programmatically assign labels. Details are presented in Section 4.1.
3. Construct the feature vector $\boldsymbol{x} \in \mathcal{X}$. In this work, we apply a feature selection method based on the Maximal Information Coefficient (MIC) [61], as discussed in Section 4.2, to construct a feature set consisting of local thermophysical quantities that include the mixture fraction, progress variable, density, local Prandtl number, and Euclidean norm of the mixture fraction gradient, *viz.*, $\boldsymbol{x} = [\widetilde{Z}, \widetilde{C}, \overline{\rho}, \widetilde{T}, Pr_{\Delta}, \|\nabla\widetilde{Z}\|_2]$.
4. Train, validate, and test the classification algorithm. In this work, a random forest classifier is used for combustion submodel assignment. Details of the algorithm are presented in Section 4.3.

4.1. Label assignment

We present a multiclass classification problem for optimal assignment of three combustion models $\mathcal{Y} = \{\text{IM, FPV, FRC}\}$. In this problem, we consider the FRC model as combustion model of highest fidelity but at the expense of highest computational cost. Hence, regions with local scalar predictions by IM and FPV models that match those of FRC can be considered optimally assigned. Therefore, we assign labels in the training set based on the normalized combustion submodel error ϵ_Q^y of quantities of interest $\alpha \in Q$ between FRC and the models of lower fidelity [19]:

$$\epsilon_Q^y = \sum_{\alpha \in Q} w_\alpha \frac{|\alpha^{\text{FRC}} - \alpha^y|}{\|\alpha^{\text{FRC}}\|_\infty} \quad \text{with} \quad y \in \{\text{FPV, IM}\}\,, \tag{2}$$

where the error for considering N quantities-of interest is a weighted linear combination of each individual submodel error. The weights for each QoI w_α is subject to the following constraints: $\sum_{\alpha \in Q}^{N} w_\alpha = 1$ and $w_\alpha \geq 0$. In this study, the use of temperature and mass fractions of CO and OH as QoIs. In the combined use of both temperature and CO mass fraction, $Q = \{\widetilde{T}, \widetilde{Y}_{\text{CO}}\}$, both QoIs are equally weighted: $w_T = 0.5$ and $w_{\text{CO}} = 0.5$. Similarly for the combined use of three QoIs $Q = \{\widetilde{T}, \widetilde{Y}_{\text{CO}}, \widetilde{Y}_{\text{OH}}\}$, all QoIs are equally weighted: $w_T = 0.33$, $w_{\text{CO}} = 0.33$, and $w_{\text{OH}} = 0.33$. Temperature $\widetilde{T}$ is chosen as a proxy to describe the combustion efficiency and engine performance. The CO mass fraction $\widetilde{Y}_{\text{CO}}$ is chosen to challenge the deficiencies of tabulation methods in capturing intermediate species [20]. OH mass fraction $\widetilde{Y}_{\text{OH}}$ is selected since radical formation is essential in combustion phenomena.

FRC data is used to reconstruct FPV and IM quantities of interest $\alpha \in Q$ by interpolating the generated flamelet tables using reconstructed values of mixture fraction and progress variable:

$$\alpha^y \approx \alpha^y_{\text{table}}(\widetilde{Z}_{\text{FRC}}, \widetilde{C}_{\text{FRC}}) \quad \text{where} \quad y \in \{\text{FPV, IM}\}\,. \tag{3}$$

The mixture fraction is computed using Bilger's definition [62], while the progress variable is computed using the sum of major combustion products, as described in Section 2.2. We must note that since α^y is reconstructed from FRC data, the resulting error metric ϵ_Q^y is an approximation of the true errors between FRC and tabulated chemistry. However, the use of this error metric is well-justified since Bilger's mixture fraction and the sum of major combustion products are robust quantities for bridging FRC and tabulated methods. Labels are assigned programmatically as demonstrated in Algorithm 1. In this algorithm, a model of higher fidelity is assigned when the QoI submodel error ϵ_Q^y exceeds a user-defined threshold θ_Q^y, with FRC chosen when all conditions for selecting FPV and IM are not met. While θ_Q^{FPV} and θ_Q^{IM} can be assigned distinct values, throughout this study we will explore cases that use the same threshold for both IM and FPV, *viz.*, $\theta_Q^{\text{IM}} = \theta_Q^{\text{FPV}} = \theta_Q$ for simplicity.

4.2. Feature selection

Adding uninformative features to the learning dataset can reduce accuracy and computational efficiency of learning algorithms [63]. Carrying out appropriate feature selection beforehand can improve the interpretability of the predictions of the trained model. To this end, feature selection can be used for identifying the most descriptive and discriminative features from the raw dataset to use as inputs for our learning algorithms. In this work, we select features from local quantities and group parameters that can characterize the reacting flow, combustion state, and turbulence.

For feature selection, we rely on the Maximal Information-based Non-parametric Exploration (MINE) tools [61] that utilize mutual information between variable pairs to ascertain the strength of relationships between variables based on instantaneous flow-field representations from a monolithic FRC simulation. MINE utilizes the Maximum Information Coefficient (MIC) to ensure (i) generality, where the association between the variables are not limited to a particular form such as linear associations, and (ii) equitability, where the effect of noise on different relationships is similar.

While Pearson's correlation has been utilized to ascertain the strength of relationships between variables in scientific applications, this does not account for any non-linear relationships. This is illustrated in Fig. 3, where Pearson's coefficient, or Pearson r, is compared to MIC for different scatter points. As can be seen in Fig. 3(a), for linear relationships with noise, both coefficients are similar. However, in Figs. 3(b) and (c), non-linear associations between variables are ignored by Pearson's correlation coefficient while MIC is able to account for such complex relationships. Mutual-information-based measures that ensure generality and equitability, like MIC, can be used to compare different features, rank them and select subsets of the most descriptive and discriminative features. Additionally, such mutual information based feature selection is model agnostic and can be used across different machine learning models, as a pre-processing step. In this vein, MIC measure has been utilized for feature selection in prior works with success [64].

Figure 4(a) and (b) show MIC scores relating 16 potential features with IM model error $\epsilon^{\text{IM}}_{\{T,\text{CO}\}}$ and FPV model error $\epsilon^{\text{FPV}}_{\{T,\text{CO}\}}$, respectively. These 16 potential features consist of thermophysical quantities and dimensionless quantities that characterize each cell within the domain. Dimensionless quantities include the local Prandtl number, $Pr_\Delta = \widetilde{\nu}/\widetilde{\alpha}$, comparing the local ratio of viscosity and thermal diffusivity, and the local Reynolds number, $Re_\Delta = \Delta|\widetilde{\boldsymbol{u}}|/\nu$, which is the ratio of inertial forces and viscous force within each cell and Δ denotes the characteristic length of each computational cell. It can be seen that the MIC scores for $\epsilon^{\text{FPV}}_{\{T,\text{CO}\}}$ are much lower than for $\epsilon^{\text{IM}}_{\{T,\text{CO}\}}$. This indicates that it is more challenging to form statistical relationships between features and FPV model errors than for IM model error. This observation is consistent with the intuition that it is much easier to identify failure of the IM models than the shortfall of the FPV model.

In the following, the top five features from both MIC tests are used to construct the feature set consisting of mix-

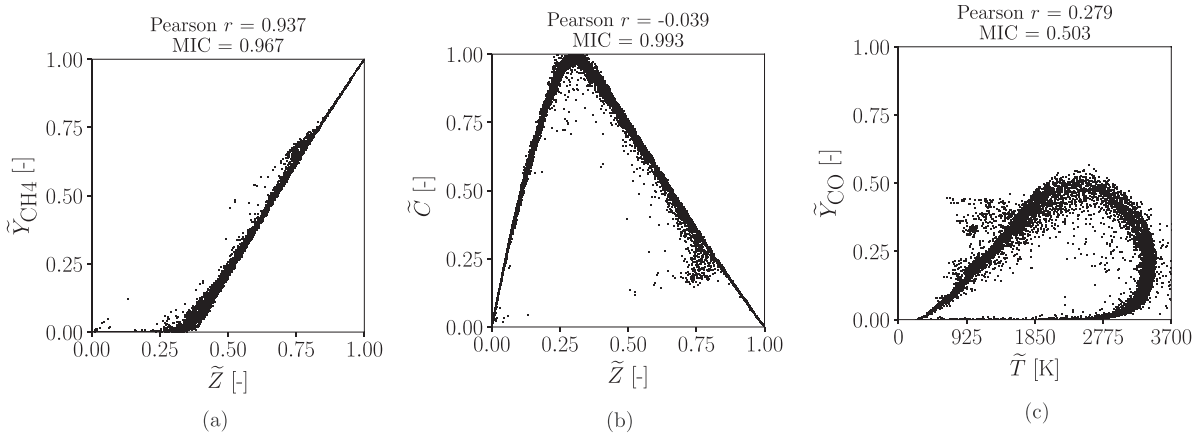


Fig. 3. Comparison between Maximum Information Coefficient (MIC) and Pearson's Correlation Coefficient (Pearson r) for (a) near-linear scatter points, and (b,c) non-linear scatter points.

ture fraction, progress variable, density, local Prandtl number, and Euclidean norm of the mixture fraction gradient: $\boldsymbol{x} = [\widetilde{Z}, \widetilde{C}, \overline{\rho}, \widetilde{T}, Pr_{\Delta}, \|\nabla\widetilde{Z}\|_2]^T$. The inclusion of Pr_{Δ} in the feature set is unexpected since Pr_{Δ} is approximately constant and has weak temperature dependence. However, given that Pr_{Δ} is slightly higher in fuel and oxidizer when compared to combustion products, small variations within flow field prove useful for the random forests. We note that the data-driven framework in this study presently restricts the construction of feature and label sets to local quantities for simplicity. More elaborate methods for incorporating spatial and temporal dependencies into present approach, through the use of convolutional neural networks [32,35], should be subject to further study.

4.3. Random forest classifier

Sections 4.1 and 4.2 detailed the procedures applied in this study for preprocessing the monolithic FRC LES data for training. During training, the classification algorithm learns a function $f : \boldsymbol{x} \mapsto y$ that associates the input vector $\boldsymbol{x} \in \mathcal{X}$, with the corresponding response $y \in \mathcal{Y}$. After training, the learning algorithm can be used to predict the optimal combustion submodel when given new sets of input vectors $\boldsymbol{x} \in \mathcal{X}$. These steps are summarized in Fig. 5.

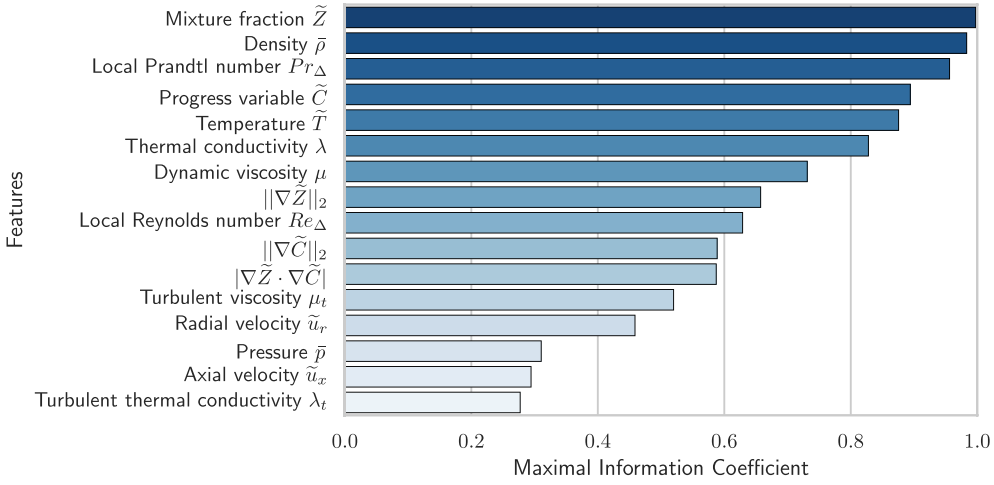
In this study, we employ the random forest as our classification algorithm. Random forests [65] consist of an ensemble of decorrelated Classification And Regression Trees (CARTs) [66]. CARTs are a machine learning approach for formulating prediction models from data by recursively partitioning the inputted feature space, and fitting a simple prediction within each final partition. As a result, the partitioning can be represented graphically as a decision tree. Such decision trees are a graph algorithm, where each node represents a selected feature or attribute, each edge represents a decision based on the properties of this feature, and the leaf nodes represent a final outcome or classification. Decision trees are non-parametric and can model arbitrarily complex relations without any *a priori* assumptions.

In a machine learning algorithm, the expected generalization error is a key characteristic, measuring the accuracy in making predictions for previously unseen data. This error can be decomposed into bias, variance and noise. The bias in the predictions is the deviation from the true value of the expectation (or mean) of the model predictions. In this context, the variance is the variability in the predictions of models. Noise is the inherent stochastic noise in the data. Decision trees are prone to overfitting. In terms of the bias-variance decomposition, these overfitted models possess low bias but high variance. Ensemble methods offer a simple amelioration by introducing random perturbations in the training procedure to produce several randomized models from the same data, and then combining the predictions of the individual models to form the ensemble prediction. The decorrelated nature of each constituent model reduces the variance of predictions while retaining the low bias.

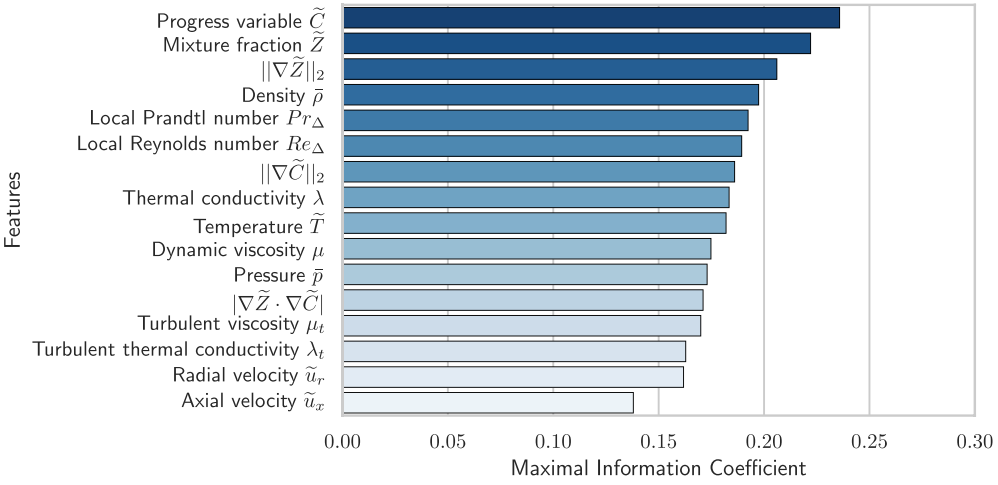
Random forests are an ensemble method, using ensembles of trees to create a forest. Here, the ensemble model is a collection of Classification And Regression Trees. The final prediction of this ensemble model is via a majority vote of trained individual trees. The key motivation is to create an ensemble model that has lower variance than the individual trees, while maintaining the low bias. It can be shown that the variance of the ensemble model is directly proportional to the correlation between individual models in the ensemble [65]. Thus, the more uncorrelated our individual models are, the lower the variance of the ensemble model. To inject this decorrelation between the individual decision trees in the Random Forest, two concepts are utilized, explicitly:

- Bagging [65]: Bagging (or Bootstrap aggregating) is an approach to create different machine learning models from the same data set. In the first step, we can generate multiple new training datasets from the original by sampling from it, uniformly and with replacement (Bootstrapping). Each of these sampled datasets can be used to train a machine learning model. The final prediction is chosen by aggregating the predictions of these individual models (aggregating). In Random forests, each individual tree gets such a bootstrap sample of the original training dataset to learn from. This ensures that every tree has to train on a different dataset and, thus imparts a level of decorrelation to the individual trained tree based models in the ensemble.
- Random subsampling over features [67]: During their training, CARTs are grown by learning splits (or partitions) at each node. Herein, the trees have to determine the best split over the entire set of features to partition the solution space. In random forests, only a small randomized subset of the total set of features is assigned to each tree during training. This introduces additional decorrelation between the trees in the ensemble.

Using Bagging in conjunction with random subsampling over the features, introduces adequate decorrelation over the individual trees in the ensemble to reduce the variance, while maintaining the low bias. In prior investigations, it has been observed that random forests outperform many other algorithms in classification over scalar inputs from structured datasets [68,69].



(a) MIC relating features with IM model error $\epsilon^{\mathrm{IM}}_{\{T,\mathrm{CO}\}}$.



(b) MIC relating features with FPV model error $\epsilon^{\mathrm{FPV}}_{\{T,\mathrm{CO}\}}$.

Fig. 4. Maximal information coefficient score for features and model error.

In the present investigation, the random forest classifier from the OpenCV library [70] is used. Classification cost scales with the number of trees, tree depth and the number of training points [66]. Hence, a random forest consisting of twenty decision trees, and maximum depth of ten nodes is employed. Additionally, 1×10^4 training points have been randomly sampled from a single LES snapshot consisting of 2×10^5 cells. A similar approach is used in other supervised learning problems [25]. We must note that the flow in the present configuration is statistically stationary, and thus training data from a single snapshot was found to be sufficient for representing the thermophysical behavior of the combustor. The number of trees, tree depth, and the number of training points are determined *a priori* by ensuring that the classification performance remains unchanged on a validation set. Training is performed once *a priori*, and requires 530 ms of walltime with 1 CPU. In *a posteriori* simulations, random forest evaluations for 2×10^5 cells at each timestep require 1 ms of wall time with 600 CPUs.

5. Results

This section assesses the random forest classifier as a method for combustion submodel assignment in data-assisted simulations. *A priori* assessment is performed first to investigate the behavior of random forests when targeting different QoIs. This is followed by an *a posteriori* assessment to study improvements in target QoIs and other quantities that result from the use of random forests in transient data-assisted simulations. Table 1 summarizes the eight cases, with different QoIs and combustion submodel error threshold values θ_{Q}, explored in both *a priori* and *a posteriori* assessment.

5.1. A priori *assessment*

A priori assessment involves using the random forest classifier to assign suitable combustion submodels in a test dataset that is created from a monolithic FRC simulation at an unseen timestep. Temperature and CO and OH mass fraction $\alpha \in \{\widetilde{T}, \widetilde{Y}_{\mathrm{CO}}, \widetilde{Y}_{\mathrm{OH}}\}$ in the test set is then used as QoI for reconstructing the true response, through the procedure described in Section 4.1, for comparison with random forest predictions. Figure 6 shows the use of this labeling approach on the training data in $\widetilde{Z} - \widetilde{C}$ composition space for $\theta_{\{T,\mathrm{CO}\}} = 0.02$ and $\theta_{\{T,\mathrm{CO}\}} = 0.05$, respectively. In both cases, IM is shown to be assigned at points where $\widetilde{C} \approx 0$, FPV is assigned mostly to conditions near the equilibrium composition. The submodel assignment reverts back to FRC in regions dominated by non-equilibrium effects and heat-losses that are not captured by the adiabatic steady-state flamelet formulation. Employing $\theta_{\{T,\mathrm{CO}\}} = 0.02$ is seen to be more stringent than em-

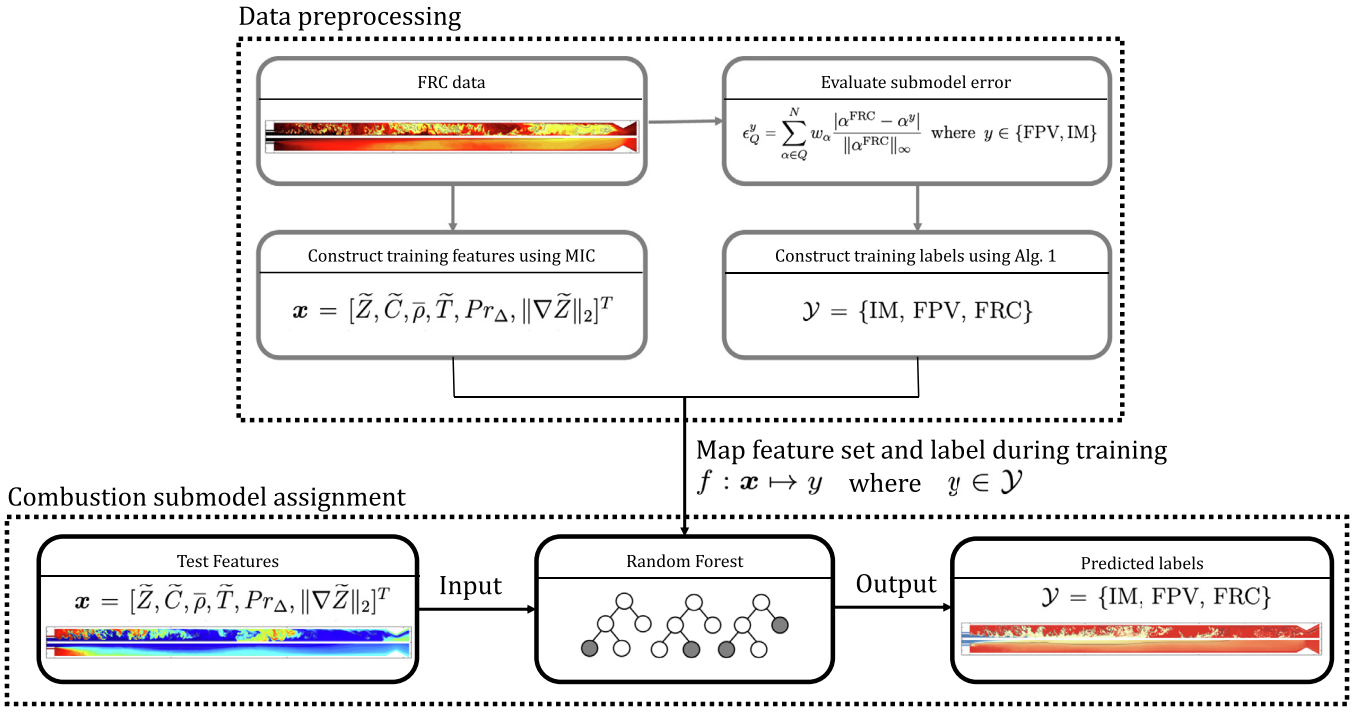
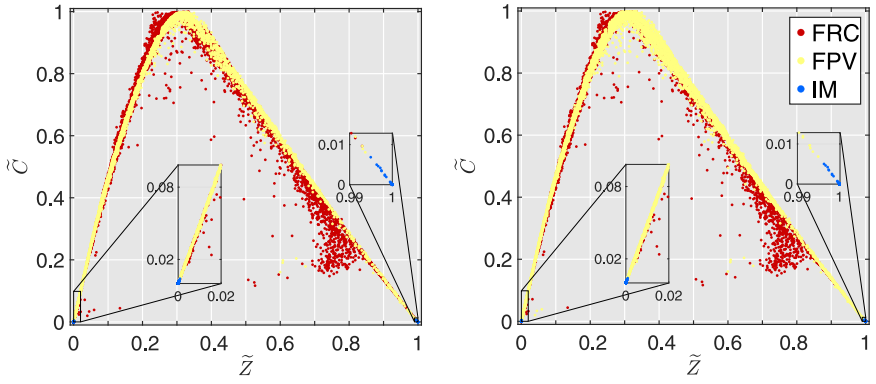


Fig. 5. Application of random forest classifier for combustion submodel assignment of a single element GOX/GCH4 rocket combustor.

Table 1
Cases investigated in the present study.

Case	θ_T=0.05	θ_T=0.02	θ_{CO}=0.05	θ_{CO}=0.02	$\theta_{\{T,\mathrm{CO}\}}$=0.05	$\theta_{\{T,\mathrm{CO}\}}$=0.02	$\theta_{\{T,\mathrm{CO,OH}\}}$=0.05	$\theta_{\{T,\mathrm{CO,OH}\}}$=0.02
QoI, Q	$\widetilde{T}$	$\widetilde{T}$	$\widetilde{Y}_{\mathrm{CO}}$	$\widetilde{Y}_{\mathrm{CO}}$	$\{\widetilde{T}, \widetilde{Y}_{\mathrm{CO}}\}$	$\{\widetilde{T}, \widetilde{Y}_{\mathrm{CO}}\}$	$\{\widetilde{T}, \widetilde{Y}_{\mathrm{CO}}, \widetilde{Y}_{\mathrm{OH}}\}$	$\{\widetilde{T}, \widetilde{Y}_{\mathrm{CO}}, \widetilde{Y}_{\mathrm{OH}}\}$
Model threshold, θ_Q	0.05	0.02	0.05	0.02	0.05	0.02	0.05	0.02
Assessment	*A priori*	*A priori*	*A priori*	*A priori*	*A priori, A posteriori*	*A priori, A posteriori*	*A priori*	*A priori*



(a) $\theta_{\{T,\mathrm{CO}\}} = 0.02$ (b) $\theta_{\{T,\mathrm{CO}\}} = 0.05$

Fig. 6. Training data for two different combustion submodel error thresholds $\theta_{\{T,\mathrm{CO}\}}$.

ploying $\theta_{\{T,\mathrm{CO}\}} = 0.05$, with a 0.18 greater fraction of scatter data on the stable branch assigned as FRC, especially for fuel-rich mixtures. It should be noted that while most out-of-flamelet regions would be assigned FRC, some regions with low reactivity and far from stoichiometry (eg. $\widetilde{Z} = 0.7$) generate smaller errors which are then be assigned FPV.

Figure 7 demonstrates the *a priori* combustion submodel assignment on an unseen FRC-simulation snapshot using the six different random forest cases summarized in Table 1. For all six cases, IM is assigned at the injector and the oxidizer core. In general, FRC is assigned at the near-wall and fuel-rich regions within the combustor where intermediate reactions are not captured well by tabulated chemistry submodels. Using temperature as QoI and a model threshold of $\theta_T = 0.05$ results in an IM assignment of 5% of the domain, 28% FRC assignment, with the rest being described by the FPV model. Constraining the temperature model threshold $\theta_T = 0.02$ results in FRC assignment in 62% of the domain, with IM assignment remaining unchanged.

Using $\widetilde{Y}_{\mathrm{CO}}$ as QoI and a model threshold of $\theta_{\mathrm{CO}} = 0.05$ results in greater (18% of the domain) IM assignment, since the CO mass fraction in most of the oxidizer core is close to zero. FRC is assigned to 34% of the domain. Reducing the CO model threshold $\theta_{\mathrm{CO}} = 0.02$ results in 47% FRC assignment, with IM assignment unchanged. Finally the combined use of both temperature and CO

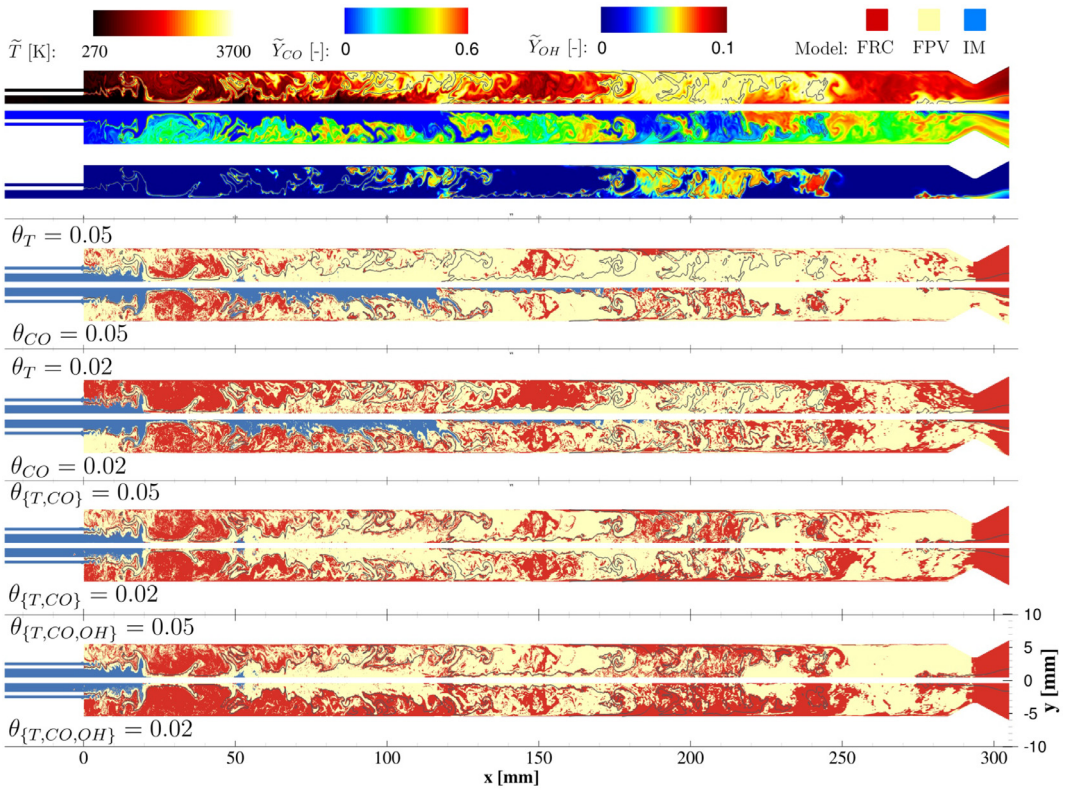


Fig. 7. *A priori* analysis, comparing combustion model assignments. Instantaneous temperature, and mass fractions of CO and OH of the test set are also presented; stoichiometric isocontour with $\widetilde{Z}_{st} = 0.2$ is shown in black.

Table 2
A priori analysis of classifier, summarizing submodel assignment and assignment accuracy.

Case	θ_T=0.05	θ_T=0.02	θ_{CO}=0.05	θ_{CO}=0.02	$\theta_{\{T,CO\}}$=0.05	$\theta_{\{T,CO\}}$=0.02	$\theta_{\{T,CO,OH\}}$=0.05	$\theta_{\{T,CO,OH\}}$=0.02
IM:FPV:FRC	5:67:28	5:33:62	18:48:34	18:35:47	6:63:31	6:42:52	6:57:37	6:24:70
True Classification	0.774	0.725	0.756	0.715	0.753	0.734	0.709	0.691

mass fraction as QoI, $Q = \{\widetilde{T}, \widetilde{Y}_{CO}\}$, results in submodel assignment with combined characteristics of employing each individual QoI. $\theta_{\{T,CO\}} = 0.05$ results in 31% FRC assignment within the domain, while $\theta_{\{T,CO\}} = 0.02$ results in 52% FRC assignment. Adding OH mass fraction to the QoI set $Q = \{\widetilde{T}, \widetilde{Y}_{CO}, \widetilde{Y}_{OH}\}$ increases the FRC assignment to 37% and 70% for thresholds $\theta_{\{T,CO,OH\}} = 0.05$ and $\theta_{\{T,CO,OH\}} = 0.02$ respectively. Results demonstrate that reducing model threshold θ_Q and increasing the number of QoIs increases submodel assignment of FRC. The submodel assignments for each case are summarized in Table 2.

Table 2 also summarizes the true classification of random forests for the eight different cases. Here, true classification is defined as the percentage of classifier assignments that correctly match the true output responses evaluated directly from simulation data. The true classification fraction range from approximately 0.7 to 0.8, which is comparable to the use of random forests on another classification problem in a flow physics context [23]. Higher true classification can be achieved through the use of complex deep learning classifiers, which requires (i) more elaborate efforts than the random forests in hyperparameter tuning and (ii) much larger datasets for good performance, and should be subject to further study.

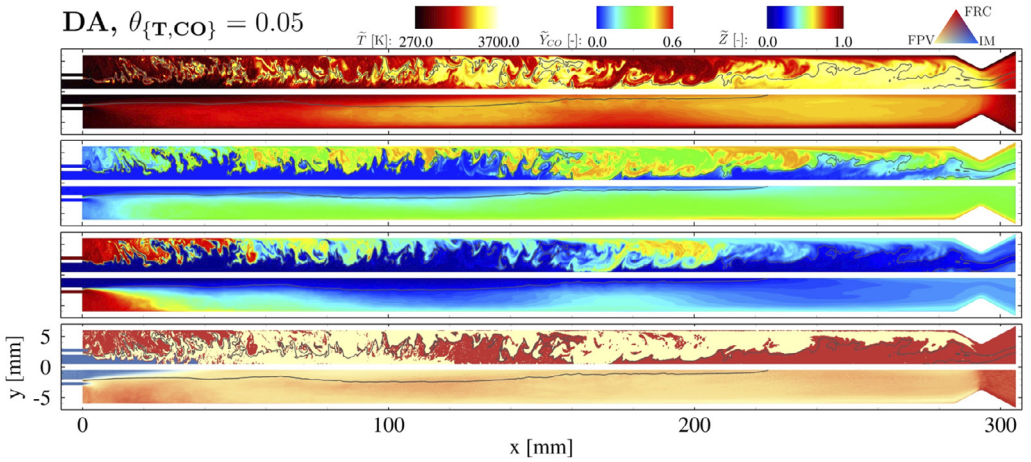
From Fig. 7, we observe that model assignment in all six cases is not spatially smooth, and that model assignment appears speckled. This is because the smoothness of classification boundaries formed within the 6-dimensional feature space is not translated when transformed to physical space. This is a common issue in classification problems involving spatial data, such as in medical imaging or image processing. Two strategies can be employed to improve spatial smoothness in classification problems [19,71]: (i) applying the classification techniques to a neighborhood of cells, or (ii) applying a spatial filter on the predicted labels and discretizing the filtered labels. In the *a posteriori* assessment in Section 5.2, we apply the latter strategy since it is better suited with the current framework that uses local quantities as QoIs and features.

These results demonstrate that the present data-assisted framework enables a fully adjustable level of simulation fidelity through the use of varying submodel error threshold values. Random forests are demonstrated to be a reasonably accurate and simple approach for the combustion submodel assignment problems.

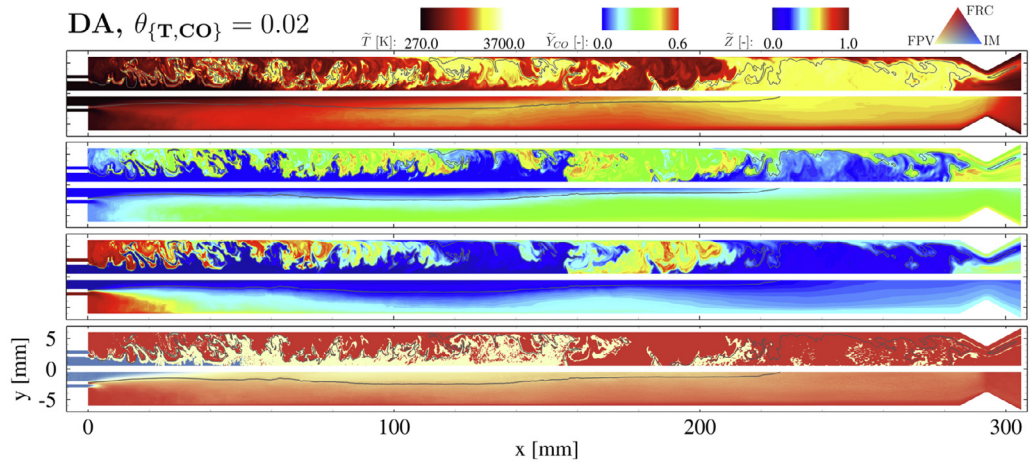
5.2. A posteriori *assessment: data-assisted LES*

Data-assisted (DA) simulations using two different model thresholds, $\theta_{\{T,CO\}} = 0.05$ and $\theta_{\{T,CO\}} = 0.02$ are performed by employing random forest classifiers in-flight during simulation runtime. The discussion from this section also includes comparisons with monolithic FRC and FPV simulations.

Figure 8(a) shows that employing model threshold $\theta_{\{T,CO\}} = 0.05$ on the DA simulation results in temperature predictions that are in good agreement with the monolithic FRC simulation, shown in Fig. 2(a). However, time-averaged results show that a thin layer of CO develops at the chamber wall at 170 mm. Additionally, a



(a) *A posteriori* DA LES with $\theta_{\{T,CO\}} = 0.05$.



(b) *A posteriori* DA LES with $\theta_{\{T,CO\}} = 0.02$.

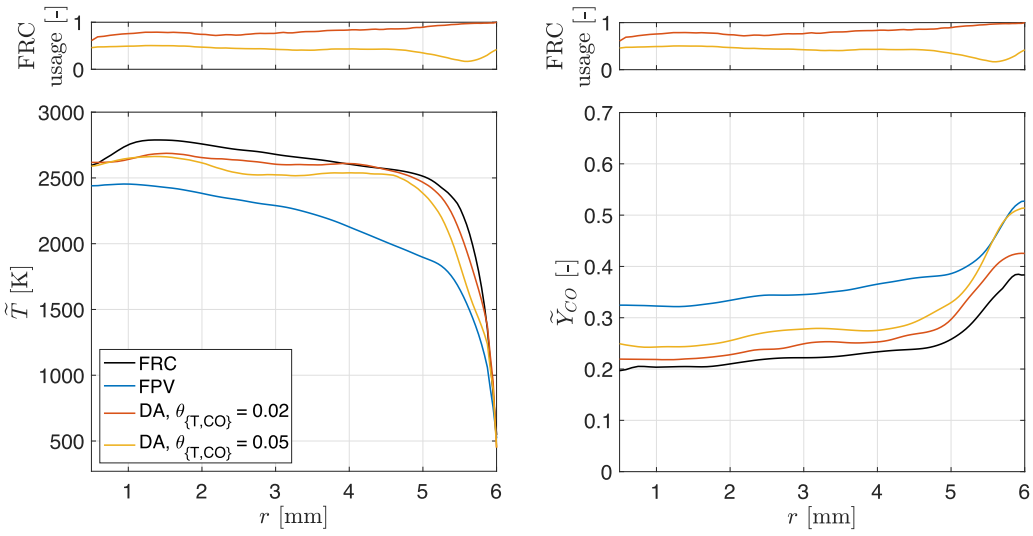
Fig. 8. Temperature, CO mass fraction, and mixture fraction fields (from top to bottom) from *a posteriori* DA LES for (a) $\theta_{\{T,CO\}} = 0.05$ and (b) $\theta_{\{T,CO\}} = 0.02$. Upper half: instantaneous fields, bottom half: time-averaged fields; stoichiometric isocontour with $\tilde{Z}_{st} = 0.2$ is shown in black.

thicker thermal boundary layer is also observed when compared to monolithic FRC simulations. Nonetheless, both species and thermal boundary layers are thinner than the monolithic FPV simulations that were presented in Fig. 2(b). Averaged FRC utilization with $\theta_{\{T,CO\}} = 0.05$ is at 34% of the domain with IM-utilization at 4%. In addition, a thin intermittent area close to the wall is also assigned FRC. This indicates that the random forest recognizes the importance of wall effects on CO and temperature but that the user-defined model error threshold $\theta_{\{T,CO\}} = 0.05$ is too large.

Figure 4(b) shows that tightening the model threshold $\theta_{\{T,CO\}} = 0.02$ results in temperature, CO, and mixture fraction fields that agree with the monolithic FRC simulation, shown in Fig. 2(a). Model assignment using this threshold results in 60% FRC utilization. Before $x = 150$ mm FRC is assigned to all fuel-rich and near-wall regions. For $x > 150$ mm, FRC is assigned to most of the domain where incomplete combustion products and intermediate species are dominant.

Figure 9 shows comparisons of radial profiles of time-averaged temperature and CO mass fraction at an axial distance of 250 mm. Effects of wall-heat loss on the monolithic FPV simulation is seen to reduce the overall temperature and thicken the thermal boundary layer, which in turn results in greater CO mass fraction. Using a model threshold of $\theta_{\{T,CO\}} = 0.05$, DA-predictions for temperature and CO mass fraction profiles away from the wall are in good agreement with monolithic FRC simulations, and averaged FRC submodel utilization ranges between 16% and 38%. At $r = 5$ mm, the random forest is able to recognize when the absolute error between temperature diminishes and thus assigns less FRC accordingly, which results in greater temperature and CO mass fraction deviation from monolithic FRC simulations. After $r = 5.7$ mm, the random forest begins to recognize the importance of near-wall effects and assigns more FRC. However, this FRC utilization is still insufficient for recreating monolithic FRC simulations. Further constraining the DA-simulation threshold to $\theta_{\{T,CO\}} = 0.02$ improves the agreement with monolithic FRC-simulations. However, small errors can still be seen even with high FRC submodel utilization that ranges from 61% to 90%.

Results from Fig. 9 show that the present data-assisted modeling approach can generate simulation results that are in agreement with monolithic FRC calculations. However errors observed are greater than the local model error threshold $\theta_{\{T,CO\}}$ used for training the random forests. This is caused by small changes in one state that can result in significant deviations in later states. This effect is illustrated by applying DA combustion modeling with local model error threshold $\theta_{\{T,CO\}} = 0.02$ on CO mass fraction, using a rich methane-air mixture ($Z = 0.55$) in a constant pressure homogeneous reactor at 20 bar and initial temperature of 1800 K, as shown in Fig. 10. In this setup, it is observed that while the random forest correctly assigns the correct model based on local model error at 5800 timesteps, the CO trajectory leads to a total error exceeding the local error threshold of 0.02 as the DA simulation no longer has knowledge of the monolithic FRC CO production



(a) Temperature (b) CO mass fraction

Fig. 9. Comparisons of time-averaged radial profiles of (a) temperature and (b) CO mass fraction between monolithic FRC, monolithic FPV, and data-assisted (DA) simulations at an axial distance $x = 250$ mm. Time-averaged utilization of FRC is included.

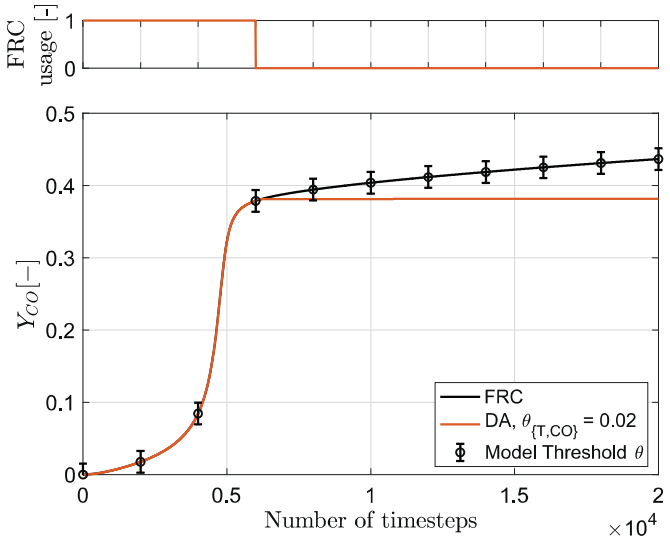


Fig. 10. FRC and DA-assisted calculation of CO mass fraction as a function of time step in a 0D homogeneous reactor.

beyond this timestep and cannot recover to the correct state. However, the benefit of the present approach is that, in the worst-case, errors made do not exceed errors made by the lowest fidelity combustion model employed.

Generating numerical predictions that match experimental wall measurements are challenging for this rocket combustor case, since these quantities are dependent on overall flow and temperature fields in a highly nonlinear system. Studies [59,72] comparing LES and RANS results have reported up to 8% deviation from wall pressure measurements. Wall heat flux predictions are more sensitive to simulation parameters, where deviations up to 75% have been reported in the same studies. While the aim of the present study is not to find simulation results that match the experimental results, LES calculations of wall pressure and wall heat flux are presented with measurements by Perakis and Haidn [57] in Fig. 11 to quantify effects of applying the DA formulation on overall combustor behavior.

Figure 11(a) shows that wall pressure predictions between monolithic FRC agree well with experimental measurements. The DA simulation with $\theta_{\{T,CO\}} = 0.02$ shows a small underprediction, but still possesses reasonable agreement with monolithic FRC. The DA simulation with $\theta_{\{T,CO\}} = 0.05$ shows a greater underprediction. Wall pressure underprediction can be caused by reduced fuel conversion [73]. This is likely the case since higher CO levels in both cases are observed in Fig. 9. Additionally, the monolithic FPV simulation also demonstrates the lowest pressure and highest CO levels.

Figure 11(b) shows that wall heat flux predictions for FRC simulation are in good agreement with experimental data after $x = 120$ mm, but with a steeper heat flux rise. This steep heat flux rise is likely due to the misrepresentation of turbulent mixing in a thin axisymmetric domain, and is also seen in other axisymmetric studies [55,56]. Tightening the model threshold $\theta_{\{T,CO\}}$ results in better convergence with monoolithic FRC calculations. The DA simulation with $\theta_{\{T,CO\}} = 0.02$ is in reasonable agreement with the FRC simulation, while the FPV simulation demonstrates the lowest heat flux due to low overall temperatures from low combustion efficiency.

Figure 12 shows FRC usage and corresponding computational cost (normalized by FRC cost) of the data-assisted simulation as a function of combustion submodel error threshold $\theta_{\{T,CO\}}$ when computed using 600 Intel Xeon (E5-2680v2) processors. Each timestep in the FPV simulation requires 50 ms of wall time to solve, while each timestep in the FRC requires a wall time of 2300 ms. When $\theta_{\{T,CO\}} = 0.50$, the classifier does not assign FRC in the entire domain, resulting in a normalized cost of 8%. This additional cost represents the overhead from the random forest evaluation and the coupling of the three combustion submodels in the same domain. Simulations performed in this study utilized 34% ($\theta_{\{T,CO\}} = 0.05$) and 60% FRC ($\theta_{\{T,CO\}} = 0.02$), which resulted in 70% and 80% of FRC cost, respectively. These results demonstrate that classification algorithms can be utilized in high-fidelity simulations to reduce computational cost. Further reductions of the computational cost is achievable by combining the method proposed in this work with regression techniques [38,39] to reduce the complexity of the finite-rate chemistry representation.

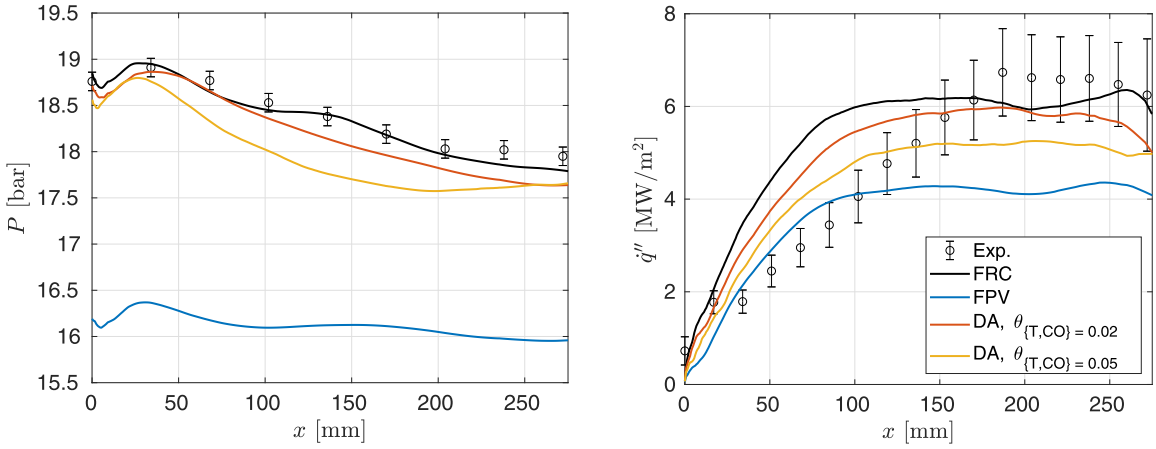


Fig. 11. Comparison of simulation results for (a) wall pressure and (b) wall heat flux calculations with experimental measurements [57].

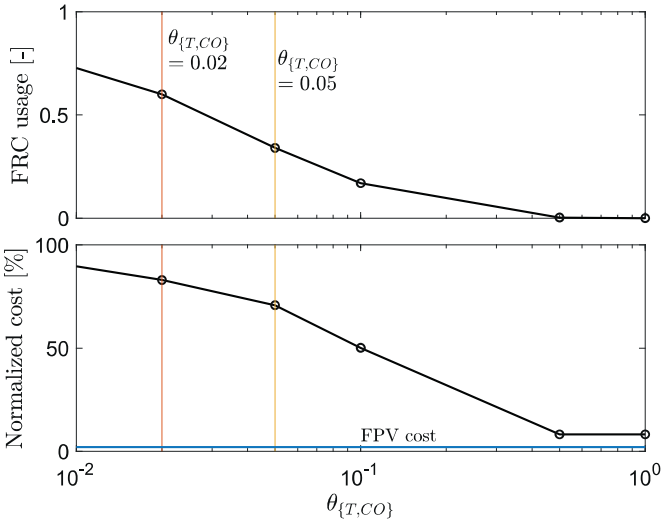


Fig. 12. FRC utilization and normalized computational cost versus combustion submodel error threshold $\theta_{\{T,CO\}}$.

5.3. Generalization

In order to demonstrate the ability of random forests to generalize, additional LES are performed on a modified configuration with three times the inlet mass flow, while keeping all other parameters constant. Figure 13 compares time-averaged temperature and CO mass fraction fields for monolithic FRC, monolithic FPV, and *a posteriori* DA LES ($\theta_{\{T,CO\}} = 0.02$) for this setup. All three LES cases in this modified configuration demonstrate a longer oxygen core than the original configuration (Fig. 2) due to higher flow

Algorithm 1: Assigning labels in the training set.

```
if ε_Q^IM < θ_Q^IM then
    use inert mixing (IM)
else if ε_Q^FPV < θ_Q^FPV then
    use tabulated chemistry (FPV)
else
    use finite-rate chemistry (FRC)
end
```

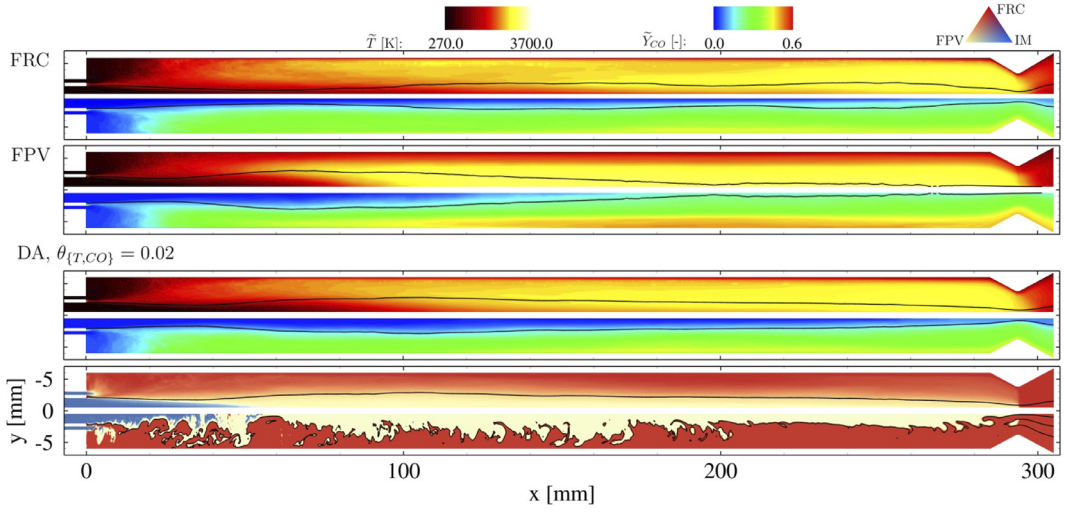


Fig. 13. Comparison of time-averaged temperature and CO mass fraction fields for monolithic FRC, monolithic FPV, and *a posteriori* DA LES ($\theta_{\{T,CO\}} = 0.02$) on a configuration with three times the inlet mass flow rate. Time-averaged and instantaneous model assignment for DA LES is shown at the bottom. Stoichiometric isocontour with $\tilde{Z}_{st} = 0.2$ is shown in black.

velocity, indicating less complete combustion. When compared to FRC, FPV overpredicts the thickness of the thermal boundary layer and CO formation. DA LES with model threshold ($\theta_{\{T,\mathrm{CO}\}} = 0.02$) predicts temperature and CO flow fields in good agreement with monolithic FRC calculations. Random forest assigns FPV to the lean side of the flame, while assigning FRC to the rich side. This is also seen in the DA case of the original configuration in Fig. 8 from 0 to 150 mm, where major combustion products have not fully formed. Model assignment using this threshold results in 51% FRC and 6% IM utilization, resulting in 77% of the FRC cost.

Results from this modified configuration demonstrate that the present data-assisted approach can be applied to different configurations as long as the training data can represent the underlying thermo-physical behavior. We note that all simulations and training data from the present study employ the same mesh. Since the random forest classifies well in this modified configuration, this method should still be effective for different mesh resolutions as long as the flow can be represented by local points of the training data. The generalizability of this method improves with increasing availability of representative data.

6. Conclusions

This study introduced a data-assisted modeling approach, employing random forest classifiers, as a method for dynamic and local combustion model assignment in reacting flow simulations. *A priori* assessment was conducted on the random forests, which were fed with six input features based on local thermofluid properties, to evaluate the behavior of the classifiers during submodel assignment when targeting different QoIs. Random forests were shown to assign three different candidate combustion models – finite-rate chemistry (FRC), flamelet progress variable (FPV) approach, and inert mixing (IM) – based on predefined QoIs with fraction of true classification ranging from approximately 0.70 to 0.80.

Two cases of *a posteriori* simulations using random forest classifiers for combustion submodel assignment during simulation runtime, were performed. Time-averaged results of temperature and CO mass fraction demonstrated that the data-assisted simulation produced species and temperature profiles in better agreement with monolithic FRC than monolithic FPV calculations. The use of the random forest with submodel error threshold of $\theta_{\{T,\mathrm{CO}\}} = 0.02$ results in significant improvements from monolithic FPV simulations in all quantities at a 20% lower cost than monolithic FRC calculations. An additional DA LES ($\theta_{\{T,\mathrm{CO}\}} = 0.02$), performed on a modified configuration with three times the inlet mass flow rate, demonstrated that the present approach can be applied to different configurations as long as the training data can represent the relevant thermo-physical behavior.

Results from *a priori* and *a posteriori* assessments demonstrated that the present data-assisted framework is adjustable and effective for the purpose of combustion model assignment, so long as high-quality data is available. While this method avoids the challenging task of constructing a mathematical model-compliance indicator [19], the present approach is not Pareto-optimized since only local submodel errors were utilized for training. Thus, additional concepts from the Pareto-efficient combustion framework can supplement the present data-assisted LES framework. Additionally, the exploration of other cost-efficient and accurate classification algorithms could improve the classification accuracy of the present data-assisted approach. In particular, ANNs with deep learning architectures have shown high accuracy in numerous classification problems. Other opportunities for extending this work include (i) the extension of the current framework to bridge local submodel error with non-local errors, (ii) the addition of non-local quantities in the feature and label set, and (iii) the consideration of a more extensive candidate combustion submodel set.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors gratefully acknowledge financial support from the Air Force Office of Scientific Research under Award No. FA9300-19-P-1502, NASA with award No. 80NSSC18C0207, and Stanford University Harold and Marcia Wagner Engineering Fellowship. Resources supporting this work are provided by the High-End Computing (HEC) Program at NASA Ames Research Center.

References

[1] T. Lu, C.K. Law, Toward accommodating realistic fuel chemistry in large-scale computations, Prog. Energy Combust. Sci. 35 (2) (2009) 192–215, doi:10.1016/j.pecs.2008.10.002.
[2] T. Turányi, Reduction of large reaction mechanisms, New J. Chem. 14 (1990) 795–803.
[3] L.E. Whitehouse, A.S. Tomlin, M.J. Pilling, Systematic reduction of complex tropospheric chemical mechanisms, part I: sensitivity and time-scale analyses, Atm. Chem. Phys. 4 (7) (2004) 2025–2056, doi:10.5194/acp-4-2025-2004.
[4] G. Li, H. Rabitz, A general analysis of exact lumping in chemical kinetics, Chem. Eng. Sci. 44 (6) (1989) 1413–1430, doi:10.1016/0009-2509(89)85014-6.
[5] R. Fournet, V. Warth, P.A. Glaude, F. Battin-Leclerc, G. Scacchi, G.M. Côme, Automatic reduction of detailed mechanisms of combustion of alkanes by chemical lumping, Int. J. Chem. Kin. 32 (1) (2000) 36–51, doi:10.1002/(SICI)1097-4601(2000)32:1.
[6] T. Lu, Y. Ju, C.K. Law, Complex CSP for chemistry reduction and analysis, Combust. Flame 126 (1) (2001) 1445–1455, doi:10.1016/S0010-2180(01)00252-8.
[7] U. Maas, S. Pope, Simplifying chemical kinetics: intrinsic low-dimensional manifolds in composition space, Combust. Flame 88 (3) (1992) 239–264, doi:10.1016/0010-2180(92)90034-M.
[8] D.A. Schwer, P. Lu, W.H. Green Jr., V. Semião, A consistent-splitting approach to computing stiff steady-state reacting flows with adaptive chemistry, Combust. Theor. Model. 7 (2) (2003) 383–399, doi:10.1088/1364-7830/7/2/310.
[9] M.A. Singer, S.B. Pope, Exploiting ISAT to solve the reaction–diffusion equation, Combust. Theor. Model. 8 (2) (2004) 361–383, doi:10.1088/1364-7830/8/2/009.
[10] S.B. Pope, Small scales, many species and the manifold challenges of turbulent combustion, Proc. Combust. Inst. 34 (2013) 1–31.
[11] S.P. Burke, T.E.W. Schumann, Diffusion flames, Ind. Eng. Chem. 20 (10) (1928) 998–1004, doi:10.1021/ie50226a005.
[12] O. Gicquel, N. Darabiha, D. Thévenin, Laminar premixed hydrogen/air counterflow flame simulations using flame prolongation of ILDM with differential diffusion, Proc. Combust. Inst. 28 (2) (2000) 1901–1908.
[13] J. van Oijen, L. de Goey, Modelling of premixed laminar flames using flamelet-generated manifolds, Combust. Sci. Technol. 161 (1) (2000) 113–137, doi:10.1080/00102200008935814.
[14] C.D. Pierce, P. Moin, Progress-variable approach for large-eddy simulation of non-premixed turbulent combustion, J. Fluid Mech. 504 (2004) 73–97, doi:10.1017/S0022112004008213.
[15] M. Ihme, C.M. Cha, H. Pitsch, Prediction of local extinction and re-ignition effects in non-premixed turbulent combustion using a flamelet/progress variable approach, Proc. Combust. Inst. 30 (2005) 793–800.
[16] Y. Liang, S.B. Pope, P. Pepiot, A pre-partitioned adaptive chemistry methodology for the efficient implementation of combustion chemistry in particle PDF methods, Combust. Flame 162 (9) (2015) 3236–3253, doi:10.1016/j.combustflame.2015.05.012.
[17] W. Xie, Z. Lu, Z. Ren, L. Hou, Dynamic adaptive chemistry via species time-scale and Jacobian-aided rate analysis, Proc. Combust. Inst. 36 (1) (2017) 645–653.
[18] S. Yang, R. Ranjan, V. Yang, S. Menon, W. Sun, Parallel on-the-fly adaptive kinetics in direct numerical simulation of turbulent premixed flame, Proc. Combust. Inst. 36 (2) (2017) 2025–2032, doi:10.1016/j.proci.2016.07.021.
[19] H. Wu, Y.C. See, Q. Wang, M. Ihme, A Pareto-efficient combustion framework with submodel assignment for predicting complex flame configurations, Combust. Flame 162 (2015) 4208–4230.
[20] H. Wu, P.C. Ma, T. Jaravel, M. Ihme, Pareto-efficient combustion modeling for improved CO-emission prediction in LES of a piloted turbulent dimethyl ether jet flame, Proc. Combust. Inst. 37 (2019) 2267–2276, doi:10.1016/j.proci.2018.08.010.
[21] Q. Douasbin, M. Ihme, C. Arndt, Pareto-efficient combustion framework for predicting transient ignition dynamics in turbulent flames: application to a pulsed jet-in-hot-coflow flame, Combust. Flame 223 (2021) 153–165.

[22] V. Dhar, Data science and prediction, Commun. ACM 56 (12) (2013) 64–73, doi:10.1145/2500499.

[23] J. Ling, J. Templeton, Evaluation of machine learning algorithms for prediction of regions of high Reynolds-averaged Navier-Stokes uncertainty, Phys. Fluids 27 (8) (2015) 085103, doi:10.1063/1.4927765.

[24] J.X. Wang, H. Xiao, Data-driven CFD modeling of turbulent flows through complex structures, Int. J. Heat Fluid Flow 62 (2016) 138–149, doi:10.1016/j.ijheatfluidflow.2016.11.007.

[25] J.-L. Wu, H. Xiao, E. Paterson, Physics-informed machine learning approach for augmenting turbulence models: a comprehensive framework, Phys. Rev. Fluids 3 (2018) 74602, doi:10.1103/PhysRevFluids.3.074602.

[26] K. Duraisamy, G. Iaccarino, H. Xiao, Turbulence modeling in the age of data, Annu. Rev. Fluid Mech. 51 (1) (2019) 357–377, doi:10.1146/annurev-fluid-010518-040547.

[27] F.C. Christo, A.R. Masri, E.M. Nebot, S.B. Pope, An integrated PDF/neural network approach for simulating turbulent reacting systems, Proc. Combust. Inst. 26 (1996) 43–48.

[28] J.A. Blasco, N. Fueyo, J.C. Larroya, C. Dopazo, J.Y. Chen, Single-step time-integrator of a methane-air chemical system using artificial neural networks, Comput. Chem. Eng. 23 (9) (1999) 1127–1133.

[29] M. Ihme, C. Schmitt, H. Pitsch, Optimal artificial neural networks and tabulation methods for chemistry representation in LES of a bluff-body swirl-stabilized flame, Proc. Combust. Inst. 32 (2009) 1527–1535, doi:10.1016/j.proci.2008.06.100.

[30] A. Kempf, F. Flemming, J. Janicka, Investigation of lengthscales, scalar dissipation, and flame orientation in a piloted diffusion flame by LES, Proc. Combust. Inst. 30 (2005) 557–565.

[31] B.A. Sen, S. Menon, Linear eddy mixing based tabulation and artificial neural networks for large eddy simulations of turbulent flames, Combust. Flame 157 (1) (2010) 62–74.

[32] C.J. Lapeyre, A. Misdariis, N. Cazard, D. Veynante, T. Poinsot, Training convolutional neural networks to estimate turbulent sub-grid scale reaction rates, Combust. Flame 203 (2019) 255–264, doi:10.1016/j.combustflame.2019.02.019.

[33] R. Ranade, T. Echekki, A framework for data-based turbulent combustion closure: a posteriori validation, Combust. Flame 210 (2019) 279–291, doi:10.1016/j.combustflame.2019.08.039.

[34] M.T. Henry de Frahan, S. Yellapantula, R. King, M.S. Day, R.W. Grout, Deep learning for presumed probability density function models, Combust. Flame 208 (2019) 436–450.

[35] A. Seltz, P. Domingo, L. Vervisch, Z.M. Nikolaou, Direct mapping from LES resolved scales to filtered-flame generated manifolds using convolutional neural networks, Combust. Flame 210 (2019) 71–82, doi:10.1016/j.combustflame.2019.08.014.

[36] S. Yao, B. Wang, A. Kronenburg, O.T. Stein, Conditional scalar dissipation rate modeling for turbulent spray flames using artificial neural networks, Proc. Combust. Inst. (2020), doi:10.1016/j.proci.2020.06.135. In press

[37] J.A. Blasco, N. Fueyo, C. Dopazo, J. Ballester, Modelling the temporal evolution of a reduced combustion chemical system with an artificial neural network, Combust. Flame 113 (1–2) (1998) 38–52, doi:10.1016/S0010-2180(97)00211-3.

[38] A. Chatzopoulos, S. Rigopoulos, A chemistry tabulation approach via rate-controlled constrained equilibrium (RCCE) and artificial neural networks (ANNs), with application to turbulent non-premixed $CH_4/H_2/N_2$ flames, Proc. Combust. Inst. 34 (1) (2013) 1465–1473, doi:10.1016/j.proci.2012.06.057.

[39] L.L. Franke, A.K. Chatzopoulos, S. Rigopoulos, Tabulation of combustion chemistry via artificial neural networks (ANNs): methodology and application to LES-PDF simulation of Sydney flame L, Combust Flame 185 (2017) 245–260, doi:10.1016/j.combustflame.2017.07.014.

[40] S. Alqahtani, T. Echekki, A data-based hybrid model for complex fuel chemistry acceleration at high temperatures, Combust Flame 223 (2021) 142–152, doi:10.1016/j.combustflame.2020.09.022.

[41] O. Owoyele, P. Kundu, M.M. Ameen, T. Echekki, S. Som, Application of deep artificial neural networks to multi-dimensional flamelet libraries and spray flames, Int. J. Engine Res. 21 (1) (2020) 151–168.

[42] S. Silvestri, M.P. Celano, O.J. Haidn, O. Knab, Comparison of single element rocket combustion chambers with round and square cross sections, 6th Euro. Conf. Aeronautics Space Sci. (EUCASS) (2015).

[43] S. Silvestri, M.P. Celano, C. Kirchberger, G. Schlieben, O. Haidn, O. Knab, Investigation on recess variation of a shear coax injector for a single element GOX-GCH4 combustion chamber, Trans. JSASS Aerosp. Tech. Jpn. 14 (ists30) (2016) 101–108.

[44] Y. Khalighi, J.W. Nichols, F. Ham, S.K. Lele, P. Moin, Unstructured large eddy simulation for prediction of noise issued from turbulent jets in various configurations, AIAA Paper 2011-2886 (2011).

[45] P.C. Ma, Y. Lv, M. Ihme, An entropy-stable hybrid scheme for simulations of transcritical real-fluid flows, J. Comput. Phys. 340 (2017) 330–357.

[46] H. Wu, P.C. Ma, M. Ihme, Efficient time-stepping techniques for simulating turbulent reactive flows with stiff chemistry, Comput. Phys. Commun. 243 (2019) 81–96, doi:10.1016/J.CPC.2019.04.016.

[47] P. Moin, K. Squires, W. Cabot, S. Lee, A dynamic subgrid-scale model for compressible turbulence and scalar transport, Phys. Fluids A 3 (11) (1991) 2746–2757, doi:10.1063/1.858164.

[48] O. Colin, F. Ducros, D. Veynante, T. Poinsot, A thickened flame model for large eddy simulations of turbulent premixed combustion, Phys. Fluids 12 (7) (2000) 1843–1863, doi:10.1063/1.870436.

[49] M. Ihme, L. Shunn, J. Zhang, Regularization of reaction progress variable for application to flamelet-based combustion models, J. Comput. Phys. 231 (2012) 7715–7721.

[50] A. Felden, E. Riber, B. Cuenot, Impact of direct integration of analytically reduced chemistry in LES of a sooting swirled non-premixed combustor, Combust. Flame 191 (2018) 270–286, doi:10.1016/j.combustflame.2018.01.005.

[51] A. Vreman, B. Albrecht, J. van Oijen, L. de Goey, R. Bastiaans, Premixed and nonpremixed generated manifolds in large-eddy simulation of Sandia flame D and F, Combust. Flame 153 (3) (2008) 394–416, doi:10.1016/j.combustflame.2008.01.009.

[52] G.P. Smith, D.M. Golden, M. Frenklach, N.W. Moriarty et. al., GRI-Mech 3.0, 2000, http://www.me.berkeley.edu/gri-mech/.

[53] D.G. Goodwin, R.L. Speth, H.K. Moffat, B.W. Weber, Cantera: an object-oriented software toolkit for chemical kinetics, thermodynamics, and transport processes, 2018, https://www.cantera.org. 10.5281/zenodo.1174508

[54] H. Pitsch, FlameMaster v3.1: a C++ computer program for 0D combustion and 1D laminar flame calculations, 1998.

[55] J. Zips, H. Müller, M. Pfitzner, Non-adiabatic tabulation methods to predict wall-heat loads in rocket combustion, AIAA Paper 2017-1469 (2017), doi:10.2514/6.2017-1469.

[56] P.E. Lapenna, R. Amaduzzi, D. Durigon, G. Indelicato, F. Nasuti, F. Creta, Simulation of a single-element GCH4/GOx rocket combustor using a non-adiabatic flamelet method, AIAA Paper (2018) 2018–4872, doi:10.2514/6.2018-4872.

[57] N. Perakis, O.J. Haidn, Inverse heat transfer method applied to capacitively cooled rocket thrust chambers, Int. J. Heat Mass Transf. (2019) 150–166, doi:10.1016/j.ijheatmasstransfer.2018.11.048.

[58] S. Kawai, J. Larsson, Dynamic non-equilibrium wall-modeling for large eddy simulation at high Reynolds numbers, Phys. Fluids 25 (1) (2013) 015105, doi:10.1063/1.4775363.

[59] N. Perakis, O.J. Haidn, M. Ihme, Investigation of CO recombination in the boundary layer of CH_4/O_2 rocket engines, Proc. Combust. Inst. 38 (2020). In press

[60] P.C. Ma, H. Wu, M. Ihme, J.-P. Hickey, Nonadiabatic flamelet formulation for predicting wall heat transfer in rocket engines, AIAA J. 56 (6) (2018) 2336–2349, doi:10.2514/1.J056539.

[61] D.N. Reshef, Y.A. Reshef, H.K. Finucane, S.R. Grossman, G. McVean, P.J. Turnbaugh, E.S. Lander, M. Mitzenmacher, P.C. Sabeti, Detecting novel associations in large data sets, Science 334 (6062) (2011) 1518–1524.

[62] R.W. Bilger, Turbulent jet diffusion flames, Prog. Energy Combust. Sci. 1 (2–3) (1976) 87–109, doi:10.1016/0360-1285(76)90022-8.

[63] J. Li, K. Cheng, S. Wang, F. Morstatter, R.P. Trevino, J. Tang, H. Liu, Feature selection: a data perspective, ACM Comput. Surv. 50 (2017), doi:10.1145/3136625.

[64] R. Ge, M. Zhou, Y. Luo, Q. Meng, G. Mai, D. Ma, G. Wang, F. Zhou, McTwo: a two-step feature selection algorithm based on maximal information coefficient, BMC Bioinform. 17 (142) (2016) 14pages.

[65] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.

[66] L. Breiman, J. Friedman, R. Olshen, C. Stone, Classification and Regression Trees, Routledge, 1984.

[67] Y. Amit, D. Geman, K. Wilder, Joint induction of shape features and tree classifiers, IEEE Trans. Pattern Anal. and Mach. Intell. 19 (11) (1997) 1300–1305.

[68] M. Fernández-Delgado, E. Cernadas, S. Barro, D. Amorim, Do we need hundreds of classifiers to solve real world classification problems? J. Mach. Learn. Res. 15 (1) (2014) 3133–3181.

[69] A.J. Wyner, M. Olson, J. Bleich, D. Mease, Explaining the success of Adaboost and random forests as interpolating classifiers, J. Mach. Learn. Res. 18 (1) (2017) 1558–1590.

[70] G. Bradski, The OpenCV Library, Dr. Dobb's J. Softw. Tools, 2000.

[71] K. Schindler, An overview and comparison of smooth labeling methods for land-cover classification, IEEE Trans. Geosci. Remote Sens. 50 (11 PART1) (2012) 4534–4545, doi:10.1109/TGRS.2012.2192741.

[72] H. Müller, J. Zips, M. Pfitzner, D. Maestro, B. Cuenot, L. Selle, R. Ranjan, P. Tudisco, S. Menon, Numerical investigation of flow and combustion in a single-element GCH4/GOX rocket combustor: a comparative LES study, AIAA Paper 2016-4997 (2016), doi:10.2514/6.2016-4997.

[73] C. Roth, O. Haidn, A. Chemnitz, T. Sattelmayer, Y. Daimon, G. Frank, H. Müller, J. Zips, M. Pfitzner, R. Keller, P. Gerlinger, D. Maestro, B. Cuenot, H. Riedmann, L. Selle, Numerical investigation of flow and combustion in a single-element GCH_4/GOX rocket combustor, AIAA Paper 2016-4995 (2016), doi:10.2514/6.2016-4995.