USING MACHINE LEARNING TO ANALYZE FACTORS DETERMINING CYCLE-TO-CYCLE VARIATION IN A SPARK-IGNITED GASOLINE ENGINE

Janardhan Kodavasal, Ahmed Abdul Moiz, Muhsin Ameen, and Sibendu Som
Energy Systems Division, Argonne National Laboratory, Argonne, IL, 60439, USA

ABSTRACT

In this work, we have applied a machine learning (ML) technique to provide insights into the causes of cycle-to-cycle variation (CCV) in a gasoline spark-ignited (SI) engine. The analysis was performed on a set of large eddy simulation (LES) calculations of a single cylinder of a four-cylinder port-fueled SI engine. The operating condition was stoichiometric, without significant knock, at a load of 16 bar brake mean effective pressure (BMEP), at an engine speed of 2500 revolutions per minute. A total of 123 cycles was simulated. Of these, 49 were run in sequence, while 74 were run in parallel. For the parallel approach, each cycle is initialized with its own synthetic turbulent field to generate CCV, as part of another work performed by us. In the current work, we used 3D information from all 123 cycles to compute flame topology and pre-ignition flow-field metrics. We then evaluated correlations between these metrics, and peak cylinder pressure (PCP) employing an ML technique called random forest. The computed metrics form the inputs to the random forest model, and PCP is the output. This model captures the effect of all inputs, as well as interactions between them owing to its decision-tree structure. The goal of this work is to demonstrate (as a first step) that ML models can implicitly learn complex relationships between pre-ignition flow-fields, flame shapes, and the eventual outcome of the cycle (whether a cycle will be a high or a low cycle).

INTRODUCTION

Computer simulation is critical in providing fundamental insights into the engine combustion process. Among the many complicated phenomena associated with engine combustion, cycle-to-cycle variation (CCV) [1–4] in engines is an undesirable phenomenon that leads to unstable operation including misfire and knock at extreme conditions. Understanding the causes of CCV is critical to developing strategies to mitigate it. CCV has been studied by various researchers in the past using both experimental and simulation methods.

Among simulation methods, the Reynolds averaged Navier-Stokes (RANS) turbulence modeling approach has been employed owing to its computational efficiency [5, 6]. The RANS approach inherently smoothes out fluctuations in velocity owing to the time-averaging done in this method, and this makes it difficult to capture the small-scale phenomena that lead to CCV. Large eddy simulation (LES) has been found to be more accurate than RANS-based methods to capture CCV in SI engines. Unlike RANS, LES directly simulates larger eddies which are influenced by the flow-field and are on the order of the grid-cell size. A sub-grid model is used to model small-scale eddies with an isotropic assumption. However, LES requires higher mesh resolutions than RANS.

Over the recent years several groups have performed LES of SI engines to develop a better fundamental understanding of CCV [7–19]. Fontanesi et al. [9] performed multi-cycle LES analysis (20 cycles) of a V-8 engine to understand the primary factors contributing to CCV. They found that inhomogeneity and variation in fuel distribution, turbulent energy, and velocity magnitude in the spark gap from cycle to cycle were major contributors to CCV. In a later work [10] based on these simulations, they report that spark plug orientation relative to the flow-field did not have a significant effect on CCV, however shifting the spark plug location towards the intake side improves the knock limit for their engine. Through a correlation analysis, they concluded that high peak cylinder pressure (PCP) cycles were characterized by flow velocities in the spark gap below an upper limit of 35 m/s, and found that fuel enrichment enhances ignitability. Truffin et al. [14] computed 30 LES cycles based on 200 cycles of experimental data and performed a linear regression analysis to understand the correlation between several factors representing the flow-field and fuel distribution, with respect to cycle indicated mean effective pressure (IMEP), in an attempt to explain which flow-field features determine high and low cycles. Tatschl et al. [15] focused on identifying the causes of CCV in a direct injection SI engine using LES. Granet et al. [16] used LES (with a thickened flame approach) to predict low and high cycles along with distinguishing an unstable point from a stable one. From such studies, it can be inferred that when the flame interacts with the surrounding flow-field, the flame topology changes, which affects the combustion pressure evolution causing a different PCP in each cycle. Thus, these aspects were taken into account in the analysis of the present work.

In a related work [18], we performed a univariate linear regression analysis to study the impact of flame topologies and velocity fields on PCP for 49 consecutive LES cycles of a single-cylinder of a four-cylinder SI port-fuel injected (PFI) engine. In the current work, we extend this analysis to obtain data from 74 parallel LES cycles since we have established in another related previous work that the overall statistics are preserved when the LES cycles are run in parallel with perturbation to the initial turbulent flow-field [20]. Additionally, in this work, we used a decision-tree-based approach called *random forest* [21] to build a model that correlates PCP to various metrics representing flame topology and the flow-field. Based on its formulation, the random forest inherently captures interactions between inputs (the metrics) without us having to explicitly formulate interaction terms as is the case with multi-variate linear regression, for example.

With the increasing availability of supercomputing and high-performance computing resources, and innovative techniques to parallelize sequential runs to achieve high throughput [20], it is now possible to run hundreds to thousands of engine simulations in parallel and generate a large number of samples or realizations [22, 23] in a matter of days. Detailed simulations also allow better analysis of novel combustion modes [24–33]. Further, improvements to code scalability allow each simulation to be run on hundreds to thousands of cores [34], enabling high-fidelity LES calculations that produce more realistic spatial and temporal data. Considering both the volume and fidelity of data that will be produced from simulations and experiments going forward, we anticipate that ML techniques will play a significant role in data analysis not just for engine design but in general for various engineering applications [35–39].

In the case of simulations for CCV we believe that ML can help correlate high and low cycles (characterized by high or low values of PCP) with representations of flame topologies and flow structures. Random forest is one of the most widely used ML algorithms, because of its good applicability on datasets ranging from small to large, better interpretability (i.e. the ability to understand the model results) compared to neural networks methods (which are more abstract and hard to understand), and its robustness arising from ensembling several *weak* learners (a weak learner is a model that is intentionally exposed to only a small subset of the data) thus not overfitting to data. An overfitted model *memorizes* the training data, and does not generalize well on new data. Further, random forest allows input data to be a mix of continuous data (eg. average velocity in a region) and categorical data (eg. spark plug manufacturer), due to its if-else structure.

In this work, we examine correlations between various factors relating to the flow-field and flame geometry with respect to peak cylinder pressure (PCP) over the 123 simulated cycles using a random forest model.

The objectives of this work are three-fold:

- Show that the parallel perturbation method generates cycles that demonstrate similar correlations as the sequential cycles, in terms of flame topologies and velocity fields that result in high/low cycles, to justify co-mingling of the datasets for analysis.
- Demonstrate the application of the random forest approach as an analysis tool to determine the relative importance of various flame topology and pre-ignition flow-field metrics in terms of influencing PCP in an SI engine.
- Based on the most important metrics, build and test a random forest model to predict high and low cycles based on flame topology and pre-ignition flow-field.

We first start with a brief description of the engine setup, followed by the CFD model setup information. We then present the results along with some discussion, and end with a summary of and conclusions from this work.

NOMENCLATURE

Abbreviations

BMEP	brake mean effective pressure
CCV	cycle-to-cycle variation
CFD	computational fluid dynamics
CR	compression ratio
FGW	fixed geometry with wastegate
GB	gigabyte
TB	terabyte
IMEP	indicated mean effective pressure
IVC	intake valve closure
LES	large eddy simulations
ML	machine learning
PCP	peak cylinder pressure
PFI	port fuel injection
RANS	Reynolds averaged Navier-Stokes
RPM	revolutions per minute
SI	spark ignition
SOI	start of injection
TDC	top dead center

ENGINE AND SIMULATION SETUP

Engine specifications

The engine modelled in this work is an inline four-cylinder SI-PFI engine. The engine specifications and operating condition are listed in Table 1. The geometric compression ratio (CR) is 9.8 and an effective CR of 9.72 is prescribed in the simulations. The operating condition studied is stoichiometric, and the fuel flow was measured to be 34.9 mg/cycle/cylinder. This condition represents a stable operating point, without partial burn or any significant knock. The brake mean effective pressure (BMEP) for this point is 16 bar, and the engine was operated at 2500 revolutions per minute (rpm).

Table 1. Engine specifications and operating condition

Engine specifications	
Engine type	SI
Number of cylinders	4
Displacement (cm^3)	1368
Bore (mm)	72
Stroke (mm)	84
Compression ratio	9.8:1
Fuel Injection System	PFI
Turbocharger	FGW

Operating condition	
Speed (rpm)	2500
BMEP (bar)	16
Relative Air to Fuel Ratio	1
Fueling (mg/cycle/cylinder)	34.9
Ignition timing	711˚

CFD model setup

The numerical simulations presented in this study were performed using the commercial CFD code CONVERGE v2.3 [40]. The code adopts a cartesian cut-cell approach to mesh generation, which is done at run-time. There is a user-supplied base mesh size, which represents the size of the largest cells in the domain (here 2.8 mm), and additionally there are provisions to increase the resolution by *embedding* at various levels over the base mesh size in specific regions of interest during certain portions of the simulation, such as in the spark gap during the spark event (0.175 mm), or at the nozzle exit during injection (0.35 mm). Additionally, there is also the ability to perform on-the-fly adaptive mesh refinement based on local gradients in fields such as temperature and velocity which reduces the grid to 0.35 mm when active in our setup. The peak cell count occurs during intake, and is 7 million cells, and the cell count at combustion top dead center is 1 million cells. The simulations were run on 96 cores, with a single MPI rank per core and no threading. The system used was an Intel Sandy Bridge Pentium Xeon cluster, called Blues at Argonne. Each node has 64 GB RAM and a clock speed of 2.6 GHz. There are 16 cores per node. A single cycle (720 degrees) takes on the order of 36 hours on this system. Each cycle generates on the order of 2 TB of data There are various models available within CONVERGE to simulate spray, liquid drop dynamics, turbulence, combustion, etc. In this study, a conventional Eulerian-Lagrangian Discrete Droplet Model [41] has been chosen. The liquid spray is modeled as a dispersed phase in the Lagrangian framework and the surrounding gas is modeled as a continuous phase in the Eulerian framework. The Kelvin-Helmholtz Rayleigh-Taylor spray break-up model is adopted to model droplet breakup and atomization [42]. A No Time Counter collision model is used to simulate the collision outcomes [25]. Turbulence is modeled using the dynamic structure LES model [44]. In this model, a sub-grid scale kinetic energy transport equation is applied to model the energy flow from the resolved scales to the sub-grid scales, based on an energy flow budget between resolved and the sub-grid scales. We use a second-order in space, and first-order in time discretization scheme.

The G-equation combustion model, which is a non-detailed-chemistry level-set based model, was used to simulate the flame kernel growth and flame propagation (combustion). Williams [45] proposed the G-equation model based on flamelet modeling assumptions. The flame front is defined as the interface between the unburned and burned gases. Peters [46] also used the level-set method and the G-equation model to simulate premixed turbulent combustion with corrugated flamelets or thin flame fronts for RANS simulations. Pitsch [47] in 2002 provided the formulation for the G-equation in the context of LES. The parameter G is a scalar and it represents the distance from the flame front. The flame front position is at G = 0, and points (or cells) in the burned region have a positive value of G, while points (or cells) in the unburned region have a negative value of G.

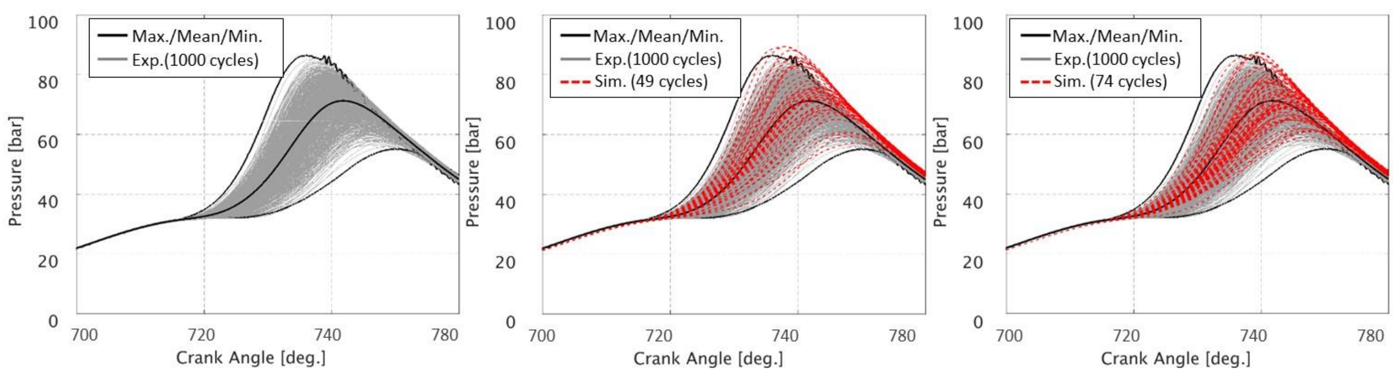

Figure 1: Experimental cycles (left), simulated consecutive cycles (middle), simulated parallel cycles (right)

RESULTS AND DISCUSSION

Figure 1 shows experimental 1000 cycles (left), consecutively run 49 LES cycles (middle), and parallel-run 74 LES cycles (right). The highs and lows of PCP were not seen to follow any order and were found to be stochastic in nature. In the case of the parallel approach, we see from Figure 1 that the spread of the cycles, along with the minimum-mean-maximum cycles, is very well reproduced compared to the consecutive cycles and experiments. More details on the parallel perturbation method are presented in Ref. [20]. This method significantly reduces overall turnaround time by running in parallel all the cycles on a large number of processors. From both our previous works with consecutive cycles [18] and parallel run cycles [20], we see that the coefficient of variance of parameters like IMEP and PCP are in good agreement with experiments.

In our related work [18], we determined that high cycles are typically characterized by flame volumes that are more squished in the Z-direction (piston-to-head direction), and more centered along the X-axis (which represents the direction of the mean flow caused by tumble). Figure 2 shows the axes orientation with respect to the CFD model, for reference.

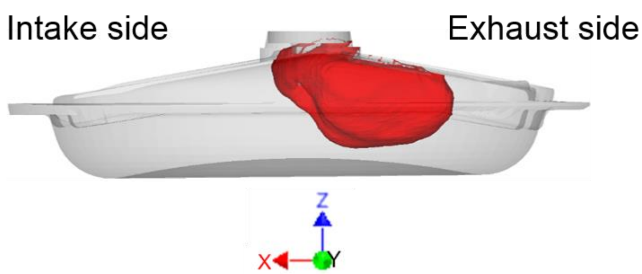

Figure 2: Snapshot showing axes orientation with respect to the CFD model

Low cycles are characterized by more elongated flame volumes in the Z-direction, and also flame volumes that are more offset in the negative X-direction (towards the exhaust). The parallel cycle approach captures these same effects, and representative high and low cycles from both the parallel approach and consecutive cycles are shown in Figure 3. In Figure 3, the flame images are shown at the same crank angle location of 722 deg (11 degrees after ignition). The choice of this angle is subjective, and just represents a point when all the cycles have a well-developed flame volume. Every cycle is at a different mass percent burned point at 722° due to cycle-to-cycle variation. The flame images are from individual cycles and are not averaged. We have evaluated several cycles to make sure the patterns seen in the figures shown here are consistent over several cycles and are indeed true patterns. In our related work [18] we show the several cycles we checked for consistency in flame volume patterns. Also in that work [18], we looked into aspects of heat transfer from the cylinder head, piston, and liner surfaces and found that high cycles have a greater heat transfer compared to low cycles and it is not that low cycles are low in magnitude due to higher heat losses.

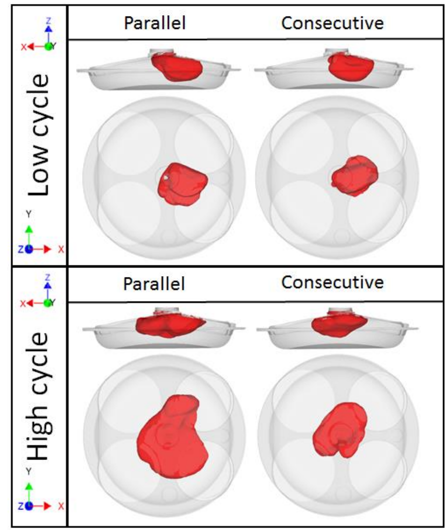

Figure 3 A single pair of one low cycle and one high cycle at 722 deg. CA showing similar physical behavior when simulated using the parallel and consecutive approaches

Even when we look at the flame volumes for the cycles at the same mass percent burned (3%), as shown in Figure 4, we still note that the flame shapes are qualitatively similar for the parallel and consecutive cycles, when looking at low and high cycles respectively from both simulation approaches.

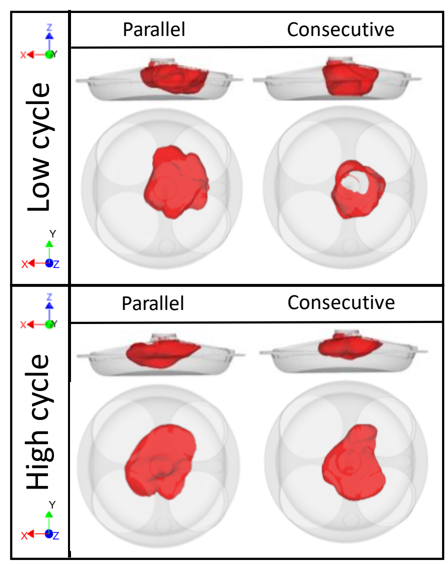

Figure 4: A single pair of one low cycle and one high cycle at a common mass percent burned of 3% showing similar physical behavior when simulated using the parallel and consecutive approaches (similar to the behavior observed with them at the same CA)

This provides us with further confidence in the validity of the parallel perturbation approach, beyond its ability to reproduce global statistics obtained from the consecutive cycles. Based on this, we combine the data from the consecutive and parallel cycles and treat them as a single dataset going forward with the analysis in this paper, since this allows us to run a more robust analysis on 123 points, versus the regression analysis we have performed on the 49 consecutive cycles in our related work [18].

A total of ten metrics representing flame topology at 722˚, as well as velocity metrics representing the pre-ignition flow-field at 710˚ were chosen, and their correlation with respect to each other, as well as PCP is shown in Figure 5, in terms of pair-wise Pearson correlation. This correlation plot is generated using the ML pipeline AutoMOMML [48]. Given two vectors X and Y, the Pearson correlation coefficient $\rho_{X,Y} \in [-1.0,\ 1.0]$ is given by the following equation:

$$\rho_{X,Y} = \frac{covariance(X,Y)}{\sigma_X \sigma_Y}, \qquad (1)$$

where σ represents standard deviation.

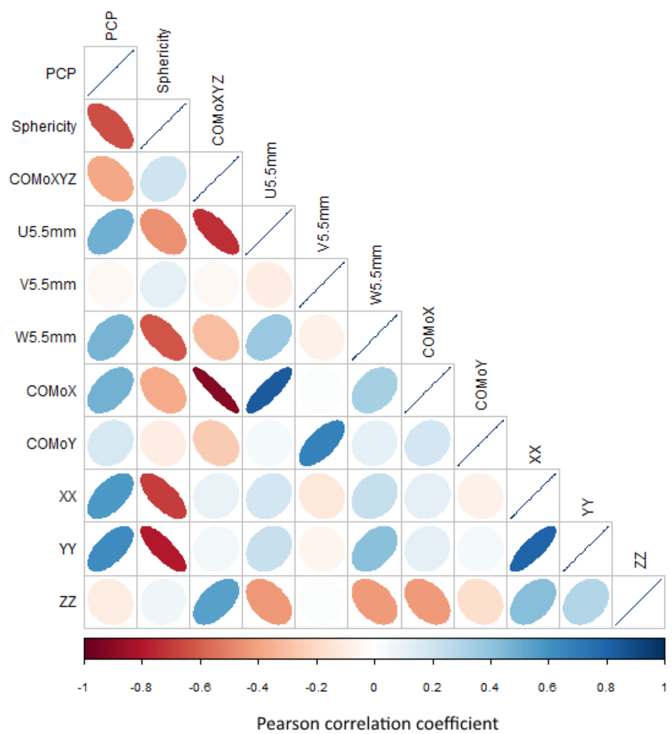

Figure 5: Pearson correlation between ten metrics chosen to represent flame topologies, pre-ignition velocity flow-fields, and PCP

In Figure 5, pairs which show a high positive correlation exhibit dark-blues shades and are more elliptical and those which show a high inverse correlation are more towards the red spectrum and are also more elliptical. Pairs which are uncorrelated are more circle-like and have more of a white shade. The shape for a perfectly correlated (both positive and negative) parameter-pair is a line. Sphericity represents the sphericity of the flame volumes at 722˚, U5.5mm, V5.5mm, and W5.5mm represent the average pre-ignition velocity, at 710˚, in the X-, Y-, and Z-directions respectively, in a 5.5 mm radius sphere centered in the spark gap. COMoX, COMoY, and COMoXYZ represent the offset in the center of mass of the flame volumes at 722˚ in the X- and Y- directions, and absolute 3-D offset respectively, with reference to the ignition location. The metrics XX, YY, and ZZ represent the spans of the flame volumes at 722˚ in the X-, Y-, and Z-direction respectively, in other words, the maximum minus minimum values of X, Y, and Z coordinates respectively of these flame volumes at 722˚.

From Figure 5 we see, for example, that there is positive correlation between U5.5mm and COMoX. Overall, U5.5mm values are negative, since the pre-ignition mean flow field (at 710˚) in the spark-region brought about by tumble, is from the intake to the exhaust. This means that a stronger pre-ignition mean flow in the negative X-direction (from intake to exhaust) in the spark-region at 710˚ causes the flame to grow eccentrically in the negative X-direction (towards the exhaust). This observation is in line with the works of Truffin et al. [14], and also Fontanesi et al. [9], and was noted in our related work [18, 19]. Similarly, we see a correlation (although weaker) between V5.5mm and COMoY. A more detailed discussion on the correlation between pre-ignition flow-fields, flame topologies, and PCP is provided in Ref. [18].

One important factor to note from Figure 5 is that PCP (which denotes whether a cycle is a high- or low-cycle), is not strongly correlated to any one of these metrics alone, indicating that a plain univariate linear regression analysis might not be completely sufficient in describing PCP as a function of any of these metrics. Even if we were to use multivariate regression with higher order and interaction terms, we would have to explicitly formulate these additional terms.

In order to implicitly account for the interactions between various metrics that influence PCP, we adopt the random forest ML method. Random forest uses an ensemble of decision trees. Each decision tree is trained on the same *number* of samples as in the training data supplied to the random forest model, however, the individual samples are drawn randomly, *with replacement* from the training data, so that every tree is trained on potentially a different "bootstrapped" sample from the training data. Further each tree is built using a random subset of the total number of features, usually equal to the square root of the total number of features (rounded up). This way, random forest models overcome the over-fitting problem where the model is overly complex and is too closely fit to a limited set of data points, resulting in poor prediction accuracy of unseen data. As a consequence of its decision tree structure, the random forest method can also output the ranks of the various features in terms of their importance in affecting the target variable, in our case, PCP. This functionality of random forest is used in the present work and the feature importance is compared with univariate linear regression; a higher feature importance should correspond to higher fit value (higher R^2) between that feature and PCP. Again it is important to note that this would just be a qualitative comparison, as univariate linear regression ignores the influence of other features on the target, and also potential non-linear interactions between features. Figure 6 shows a sample decision tree, which is constructed based on a sub-sample consisting of four features (or metrics representing flame shape and pre-ignition velocities), trained on a bootstrapped sub-sample of the 123 cycles of data. Each decision tree is considered a “voter”.

The random forest model is an ensemble of multiple decision trees or voters. Given features corresponding to a new cycle, that none of the decision trees in the random forest model has been trained on, the random forest averages the estimations (or votes) from all the decision trees to make a prediction on PCP.

In constructing each decision tree, the information space (based on the samples the tree is trained on) is split into many domains and the number of splits is decided by the decision tree algorithm based on the concept of information entropy. In other words, splitting of the feature matrix of the data set is stopped when the algorithm cannot add any more information to each subset. In Figure 6, each branching point (node) is a split at which a yes/no (or if/else) decision is made, according to which the information space is divided. Splits are done starting from the most important features, in other words features over which there is maximum variance in the data. The colored nodes in Figure 6, contain the averaged value of PCP in that split domain and the percentage of data points in that domain. Given a set of values for features, one can traverse through this decision tree to arrive at a leaf node (terminal node), which is what this decision tree thinks the value of PCP will be. This is considered the “vote” of this particular decision tree for a given set of new values for the features. The random forest works by averaging the votes over all its constituent decision trees.

KNIME [49] is an open-access data analytics, reporting, and integration platform. We used this platform to create a pipeline going from the input data (123 cycles) to the ultimate random forest model, along with model performance evaluation. The modular structure of this workflow is shown in Figure 7. At the outset, the pipeline reads in the data (123 rows representing the simulated cycles, with each row containing the values of the features we choose to build the model based on, as well as the

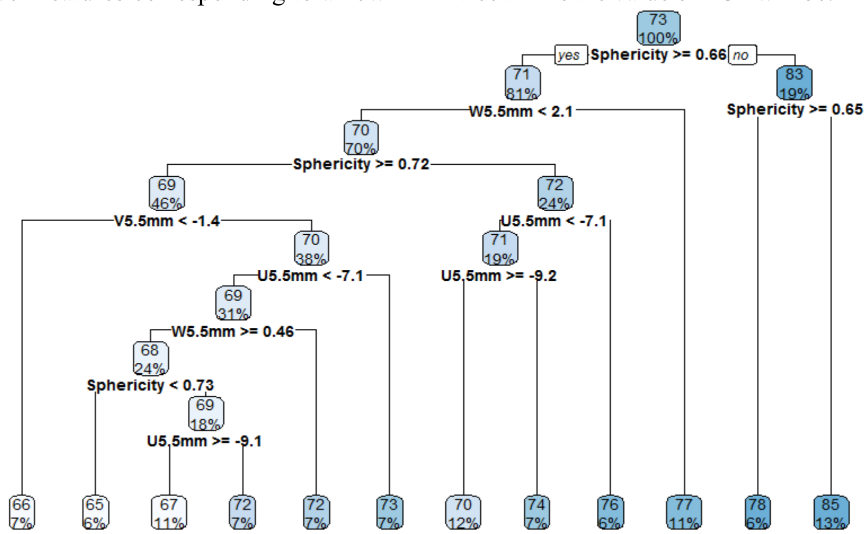

Figure 6: Sample decision tree showing the partitioning of the data based on a random subset of four features of the ten input features (4 = √10 rounded up) representing flame shape and pre-ignition velocities, using an if/else structure; the random forest is an ensemble of multiple decision trees

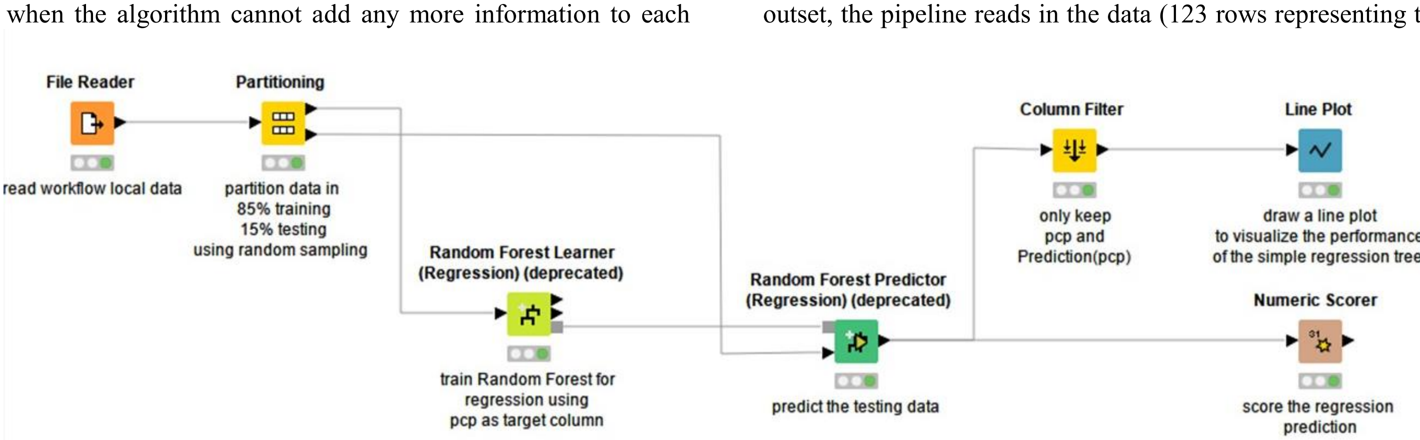

Figure 7: KNIME workflow for building a random forest model

PCP value), and then randomly partitions the data into a training set (85% of the cycles), and test set (the other 15% of the cycles).

The test set is not used to train any of the decision trees in the random forest model. Each decision tree of the random forest model is trained on a random bootstrapped subsample of the 85% training data, and a random sample of the chosen features (inputs of metrics) considered.

The random forest predictor module then predicts the PCP for the 15% test data, using the test data, and the trained random forest model from the previous step as inputs. The line plot module is used to plot the predicted PCP versus the actual PCP for the test dataset (the Column Filter is just used to keep only the required elements to be plotted). The numeric scorer is used to "score" the accuracy of the prediction in terms of a percentage error in the predictions on the test dataset. The random forest model also outputs a ranking of features (or inputs) based on their relative importance in predicting PCP, and this ranking is shown in Figure 8, when all the ten features are used to build this model. Note that there is no direct correlation between the ranking of the splits in Figure 6 which shows a single example tree, and the ranking shown in Figure 8, as the ranking in Figure 8 is based on a collection of several trees, and gives the true representation of feature importance. Based on our related work [18] and after performing the feature importance study shown in Figure 8 using all ten features, as a next step, we chose four features—Sphericity, U5.5mm, W5.5mm, and COMoX as input metrics to build a random forest predictor. The reason we don't choose all the metrics is because several metrics are correlated, such as COMoXYZ, and COMoX. Further, the metrics XX, YY (in spite of ranking highly in the feature importance) are highly correlated with the mass percent burned of individual cycles at 722˚ (since cycles with high PCP would have burned to a greater extent at 722˚, and would have a higher value of XX and YY due to this), so it is not really fair to include these as predictors. It is also generally a good practice when building ML models to reduce complexity in terms of features to avoid potential overfitting. ZZ was not included as a predictor, since there are two competing effects on ZZ at 722˚ as shown in our related work [18]—on the one hand ZZ would tend to be higher for cycles that have a bigger flame volume, and on the other hand, ZZ is also preferentially lower for a given mass fraction burned due to the squishing effect of the flame volume associated with high cycles.

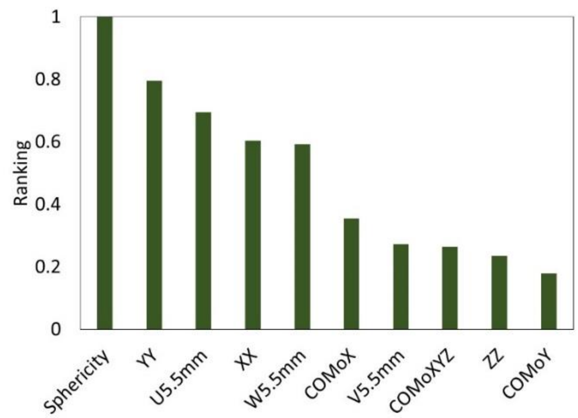

Figure 8: Feature (or input) importance for predicting PCP using random forest

Figure 9 shows the univariate linear correlations for these four important features with respect to PCP. The numbers in the plot across the solid line correspond to the cycle number (and the respective cycle data is plotted on X- and Y-axis). We note that all these features have a relatively strong correlation to PCP. The scatter about the linear fit line is a consequence of the univariate linear model capturing only one feature at a time, which is not the case with the random forest approach, which accounts for all the features, as well as interaction between features. Using the predictor functionality of the pipeline, the 15% dataset was predicted based on using these four input features for training on the 85% dataset.

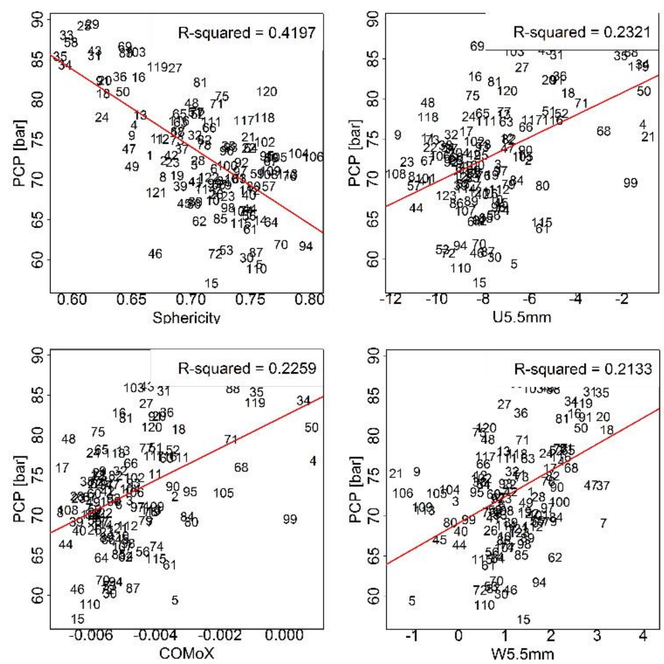

Figure 9: Linear fit of PCP to high-ranking inputs, with scatter around the fit shown

Figure 10 shows the actual PCP overlaid on the PCP predicted on the 15% test data. The X-axis shows the cycle numbers from the randomly chosen 15% test dataset. The highs and lows are well-predicted for these cycles, and the R^2 value of the fit is 0.59. This indicates that the ML model has sufficiently "learned" the correlations between flame topologies and pre-ignition velocity flow-fields, and PCP.

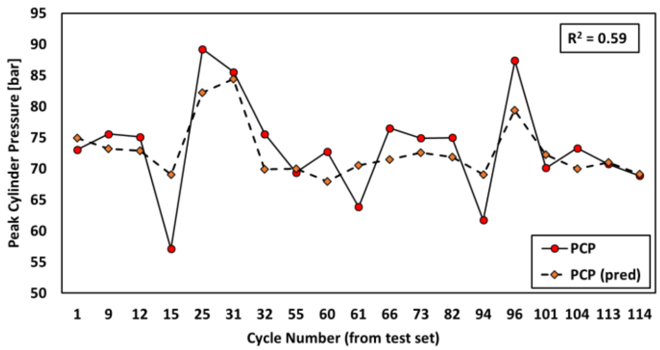

Figure 10: Predictions on the 15% test set considering four metrics—sphericity, U5.5mm, W5.5mm, and COMoX

At this stage, it is worth highlighting that to generate these metrics we require expensive LES, so as of now the random forest model constructed in this work is not meant to completely replace the more computationally expensive CFD simulations, rather the objective here is to demonstrate that ML models can implicitly learn complex relationships between flow-fields, flame shapes, and the eventual outcome (in this case PCP) of the cycle. One eventual goal would be the construction of an ML model that could (given enough cycles of data) learn the correlations between initial flow-fields at say intake valve closure (IVC) and high/low cycles, thus being able to predict PCP purely based on the initial conditions.

As a first step in this direction, for the dataset analyzed, the prediction capability of this random forest model was explored by choosing as inputs two important features that represent the pre-ignition flow-field in the spark region. These features can be potentially obtained from non-reacting simulations up to ignition. These are the X- and Z- direction velocities at 710°—U5.5mm and W5.5mm. One could envision that the random forest model could then be used in the place of expensive post-ignition simulation to predict whether a cycle would be a high or low cycle, purely based on the pre-ignition flow-field. Based on running multiple non-reacting simulations up to ignition using the parallel perturbation method, one could estimate the coefficient of variation in PCP for a given engine design or operating condition. Figure 11 shows the predictions from the random forest model (on the 15% test data) using only these two features as inputs, trained on 85% of the 123-cycle data. Figure 11 shows a weaker prediction when we choose these parameters (with an R^2 of 0.36); however overall highs and lows are still qualitatively predicted. We intentionally did not include COMoX and sphericity in this model, as to obtain these quantities, we would in any case need to do the reacting portion of the simulation.

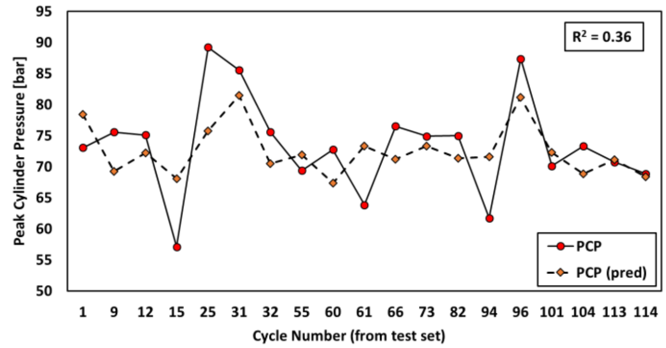

Figure 11: Predictions on the 15% test set considering two pre-ignition flow metrics—U5.5mm, and W5.5mm.

CONCLUSIONS

In this work we used an ML model (random forest) to learn the relationship between metrics representing flame topology and pre-ignition flow-fields, and PCP using a dataset comprised of multiple cycles of LES calculations performed on a spark-ignited gasoline engine with cycle-to-cycle variation.

1. We demonstrated the validity of comingling data from parallel-run and serial-run LES cycles to perform an analysis of flame topology and pre-ignition velocity fields in terms of their effect on PCP.
2. Using the combined dataset of 123 cycles we built a random forest model that learned relationships between flame topology and pre-ignition velocities near the spark, and PCP.
3. Based on the feature importance obtained from the random forest model built using ten metrics, and analysis from our related work [18], we reduced model complexity to four input features: sphericity, COMoX, U5.5mm, and W5.5mm. The model built using these four metrics was able to qualitatively capture high and low cycles in a test dataset.
4. We also built a version of the model where we only included the pre-ignition features U5.5mm and W5.5mm as inputs. The reasoning behind this is that these features can be generated by just running a non-reacting perturbed simulation up to ignition timing, as compared to sphericity and COMoX, which would require the combusting portion of the simulation to be run as well. This model, though weaker in prediction compared to when we include sphericity and COMoX, does a reasonable job in qualitatively predicting high and low cycles. It is envisioned that with more training data, the random forest model can better learn the relationship between pre-ignition flow-fields and PCP.

This work is a proof-of-concept of the capability of ML models to learn complex correlations from 3D CFD data. The ultimate goal would be to develop an ML model trained on tens to hundreds of thousands of CFD simulations spanning a sparse sampling of a large design and operating space with features such as engine design and operating parameters, initial intake and in-cylinder flow-fields, that could predict cycle-to-cycle variation as a function of initial conditions, operating conditions, and design features. Such a model could then be used to probe the multidimensional design and operating space at a much higher resolution (than the sparsely sampled expensive CFD runs it was trained on) to discover optimum design and operating strategies.

ACKNOWLEDGMENTS

The submitted manuscript has been created by UChicago Argonne, LLC, Operator of Argonne National Laboratory ("Argonne"). Argonne, a U.S. Department of Energy Office of Science laboratory, is operated under Contract No. DE-AC02-06CH11357. The U.S. Government retains for itself, and others acting on its behalf, a paid-up nonexclusive, irrevocable worldwide license in said article to reproduce, prepare derivative works, distribute copies to the public, and perform publicly and display publicly, by or on behalf of the Government. This research was funded by DOE's Office of Vehicle Technologies, Office of Energy Efficiency and Renewable Energy under Contract No. DE-AC02-06CH11357. The authors wish to thank Gurpreet Singh and Leo Breton, program managers at DOE, for their support. We gratefully acknowledge the computing resources provided on "Blues," a high-performance computing cluster operated by the Laboratory Computing Resource Center at Argonne National Laboratory. We also acknowledge Dr. Prasanna Balaprakash from the Mathematics and Computer Science division at Argonne National Laboratory for advice on machine learning, and Dr. Riccardo Scarcelli, also from Argonne, for CFD case setup advice. We thank Mohsen Mirzaeian and Prof. Federico Millo of Politecnico di Torino for providing experimental engine data, and Drs. Syed Wahiduzzaman, Navin Fogla, and Michael Bybee of Gamma Technologies Inc. for valuable discussions on flame topology and flow-field effects on CCV. We acknowledge Convergent Science for CFD software licenses and application support.

REFERENCES

1. Patterson, D. J., 1966. "Cylinder pressure variations, a fundamental combustion problem". *SAE Technical Paper*, *Paper No. 660129*.
2. Ozdor, N., Dulger, E. and Sher, E., 1994. "Cyclic variability in spark-ignition engines. A literature survey". *SAE Technical Paper*, *Paper No. 940987.*
3. Srinivasan, K. K., Krishnan, S. R. and Qi, Y., 2013. "Cyclic Combustion Variations in Dual Fuel Partially Premized Pilot-Ignited Natural Gas Engines". *Journal of Energy Resources Technology,* ***136***(1), 012003-012003-10.
4. Zhou, L., Shao, A., Hua, J., Wei, H. and Feng, D., 2018. "Effect of retarded injection timing on knock resistance and cycle to cycle variation in GDI engine". *Journal of Energy Resources Technology,* ***Accepted Manuscript****,* doi: 10.1115/1.4039322.
5. Scarcelli, R., Sevik, J., Wallner, T., Richards, K., Pomraning, E. and Senecal, P. K., 2015. "Capturing cyclic variability in EGR dilute SI combustion using multi-cycle RANS". In Proceedings of 2015 ASME Internal Combustion Engine Division Fall Technical Conference.
6. Scarcelli, R., Richards, K., Pomraning, E., Senecal, P. K., Wallner, T. and Sevik, J., 2016. "Cycle-to-cycle variations in multi-cycle engine RANS simulations". *SAE Technical Paper, Paper No. 2016-01-0593.*
7. Moureau, V., Barton, I., Angelberger, C. and Poinsot, T., 2004. "Towards large eddy simulation in internal-combustion engines: simulation of a compressed tumble flow". *SAE Technical Paper*, *Paper No. 2004-01-1995.*
8. Vermorel, O., Richard, S., Colin, O., Angelberger, C., Benkenida, A. and Veynante, D., 2009. "Towards the understanding of cyclic variability in a spark ignited engine using multi-cycle LES". *Combustion and Flame,* ***156***(8), pp. 1525–1541.
9. Fontanesi, S., Paltrinieri, S., Tiberi, A. and D'Adamo, A., 2013. "LES multi-cycle analysis of a high performance GDI engine". *SAE Technical Paper*, *Paper No. 2013-01-1080.*
10. Fontanesi, S., d'Adamo, A. and Rutland, C. J., 2015. "Large-Eddy simulation analysis of spark configuration effect on cycle-to-cycle variability of combustion and knock". *International Journal of Engine Research,* ***16***(3), pp. 403–418.
11. d'Adamo, A., Breda, S., Fontanesi, S. and Cantore, G., 2015. "LES modelling of spark-ignition cycle-to-cycle variability on a highly downsized DISI engine". *SAE International Journal of Fuels and Lubricants,* ***8***(5), pp. 2029–2041.
12. Enaux, B., Granet, V., Vermorel, O., Lacour, C., Pera, C., Angelberger, C. and Poinsot, T., 2011. "LES study of cycle-to-cycle variations in a spark ignition engine". *Proceedings of the Combustion Institute,* ***33***(2), pp. 3115–3122.
13. Koch, J., Schmitt, M., Wright, Y. M., Steurs, K. and Boulouchos, K., 2014. "LES multi-cycle analysis of the combustion process in a small SI engine". *SAE International Journal of Fuels and Lubricants,* ***7***(1), pp. 269–285.
14. Truffin, K., Angelberger, C., Richard, S. and Pera, C., 2015. "Using large-eddy simulation and multivariate analysis to understand the sources of combustion cyclic variability in a spark-ignition engine". *Combustion and Flame,* ***162***(12), pp. 4371–4390.
15. Tatschl, R., Bogensperger, M., Pavlovic, Z., Priesching, P., Schuemie, H., Vitek, O. and Macek, J., 2013. "LES simulation of flame propagation in a direct-injection SI-engine to identify the causes of cycle-to-cycle combustion variations". *SAE Technical Paper*, *Paper No. 2013-01-1084.*
16. Granet, V., Vermorel, O., Lacour, C., Enaux, B., Dugué, V. and Poinsot, T., 2012. "Large-Eddy Simulation and experimental study of cycle-to-cycle variations of stable and unstable operating points in a spark ignition engine". *Combustion and Flame,* ***159***(4), pp. 1562–1575.
17. Ameen, M. M., Yang, X., Kuo, T. W., Xue, Q. and Som, S., 2015. "LES for Simulating the Gas Exchange Process in a Spark Ignition Engine". In Proceedings of 2015 ASME Internal Combustion Engine Division Fall Technical Conference.
18. Zhao, L., Moiz, A. A., Som, S., Fogla, N., Bybee, M., Wahiduzzaman, S., Mirzaeian, M., 2017. "Examining the

role of flame topologies and in-cylinder flow fields on cyclic variability in spark-ignited engines using large eddy simulation". *International Journal of Engine Research.* https://doi.org/10.1177/1468087417732447.

19. Zhao, L., Moiz, A. A., Som, S., Fogla, N., Bybee, M., Wahiduzzaman, S., Mirzaeian, M., Millo, F., Kodavasal, J., 2017. "Multi-cycle large eddy simulation to capture cycle-to-cycle variation (CCV) in spark-ignited (SI) engines". In Proceedings: *10th US National Combustion Meeting, College Park, MD.*
20. Ameen, M. M., Yang, X, Kuo, T. W. and Som, S., 2017. "Parallel methodology to capture cyclic variability in motored engines". *International Journal of Engine Research,* ***18***(4), pp. 366–377.
21. Ho, T. K., 1995. "Random Decision Forests". In Proceedings of 3rd International Conference on Document Analysis and Recognition.
22. Kodavasal, J., Pei, Y., Harms, K., Ciatti, S., Wagner, A., Senecal, P. K., García, M. and Som, S., 2016. "Global Sensitivity Analysis of a Gasoline Compression Ignition Engine Simulation with Multiple Targets on an IBM Blue Gene/Q Supercomputer". *SAE Technical Paper, Paper No. 2016-01-0602*.
23. Kodavasal, J., Van Dam, N., Pei, Y., Harms, K., Maheshwari, K., Wagner, A., García, M., Ciatti, S., Senecal, P. K. and Som, S., 2016. "Sensitivity Analysis on key CFD model inputs for Gasoline Compression Ignition on an IBM Blue Gene/Q supercomputer". In Proceedings of THIESEL 2016 Conference on Thermo- and Fluid Dynamic Processes in Direct injection Engines.
24. Kodavasal, J., Kolodziej, C. P., Ciatti, S. A. and Som, S., 2015. "Computational Fluid Dynamics Simulation of Gasoline Compression Ignition". *Journal of Energy Resources Technology,* ***137(3)***, 032212-032212-13.
25. Kodavasal, J., Kolodziej, C. P., Ciatti, S. A. and Som, S., 2017. "Effects of injection parameters, boost, and swirl ratio on gasoline compression ignition operation at idle and low-load conditions". *International Journal of Engine Research,* ***18***(8), pp. 824–836.
26. Kolodziej, C., Kodavasal, J., Ciatti, S., Som, S., Shidore, N. and Delhom, J., 2015. "Achieving stable engine operation of gasoline compression ignition using 87 AKI gasoline down to idle". *SAE Technical Paper, Paper No. 2015-01-0832.*
27. Kodavasal, J. and Som, S., 2018. "Gasoline Compression Ignition – A Simulation-Based Perspective". In: Srivastava D., Agarwal A., Datta A., Maurya R. (eds) *Advances in Internal Combustion Engine Research.* Energy, Environment, and Sustainability. Springer, Singapore.
28. Kodavasal, J., Lavoie, G. A., Assanis, D. N. and Martz, J. B., 2014. "The effects of thermal and compositional stratification on the ignition and duration of homogeneous charge compression ignition combustion". *Combustion and Flame,* ***162***(2), pp. 451–461.
29. Kodavasal, J., Lavoie, G. A., Assanis, D. N. and Martz, J. B., 2015. "The effect of diluent composition on homogeneous charge compression ignition auto-ignition and combustion duration". *Proceedings of the Combustion Institute,* ***35***(3), pp. 3019–3026.
30. Kodavasal, J., Lavoie, G. A., Assanis, D. N. and Martz, J. B., 2015. "Reaction-space analysis of homogeneous charge compression ignition combustion with varying levels of fuel stratification under positive and negative valve overlap conditions". *International Journal of Engine Research,* ***17***(7), pp. 776–794.
31. Kodavasal, J., 2013. "Effect of Charge Preparation Strategy on HCCI Combustion". *PhD Thesis*, University of Michigan.
32. Kodavasal, J., Keum, S. H. and Babajimopoulos, A., 2011. "An extended multi-zone combustion model for PCI simulation". *Combustion Theory and Modelling,* ***15***(6), pp. 893–910.
33. Kodavasal, J., McNenly, M. J., Babajimopoulos, A., Aceves, S. M., Assanis, D. N., Havstad, M. A. and Flowers, D. L., 2013. "An accelerated multi-zone model for engine cycle simulation of homogeneous charge compression ignition combustion". *International Journal of Engine Research,* ***14***(5), pp. 416–433.
34. Kodavasal, J., Harms, K., Srivastava, P., Som, S., Quan, S., Richards, K. and García, M., 2016. "Development of a Stiffness-Based Chemistry Load Balancing Scheme, and Optimization of Input/Output and Communication, to Enable Massively Parallel High-Fidelity Internal Combustion Engine Simulations". *Journal of Energy Resources Technology,* ***138*****,** 052203-1–11.
35. Sorgun, M., Ozbayoglu, A. M. and Ozbayoglu, M. E., 2014. "Support Vector Regression and Computational Fluid Dynamics Modeling of Newtonian and Non-Newtonian Fluids in Annulus With Pipe Rotation". *Journal of Energy Resources Technology,* ***137(3),*** 032901-032901-5.
36. Moiz., A. A., Pal, P., Probst, D., Pei, Y., Zhang, Y., Som, S. and Kodavasal, J., 2018. "A Machine Learning-Genetic Algorithm Approach for Rapid Virtual Optimization Using High-Performance Computing". *SAE Technical Paper, Paper No. 2018-01-0190.*
37. Rostami, H. and Manshad, A. K., 2014. "A New Support Vector Machine and Artificial Neural Networks for Prediction of Stuck Pipe in Drilling of Oil Fields". *Journal of Energy Resources Technology,* ***136(2),*** 024502-024502-4.
38. El-Emam, R. S. and Dincer, I., 2016. "Assessment and Evolutionary Based Multi-Objective Optimization of a Novel Renewable-Based Polygeneration Energy System". *Journal of Energy Resources Technology,* ***139(1),*** 012003-012003-13.
39. Van, S. L. and Chon, B. H., 2017. "Effective Prediction and Management of a CO_2 Flooding Process for Enhancing Oil Recovery Using Artificial Neural Networks". *Journal of Energy Resources Technology,* ***140(3),*** 032906-032906-14.
40. Richards, K. J., Senecal, P. K. and Pomraning, E., 2017. CONVERGE v2.3 Manual, Convergent Science, Inc., Madison, WI.

41. Reitz, R. and Diwakar, R., 1987. “Structure of High-Pressure Fuel Sprays”. *SAE Technical Paper, Paper No. 870598*.
42. Senecal, P., Richards, K., Pomraning, E., Yang, T., Dai, M. Z., McDavid, R. M., Patterson, M. A., Hou, S. and Shethaji, T., 2007. “A New Parallel Cut-Cell Cartesian CFD Code for Rapid Grid Generation Applied to In-Cylinder Diesel Engine Simulations”. *SAE Technical Paper, Paper No. 2007-01-0159*.
43. Schmidt, D. P. and Rutland, C. J., 2000. “A new droplet collision algorithm”. *Journal of Computational Physics,* ***164***(1), pp. 62–80.
44. Pomraning, E. and Rutland, C. J., 2002. “Dynamic One-Equation Nonviscosity Large-Eddy Simulation Model”. *AIAA Journal*, ***40***(4), pp. 689–701.
45. Williams, F., 1985. “Turbulent combustion”. *The Mathematics of Combustion (ed. J. D. Buckmaster)*, pp. 197–1318. Society for Industrial and Applied Mathematics.
46. Peters, N., 2000. “Turbulent Combustion”. Cambridge University Press.
47. Pitsch, H., 2002. “A *G*-equation formulation for large-eddy simulation of premixed turbulent combustion”. *Center for Turbulence Research Annual Research Briefs 2002*.
48. Balaprakash, P., Tiwari, A., Wild, S. M., Carrington, L. and Hovland, P. D., 2016. “AutoMOMML: Automatic Multi-objective Modeling with Machine Learning”. ISC High Performance 2016, LNCS 9697, pp. 219–239.
49. Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K. and Wiswedel, B., 2007. “The Konstanz Information Miner”. *Studies in Classification, Data Analysis, and Knowledge Organization*, Springer.