Contents lists available at ScienceDirect

Combustion and Flame

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

Interpretable data-driven methods for subgrid-scale closure in LES for transcritical LOX/GCH4 combustion

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

[a] Department of Mechanical Engineering, Stanford University, Stanford, CA 94305, USA
[b] SLAC National Accelerator Laboratory, Menlo Park, CA 94025, USA

ARTICLE INFO

Article history:
Received 1 March 2021
Revised 11 August 2021
Accepted 15 September 2021
Available online xxx

Keywords:
Turbulence modeling
Direct numerical simulation
Transcritical combustion
Random forests
Machine learning
Explainable artificial intelligence

ABSTRACT

Although data-driven methods have shown high accuracy as closure models for simulating turbulent flames, these models are often criticized for lack of physical interpretability, wherein they provide answers but no insight into their underlying rationale. In this work, we show that two interpretable machine learning algorithms, namely the random forest regressor and the sparse symbolic regression, can offer opportunities for discovering analytic expressions for subgrid-scale (SGS) terms of a turbulent transcritical flame. These transcritical conditions are found in various combustion systems that operate under high pressures that surpass the thermodynamic critical limit of fuel-oxidizer mixtures, and require the consideration of complex fluid behaviors that can cast doubts on the validity of existing subgrid-scale (SGS) models in large-eddy simulations. To this end, direct numerical simulations (DNS) of transcritical liquid-oxygen/gaseous-methane (LOX/GCH4) inert and reacting flows are performed. Using this data, *a priori* analysis is performed on the Favre-filtered DNS data to compare the accuracy of random forest SGS-models with conventional physics-based SGS-models. SGS stresses calculated with the gradient model are shown to have good agreement with the exact terms extracted from filtered DNS. Results demonstrate that random forests can perform as effectively as algebraic models when modeling subgrid stresses, when trained on a sufficiently representative database and with a suitable choice of the feature set. The employment of the random forest feature importance score is shown to enable the discovery of a analytic model for subgrid-scale stresses through sparse symbolic regression. The generalizability of random forest and sparse symbolic regression is demonstrated by modeling the subgrid-scale temperature, a term that arises from filtering the non-linear real-fluid equation-of-state, with good accuracy.

1. Introduction

The development of accurate computational tools is essential for studying turbulent flames within high pressure combustors. Large-eddy simulations (LES) provide a feasible computational approach in capturing the behavior of flows within practical combustors. However, high pressure combustors in rockets and diesel engines operate under conditions that exceed the thermodynamic critical limits of both fuel and oxidizer. Consequently, these conditions generate trans- and supercritical flows – with complex behaviors that pose challenges for numerical modeling and simulations [1]. In many simulations, Lagrangian droplet methods are typically employed for simulating, which assumes the presence of liquid and gas phases. Simulations employing the Lagrangian droplet method have been shown to have good agreement with experimental measurements [2,3]. However, these models often involve careful selection of the breakup and evaporation models along with parameter tuning.

Another approach for investigating high-pressure flows involves the use of the diffuse-interface method. In contrast to sharp interface techniques, where interfaces are explicitly tracked or resolved in the computational domain, this method artificially diffuses the interface and treats the entire flow with a single real-fluid state equation. While LES incorporating real-fluids effects have successfully been employed to simulate transcritical combustion [4–6], many of these simulations employ existing subgrid-scale (SGS) models that were developed for applications in subcritical pressure conditions [7,8]. As a consequence, the application of these models to non-ideal flow regimes introduces uncertainties. One method for evaluating SGS models involves *a priori* analysis, where modeled SGS terms are compared with exact unclosed terms extracted from filtered DNS. Selle et al. [9] performed *a priori* analysis on a three-dimensional DNS database of supercritical binary mixtures in turbulent mixing layers to demonstrate that the Smagorin-

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

https://doi.org/10.1016/j.combustflame.2021.111758
0010-2180/© 2021 The Combustion Institute. Published by Elsevier Inc. All rights reserved.

sky model [10] performed poorly when predicting SGS stresses, while the gradient [8] and scale-similar [11] models performed well. In the same work, the consideration of previously neglected unclosed terms for pressure and heat flux were shown to be essential under supercritical conditions. Unnikrishnan et al. [12] performed *a priori* analysis on two-dimensional DNS of a transcritical reacting liquid-oxygen/gaseous-methane (LOX/GCH4) mixture to demonstrate that the mixed SGS model incorporating the dynamic Smagorinsky [13] was three times more accurate than the sole use of the dynamic Smagorinsky.

One approach for developing closure models in turbulent reacting flows involves the use of data-driven methods [14]. *A priori* studies have been performed to demonstrate that neural-networks can provide accurate closure for turbulent combustion models [15–17]. Henry de Frahan et al. [18] demonstrated that deep learning models can generate as accurate results as random forests with 25-fold improvement in computational costs when predicting the subfilter probability density function. Ranade and Echekki [19] conducted an *a posteriori* study to show that deep learning models can be trained with experimental data to generate closure models for chemical scalars in Reynolds-averaged Navier-Stokes (RANS) simulations of turbulent jet flames. These deep learning models are highly accurate and flexible approaches to data-driven modeling. The employment of convolutional architectures enable an automatic end-to-end processing of unstructured spatial data, which is useful when treating simulation data, while the manipulation of its loss functions and network structure enables the inclusion of physics-specific information [20,21]. However, these methods tend to be uninterpretable, thereby offering little insight towards the discovery of physical properties.

Computational studies of high-pressure non-premixed flames were pioneered by Bray et. al. [2,3,22]. One work [22] examined the effect of different Damköhler numbers (Da) on autoignition in high-pressure non-premixed flames under decaying homogeneous isotropic turbulence. In the present work, we perform DNS calculations of inert and reacting LOX/GCH4 non-premixed mixtures in the presence of decaying turbulence, under different Da-conditions, in order to evaluate algebraic and data-driven models for predicting unclosed SGS terms for high pressure applications. Within this context, the present study has the following objectives:

- To identify and quantify limitations of conventional algebraic SGS stresses in transcritical flows.
- To utilize interpretable machine learning algorithms, namely the random forest regressor and the sparse symbolic regression, in constructing data-informed models for SGS stresses.
- To apply these machine learning methods towards modeling additional SGS terms that can arise from real-fluid effects.

The mathematical models for simulating the turbulent transcritical flows are presented in Section 2. Details regarding the DNS configuration are discussed in Section 3. Section 4 describes the SGS models and data-driven methods employed in the present work. Results from this *a priori* study are discussed in Section 5, before offering concluding remarks in Section 6.

2. Mathematical models

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

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

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

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

$$\partial_t (\rho Y_k) + \nabla \cdot (\rho \boldsymbol{u} Y_k) = -\nabla \cdot \boldsymbol{j}_v + \dot{\omega}_k \quad \text{where} \quad k = 1, 2, \ldots, N_s - 1 \tag{1d}$$

with density ρ, velocity vector $\boldsymbol{u}$, pressure p, specific total energy e_t, stress tensor $\boldsymbol{\tau}$, and heat flux $\boldsymbol{q}$. Y_k, $\boldsymbol{j}_k$, and $\dot{\omega}_k$ are the mass fraction, diffusion flux, and source term for the kth species, while subscript v denotes viscous quantities.

In the *a priori* analysis carried out in this study, a top-hat filter H with a desired filter size $\overline{\Delta}$ is applied on a arbitrary quantity ϕ from the DNS data through a volume integral:

$$\overline{\phi(\boldsymbol{x})} = \int_V \phi(\boldsymbol{x}) H(\boldsymbol{x} - \boldsymbol{y}, \overline{\Delta}) d\boldsymbol{y} \tag{2}$$

with:

$$H(\boldsymbol{x} - \boldsymbol{y}, \overline{\Delta}) = \begin{cases} \frac{1}{\overline{\Delta}} & \text{for } |\boldsymbol{x} - \boldsymbol{y}| \leq \overline{\Delta}/2, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

and Favre-averaged quantity:

$$\widetilde{\phi} = \frac{\overline{\rho\phi}}{\overline{\rho}} \tag{4}$$

where $\overline{\cdot}$ denotes a filtered quantity and $\widetilde{\cdot}$ is a Favre-filtered quantity. After filtering, the governing equations become:

$$\partial_t \overline{\rho} + \nabla \cdot (\overline{\rho} \widetilde{\boldsymbol{u}}) = 0 \tag{5a}$$

$$\partial_t (\overline{\rho} \widetilde{\boldsymbol{u}}) + \nabla \cdot (\overline{\rho} \widetilde{\boldsymbol{u}} \widetilde{\boldsymbol{u}}) = -\nabla \overline{p} + \nabla \cdot (\overline{\boldsymbol{\tau}}_v + \boldsymbol{\tau}^{sgs}) \tag{5b}$$

$$\partial_t (\overline{\rho} \widetilde{e}_t) + \nabla \cdot [\widetilde{\boldsymbol{u}}(\overline{\rho} \widetilde{e}_t + \overline{p})] = -\nabla \cdot (\overline{\boldsymbol{q}}_v + \boldsymbol{q}^{sgs}) + \nabla \cdot [(\overline{\boldsymbol{\tau}}_v + \boldsymbol{\tau}^{sgs}) \cdot \widetilde{\boldsymbol{u}}] \tag{5c}$$

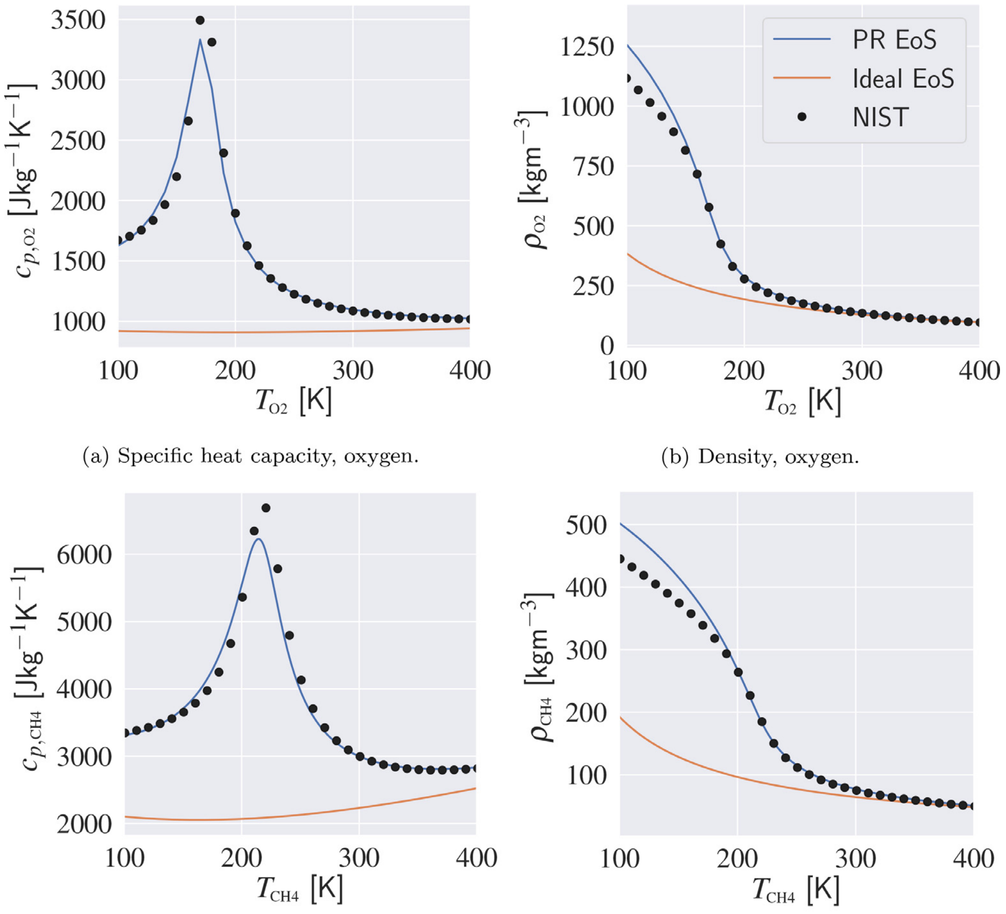
$$\partial_t (\overline{\rho} \widetilde{Y}_k) + \nabla \cdot (\overline{\rho} \widetilde{\boldsymbol{u}} \widetilde{Y}_k) = -\nabla \cdot (\overline{\boldsymbol{j}}_v + \boldsymbol{j}^{sgs}) + \overline{\dot{\omega}}_k \quad \text{where} \quad k = 1, 2, \ldots, N_s - 1 \tag{5d}$$

with superscript *sgs* denoting subgrid-scale quantities. For the flow-conditions considered in this study, a filter width of $16\overline{\Delta}$ is equivalent to the integral lengthscale. Since LES should resolve the inertial subrange, we employ a maximum filter width of $8\overline{\Delta}$ in the present study. Exact subgrid-scale quantities can then be extracted directly from the filtered DNS or approximated through the models described in Section 4.

The Peng-Robinson (PR) cubic equations-of-states [23] (EoS) is employed to model real-fluid thermodynamics under transcritical conditions:

$$p = \frac{\rho RT}{1 - b\rho} - \frac{a\rho^2}{1 + 2b\rho - b^2\rho^2} \tag{6}$$

with mixture-specific gas constant R. The coefficients a and b account for effects of intermolecular forces and volumetric displacement, and are dependent on temperature and composition [24]. Since oxygen and methane mixtures are a miscible system, where the effects of phase separation are not encountered due to the similarity of the critical states and molecular properties, this transcritical configuration can be represented by a cubic EoS [25]. Details regarding the evaluation of specific heat capacity, internal energy, and partial enthalpy from the Peng-Robinson state equation is described in Ma et al. [26]. Fig. 1 compares Peng-Robinson and ideal EoS for methane and oxygen. At the initial conditions of 120 K and 300 K for O_2 and CH_4, specific heat capacity evaluated from the PR EoS is in excellent agreement with NIST data. However, it can be seen that the PR EoS overpredicts the oxidizer density but provides accurate results for the fuel. Since this study is primarily focusing on the development and assessment of a data-driven modeling framework for the constructing SGS-closures, we believe that this discrepancy is acceptable for the present study.



(a) Specific heat capacity, oxygen.

(b) Density, oxygen.

(c) Specific heat capacity, methane.

(d) Density, methane.

Fig. 1. Comparison of Peng-Robinson (PR) and ideal equations-of-states (EoS) for (a,b) oxygen and (c,d) methane with NIST [27] data at $p = 10$ MPa.

In this study, the two-step five-species CH4-BFER mechanism [28] is employed, which was applied to investigate a supercritical gas-turbine combustor at 20 MPa in another study [29]. In DNS of trans- and supercritical combustion, reduced chemical mechanisms [30,31] have been employed to circumvent large computational costs incurred by solving non-ideal state equations. Takahashi's high-pressure correction [32] is used to evaluate the binary diffusion coefficients. Since only two species are used in the inert simulations, the binary diffusion coefficients are exact. Thermal conductivity and dynamic viscosity are evaluated using Chung's method with high-pressure correction [33]. For multi-species mixtures in the reacting cases, Chung's pressure correction is known to produce oscillations [26,34], especially for dynamic viscosity. Hence, transport properties of the mixture are evaluated through mole-fraction-averaging, after employing Chung's method on each individual species. A similar approach has been applied in prior studies [5,12].

Simulations are performed by employing an unstructured compressible finite-volume solver [26,35,36]. 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 variations using a threshold-based sensor to describe sharp interfaces present in transcritical flows. Due to the density gradients present at trans- and supercritical conditions, an entropy-stable flux correction technique [26] is used to dampen non-linear instabilities in the numerical scheme. The double-flux method by Ma et al. [26] is used with a dynamic sensor to eliminate spurious pressure oscillations. 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 scheme [37] for advancing the chemical source terms.

3. DNS configuration

Inert and reacting direct numerical simulations are performed on a three-dimensional cubic domain, with length L, a mixture of LOX/GCH4 shown in Fig. 2. In this setup, a spherical liquid oxygen core, with a radius $r = 0.25L$, is initialized in gaseous methane, where the radial profile of the initial condition is chosen to match inert and reacting steady one-dimensional Cantera [38] counterflow diffusion flame calculations, solved in mixture-fraction space and incorporating the Peng-Robinson equation-of-state, under the same fuel and oxidizer conditions. For the reacting cases, the initial temperature and composition profile corresponds to maximum strain rates (from one-dimensional flames) of 2×10^5 s^{-1} and 2×10^6 s^{-1} for cases Da 780 and Da 10, respectively. Fuel and oxidizer temperatures are set to $T_{\mathrm{CH4}} = 300$ K and $T_{\mathrm{O2}} = 120$ K, respectively, while the pressure is set at 10 MPa. The laminar flame speed S_L of a stoichiometric premixed flame of $S_L = 0.306$ ms^{-1} is evaluated through Cantera [38] at a pressure of 10 MPa and initial temperature of 210 K (the average of fuel and oxidizer temperature). Note that the critical temperature T_c and pressure P_c for oxidizer and fuel are $T_{c,\mathrm{O2}} = 154.6$ K and $P_{c,\mathrm{O2}} = 5.04$ MPa, and $T_{c,\mathrm{CH4}} = 190.6$ K and $P_{c,\mathrm{CH4}} = 4.60$ MPa, respectively.

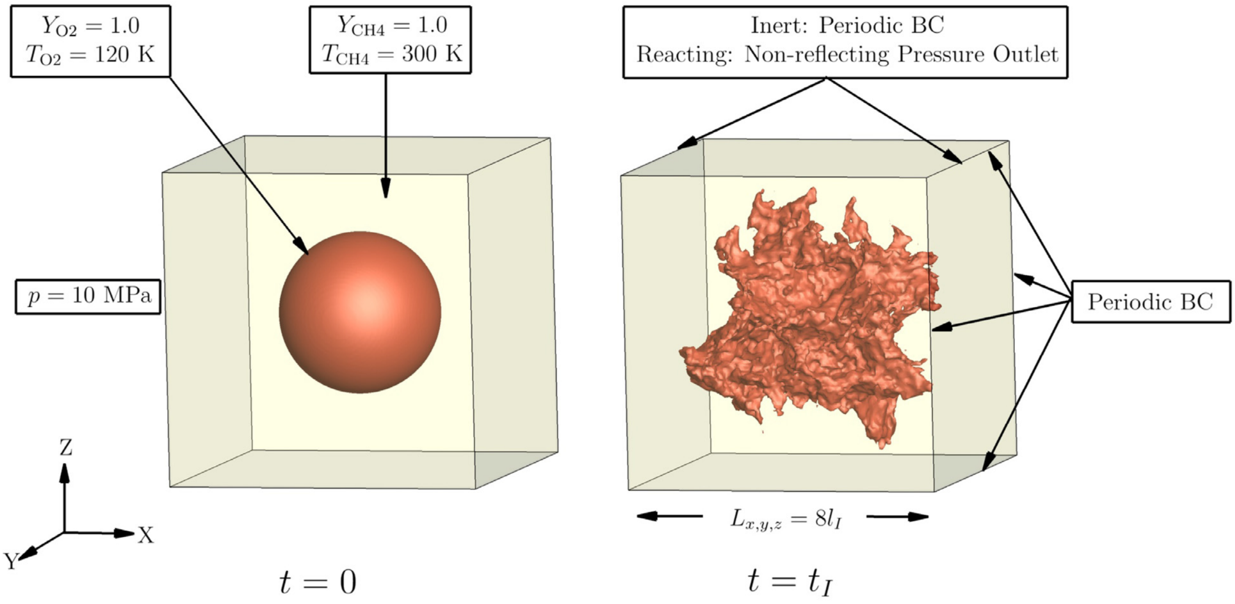


Fig. 2. DNS investigated at initial time $t = 0$ and one eddy turnover time $t = t_I$. Isosurface shows stoichiometric mixture fraction $Z = 0.2$ for the inert case.

These operating conditions are chosen to match practical LOX/GCH4 combustors, and were investigated in previous studies [39,40]. Periodic boundary conditions are used for all boundaries for the inert case. For the reacting cases, non-reflecting pressure outlets are used in both boundaries in the x-direction, while the remaining boundaries are periodic.

The initial velocity profile was generated with a synthetic isotropic turbulence generator by Saad et al. [41] with zero mean velocity, based on the von Kármán-Pao energy spectrum:

$$E(\kappa) = \alpha \frac{u'^2}{\kappa_I} \frac{(\kappa/\kappa_I)^4}{[1 + (\kappa/\kappa_I)]^{17/6}} \exp\left[-2\left(\frac{\kappa}{\kappa_\eta}\right)^2\right], \tag{7a}$$

$$\alpha = 1.453, \tag{7b}$$

$$\kappa_I = 0.746834/l_I, \tag{7c}$$

where u' is the RMS velocity, κ is the wave number, and κ_η the Kolmogorov wave number. The chosen scaling constant α and large-eddy wavenumber κ_I are typical for isotropic turbulence [42]. In all cases, the integral lengthscale l_I and root-mean-squared (RMS) velocity fluctuation u' have been chosen to produce a turbulent Reynolds number Re_t of 80, which has been computed with the averaged kinematic viscosity of oxygen and methane at 120 K and 300 K, respectively.

In the reacting cases, two different Damköhler numbers, Da, of 780 and 10 are investigated, corresponding to flamelet and unsteady regimes [43], respectively. The Damköhler number is given by the ratio of physical timescale t_{conv} and chemical timescale t_{chem}:

$$\mathrm{Da} = \frac{t_{conv}}{t_{chem}}, \tag{8}$$

where $t_{chem} = 0.412$ μs is approximated from the extinction strain rate of a one-dimensional counterflow diffusion flame of a LOX/GCH4 mixture under similar conditions, and physical time is evaluated from the eddy turnover time $t_{conv} = t_I$. Fig. 3a shows that the mean temperature $\langle T \rangle$ is lower when Da = 10 than when Da = 780, due the presence of local extinction. This is also reflected in Fig. 3b where the consumption of CH_4 is slower in the case Da = 10 than the case Da = 780. This decrease in temperature and composition also results in a slower decay of the turbulence, as shown by the mean turbulent kinetic energy ⟨TKE⟩, normalized by the initial TKE, shown in Fig. 3c.

An additional inert simulation with ideal gas law is performed to demonstrate real-fluid effects on subgrid-scale terms that can arise from the non-linearities of the Peng-Robinson EoS. For this ideal configuration, atmospheric conditions $p = 101.325$ kPa at room temperature are employed, with T_{CH4} and T_{O2} at 300 K.

In this study, analysis is performed on all cases after $t = \mathrm{argmax}(t_I, t_{chem})$, which is typically done for DNS of combustion under decaying turbulence in order to ensure the flow fields are independent of initialization [44]. Instantaneous flow fields for axial velocity component u_1, mixture fraction Z, and mixture-fraction conditioned temperature T for the reacting cases at $t = 0$ and $t = t_I$ are shown in Fig. 4.

Table 1 summarizes the DNS cases examined in this study. The domain lengths in all direction were chosen to be eight times the size of the integral lengthscale l_I to minimize effects of the boundary conditions. The cell size Δ is prescribed on the order of the Kolmogorov lengthscale, ensuring that all lengthscales are resolved. In addition, a mesh refinement study was performed, where the energy spectra of velocity was found to converge between 128^3 and 256^3. Simulations for all three cases are advanced with an acoustic CFL number of unity, corresponding to timesteps of 2.5 and 0.5 ns for cases Da = 780 and Da = 10, respectively. The simulations were performed using 960 Intel Xeon (E5-2698 v3) processors, and 2.3 μs and 0.6 μs of physical time could be completed in about an hour wall clock time for cases Da = 780 and Da = 10, respectively.

4. Subgrid-scale models and data-driven methods

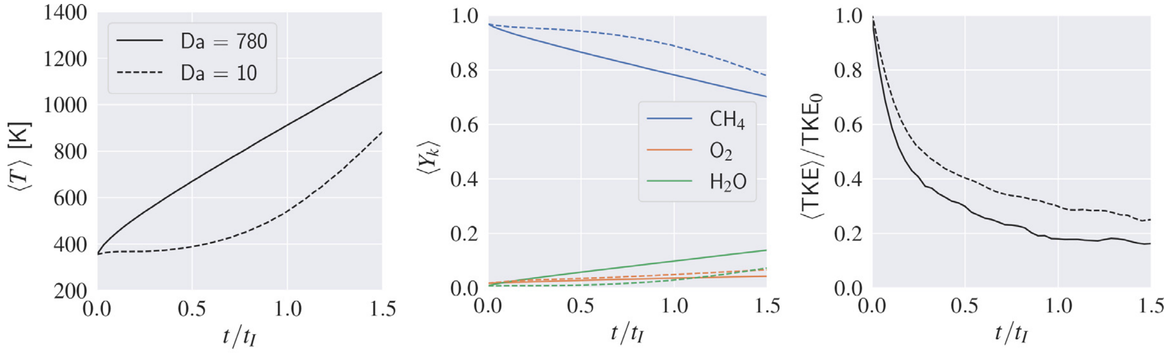
4.1. Real-fluid effects

We investigate the effects of additional non-linearities from the real-fluid equation-of-state, by employing a analysis similar to Huo and Yang [45] that they applied to model SGS density. Eq. (6) can be reexpressed with a compressibility factor ζ:

$$p = \rho R T \zeta \tag{9}$$

By rearranging and Favre-filtering, we obtain:

$$\widetilde{T} = p \cdot \widetilde{(\rho R \zeta)}^{-1} \tag{10}$$



(a) Temperature.

(b) Mass fraction.

(c) Turbulent kinetic energy.

Fig. 3. Temporal evolution of global temperature T, mass fraction Y_k, and normalized turbulent kinetic energy TKE for two reacting cases.

Table 1
Summary of DNS cases.

Case	$N_{x,y,z}$	$L_{x,y,z}$ [μm]	Re_t	l_I [μm]	η_k [μm]	Δ [μm]	t_I [μs]	u' [ms^{-1}]
Inert	128	500	80	62.5	2.32	3.91	286	0.22
Da = 780	128	500	80	62.5	2.32	3.91	286	0.22
Da = 10	128	60	80	7.50	0.278	0.469	4.12	1.80
Ideal EoS	128	500	80	62.5	2.32	3.91	3	20.67

However in the present LES solver, the Favre-filtered temperature is obtained by inputting filtered quantities into the real-fluid EoS:

$$\widetilde{T} = \overline{p}\cdot[\overline{\rho}\overline{R}\bar{\zeta}(\overline{\rho},\overline{p},\widetilde{Y})]^{-1} + \left[p\cdot\widetilde{(\rho R\zeta)}^{-1} - \overline{p}\cdot(\overline{\rho}\overline{R}\bar{\zeta})^{-1}\right] \tag{11a}$$

$$\widetilde{T} = \widetilde{T}_{LES}(\overline{\rho},\overline{p},\widetilde{Y}) + T_{sgs} \tag{11b}$$

which gives rise to a subgrid-scale temperature T_{sgs}, *i.e.* the second term on the right-hand-side.

T_{sgs} is typically neglected in ideal-gas configurations. This is often an acceptable assumption as shown by the ideal EoS case in Fig. 5. In the transcritical inert case, $|T_{sgs}|/\widetilde{T}$ of approximately 0.05 is observed, which is similar with observations from another study [9]. However, T_{sgs} becomes non-negligible for the transcritical reacting cases, where $|T_{sgs}|/\widetilde{T}$ exceeds values of 0.1 in the reacting regions, where multi-species compositions are present, and regions with high density gradient. Non-negligible SGS EoS terms are also reported by other studies [45,46]. This added significance of T_{sgs} arises from applying the filtering operation on density and multi-species mass fractions, and then feeding the filtered quantities into a highly non-linear equation.

Amplified non-linearities in transcritical reacting flow present an additional source of uncertainty in all SGS modeling. To investigate this, we employ conventional algebraic and novel data-driven methods for predicting the subgrid-scale fluxes from the LES momentum equation (Eq. (5b)):

$$\tau_{ij}^{sgs} = \overline{\rho}(\widetilde{u_i u_j} - \widetilde{u}_i\widetilde{u}_j) \tag{12}$$

Two algebraic SGS models, namely the Vreman and the gradient model, as well as random forest regressors are evaluated. Additionally, we demonstrate the employment of random forest feature importance scores for assisting the discovery of algebraic SGS stress models by sparse symbolic regression. Since algebraic models for SGS temperature (Eq. (11)) have not been developed, we then evaluate the ability of an interpretable machine learning algorithm in modeling T_{sgs}.

4.2. Algebraic SGS stress models

The Vreman SGS model [7] is derived from the eddy-viscosity hypothesis:

$$\tau_{ij}^{sgs,v} \simeq -2\overline{\rho}\nu_{SGS}\widetilde{S}_{ij} + \frac{1}{3}\tau_{kk}\delta_{ij}, \tag{13}$$

where S_{ij} is the velocity strain tensor, and δ_{ij} is the Kronecker delta. The eddy viscosity ν_{SGS} is evaluated for a filter width $\overline{\Delta}$ as follows:

$$\nu_{SGS} = C_\nu\sqrt{\frac{B}{a_{ij}a_{ij}}}, \tag{14a}$$

$$a_{ij} = \frac{\partial\widetilde{u}_i}{\partial x_j}, \tag{14b}$$

$$B = \beta_{11}\beta_{22} - \beta_{12}^2 + \beta_{11}\beta_{33} - \beta_{13}^2 + \beta_{22}\beta_{33} - \beta_{23}^2, \tag{14c}$$

$$\beta_{ij} = \overline{\Delta}^2 a_{ki}a_{kj}, \tag{14d}$$

where a Vreman coefficient C_ν of 0.07 is typically used in isotropic turbulence [7].

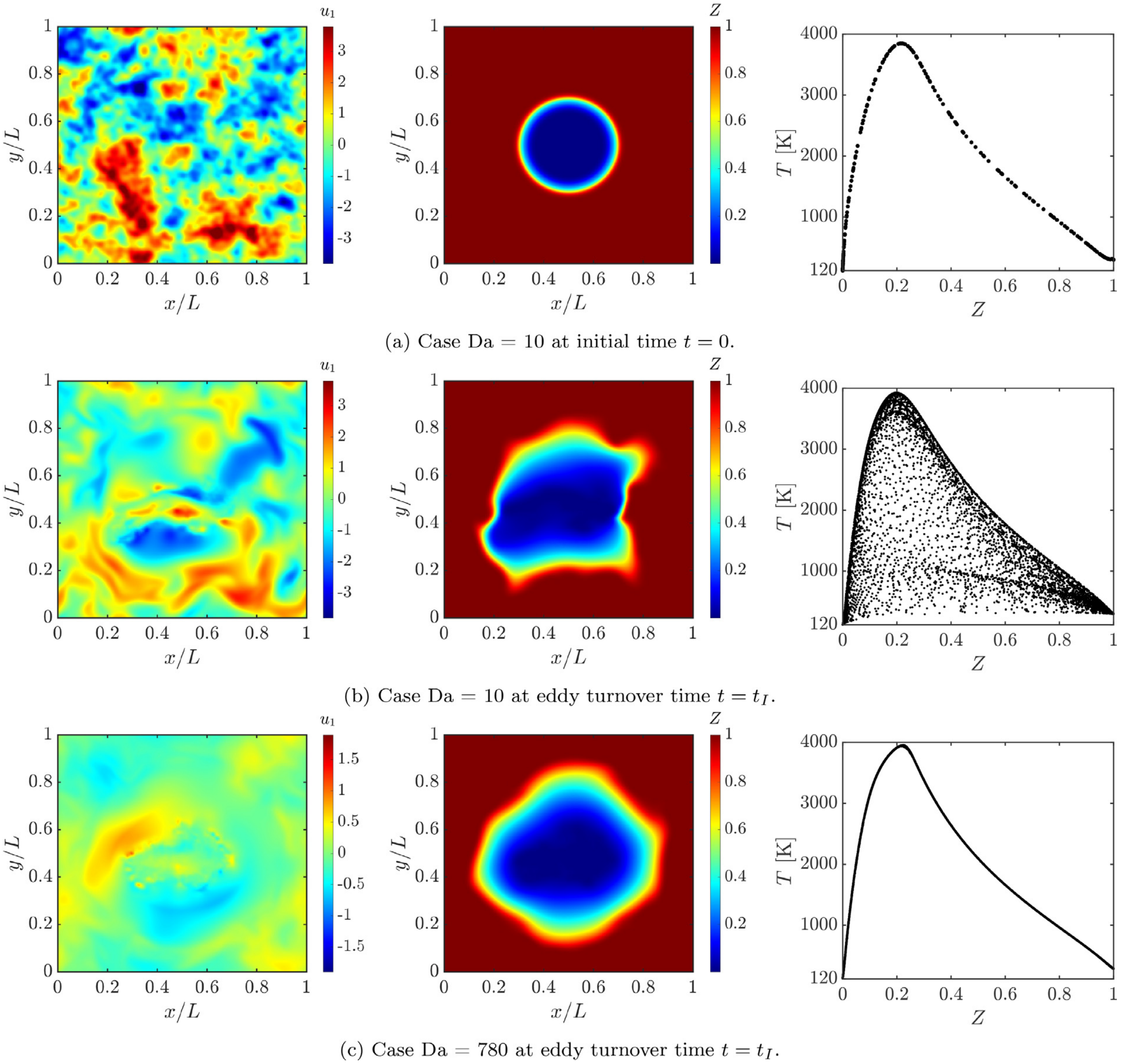
The gradient model by Clark et al. [8] is extracted from the first term in the Taylor series expansion of the filtering operation, and is given by:

$$\tau_{ij}^{sgs,g} \approx \overline{\rho}C_g\overline{\Delta}^2\frac{\partial\widetilde{u}_i}{\partial x_k}\frac{\partial\widetilde{u}_j}{\partial x_k}, \tag{15}$$

where a coefficient C_g of 1/12 is typically used when a top-hat filter is employed [8]. In the present study, we will evaluate both models and compare results against DNS data and a data-driven approach.

4.3. Random forest regressor

In this study, we employ the random forest as our regression algorithm for predicting SGS stresses and SGS temperature. Table 2 summarizes the input/features, outputs, and data for the random forests employed in this study. All random forests are trained with



(a) Case Da = 10 at initial time $t = 0$.

(b) Case Da = 10 at eddy turnover time $t = t_I$.

(c) Case Da = 780 at eddy turnover time $t = t_I$.

Fig. 4. Axial velocity u_1, mixture fraction Z, and conditional temperature T for the reacting cases at transverse location $z = 0$.

Table 2
Random forests employed in this study.

Random forest	RF_INFORM	RF_BLIND	RF_INERT	RF_DA780	RF_DA10	RF_TSGS
Training data ($t = t_I$)	Inert, Da = 780, Da = 10	Inert, Da = 780, Da = 10	Inert	Da = 780	Da = 10	Da = 780, Da = 10
Testing data ($t = 1.5t_I$)	Inert, Da = 780, Da = 10					Da = 780, Da = 10
Features (Input)	$\widetilde{S}_{ij}$, $\widetilde{S}_{ik}\widetilde{S}_{kj}$, $\widetilde{R}_{ik}\widetilde{R}_{kj}$, $\widetilde{S_{ik}}\widetilde{R_{kj}} - \widetilde{R_{ik}}\widetilde{S_{kj}}$	$\widetilde{u}_i$, $\frac{\partial \widetilde{u}_i}{\partial x_j}$, $\frac{\partial^2 \widetilde{u}_i}{\partial x_j x_k}$				T_{LES}, $\frac{\partial T_{LES}}{\partial x_j}$, $\frac{\partial^2 T_{LES}}{\partial x_j x_k}$
Output	τ_{ij}^{sgs}					T_{sgs}

snapshots at one eddy turnover time $t = t_I$ and tested on the three cases at $t = 1.5t_I$.

For SGS stresses, two different sets of feature, or inputs, are employed to train the random forests. One feature set corresponds to a domain-blind random forest RF_BLIND, consisting only of velocity, and the first and second spatial derivatives of velocity. The other set considers Galilean invariant basis functions constructed from strain $\widetilde{S}_{ij}$ and rotation $\widetilde{R}_{ij}$ tensors as features, shown to predict anisotropy well in a previous study [47]. These Galilean invariant features are used to train the random forest RF_INFORM. In order to investigate the generalizability of random forests in the absence of a vast representative dataset, we evaluate the predictive performance of three additional random forests RF_INERT, RF_DA780, and RF_DA10, which are trained solely from the inert, Da = 780, and Da = 10 cases, respectively.

In addition, we also examine the performance of random forest in predicting thermodynamic quantities. Since SGS temperature is significant for reacting transcritical cases, training and testing data for RF_TSGS are taken from the two transcritical reacting cases.

Random forests [48] consist of an ensemble of decorrelated Classification and Regression Trees (CARTs) [49]. CARTs are a machine learning approach for formulating prediction models from

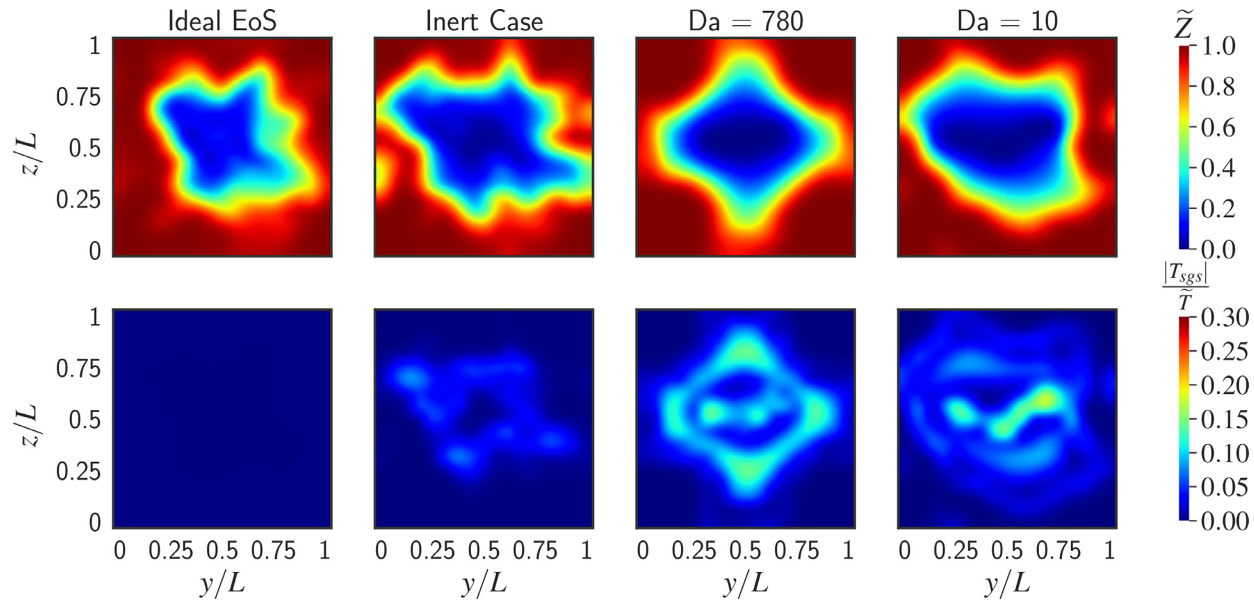


Fig. 5. Comparisons of filtered mixture fraction $\widetilde{Z}$ and magnitude of normalized subgrid-scale temperature $|T_{sgs}|/\widetilde{T}$ between an ideal-gas case and three transcritical cases. A filter width of $\overline{\Delta} = 8\Delta$ is employed.

data by recursively partitioning the inputted feature space, and fitting a simple prediction within each final partition. The partitioning of the feature space can be represented as a decision trees. Decision trees are supervised graph based model wherein the tree consists of nodes and edges. The internal (or non-terminal) nodes of the tree represent splits based on learned partitions of the feature space. Each leaf (or terminal) node is associated with a numerical value for regression trees (as opposed to categorical targets for classification trees).

During the training phase, the structure of the decision tree and the partitions associated with each node are inferred. During each step of this phase, exhaustive sets of splits over different input features are evaluated. The split leading to maximal decrease in prediction variance is selected at the associated node. The procedure continues to make recursive splits based on dataset until it has reduced the overall variance below a given threshold or upon reaching a given stopping parameter (for instance, upon reaching a maximum depth of the tree).

During the prediction phase, a new sample is traversed down the tree from the root node to a leaf node, wherein its path is determined based on the partition at each internal node. Once a leaf node is reached, the numerical value associated with the specific leaf node is outputted as the prediction for the sample.

Decision trees are non-parametric and can model arbitrarily complex relations without any *a priori* assumptions, but are prone to overfitting due to the greedy nature of the inference algorithm. Thus, decision trees have low bias but high variance. This issue can be addressed by combining the decision trees into an ensemble, which results in a random forest.

Random Forests are an ensemble learning algorithm, wherein the predictions of ensembles of decorrelated decision trees are aggregated so as to give a final meta-model with low bias and low variance. The decorrelation amongst the individual trees in the ensemble is achieved using bootstrapping [48] in conjunction with feature bagging [50] during the training of each decision tree. The final prediction of the resulting ensemble model is by averaging the predictions of all trained individual trees (or equivalently, via aggregation).

While accurate predictions are an important goal for machine learning models, in many fields of application it is just as important to derive understanding from the model. At the basic level, this can be embodied via feature importances, wherein the trained model also provides information regarding importances of different input features to the final prediction. Such measures of model interpretability provide insight into the underlying rationale learned by the model during training and can lead to high confidence in the model. Similarly, such interpretability measures can lead to data-driven discovery of new relationships between the input features and targets. Random forests provide the Mean Decrease Impurity (MDI) importance measure for all the features in the input set [51]. Here, the importance of a feature is given by aggregating the weighted decrease in variance for all the nodes where the specific feature is used as the criterion for partitioning the feature space.

In the present investigation, the random forest regressor implementation from the Scikit-learn library [52] is used. Here, a random forest consisting of fifty decision trees is employed. The hyperparameters of the random forest are determined using a random grid search approach with a 3-fold cross-validation set. Training is performed once *a priori*, and requires 88s of walltime with 8 CPUs, when trained on data coarsened for three different filter sizes from a single timestep. Prediction time for a 64^3 dataset requires 2.4 s on a single CPU.

4.4. Sparse symbolic regression for model discovery

A linear model $\hat{f}_i$ for m number of samples is typically expressed as the weighted sum of independent quantities X:

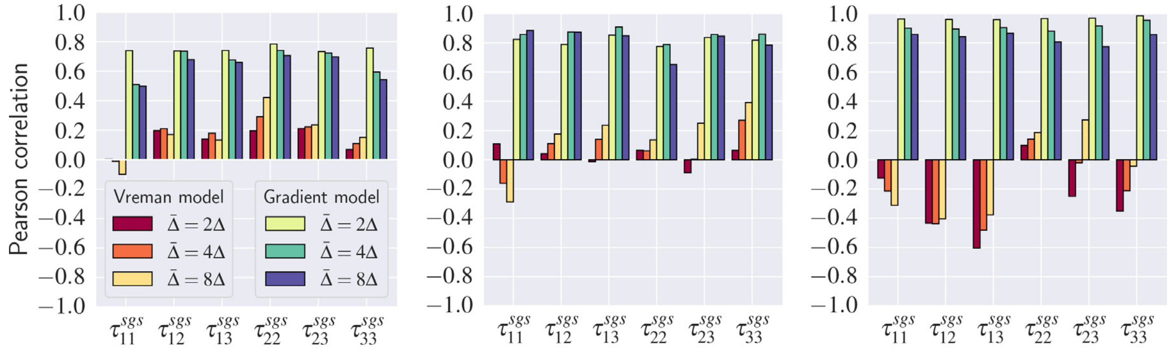
$$\hat{f}_i = \sum_{j=1}^{n} \beta_j X_{ij} \quad 1 \le i \le m \tag{16}$$

where n is the number of model coefficients β being employed.

When samples of the ground truth $\mathbf{f}$, *i.e.* the target for predictive modeling, are available, the model coefficients can be found with the l_1-norm regularized least squares or *lasso* method [53]:

$$\min_{\boldsymbol{\beta}} \left\{ \frac{1}{m} ||\mathbf{f} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_1 \right\}, \tag{17}$$

where λ is a regularization parameter for controlling the tradeoff between the least squares fit and the l_1-norm. Since, the optimization scheme in Eq. 17 also minimizes the l_1-norm of the model



(a) Inert case.

(b) Da = 780.

(c) Da = 10.

Fig. 6. Pearson correlations between exact and algebraically modeled SGS stresses for three different filter widths $\bar{\Delta}$.

coefficients $||\boldsymbol{\beta}||_1$, lasso encourages sparsity, *i.e.* reduces the number of terms in the linear model, as zero-valued model coefficients are preferred.

In the context of discovering subgrid-scale models, non-linearities can be introduced by replacing $\mathbf{X}$ with non-linear functions $G(\mathbf{X})$ of the original variables. In this study, we construct a model with non-linear variables by evaluating d-order polynomial functions:

$$\hat{\boldsymbol{f}} = G^d(\mathbf{X})\boldsymbol{\beta}, \tag{18a}$$

$$G^d(\mathbf{X}) = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1n} & X_{11}^2 & X_{11}X_{12} & \cdots & X_{1n}^d \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{m1} & X_{mn} & \cdots & X_{mn} & X_{m1}^2 & X_{m1}X_{m2} & \cdots & X_{mn}^d \end{bmatrix}, \tag{18b}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 & \beta_1 & \cdots & \beta_k \end{bmatrix}^T, \quad k = \sum_{i=1}^{d} n^i. \tag{18c}$$

Eq. (18) shows that the dimensionality of this approach scales to the order of polynomial functions $\mathcal{O}(mn^d)$. Hence, the number of candidate variables must be reduced for this method to remain tractable. In this work, we employ the random forest feature importance score to reduce the number of candidate variables.

5. Results

5.1. Algebraic SGS stress models

A priori analysis is performed by comparing SGS stresses τ_{ij}^{sgs} computed from filtered DNS, with SGS stress modeled by the Vreman model (Eq. (13)) and Clark's gradient model (Eq. (15)). The performance of these SGS models is evaluated through the Pearson correlation coefficient, which measures the linear correlation between two variables. A Pearson correlation of 1 and −1 corresponds to perfectly positive and negative linear relationships, respectively, whereas a correlation of 0 indicates a negligible linear relationship.

Figure 6 presents the resulting Pearson correlation between exact and algebraically modeled SGS stresses for three different filter widths $\bar{\Delta}$ for all three DNS cases specified in Table 1, at time $t = 1.5t_I$. For all three cases and filter sizes, negative correlations and weak positive correlations ranging from approximately −0.6 to 0.4 are observed for the Vreman model. Negative correlations suggest deviations from the eddy-viscosity hypothesis, which causes the Vreman model to be ineffective. In all three cases and three filter sizes, strong positive correlations, ranging from 0.5 to 0.95, suggest that the gradient model is highly suitable for modeling SGS stresses in transcritical inert and reacting flows.

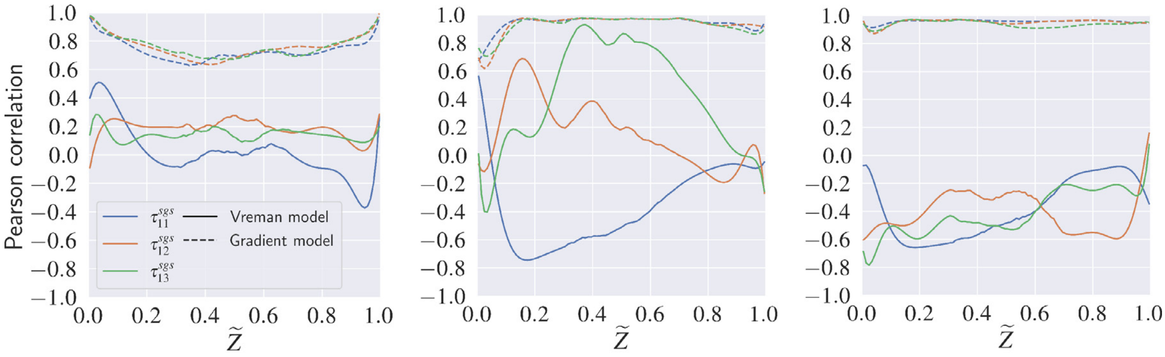
The effectiveness of the Vreman and gradient models are further assessed by examining the conditional Pearson correlation for τ_{1i}^{sgs} with respect to the mixture fraction $\widetilde{Z}$ at filter size $\bar{\Delta} = 2\Delta$. The mixture fraction for the reacting cases have been evaluated using Bilger's definition. Figure 7a shows that weak correlations ranging from −0.4 to 0.5 are observed throughout the inert case. In both reacting cases in Fig. 7b, the deviations from eddy-viscosity is much larger than the inert case, as denoted by the presence of highly negative correlations (−0.8) in the Vreman model. In the inert case, the gradient model has the highest correlation of approximately 1.0 in pure methane and pure oxygen, and the lowest correlation of 0.6 when $\widetilde{Z} = 0.5$. For the case Da = 780, the gradient model has the lowest correlation (0.7) close to the pure oxygen stream, with the correlation steadily increasing as the mixture approaches stoichiometry ($Z_{st} = 0.2$), after which the correlations remain high (0.85–1.0). For the case Da = 10, the correlations for the gradient model are high (0.8–1.0) throughout the entire mixture.

The accuracy of the gradient model in predicting the magnitude of SGS stresses is evaluated by examining the least squares fit between the exact and modeled SGS stresses. A slope greater than unity indicates underprediction of the modeled SGS stresses, while a slope less than unity indicates overprediction. Figure 8 shows that the slopes from the gradient model range from 1 to 4.5. The average of the slopes is 1.98, which suggests that the gradient model with a constant coefficient should employ $C_g = 1/6$, instead of the typical $C_g = 1/12$. However, since a wide range of coefficients are observed, a dynamic gradient model scheme is likely more suited in *a posteriori* simulations. This is confirmed by results from *a posteriori* evaluations of the dynamic gradient model from transcritical inert DNS [54].

5.2. Random forest SGS stress models

The *a priori* analysis performed in Section 5.1 is repeated in this section for the SGS stresses modeled by random forest regressors. Figure 9 presents the Pearson correlation between exact SGS stresses and the SGS stresses modeled by the random forests RF_BLIND and RF_INFORM. Details regarding the input, output, and training of these two random forests are described in Table 2. Figure 9a shows that strong correlations (0.4–0.95) are observed when the random forest is trained with an uninformed approach, which is similar to the gradient model and higher than the Vreman model in Fig. 6. Figure 9b demonstrates that the employment of invariant basis functions as features decreases the range of correlations (0.35–0.9) by 0.05. This small decrease is likely caused by the additional constraints placed on the random forest when forming a hypothesis space.

Figure 10 presents the Pearson correlations between exact and random forest SGS stresses τ_{1i}^{sgs} conditioned to mixture fraction $\widetilde{Z}$



(a) Inert case.

(b) Da = 780.

(c) Da = 10.

Fig. 7. Conditional Pearson correlations with respect to mixture fraction $\widetilde{Z}$ between exact and algebraically modeled SGS stresses τ_{1i}^{sgs} for a single filter width $\overline{\Delta} = 2\Delta$.

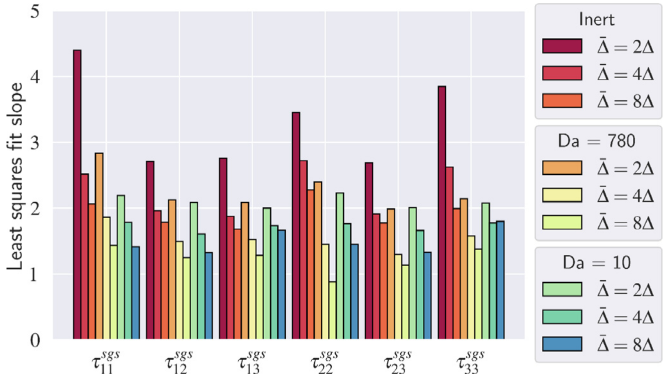


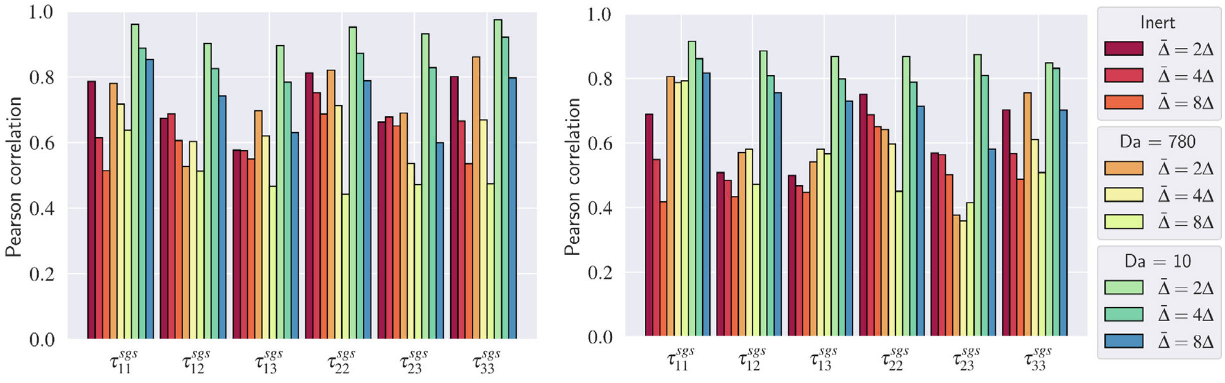
Fig. 8. Slopes from a least squares fit of exact and gradient modeled SGS stress for three different filter widths $\overline{\Delta}$.

at $\overline{\Delta} = 2\Delta$. In the inert case, shown in Fig. 10a, highest correlation from RF_BLIND of approximately 0.95 is observed in pure methane and pure oxygen, and lowest correlation of 0.5 when $Z = 0.5$. For the case Da = 780 in Fig. 10b, RF_BLIND possesses the lowest correlation (0.7) close to the oxygen stream, with the correlation steadily increasing as the mixture approaches stoichiometric conditions ($\widetilde{Z}_{st} = 0.2$), after which the correlations remain high (0.85 to 1.0). For the case Da = 10, shown in Fig. 10c, the correlations for the gradient model are high (0.8–1.0) throughout the entire mixture. The conditional Pearson correlation produced from RF_BLIND in all three cases are similar qualitatively and quantitatively to correlations from the gradient model in Fig. 7. This suggests that RF_BLIND has approximated a function similar to the gradient model, even when trained solely on exact SGS stresses and without any prior knowledge of the gradient Model. The correlations from RF_INFORM share similar qualitative behaviors as the correlations from RF_BLIND, but with up to a 0.2 lower values.

Figure 11 presents slopes from a least squares fit between the exact and the random forest SGS stresses. Figure 11a shows that the slopes from RF_BLIND range from 0.25 to 1.6, with an average slope of 0.96, which demonstrates excellent agreement between modeled and exact magnitudes of SGS stresses. The employment of invariant features leads to lower slopes (0.25 to 1.35), with an average slope of 0.867, as presented in Fig. 11b. The use of the invariant feature set not only leads to lower correlations but also to an overprediction in magnitudes of SGS stresses.

Figure 12 compares instantaneous fields for the exact and modeled SGS stress $\tau_{12}^{sgs}/\overline{\rho}$ at filter width $\overline{\Delta} = 4\Delta$. In the inert case, both SGS stresses from the gradient model and RF_BLIND are in good agreement with the exact term. For Da = 780, the gradient model is in better agreement with the exact term than RF_BLIND. This is further supported by the difference in Pearson correlation for this particular case shown by the gradient model (0.9) and RF_BLIND (0.6) in Fig., respectively. For Da = 10, RF_BLIND predicts the magnitude of the SGS stress better than the gradient model, which is also observed in the slopes shown by RF_BLIND (0.9) and the gradient model (1.5) shown in Fig..

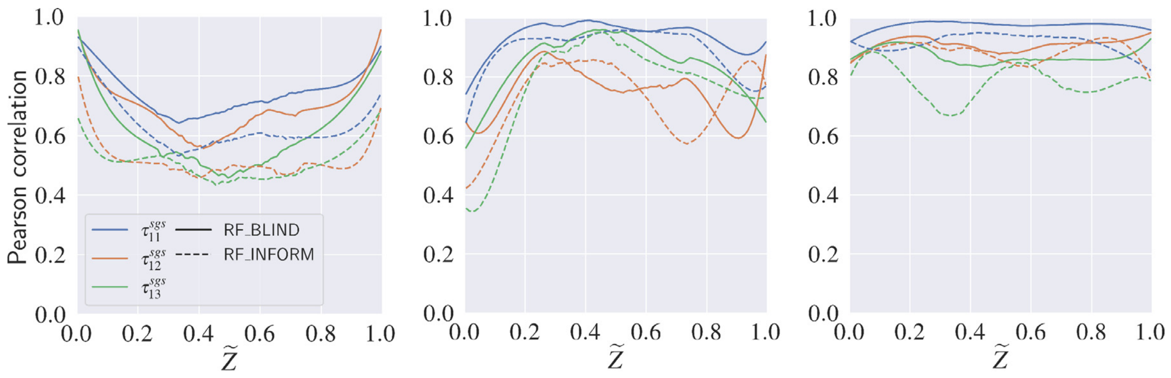
Figure 13 presents Pearson correlations from examining the generalizability of random forests in the presence of limited data. As presented in Table 2, we employ three different random forest regressors, each trained on only one DNS case, and examine their performance when tested on the two remaining cases. Ran-



(a) RF_BLIND.

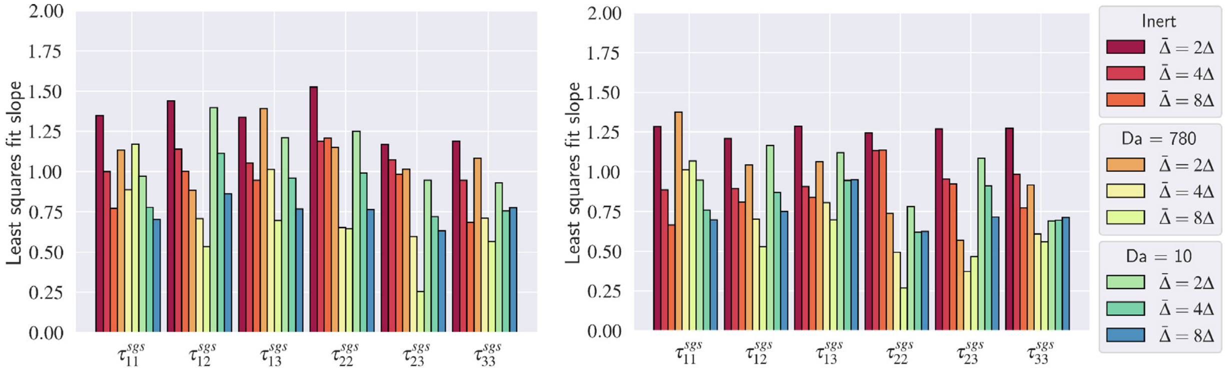
(b) RF_INFORM.

Fig. 9. Pearson correlation between exact and random forest modeled SGS stresses for three different filter widths $\overline{\Delta}$.



(a) Inert case. (b) Da = 780. (c) Da = 10.

Fig. 10. Conditional Pearson correlations as a function of mixture fraction $\widetilde{Z}$ between exact and random forest modeled SGS stresses τ_{1i}^{sgs} for a single filter width $\overline{\Delta} = 2\Delta$.



(a) RF_BLIND. (b) RF_INFORM.

Fig. 11. Slopes from a least squares fit of exact and random forest modeled SGS stress for three different filter widths $\overline{\Delta}$.

dom forest RF_INERT demonstrates a similar range of correlations (0.5–0.85) to RF_ALL when tested on the inert case with a filter size $\overline{\Delta} = 2\Delta$. However, lower ranges are observed for RF_INERT when tested on the cases Da = 780 (0.4–0.75) and Da = 10 (0.5–0.9). RF_DA780 also possesses a similar correlation as RF_BLIND when tested on case Da = 780 (0.5–0.9), but worse correlations when tested on the inert case (0.4–0.8) and case Da = 10 (0.8–0.9). Lastly, RF_DA10 performs similarly to RF_BLIND when tested on Da = 10 (0.85–0.95) but performs worse when tested on the inert (0.5–0.8) and Da = 780 (0.55–0.8) cases. These three random forests perform as well as RF_BLIND on a test set that is represented well by the training set. However, the effectiveness of random forests decreases when modeling on out-of-sample distributions. Nevertheless, these out-of-sample predictions are more accurate than the Vreman model, thus demonstrating a appreciable degree of generalizability.

5.3. Data-driven discovery of SGS stress model

In this section, we examine how the interpretability of random forests can be employed as a tool for model discovery.

Figure 14 presents feature importance scores extracted from RF_BLIND for τ_{1i}^{sgs}. For all three SGS stresses τ_{1i}^{sgs} shown, the highest scores are from $\partial\tilde{u}_1/\partial x_k$ and $\partial\tilde{u}_i/\partial x_k$ for three spatial dimensions. We employ this observation to formulate a sparse symbolic regression problem (see Eq.):

$$\frac{\tau_{ij}^{sgs}}{\overline{\rho} u'^2} = f_{ij}\left[G^{d=2}\left(\frac{\overline{\Delta}}{u'}\frac{\partial\tilde{u}_i}{\partial x_k}, \frac{\overline{\Delta}}{u'}\frac{\partial\tilde{u}_j}{\partial x_k}\right)\right] \tag{19}$$

where the independent variables consist of 2nd-order polynomial functions of the non-dimensionalized selected features. Eq. (19) is non-dimensionalized by density, filter width and initial root-mean-squared velocity to ensure dimensional consistency in the final model. This is essential for improving the dimensionality of this sparse symbolic regression problem. Since the dimensionality scales with n^d for n number of candidate variables, as discussed in Section 4.4, the employment of the feature importance score for reducing 30 candidate variables to six candidate variables results in a 25-fold reduction in dimensionality.

The following equations present the SGS model that resulted from applying sparse symbolic regression:

$$\tau_{11}^{sgs} \simeq \overline{\rho}\overline{\Delta}^2\left(0.116\frac{\partial\tilde{u}_1}{\partial x_1}\frac{\partial\tilde{u}_1}{\partial x_1} + 0.191\frac{\partial\tilde{u}_1}{\partial x_2}\frac{\partial\tilde{u}_1}{\partial x_2} + 0.207\frac{\partial\tilde{u}_1}{\partial x_3}\frac{\partial\tilde{u}_1}{\partial x_3}\right) \tag{20a}$$

$$\tau_{12}^{sgs} \simeq \overline{\rho}\overline{\Delta}^2\left(0.113\frac{\partial\tilde{u}_1}{\partial x_1}\frac{\partial\tilde{u}_2}{\partial x_1} + 0.102\frac{\partial\tilde{u}_1}{\partial x_2}\frac{\partial\tilde{u}_2}{\partial x_2} + 0.134\frac{\partial\tilde{u}_1}{\partial x_3}\frac{\partial\tilde{u}_2}{\partial x_3}\right) \tag{20b}$$

$$\tau_{13}^{sgs} \simeq \overline{\rho}\overline{\Delta}^2\left(0.119\frac{\partial\tilde{u}_1}{\partial x_1}\frac{\partial\tilde{u}_3}{\partial x_1} + 0.117\frac{\partial\tilde{u}_1}{\partial x_2}\frac{\partial\tilde{u}_3}{\partial x_2} + 0.109\frac{\partial\tilde{u}_1}{\partial x_3}\frac{\partial\tilde{u}_3}{\partial x_3}\right) \tag{20c}$$

$$\tau_{22}^{sgs} \simeq \overline{\rho}\overline{\Delta}^2\left(0.215\frac{\partial\tilde{u}_2}{\partial x_1}\frac{\partial\tilde{u}_2}{\partial x_1} + 0.135\frac{\partial\tilde{u}_2}{\partial x_2}\frac{\partial\tilde{u}_2}{\partial x_2} + 0.164\frac{\partial\tilde{u}_2}{\partial x_3}\frac{\partial\tilde{u}_2}{\partial x_3}\right) \tag{20d}$$

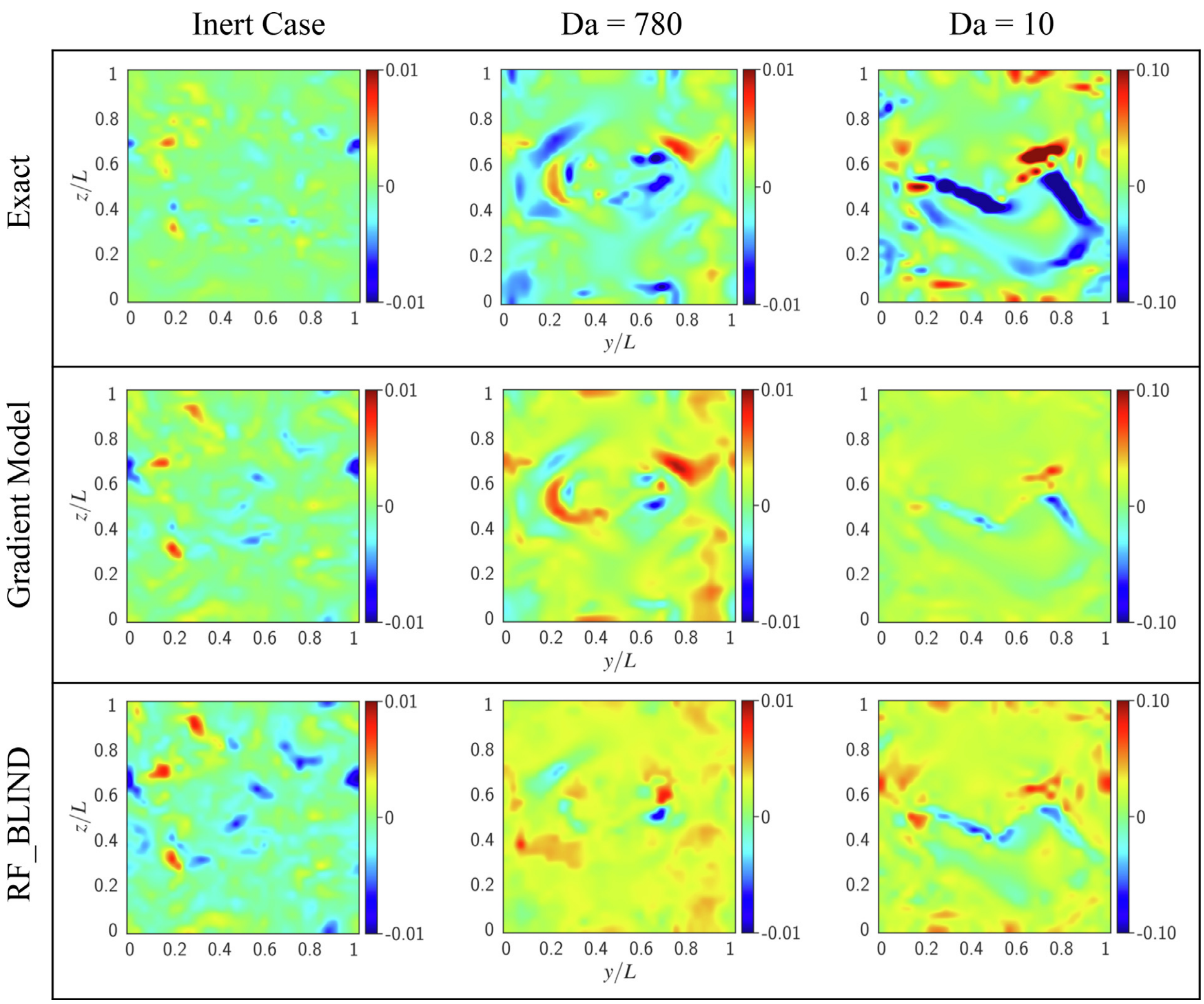


Fig. 12. Comparison of exact and modeled SGS stress $\tau_{12}^{sgs}/\overline{\rho}$ [m^{2}s^{-2}] at filter width $\overline{\Delta} = 4\Delta$ at axial location $x = 0$.

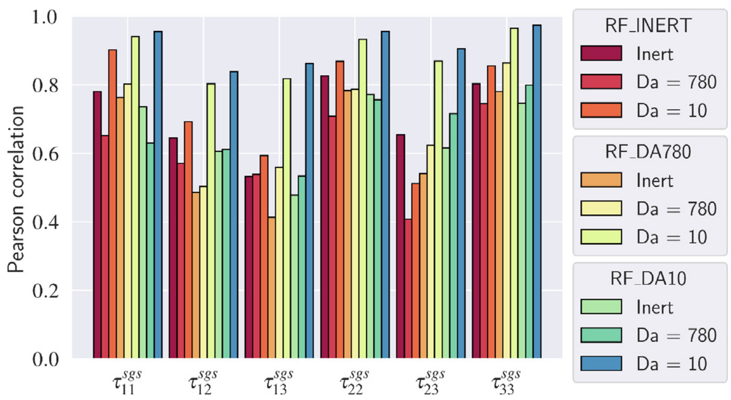


Fig. 13. Pearson correlations between exact and random forest modeled SGS stresses, from three different random forest regressors, for a single filter width $\overline{\Delta} = 2\Delta$.

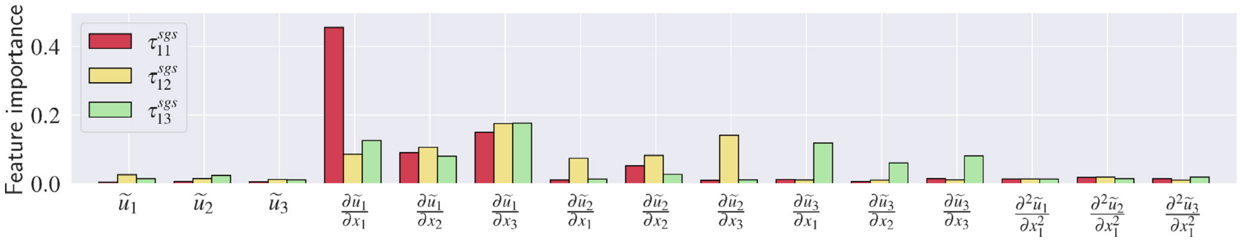
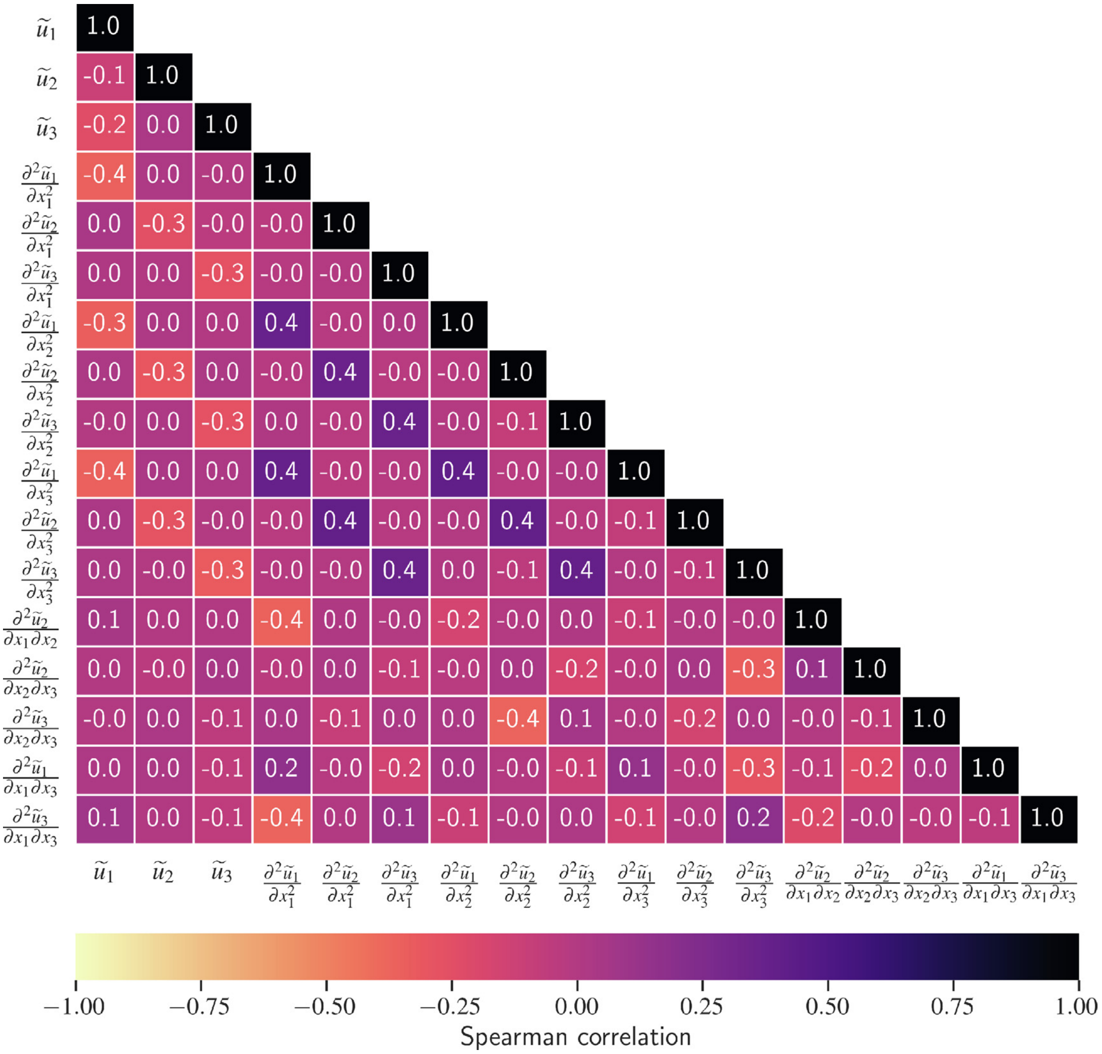


Fig. 14. Fifteen feature importance scores from RF_BLIND. The other fifteen features, with importance scores less than 0.02, are not shown for brevity.



	$\widetilde{u}_1$	$\widetilde{u}_2$	$\widetilde{u}_3$	$\frac{\partial^2\widetilde{u}_1}{\partial x_1^2}$	$\frac{\partial^2\widetilde{u}_2}{\partial x_1^2}$	$\frac{\partial^2\widetilde{u}_3}{\partial x_1^2}$	$\frac{\partial^2\widetilde{u}_1}{\partial x_2^2}$	$\frac{\partial^2\widetilde{u}_2}{\partial x_2^2}$	$\frac{\partial^2\widetilde{u}_3}{\partial x_2^2}$	$\frac{\partial^2\widetilde{u}_1}{\partial x_3^2}$	$\frac{\partial^2\widetilde{u}_2}{\partial x_3^2}$	$\frac{\partial^2\widetilde{u}_3}{\partial x_3^2}$	$\frac{\partial^2\widetilde{u}_2}{\partial x_1\partial x_2}$	$\frac{\partial^2\widetilde{u}_2}{\partial x_2\partial x_3}$	$\frac{\partial^2\widetilde{u}_3}{\partial x_2\partial x_3}$	$\frac{\partial^2\widetilde{u}_1}{\partial x_1\partial x_3}$	$\frac{\partial^2\widetilde{u}_3}{\partial x_1\partial x_3}$
$\widetilde{u}_1$	1.0																
$\widetilde{u}_2$	-0.1	1.0															
$\widetilde{u}_3$	-0.2	0.0	1.0														
$\frac{\partial^2\widetilde{u}_1}{\partial x_1^2}$	-0.4	0.0	-0.0	1.0													
$\frac{\partial^2\widetilde{u}_2}{\partial x_1^2}$	0.0	-0.3	-0.0	-0.0	1.0												
$\frac{\partial^2\widetilde{u}_3}{\partial x_1^2}$	0.0	0.0	-0.3	-0.0	-0.0	1.0											
$\frac{\partial^2\widetilde{u}_1}{\partial x_2^2}$	-0.3	0.0	0.0	0.4	-0.0	0.0	1.0										
$\frac{\partial^2\widetilde{u}_2}{\partial x_2^2}$	0.0	-0.3	0.0	-0.0	0.4	-0.0	-0.0	1.0									
$\frac{\partial^2\widetilde{u}_3}{\partial x_2^2}$	-0.0	0.0	-0.3	0.0	-0.0	0.4	-0.0	-0.1	1.0								
$\frac{\partial^2\widetilde{u}_1}{\partial x_3^2}$	-0.4	0.0	0.0	0.4	-0.0	-0.0	0.4	-0.0	-0.0	1.0							
$\frac{\partial^2\widetilde{u}_2}{\partial x_3^2}$	0.0	-0.3	-0.0	-0.0	0.4	-0.0	-0.0	0.4	-0.0	-0.1	1.0						
$\frac{\partial^2\widetilde{u}_3}{\partial x_3^2}$	0.0	-0.0	-0.3	-0.0	-0.0	0.4	0.0	-0.1	0.4	-0.0	-0.1	1.0					
$\frac{\partial^2\widetilde{u}_2}{\partial x_1\partial x_2}$	0.1	0.0	0.0	-0.4	0.0	-0.0	-0.2	-0.0	0.0	-0.1	-0.0	-0.0	1.0				
$\frac{\partial^2\widetilde{u}_2}{\partial x_2\partial x_3}$	0.0	-0.0	0.0	-0.0	0.0	-0.1	-0.0	0.0	-0.2	-0.0	0.0	-0.3	0.1	1.0			
$\frac{\partial^2\widetilde{u}_3}{\partial x_2\partial x_3}$	-0.0	0.0	-0.1	0.0	-0.1	0.0	0.0	-0.4	0.1	-0.0	-0.2	0.0	-0.0	-0.1	1.0		
$\frac{\partial^2\widetilde{u}_1}{\partial x_1\partial x_3}$	0.0	0.0	-0.1	0.2	-0.0	-0.2	0.0	-0.0	-0.1	0.1	-0.0	-0.3	-0.1	-0.2	0.0	1.0	
$\frac{\partial^2\widetilde{u}_3}{\partial x_1\partial x_3}$	0.1	0.0	-0.1	-0.4	0.0	0.1	-0.1	-0.0	0.0	-0.1	-0.0	0.2	-0.2	-0.0	-0.0	-0.1	1.0

Fig. 15. Spearman correlation matrix for selected features from RF_BLIND. Features with correlations less than 0.2 are not shown for brevity.

$$\tau_{23}^{sgs} \simeq \overline{\rho}\overline{\Delta}^2\left(0.123\frac{\partial\widetilde{u}_2}{\partial x_1}\frac{\partial\widetilde{u}_3}{\partial x_1}+0.116\frac{\partial\widetilde{u}_2}{\partial x_2}\frac{\partial\widetilde{u}_3}{\partial x_2}+0.134\frac{\partial\widetilde{u}_2}{\partial x_3}\frac{\partial\widetilde{u}_3}{\partial x_3}\right) \tag{20e}$$

$$\tau_{33}^{sgs} \simeq \overline{\rho}\overline{\Delta}^2\left(0.251\frac{\partial\widetilde{u}_3}{\partial x_1}\frac{\partial\widetilde{u}_3}{\partial x_1}+0.177\frac{\partial\widetilde{u}_3}{\partial x_2}\frac{\partial\widetilde{u}_3}{\partial x_2}+0.124\frac{\partial\widetilde{u}_3}{\partial x_3}\frac{\partial\widetilde{u}_3}{\partial x_3}\right) \tag{20f}$$

The resulting model can be rewritten as:

$$\tau_{ij}^{sgs} \simeq \overline{\rho}\overline{\Delta}^2\left(C_1\frac{\partial\widetilde{u}_i}{\partial x_1}\frac{\partial\widetilde{u}_j}{\partial x_1}+C_2\frac{\partial\widetilde{u}_i}{\partial x_2}\frac{\partial\widetilde{u}_j}{\partial x_2}+C_3\frac{\partial\widetilde{u}_i}{\partial x_3}\frac{\partial\widetilde{u}_j}{\partial x_3}\right) \tag{21}$$

where the resulting model coefficients $C_{\{1,2,3\}}$ range from 0.102 to 0.251. Eq. (21) is similar in form to the gradient model (Eq. (15)), but possesses three model coefficients instead of one. By observing that $C_{\{1,2,3\}}$ are of the same order of magnitudes, and collapsing the three coefficients by evaluating the average model coefficients, we recover the gradient model:

$$\tau_{ij}^{sgs} \simeq \overline{\rho}C_x\overline{\Delta}^2\frac{\partial\widetilde{u}_i}{\partial x_k}\frac{\partial\widetilde{u}_j}{\partial x_k} \tag{22}$$

where the model coefficient $C_x = 0.147$ is similar in value to the suggested model coefficient of 0.167 from Section 5.1. This result demonstrates that the employment of sparse symbolic regression, in conjunction with random forest feature importance can be employed to discover an algebraic expression, similar to the effective gradient model, for modeling subgrid-scale stresses in transcritical flows.

Since the present method relies on the random forest feature importance score, a statistical test must be employed to test for the effects of significant correlation amongst the features. If multiple features in the modeling basis are significantly correlated, they act as exchangeable surrogates for each other during the calculation of feature importance scores. This is similar to the phenomenon of multicollinearity in classical statistics [55]. Under such conditions, metrics such as the MDI are susceptible to correlation bias, and can generate erroneous importance scores [56,57]. As a note, almost all algorithms for estimating feature importance, including SHAP (Shapley additive explanations) [58] exhibit such correlation bias. As an alternative, Principal Component Analysis may be utilized to engender orthogonal bases for new features that are independent. However, these derived features are often difficult to ascribe physical meanings to, obfuscating their utility toward interpretability.

We utilize the Spearman correlation as a statistical test for evaluating the correlation amongst the features in the modeling basis. While the Pearson correlation is a statistical tool used for evaluating linear relationships, the Spearman correlation evaluates the monotonicity of variables in both linear and non-linear functions,

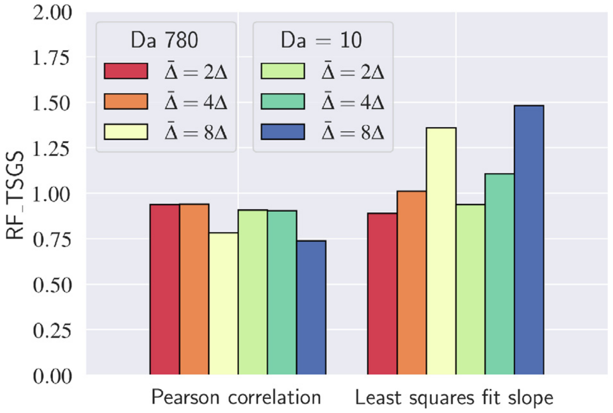


Fig. 16. Pearson correlation and slope from least squares fit between exact and random forest-modeled SGS temperature, for three different filter widths $\overline{\Delta}$.

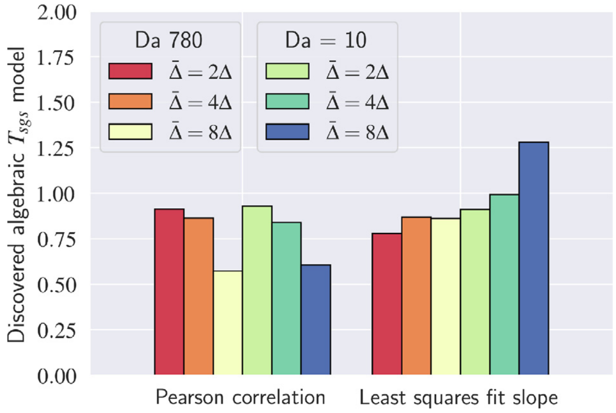


Fig. 17. Pearson correlation and slope from least squares fit between exact and algebraic-modeled SGS temperature.

i.e., whether the increasing or decreasing trend is being preserved. Spearman correlations of 1 and −1 correspond to a perfect monotonic relationship, while 0 corresponds to a negligible monotonic relationship. Figure 15 shows that Spearman correlations between different features from RF_BLIND are weak (between −0.4 and 0.4), which indicates that the feature importance score are not spurious.

5.4. SGS temperature models

In this section we extend the application of data-driven methods towards modeling SGS temperature. Figure 16 presents the Pearson correlation and slope from least squares fit between exact and random forest-modeled SGS temperature. High correlations (0.7–0.9) and slopes ranging from 0.7 to 1.5 are observed for all three filter widths, indicating good performance from the random forest SGS temperature model.

Unlike the random forests for modeling SGS stresses in Section 5.2, the feature importance scores from RF_TSGS do not provide physical insight due to the issue of multicollinearity, as T_{LES} and its gradients are used as features. In a reacting configuration, large temperature gradients are usually observed in a certain temperature range, and thus both these quantities can be significantly correlated. Nevertheless, a sparse symbolic regression problem can still be formulated without reducing the number of independent variables as the feature set for T_{sgs} is three times smaller than the feature set for τ_{ij}^{sgs}. We repeat the sparse symbolic regression procedure from Section 5.3:

$$T^{sgs} = f\left[G^{d=2}\left(\widetilde{T}_{LES}, l_{char}\frac{\partial \widetilde{T}_{LES}}{\partial x_k}, l_{char}^2\frac{\partial^2 \widetilde{T}_{LES}}{\partial x_k \partial x_k}\right)\right] \tag{23}$$

where the independent variables consist of 2nd-order polynomial functions of the features from RF_TSGS. Note that the independent variables are ensured to be dimensionally consistent with T_{sgs} by multiplying the gradients with a characteristic lengthscale l_{char}. This characteristic lengthscale can be chosen either as the filter width $\overline{\Delta}$ or a flame thickness δ_f. In the present study, δ_f can be extracted from the DNS by dividing the difference between flame and inert temperature by the maximum temperature gradient.

The following equations present the SGS temperature model that resulted from applying sparse symbolic regression:

$$T_{sgs} = \frac{\delta_F^2}{\widetilde{T}_{LES}}\left[0.00082\left(\frac{\partial \widetilde{T}_{LES}}{\partial x_1}\right)^2 + 0.00109\left(\frac{\partial \widetilde{T}_{LES}}{\partial x_2}\right)^2 + 0.00109\left(\frac{\partial \widetilde{T}_{LES}}{\partial x_3}\right)^2\right] \tag{24}$$

where $l_{char} = \delta_f$ has been chosen since a better fit is obtained when performing a least squares fit between the exact and modeled SGS temperature. By taking the average of the model coefficients, we obtain the algebraic expression:

$$T_{sgs} = \frac{C_T \delta_F^2}{\widetilde{T}_{LES}}\left(\frac{\partial \widetilde{T}_{LES}}{\partial x_k}\right)^2 \tag{25}$$

where $C_T = 0.001$.

Figure 17 presents the Pearson correlation and slope from least squares fit between exact and SGS temperature from the discovered algebraic T_{sgs} model. High correlations of approximately 0.9 are observed for $\overline{\Delta} = 2$ and $\overline{\Delta} = 4$, while a reasonable correlation of approximately 0.5 is seen for $\overline{\Delta} = 8$. The lower correlation compared to RF_TSGS is likely caused by the presence of the l_1-norm in Eq. (17), which encourages less significant terms to vanish from discovered model. Least squares fit slopes ranging from 0.8 to 1.3 are observed for all three filter widths.

6. Conclusions

DNS of inert and reacting transcritical LOX/GCH4 non-premixed mixtures under decaying turbulence were performed. Pressure and temperature were chosen to correspond to conditions in rocket combustors to examine conditions for which commonly-employed SGS are less matured. *A priori* analysis was conducted by comparing exact subgrid-scale stresses from Favre-filtered DNS data with algebraic and data-driven SGS models.

A priori analysis showed that the SGS stresses evaluated by Vreman SGS model correlated poorly with the corresponding exact terms. In contrast, good correlations are seen from the gradient SGS model. Results demonstrated a wide range of magnitude errors in the gradient model, which suggests that a dynamic gradient model approach is suited in *a posteriori* simulations. Random forests demonstrated high correlations when trained on datasets which are representative of the test sets, with reasonable predictions for the magnitude of subgrid-scale stresses. However, correlations were shown to decrease significantly when tested out-of-sample.

Sparse symbolic regression was performed to discover an algebraic expression for SGS stresses from non-linear transformations of velocity and its derivatives. The interpretability of random forests was demonstrated to reduce the dimensionality of the sparse symbolic regression problem by 25 times, by employing the feature importance score for variable selection. The derived algebraic expression was shown to be similar to the gradient model.

Sparse symbolic regression was also performed to evaluate subgrid-scale temperature, a term which emerges from filtering the non-linear real-fluids equation-of-state. The discovered algebraic expression demonstrated reasonable correlations and magnitudes when predicting subgrid-scale temperature. A random forest SGS temperature model was shown to perform better than the algebraic model.

Results demonstrate that random forests can perform as effectively or better as suitable algebraic models when modeling subgrid stresses, if trained on a sufficiently representative database. However, in the absence of such a database, this good performance is not replicated. Nevertheless, while the employment of random forests can provide insight into the discovery of subgrid-scale models through the feature importance score, as long as features are not significantly correlated. The present study should be complemented with an *a posteriori* study, and extended to other SGS closure terms that form chemical source terms and SGS scalar fluxes, to generate further insight.

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 NASA with award No. 80NSS C18C0207, the Department of Energy, National Nuclear Security Administration under award No. DE-NA0003968, and the German Research Foundation (Deutsche Forschungsgemeinschaft – DFG) in the framework of the Sonderforschungsbereich Transregio 40. Resources supporting this work are provided by the National Energy Research Scientific Computing Center, a U.S. Department of Energy Office of Science User Facility operated under contract No. DE-AC02-05CH11231.

References

[1] J. Bellan, Future challenges in the modelling and simulations of high-pressure flows, Combust. Sci. Technol. 192 (7) (2020) 1199–1218, doi:10.1080/00102202.2020.1719404.

[2] C. Chang, Y. Zhang, K.N.C. Bray, B. Rogg, Modelling and simulation of autoignition under simulated diesel-engine conditions, Combust. Sci. Technol. 113 (1) (1996) 205–219, doi:10.1080/00102209608935495.

[3] M. Zhu, K.N.C. Bray, B. Rogg, PDF modelling of spray autoignition in high pressure turbulent flows, Combust. Sci. Technol. 120 (1–6) (1996) 357–379, doi:10.1080/00102209608935581.

[4] P.C. Ma, H. Wu, T. Jaravel, L. Bravo, M. Ihme, Large-eddy simulations of transcritical injection and auto-ignition using diffuse-interface method and finite-rate chemistry, Proc. Combust. Inst. 37 (3) (2019) 3303–3310.

[5] J.C. Oefelein, Advances in modeling supercritical fluid behavior and combustion in high-pressure propulsion systems, AIAA Pap. 2019-0634 (2019), doi:10.2514/6.2019-0634.

[6] W.T. Chung, P.C. Ma, M. Ihme, Examination of diesel spray combustion in supercritical ambient fluid using large-eddy simulations, Int. J. Engine Res. 21 (1) (2020) 122–133, doi:10.1177/1468087419868388.

[7] A.W. Vreman, An eddy-viscosity subgrid-scale model for turbulent shear flow: algebraic theory and applications, Phys. Fluids 16 (10) (2004) 3670–3681.

[8] R.A. Clark, J.H. Ferziger, W.C. Reynolds, Evaluation of subgrid-scale models using an accurately simulated turbulent flow, J. Fluid Mech. 91 (1) (1979) 1–16.

[9] L.C. Selle, N.A. Okong'o, J. Bellan, K.G. Harstad, Modelling of subgrid-scale phenomena in supercritical transitional mixing layers: an a priori study, J. Fluid Mech. 593 (2007) 57–91.

[10] J. Smagorinsky, General circulation experiments with the primitive equations, Mon. Weather Rev. 91 (1963) 99–164.

[11] J. Bardina, J. Ferziger, W.C. Reynolds, Improved subgrid-scale models for large-eddy simulation, AIAA Pap. 1980-1357 (1980), doi:10.2514/6.1980-1357.

[12] U. Unnikrishnan, J.C. Oefelein, V. Yang, A priori analysis of subfilter scalar covariance fields in turbulent reacting LOX-CH4 mixing layers, AIAA Pap. 2019-1495 (2019), doi:10.2514/6.2019-1495.

[13] 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.

[14] M. Ihme, W.T. Chung, A.A. Mishra, Machine learning for combustion, under review Prog. Energy Combust. Sci. (2021).

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] J. Ling, A. Kurzawski, J. Templeton, Reynolds averaged turbulence modelling using deep neural networks with embedded invariance, J. Fluid Mech. 807 (2016) 155166.

[21] M. Raissi, P. Perdikaris, G. Karniadakis, Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations, J. Comput. Phys. 378 (2019) 686–707.

[22] K.N. Lakshmisha, B. Rogg, K.N.C. Bray, PDF modelling of autoignition in non-premixed turbulent flows, Combust. Sci. Technol. 105 (4–6) (1995) 229–243, doi:10.1080/00102209508907752.

[23] D.-Y. Peng, D.B. Robinson, A new two-constant equation of state, Ind. Eng. Chem. Fundam. 15 (1) (1976) 59–64, doi:10.1021/i160057a011.

[24] B.E. Poling, J.M. Prausnitz, J.P. O'Connell, The Properties of Gases and Liquids, McGraw-Hill, 2001.

[25] J. Zips, C. Traxinger, M. Pfitzner, Thermodynamic analysis and large-eddy simulations of LOx-CH_4 and LOx-H_2 flames at high pressure, AIAA Pap. 2018–4765 (2018).

[26] 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.

[27] E.W. Lemmon, M.O. McLinden, D.G. Friend, Thermophysical Properties of Fluid Systems in NIST Chemistry WebBook, NIST Standard Reference Database Number 69, National Institute of Standards and Technology, 2018.

[28] B. Franzelli, E. Riber, L.Y. Gicquel, T. Poinsot, Large eddy simulation of combustion instabilities in a lean partially premixed swirled flame, Combust. Flame 159 (2) (2012) 621–637, doi:10.1016/j.combustflame.2011.08.004.

[29] S.T. Chong, Y. Tang, M. Hassanaly, V. Raman, Turbulent mixing and combustion of supercritical jets, AIAA Pap. 2017-0141 (2017), doi:10.2514/6.2017-0141.

[30] J. Bellan, Direct numerical simulation of a high-pressure turbulent reacting temporal mixing layer, Combust. Flame 176 (2017) 245–262, doi:10.1016/j.combustflame.2016.09.026.

[31] W.K. Bushe, C. Devaud, J. Bellan, A priori evaluation of the double-conditioned conditional source-term estimation model for high-pressure heptane turbulent combustion using DNS data obtained with one-step chemistry, Combust. Flame 217 (2020) 131–151, doi:10.1016/j.combustflame.2020.03.015.

[32] S. Takahashi, Preparation of a generalized chart for the diffusion coefficients of gases at high pressures, J. Chem. Eng. Jpn. 7 (6) (1975) 417–420.

[33] T.H. Chung, M. Ajlan, L.L. Lee, K.E. Starling, Generalized multiparameter correlation for nonpolar and polar fluid transport properties, Ind. Eng. Chem. Res 27 (4) (1988) 671–679.

[34] A.M. Ruiz, G. Lacaze, J.C. Oefelein, R. Mari, B. Cuenot, L. Selle, T. Poinsot, Numerical benchmark for high-Reynolds-number supercritical flows with large density gradients, AIAA J. 54 (5) (2016) 1445–1460, doi:10.2514/1.J053931.

[35] 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 Pap. 2011–2886 (2011).

[36] W.T. Chung, A.A. Mishra, N. Perakis, M. Ihme, Data-assisted combustion simulations with dynamic submodel assignment using random forests, Combust. Flame 227 (2021) 172185, doi:10.1016/j.combustflame.2020.12.041.

[37] 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.

[38] 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

[39] H. Huo, V. Yang, Supercritical LOX/methane combustion of a shear coaxial injector, AIAA Pap. 2011-326 (2011).

[40] U. Unnikrishnan, J.C. Oefelein, V. Yang, Direct numerical simulation of a turbulent reacting liquid-oxygen/methane mixing layer at supercritical pressure, AIAA Pap. 2018–4564 (2018), doi:10.2514/6.2018-4564.

[41] T. Saad, D. Cline, R. Stoll, J.C. Sutherland, Scalable tools for generating synthetic isotropic turbulence with arbitrary spectra, AIAA J. 55 (1) (2017) 327–331.

[42] C. Bailly, D. Juves, A stochastic approach to compute subsonic-noise using linearized Euler's equations, AIAA Pap. 1999-1872 (1999).

[43] F. Williams, Descriptions of nonpremixed turbulent combustion, AIAA Pap. 2006-1505 (2006).

[44] M. Schoepplein, J. Weatheritt, R. Sandberg, M. Talei, M. Klein, Application of an evolutionary algorithm to LES modelling of turbulent transport in premixed flames, J. Comput. Phys. 374 (2018) 1166–1179, doi:10.1016/j.jcp.2018.08.016.

[45] H. Huo, V. Yang, Subgrid-scale models for large-eddy simulation of supercritical combustion, AIAA Pap. 2013-706 (2013).

[46] U. Unnikrishnan, X. Wang, S. Yang, V. Yang, Subgrid scale modeling of the equation of state for turbulent flows under supercritical conditions, AIAA Pap. 2017–4855 (2017).

[47] J. Ling, R. Jones, J. Templeton, Machine learning strategies for systems with invariance properties, J. Comput. Phys. 318 (2016) 22–35, doi:10.1016/j.jcp.2016.05.003.

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

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

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

[51] G. Louppe, L. Wehenkel, A. Sutera, P. Geurts, Understanding variable importances in forests of randomized trees, Adv. Neural Inf. Process. Syst. 26 (2013).

[52] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[53] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Ser. B Stat. Methodol. 58 (1) (1996) 267–288, doi:10.2307/2346178.

[54] E.S. Taskinoglu, J. Bellan, A posteriori study using a DNS database describing fluid disintegration and binary-species mixing under supercritical pressure: heptane and nitrogen, J. Fluid Mech. 645 (2010) 211254, doi:10.1017/S0022112009992606.

[55] T.K. Kumar, et al., Multicollinearity in regression analysis, Rev. Econ. Stat 57 (3) (1975) 365–366.

[56] A. Altmann, L. Toloşi, O. Sander, T. Lengauer, Permutation importance: a corrected feature importance measure, Bioinformatics 26 (10) (2010) 1340–1347.

[57] L. Toloşi, T. Lengauer, Classification with correlated features: unreliability of feature ranking and solutions, Bioinformatics 27 (14) (2011) 1986–1994.

[58] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst. 30 (2017) 4765–4774.