Fuel 246 (2019) 187–195

Contents lists available at ScienceDirect

Fuel

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

Full Length Article

Data-driven approach to fill in data gaps for life cycle inventory of dual fuel technology

Fanxu Meng*, Carolyn LaFleur, Asanga Wijesinghe, John Colvin

Houston Advanced Research Center (HARC), 8801 Gosling Rd, The Woodlands, TX 77381, USA

ARTICLE INFO

Keywords:
Life cycle assessment (LCA)
Life cycle inventory (LCI)
Emission
Data-driven multiple linear regression
Cross-validation
Uncertainty
Dual fuel engine
OPGEE

ABSTRACT

Life cycle assessment (LCA) is a powerful tool and schematic model to evaluate an emerging technology in the oil and gas industry. In the oilfield operations such as drilling and hydraulic fracturing, dual fuel (DF) diesel engines utilize natural gas and diesel fuel simultaneously, thereby reducing diesel fuel consumption and offering certain emissions reductions. Emissions from DF engines can vary greatly depending on fuel consumption and how the engine is operated. In this study, linear regression and cross validation are applied to analyze field testing data of dual fuel engines. This study enables rapidly prediction of emissions and filling of data gaps for fuel efficiency, substitution ratio and other parameters to optimize DF engine operations. Greenhouse Gas (GHG) emissions are predicted by linear regression with uncertainty of 2.7% based on power, methane loss, natural gas and diesel consumption. Exhaust after-treatment system (ATS) adds complexity and will significantly increase prediction uncertainty for carbon monoxide (CO) and non-methane hydrocarbon plus oxides of nitrogen (NMHC + NOx). This model could potentially be integrated into an engineering-based LCA tool – such as the Oil Production Greenhouse Gas Emissions Estimator (OPGEE) to predict environmental impacts for a range of energy consumptions. This is an innovative application of multiple linear regression and cross-validation to analyze dual fuel technology for the purpose of LCA.

1. Introduction

Dual fuel (DF) technology exhibits advantages by utilizing supplementary fuels such as natural gas [1,2] and hydrogen [3] to supplant some of the diesel fuel, offering economic and environmental benefits [4]. The two fuels are co-combusted simultaneously in varying proportions depending upon operating conditions. In the oil and gas industry, heavy duty stationary DF technology is gaining acceptance for drilling and hydraulic fracturing operations. In some situations, natural gas fuel can be supplied from nearby wells, reducing diesel fuel consumption for fuel cost savings. In most DF engines, natural gas fuel in vapor phase is injected into the engine air intake. The gas fuel is blended with air and supplied to the engine. The engine can operate with the addition of gas fuel (dual fuel mode) or on diesel fuel alone (diesel-only mode). Natural gas is more easily available in oil gas fields to power operations near the point of production. This avoids the work to convey, refine and transport diesel fuel to oilfields and consequently reduce energy inputs and truck traffic. Current independent field tests [5–8] of natural gas – diesel dual fuel technology show exclusive findings on emissions and economy. Harmonization is urgently needed to integrate data from different sources into a single structured database [9]. A successful technical harmonization usually uses engineering equations and fill in the missing value that were not originally reported by the literature [10].

Life cycle assessment (LCA) is a powerful tool [11,12] to support decision-making in consideration of power needs, environmental and human health effects, and other factors. LCA can evaluate environmental impacts for a range of operating conditions from a life cycle perspective. Life cycle inventory (LCI) is the basic unit for assembly of LCA. An LCI usually contains the functional unit, input energy and material, output materials and emissions, etc. Ian et al conducted study on life cycle GHG emissions of Marcellus shale gas [13] and Bakken tight oil [14]. The study concluded that life cycle GHG emissions are significantly sensitive to power generation. These studies focused on quantification of propagated uncertainty to the final output of an LCA model. Hauck modeled parameter uncertainty via Monte Carlo simulation. This revealed that technological differences in gas power are the principal cause for life cycle GHG emissions, primarily due to variability in plant efficiencies [15]. As to the dual fuel engine for power generation, emissions are related to operational parameters such as engine loading, speed (rpm), fuel consumption and after-treatment system (ATS, if applicable). The variability of those input parameters

* Corresponding author.
E-mail address: fmeng@harcresearch.org (F. Meng).

https://doi.org/10.1016/j.fuel.2019.02.124
Received 25 November 2018; Received in revised form 22 February 2019; Accepted 25 February 2019
Available online 28 February 2019
0016-2361/

Nomenclature

ATS	after-treatment system
CO	carbon monoxide
CO_2e	carbon dioxide equivalent
DF	dual fuel
DLE	diesel liter equivalent
DO	diesel only
DR	drilling
FTIR	Fourier-transform infrared spectroscopy
GHG	greenhouse gas
GWP	global warming potentials
HF	hydraulic fracturing
LCA	life cycle assessment
LCI	life cycle inventory
LHV	lower heating value
MARE	Mean Absolute Relative Error
NG	natural gas
NMHC	non-methane hydrocarbon
NOx	oxides of nitrogen
OPGEE	Oil Production Greenhouse Gas Emissions Estimator
PRELIM	Petroleum Refinery Life-Cycle Inventory Model
RMSE	Root Mean Squared Error
ZECE	Zero emission conversion efficiency

suggests considering both average values and uncertainties in order to scientifically conduct a high-quality LCA. Monte Carlo analysis is usually used to simulate and show a range of possible outcomes beyond the average value of the entire LCA system [16]. It captures the inherent variability of parameters and a single LCI. The probability distribution is necessarily assigned to each input variable followed by laborious simulation.

Aklouche et al. [17] recently developed a predictive model and analyzed the DF mode of combustion with different gaseous fuels at various loads. Di Blasio [18] and Belgiorno [19] conducted systematic parametric studies and gave valuable recommendations on optimized dual fuel engine calibration for efficiency and emissions reduction. These research studies provided solid and detailed explanation of the physics involved, enabling deeper understanding and evaluation of dual fuel technology. These have provided a basis for development of LCA with available experimental data.

To simplify and expedite analysis and evaluation, data-driven modeling is a powerful tool to estimate the uncertainty of life cycle inventory among different data sources [20]. Identifying patterns from data sets with supervised machine learning facilitates estimation of uncertainty. Cross validation is usually conducted to assure that the model function well on predicting with untrained data [21]. Successful modeling in chemical manufacturing [22,23] and the oil and gas industry [24,25] have been reported. Combining data-driven analysis with understandings of physical systems and experimental data can provide new insights for energy production processes, such as low salinity water flooding [26] and oil-water relative permeability [27] problems in enhanced oil recovery. Data-driven methods are also becoming popular and useful to analyze climate-related risk in energy system and the oil and gas industry. Masnadi et al. [28] estimated emissions from 8.966 on-stream oil files in 90 countries and reported national volume-weighted average crude oil upstream GHG intensities with uncertainty analysis. That study underscored the criticality of uncertainty analysis because of the poorly-detected emissions of fugitive methane, venting and flaring. Orellana et al. [29] presented a statistically enhanced life cycle based model to better understand the variability and uncertainty associated with oil sands GHG emissions. Integration of DF technology modeling with existing LCA tools is critical and urgent so that researchers can consider this emerging technology and convert information for decision-makers to the scientific community and public in a better and timely manner.

This study is an innovative application of data-driven modeling to analyze dual fuel technology for the purpose of LCA. This proof of concept analysis is based on 114 observations of field test data [5–8]. Each observation contains specific features of a DF engine operation and emissions. A regression model was finally constructed to predict emissions of a dual fuel engine from its features. The predicted results and scores by Repeated Random Sub-sampling Validation from 1000 repeated random experiments were presented. The scores include Mean Absolute Relative Error (MARE), Root Mean Squared Error (RMSE) and cross-validated R^2. Cross-validated R^2 is proved to be the best indicator of the prediction power of model. The average relative errors between

Table 1
Feature List.

Feature	Unit	Data Type	Harmonized Calculation
Source	N/A	Category	Which field test is the observation data from
Activity	N/A	Category	HF-hydraulic fracturing; DR-drilling
Engine Maker	N/A	Category	The manufacturer of the engine
Engine Model	N/A	Category	The model of the engine
Rated Speed	RPM	Numeric	The rated speed of the engine
Rated Power	kW	Numeric	Power output at 100% loading
Power	kW	Numeric	Actual power output during the test
NG Lower Heating Value	Btu/cf	Numeric	lower heating value of NG used during the test
Total Fuel Consumption	DLE/kWh	Numeric	Total fuel consumption in DLE*
Diesel Consumption	DLE/kWh	Numeric	Total diesel consumption in DLE
NG Consumption	DLE/kWh	Numeric	Total NG consumption in DLE, including converted and loss
Engine Load	1	Numeric	Engine loading during the test
Fuel Efficiency (ZECE)**	1	Numeric	power out / (NG power in - CH4 loss + diesel power in)
Diesel Displacement	1	Numeric	1-DF diesel rate / DO diesel rate
Substitution Ratio (Corrected)	1	Numeric	(NG power in - CH4 loss) / Total Fuel in
Substitution Ratio (Industry)	1	Numeric	NG power in / Total Fuel in
Methane Loss	1	Numeric	Methane out / NG in
Aftertreat System	N/A	Category	After-treatment system, Y- after DOC; N- before DOC or DOC is now applied during test
GHG Emission	CO_2e kg/kWh	Numeric	GHG emission includes CO_2, CH_4 (GWP = 25) and N2O (GWP = 298)
NMHC + NOx Emission	g/kWh	Numeric	NMHC emissions + NOx emissions
CO Emission	g/kWh	Numeric	CO emission

* DLE - Diesel liter equivalent;
** ZECE - Zero emission conversion efficiency [8].

predicted values and actual values mostly range from 2.7% to 10.4%. The emissions factors and framework were developed for calculation of the LCI suitable for integration with an engineering-based LCA tool – Oil Production Greenhouse Gas Emissions Estimator (OPGEE).

2. Methods

2.1. Dual fuel field testing and data

Researchers may conduct field testing of natural gas – diesel DF engines based on protocols designed to capture specific information, but a typical test design is described here [5,30]. Briefly, the tests of engine emissions and performances were conducted for the diesel-only baseline and for DF operation. Along with data from the engine control unit, natural gas and diesel fuel flow rates were directly and continuously measured. Emissions were simultaneously measured with a Fourier-transform infrared spectroscopy (FTIR) gas analyzer. Data collected at variations of engine loads and speed (rpms) were automatically recorded. The regression model uses 114 observations reported in field tests [5–8] of DF engines used in drilling and hydraulic fracturing service. Based upon information collected from users of dual fuel engines only those conditions in which 40% or more of the diesel fuel normally used is substituted with natural gas are of operational significance. Because of this, only observations with a diesel displacement greater than 40% were included in this study. This assures the database was not influenced by conditions that reflect an unusual operating conditions using only a small amount of NG. If those unusual operating conditions are included, it would diminish the value of the "Dual Fuel" model for predicting performance and benefits of DF technology. The calculations among difference sources were harmonized and summarized in significant features as shown in Table 1. Additional factors can affect DF engine operation, such as compression ratio, injection parameters, air fuel ratio and exhaust gas recirculation technology. These effects are described in systematic studies with detailed explanation of the physics involved [18,19], though usually were not reported in the active drilling fields nor for the heavy duty engines.

2.2. Multiple linear regression

A multiple linear regression is used to predict the value of a response variable based on a linear relation to multiple predictor variables [31].

$$\hat{y} = \beta_0 + \beta \bullet X$$

where $\hat{y}$ is a vector of estimated value, X is the matrix where each column contains individual predictor variables, β is a vector of coefficients of the model, and β_0 is the intercept.

Any of the features in Table 1 can be the response variable alternatively and be related to the rest of features. Features of fuel efficiency, diesel displacement, natural gas (NG) substitution ratio, emissions of greenhouse gas (GHG), non-methane hydrocarbon (NMHC) plus nitrogen oxides (NOx) and carbon monoxide (CO) are selected as the major response variables. GHG is the sum of CO_2, CH_4 and N_2O, after weighted by Global Warming Potentials (GWPs). The GWPs for CH_4 and N_2O are 25 and 298 respectively, which are consistent with the observation sources. It should be beneficial to report different impacts associated with NOx and NMHC species respectively in life cycle impact assessment methods (e.g., EPA's TRACI). However, considering the particular applications of heavy duty stationary engines in the oilfield and available data in the references, NOx and NMHC emissions are combined in this study. Notably, if some features could be directly used to calculate the predicted value, these features were not all included in the prediction of that value. For example, the fuel efficiency could be directly calculated by the method of dividing power output by fuel input, so fuel input and power output were not both included to predict the fuel efficiency. Python is used to estimate the values of regression coefficients by minimize the total residual error, which is also called Least Squares technique.

Supervised learning is used to learn the parameters of a prediction function from a training data set and to evaluate the model with those learned parameters on the test data set. The test data set usually does not contain repeated data to the training set). In this way, supervised learning was performed to overcome the overfitting issue. The data are commonly split, using 70–80% of the data to train the model and the remaining 20–30% used to evaluate performance [32]. For each split, the model is fit to the training set, and evaluation such as predictive accuracy is assessed using the test set. In our study, an 80/20 split yielded satisfactory results. The entire cycle of data split, training and evaluation was considered as one experiment. If necessary, prediction intervals can be calculated by the standard error in the model fit and the deviation of each predictor from its mean value.

2.3. Model validation

Coefficient of Determination R^2 for the training set and the test set were analyzed for each experiment, showing the goodness of fit. Notably, those coefficients are not necessarily related to the model prediction power [33].

$$\text{Coefficient of determination } R^2 = 1 - \frac{\sum (\text{Observation Value} - \text{Calculated Value})^2}{\sum (\text{Observation Value} - \text{Mean of Observation Values})^2}$$

where the Calculated Value is calculated from the model fitted by the samples and the predictor variables in the same set of Observations; Observation Value is the response variable of each sample in the set and the Mean of Observation Values is the mean value of response variable of all samples in the set.

Alternatively, the cross-validated R^2 is used to assess the predictive power of the model. In this procedure, a model is fitted with a training set. This fitted model is used to predict the response variables of the test set. If cross-validated R^2 has a value of 1, model gives a perfect prediction. If cross-validated R^2 has a negative value, the model is not better using the average of that response variable in the database. In this way, the problems such as over-fitting and selection bias can be avoided.

$$\text{Cross\_validated } R^2 = 1 - \frac{\sum (\text{Test Value} - \text{Predicted Value})^2}{\sum (\text{Test Value} - \text{Mean of Test Values})^2}$$

where the Predicted Value is calculated from the model fitted by the training set and the predictor variables in the test set, to evaluate the response variable in test set in a single experiment; The Test Value is the response variable of each sample in the test set in a single experiment and the Mean of Test Values is the mean value of response variables of all samples in the test set.

Given that the number of observations is no more than 114, Repeated Random Sub-sampling Validation, also known as Monte Carlo cross-validation, is used in this study to score the prediction. This approach randomly splits the dataset into training and test sets and completes the whole cycle of training and evaluation in this experiment. The results are then averaged over all the splits (experiments). The advantage of this method (especially over k-fold cross validation) is that the proportion of the training/validation split is independent on the number of folds. The disadvantage is that some observations may never be selected in the training set or test set. As the number of random splits (experiments) approaches infinity, the result of repeated random sub-sampling validation tends to diminish this disadvantage. In this study, 1000 random splits (experiments) were conducted.

Mean Absolute Relative Error (MARE) and Root Mean Squared Error (RMSE) were also used to illustrate the predictive power. MARE shows the relative (percentage) differences between the predicted values and actual test values. RMSE, which is also referred as Standard Deviation

Error in Prediction (SDEP or SEP) [33], shows absolute differences between predicted values and actual test values.

$$MARE = \frac{\sum \left| \frac{Predicted\ Value - Test\ Value}{Test\ Value} \right|}{Number\ of\ Test\ Value}$$

$$RMSE = \sqrt{\frac{\sum (Predicted\ Value - Test\ Value)^2}{Number\ of\ Test\ Value}}$$

where the Number of Test Value is the number of samples split to the test set.

Leave One Out Cross Validation (LOOCV) is used to predict response variables and compare experiment data. It is a simple cross-validation where each training set is created by taking all samples except one, which is the data for the test set. This cross-validation procedure is run iteratively until all samples have ever been assigned to the test set. Thus, for *n* samples, we will have *n* experiments. Each experiment will a test set with only one observation and a training set including the rest of observations.

3. Results and discussion

3.1. Data exploration

Data exploration was initiated with summary and description. Summary statistics for mean, minimum, maximum, median and standard deviation in Table 2 were calculated for numeric features.

The effects of activity (hydraulic fracturing or drilling) on fuel efficiency, diesel displacement and substitution ratio (see Table 1 for definitions) are shown in Fig. 1. During hydraulic fracturing tasks substitution ratio and diesel displacement are greater, though fuel efficiency is slightly reduced. These findings of diesel displacement and substitution ratio indicate that DF engines will use relatively more natural gas when operating in hydraulic fracturing service as compared with drilling. This explains the slightly lower fuel efficiency observed in hydraulic fracturing. The engines have methane loss out of the cylinders and consequently reduce the overall fuel efficiency.

The effects of exhaust after-treatment system (ATS) on emissions from the DF engines are shown in Fig. 2. The ATS considered here mainly refers to diesel oxidation catalyst (DOC). There are 64 samples with DOC technology. ATS is designed to largely reduce NMHC + NOx and CO emissions, but does not significantly affect the average value of GHG emission (Fig. 2). NMHC and CO will be oxidized to CO_2 by ATS. However, since the produced CO_2 by ATS constitutes a much lower portion than CO_2 by fuel combustion in the exhaust, GHG emissions are not significantly changed by ATS.

3.2. Emission prediction from fuel consumption

Predictor and response variables are selected according to the criteria that reflect physical meaning and practical application. For example, fuel consumption is usually used to predict emissions from an engine generating a specific power. The input elements (i.e. C, H, O, N) in the hydrocarbon fuel and air are converted in the form of exhaust gases (i.e. CO_2, CO, NMHC, NOx) according to the material balance. For DF technology, the predictor variables including power, natural gas consumption and diesel consumption are used to fit multiple linear regressions to predict emissions of GHG, NMHC + NOx and CO (Table 3). According to the data exploration, activity and ATS show major effects on those emissions and should also be considered if the engine is equipped with an ATS.

Prediction of CO without ATS based on multiple linear regression results in the best cross-validated R^2 of 0.865 and MARE of 4.4% among all those emission predictions. GHG predictions shows slightly less cross-validated R^2 of 0.818 and MARE of 6.3%. Those results indicate that linear regression is practical to predict GHG and CO emissions from diesel consumption, natural gas consumption, engine power and activity. Prediction of NMHC + NOx shows less accuracy. This is not unexpected because the combination of two components adds greater uncertainty. The multiple linear regression performs well in predicting GHG, CO and NMHC + NOx emissions. This result supports the potential and fundament for evaluation of these dual fuel emissions parameters in the LCI by multiplying emissions coefficients or factors related fuel consumption. Interestingly, prediction of NMHC + NOx indicates better coefficient of determination R^2 in training set than model for GHG. It is because this model contains fewer observations in the training set than model for GHG and accordingly increases the chance to fit in the data in training set. This demonstrates cross-validated R^2 is a better indicator of predictive power. Otherwise, number of observations and predictors have to be shown with coefficient of determination R^2 to reasonably compare the power of prediction.

Notably, prediction of CO with ATS has 0.384 cross-validated R^2 and 27.0% MARE, which indicates the model is less effective because significant uncertainty is added by ATS. It was challenging to predict CO emission from downstream of ATS, because ATS performance is hard to evaluate. For example, if the efficiency of ATS increases from 98.5% to 99%, the remaining CO will decrease from 1.5% to 1%. A slight fluctuation of the ATS performance will result in a 50% reduction in CO emissions. ATS performance is tightly related to features such as operating temperature, catalysts materials, service life of the catalysts, etc. Better understanding of these features would enhance prediction. Di Blasio [18] highlighted the need for better characterization of ATS to resolve issues related to emissions of methane hydrocarbon and NOx. It is inferred that this is the case in the future when we evaluate emissions

Table 2
Summary Statistics.

Feature	Unit	Count	Mean	std	min	25%	50%	75%	max
Rated Speed	RPM	114	1514	297	1200	1200	1500	1800	1950
Rated Power	kW	114	1277	220	1050	1101	1133	1382	1678
Power	kW	114	766	272	210	590	704	951	1423
NG Lower Heating Value	Btu/cf	114	979	39	943	943	982	1031	1055
Total Fuel Consumption	DLE/kWh	73	0.418	0.044	0.305	0.397	0.422	0.445	0.551
Diesel Consumption	DLE/kWh	73	0.125	0.036	0.090	0.101	0.113	0.138	0.243
NG Consumption	DLE/kWh	73	0.292	0.055	0.168	0.264	0.301	0.330	0.423
Engine Load	1	114	0.56	0.15	0.20	0.45	0.55	0.65	0.85
Fuel Efficiency (ZECE)	1	114	0.266	0.021	0.202	0.254	0.268	0.275	0.331
Diesel Displacement	1	73	0.604	0.096	0.400	0.540	0.637	0.688	0.728
Substitution Ratio (Corrected)	1	73	0.547	0.058	0.386	0.526	0.562	0.585	0.660
Substitution Ratio (Industry)	1	73	0.698	0.090	0.489	0.663	0.731	0.765	0.790
Methane Loss	1	73	0.213	0.062	0.060	0.186	0.226	0.257	0.309
GHG Emission	CO_2e kg/kWh	73	1.880	0.453	0.837	1.641	1.956	2.176	3.082
NMHC + NOx Emission	g/kWh	73	7.763	3.450	2.395	5.496	6.596	9.612	17.549
CO Emission	g/kWh	73	12.417	12.200	0.050	0.197	13.916	25.384	32.298

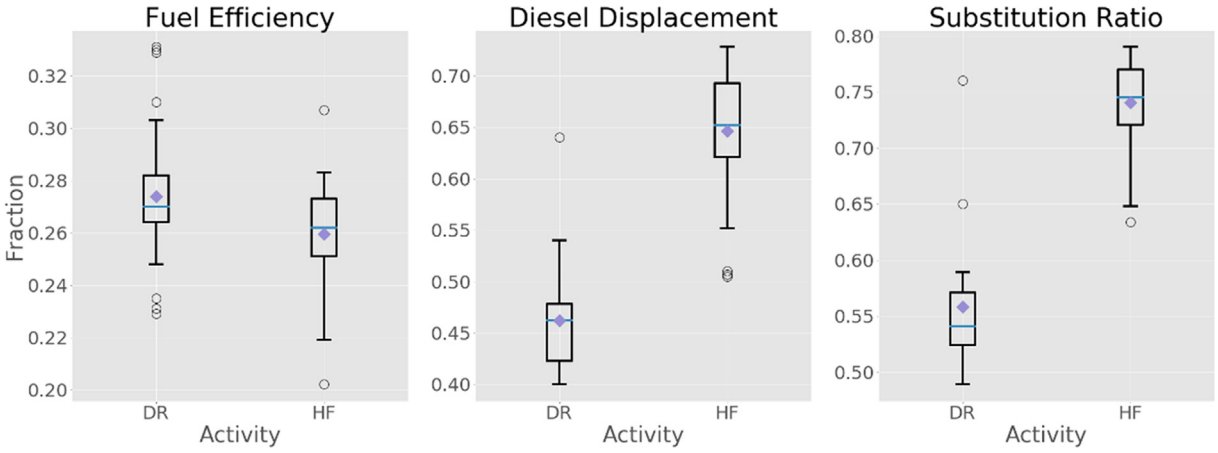

Fig. 1. Efficiency and Substitution Ratio by Activity.

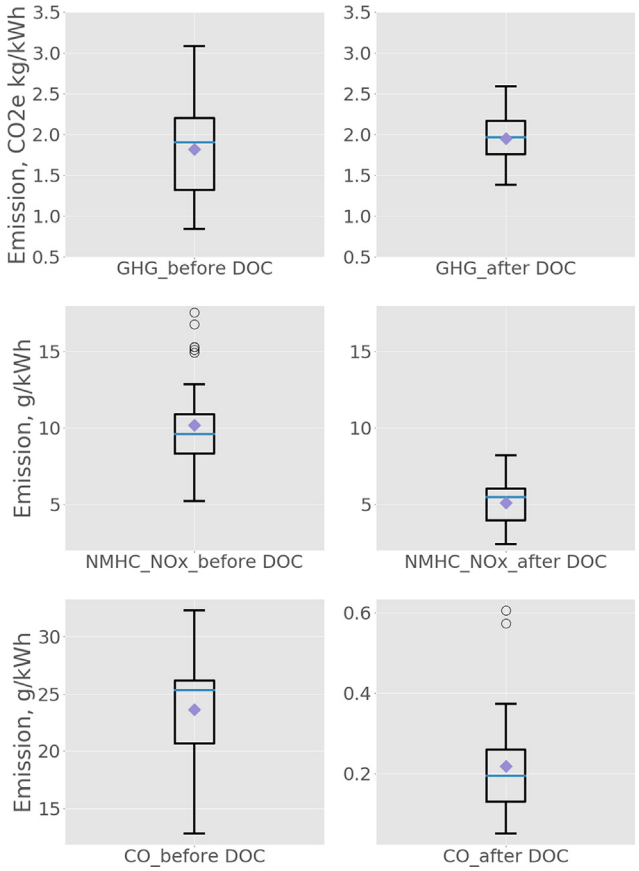

Fig. 2. The Effects of ATS to Emissions.

which ATS can effectively reduce. An alternative way of estimating emissions when engine equipment includes an ATS is by combining emissions and performance data without ATS and the known performance of ATS. For example, if an emission without ATS is 10 g/kWh and ATS efficiency is 95%, the emission from downstream of ATS is 0.5 g/kWh. Overall uncertainty will depend up on emissions data uncertainty and ATS performance. Predicting NMHC + NOx emissions is also a challenge as collection and measurement of these substances is complex. Furthermore, many of these components are intermittent and are tightly related to operational change over time, even in stationary scenarios. The cross-validated R^2 of NMHC + NOx after ATS is negative, which means the model is not improved by using the average of NMHC + NOx emissions in the test set.

Natural gas consumption in the above prediction is the natural gas flow to the engine, without considering the loss of the methane after the engine combustion. This may result in a relatively worse (larger) MARE due to the partial loss of the carbon balance. Methane loss is added into the multiple linear regression and the prediction performance is evaluated again in Table 4. It shows the MARE gets better to 2.7% and the cross-validated R^2 gets much closer to 1, which indicates a better prediction power by the model. This is enabled by a better carbon balance after methane loss is considered. Methane loss is an important factor to be considered in prediction of GHG emissions. However, measurement of methane loss is difficult before other information such as engine speed, load, power output and fuel consumption are collected. Though it may not be practical to include methane loss in model predictors, inclusion of methane loss is definitely beneficial for more accurate prediction. Nevertheless, the prediction of CO and NMHM + NOx becomes slightly worse when the cross-validated R^2 is checked.

The coefficient of determination R^2 of each prediction is better with the inclusion of methane loss and the chance to fit is increased. This again highlights that cross-validated R^2 is a reasonable indicator rather than coefficient of determination R^2 to show the power of prediction.

Table 3
Prediction Performance Summary – Emissions.

Response Variable	Unit	Number of Observations	Predictor Variables	Mean Value*	Prediction MARE**	Prediction RMSE***	Training Set R^2	Test Set R^2	Cross-validated R^2
GHG emission	CO_2e kg/kWh	73	Activity, Power, Diesel and Natural gas consumption	1.880	6.3%	0.162	0.887	0.908	0.818
NMHC + NOx w/o ATS	g/kWh	38		10.19	10.1%	1.14	0.908	0.952	0.712
NMHC + NOx w/ ATS	g/kWh	35		5.123	24.7%	1.400	0.332	0.735	−0.525
CO w/o ATS	g/kWh	38		23.65	4.4%	1.27	0.949	0.981	0.865
CO w/ ATS	g/kWh	35		0.217	27.0%	0.074	0.782	0.903	0.384

* Mean value of all samples in Summary Statistics;
** MARE: Mean Absolute Relative Error;
*** RMSE: Root Mean Squared Error.

Table 4
Prediction (with Methane Loss) Performance Summary – Emissions.

Response Variable	Unit	Number of Observations	Predictor Variables	Mean Value*	Prediction MARE**	Prediction RMSE***	Training Set R^2	Test Set R^2	Cross-validated R^2
GHG emission	CO_2e kg/kWh	73	Activity, Power, Methane loss, Diesel and Natural gas consumption	1.880	2.7%	0.062	0.986	0.994	0.977
NMHC + NOx w/o ATS	g/kWh	38		10.19	9.0%	1.10	0.932	0.982	0.708
CO w/o ATS	g/kWh	38		23.65	4.7%	1.34	0.954	0.989	0.838

* Mean value in Summary Statistics;
** MARE: Mean Absolute Relative Error;
*** RMSE: Root Mean Squared Error.

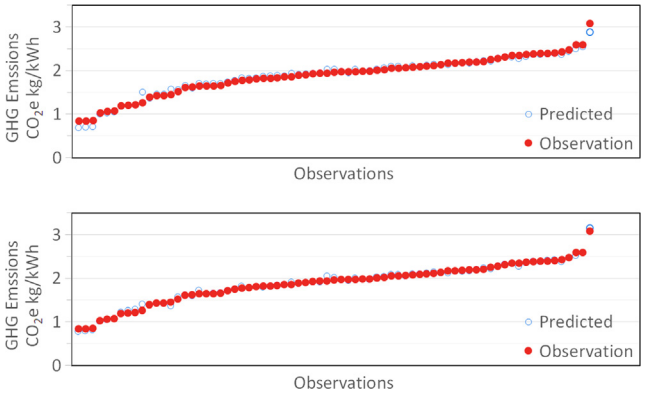

Fig. 3. Compare Experimental and Predicted GHG Emissions Data.

To further examine the effects of fuel consumption, the square of diesel consumption and natural gas consumption are added. Prediction performance is summarized in Table S1 in the Support Information (SI). Among differing predictions, better results are founded for GHG prediction by Activity, Power, Diesel consumption, Natural gas consumption, Square of diesel consumption, Square of natural gas consumption and Methane loss. This prediction has a 1.7% MARE and a 0.991 Cross-validated R^2, better than those from prediction without square of diesel consumption and natural gas consumption. However, all other predictions involving square of diesel consumption and natural gas consumption do not show improved prediction. Again, with better coefficient of determination R^2 the model can better fit the data set because additional predictors are involved. Fig. 3 compares the experimental and predicted GHG emissions data. From this, it appears that predictions of GHG emissions involving Square of diesel consumption and Square of natural gas consumption (bottom) yield better results. Comparisons are also conducted for CO and NMHC + NOx, showed in Figs. S1 and S2 in the SI.

(Top) Predictors: Activity, Power, Methane loss, Diesel and Natural gas consumption; (Bottom) Predictors: Activity, Power, Methane loss, Diesel consumption, Natural gas consumption, Square of diesel consumption and Square of natural gas consumption.

3.3. Model application - integrate dual fuel LCIs with OPGEE

A descriptive discussion is provided here to illustrate integration of our LCIs with existing LCA tools, such as OPGEE [34]. OPGEE is an open-sourced and engineering based LCA tool that estimates energy use and GHG emissions from the key process stages of upstream oil and gas operations [35,36], such as exploration, drilling & development, production, processing, and transport of crude petroleum, etc. OPGEE focuses carbon intensity of oils to overcome the methodological and data challenges associated with considering climate-related risk in oil investments [28]. It includes various types of fuel in different stages and processes. However, dual fuel technology is not included in OPGEE. Simultaneous combustion of natural gas and diesel in DF engine results in different emission factors from those factors by only natural gas or diesel. Recent studies used OPGEE to analyze the life cycle of global oil demand and supply [37], and the emissions associated with oil supply in China [38]. OPGEE also provides a valuable model to estimate the effects of energy resources other than fossil fuels. Wang, et al. [39] aligned primary energy demands with solar resources relating with OPGEE. Processes in OPGEE model demand for electricity and thermal energy that could be offset by solar sources. The PRELIM [40] model is referenced for the estimations related to downstream refineries processes. Combining the factors derived from OPGEE and PRELIM, investigators examine the potential for solar energy in global oil operations for upstream and refinery sectors.

It is practical to use OPGEE as the platform and to integrate our

multiple linear regression model to include LCA of dual fuel technology in oil and gas operation. For each stage (e.g. totaling m stages) of upstream operations in OPGEE, there are various types of fuel (e.g. totaling n types). The consumption of each fuel in each stage is the sum of fuel used by multiple equipment/processes in that stage. Therefore, energy consumptions (in the unit of MMBtu) of all n types of fuel appears as a $1 \times n$ array in each stage. OPGEE has emissions factors (in g/MMBtu - LHV) for each type of fuel that converts the fuel energy consumption (in MMBtu) to GHG emissions (g CO_2e). Those factors for different types of fuel also appears as an $1 \times n$ array in each stage. The dot product of energy consumption of all n types of fuel ($1 \times n$ array) and the emissions factors of all n types of fuel ($1 \times n$ array) result in the total GHG emission in one stage.

The consumptions (in MMBtu) of diesel and natural gas by the dual fuel engine should not be added to the above calculation. Therefore, the consumptions of diesel and natural gas should be listed separately as a new type ($n + 1$) of fuel if applied in OGPEE. The emission factors converting dual fuel consumption to emissions are obtained by multiple linear regression and prediction, listed as a new factor ($n + 1$). Those emission factors do exist according to the prediction power evaluation from the above dual fuel engine model related to diesel and natural gas fuel consumption. Therefore a dot product of a $1 \times n$ array will give the total emissions in the stage where the dual fuel engine is used.

The flowchart (Fig. 4) shows how the LCA researchers can utilize data-driven multiple linear regression to estimate emissions in life cycle inventory (LCI) for DF engine. The emissions can be predicted by collecting or assuming operational parameters such as activity and fuel consumption. Activity as either hydraulic fracturing (HF) or drilling (DR) is also input. The uncertainty is marked the same as the MARE from supervised learning in the mode of customized prediction.

3.4. Prediction from selected predictors to fill data gap

Table 5 shows the best MARE and RMSE of predictions from a series of selected features. A 4.3% MARE was obtained for predicted fuel efficiency by the multiple linear regression model. However, the 0.386 cross-validated R^2 indicates the prediction does not perform very well. It is because the fuel efficiencies distribute within a narrow range (Table 2, mean = 0.266, std = 0.021). It is therefore a challenge for the prediction model to beat the average of all samples. Predictions of diesel displacement and substitution ratio perform better, as the cross-validated R^2 are 0.588 and 0.676. Those correlations are explained from the physical perspective of engine operation. Major operational parameters (Predictors: Activity, RPM and Engine load) determine the key performances of the dual fuel engine (Response variables: diesel displacement and substitution ratio). Despite restrictions in this multiple linear regression model, it shows an improvement of predicting a response variable compared to using averaged data. This is useful to fill in missing data using sufficient known features and data in each observation.

4. Conclusion

This proof of concept practice demonstrates the power and potentiality to predict both value and uncertainty of life cycle inventory for DF technology. The average relative errors between predicted values and actual values mostly range from 2.7% to 10.1% based on the 114 observations available. The best prediction of GHG has a 1.7% MARE and a 0.991 cross-validated R^2. The ATS performance characteristics need further study for greater accuracy in prediction of emissions when the engine is equipped with ATS. If available, the incorporation of methane loss data enhances GHG prediction. It is possible to integrate our method and data in existing LCA tool such as OPGEE. Collecting various data sets from additional sources and optimizing feature selection can further improve the results.

Another objective of this work is to identify potential challenges to applying data-driven regression to model the correlations and uncertainties in life cycle inventory: 1) Data cleaning requires extensive work to harmonize the data from different sources to comply with consistent definitions and calculations. Basic principles and engineering equations are useful to fill in the missing values with calculations. 2) Defining the suitable features is key to building the supervised learning

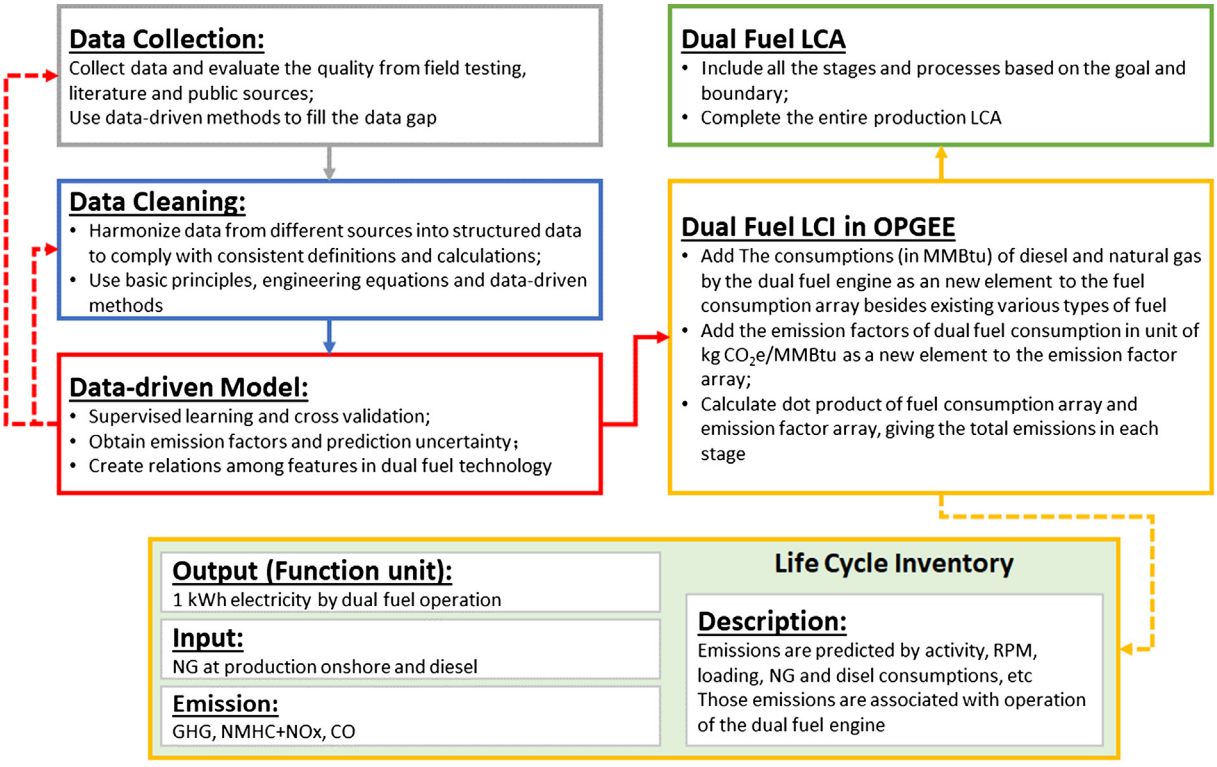

Fig. 4. Flowchart to Apply Data-driven Multiple Linear Regression to Life Cycle Inventory of Dual Fuel Technology.

Table 5
Prediction Performance Summary – Efficiency (Predictions from Selected Features).

Response Variable	Number of Observations	Predictor Variables	Mean Value*	Prediction MARE**	Prediction RMSE***	Training Set R^2	Test Set R^2	Cross-validated R^2
Fuel efficiency	114	Activity, Engine maker, Engine model, RPM, Power, Natural gas heat value and Engine load	0.266	4.3%	0.015	0.551	0.646	0.386
Diesel displacement	73	Activity, RPM and Engine load	0.604	7.6%	0.056	0.689	0.734	0.588
Substitution ratio (Industry)	73	Activity, RPM and Engine load	0.698	4.8%	0.044	0.771	0.805	0.676

* Mean value of all samples in Summary Statistics;
** MARE: Mean Absolute Relative Error;
*** RMSE: Root Mean Squared Error, (also referred as standard deviation error in prediction, SDEP).

model and optimized / customized results. A successful practice requires deep understanding of data-driven theory and technology, and how to incorporate research and business values for the users. 3) Ongoing studies and collection of representative data will support improved prediction of values and uncertainties in the future.

Our attempt shows how to apply data-driven multiple linear regression for confident decisions on adopting dual fuel technology as well as many other innovative technologies. The proposed flowchart in Fig. 4 shows the basic steps and key points in development of this study. In our future research, we shall examine LCA applications for other energy sector operations for comparison. We shall illustrate the potential economic and environmental tradeoffs with LCA perspective that might exist. It is important to combine data-driven approach with thorough understanding of parametric studies and the physics of DF engine operation. Engine operational data helps harmonizing observations and features beyond those in this study, such as compression ratio, injection parameters, air fuel ratio and exhaust gas recirculation technology, etc. To fully realize the potential benefits of this approach will requires the joint efforts of industry and academic partners.

Acknowledgements

The authors would like to thank the Program Activity Funding from the Houston Advanced Research Center (HARC) and Environmentally Friendly Drilling (EFD) Program.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.fuel.2019.02.124.

References

[1] Li Y, Li HL, Guo HS. A numerical investigation on NO2 formation reaction pathway in a natural gas-diesel dual fuel engine. Combust Flame 2018;190:337–48.
[2] Li Y, Li H, Guo H, et al. A numerical investigation on methane combustion and emissions from a natural gas-diesel dual fuel engine using CFD model. Appl Energy 2017;205:153–62.
[3] Li H, Liu S, Liew C, et al. An investigation on the mechanism of the increased NO2 emissions from H2-diesel dual fuel engine. Int J Hydrogen Energy 2018;43(7):3837–44.
[4] Wei L, Geng P. A review on natural gas/diesel dual fuel combustion, emissions and performance. Fuel Process Technol 2016;142:264–78.
[5] Wijesinghe A, LaFleur C, Meng F, et al. Fuel economy and emission characteristics of a high horsepower natural gas/diesel dual-fuel engine in oil & gas operations. IADC/SPE drilling conference and exhibition, society of petroleum engineers: Fort Worth, Texas, USA. 2018.
[6] Coastal Impacts Technology Program - Final Report, 2016.
[7] Johnson D, Heltzel R, Nix A, et al. Regulated gaseous emissions from in-use high horsepower drilling and hydraulic fracturing engines. J Pollut Eff Control 2017;5(2):187.
[8] Johnson DR, Heltzel R, Nix AC, et al. Greenhouse gas emissions and fuel efficiency of in-use high horsepower diesel, dual fuel, and natural gas engines for unconventional well development. Appl Energy 2017;206:739–50.
[9] O'Donoughue PR, Heath GA, Dolan SL, et al. Life cycle greenhouse gas emissions of electricity generated from conventionally produced natural gas systematic review and harmonization. J Ind Ecol 2014;18(1):125–44.
[10] Tu Q, Eckelman M, Zimmerman J. Meta-analysis and harmonization of life cycle assessment studies for algae biofuels. Environ Sci Technol 2017;51(17):9419–32.
[11] Masanet E, Chang Y, Gopal AR, et al. Life-cycle assessment of electric power systems. Annu Rev Environ Resour 2013;38:107–36.
[12] Meng F, Dillingham G. Life cycle analysis of natural gas-fired distributed combined heat and power versus centralized power plant. Energy Fuels 2018;32(11):11731–41.
[13] Laurenzi IJ, Jersey GR. Life cycle greenhouse gas emissions and freshwater consumption of Marcellus shale gas. Environ Sci Technol 2013;47(9):4896–903.
[14] Laurenzi IJ, Bergerson JA, Motazedi K. Life cycle greenhouse gas emissions and freshwater consumption associated with Bakken tight oil. Proc Natl Acad Sci USA 2016;113(48):E7672–80.
[15] Hauck M, Steinmann ZJN, Laurenzi IJ, et al. How to quantify uncertainty and variability in life cycle assessment: the case of greenhouse gas emissions of gas power generation in the US. Environ Res Lett 2014;9(7):7.
[16] Miller SA, Landis AE, Theis TL. Use of Monte Carlo analysis to characterize nitrogen fluxes in agroecosystems. Environ Sci Technol 2006;40(7):2324–32.
[17] Aklouche FZ, Loubar K, Bentebbiche A, Awad S, Tazerout M. Predictive model of

the diesel engine operating in dual-fuel mode fuelled with different gaseous fuels. Fuel 2018;220:599–606.
[18] Di Blasio G, Belgiorno G, Beatrice C. Effects on performances, emissions and particle size distributions of a dual fuel (methane-diesel) light-duty engine varying the compression ratio. Appl Energy 2017;204:726–40.
[19] Belgiorno G, Di Blasio G, Beatrice C. Parametric study and optimization of the main engine calibration parameters and compression ratio of a methane-diesel dual fuel engine. Fuel 2018;222:821–40.
[20] Xu M, Cai H, Liang S. Big data and industrial ecology. J Ind Ecol 2015;19(2, SI):205–10.
[21] Pan I, Pandey DS. Incorporating uncertainty in data driven regression models of fluidized bed gasification: a Bayesian approach. Fuel Process Technol 2016;142:305–14.
[22] Cashman SA, Meyer DE, Edelen AN, et al. Mining available data from the united states environmental protection agency to support rapid life cycle inventory modeling of chemical manufacturing. Environ Sci Technol 2016;50(17):9013–25.
[23] Mittal VK, Bailin SC, Gonzalez MA, et al. Toward automated inventory modeling in life cycle assessment: the utility of semantic data modeling to predict real-world chemical production. ACS Sustain. Chem. Eng. 2018;6(2):1961–76.
[24] Busby D, Pivot F, Tadjer A. Use of data analytics to improve well placement optimization under uncertainty. Abu Dhabi International Petroleum Exhibition & Conference. Abu Dhabi, UAE: Society of Petroleum Engineers; 2017.
[25] Mishra S, Lin L. Application of data analytics for production optimization in unconventional reservoirs: a critical review. SPE/AAPG/SEG unconventional resources technology conference, unconventional resources technology conference: Austin, Texas, USA. 2017.
[26] Wang L, Fu X. Data-driven analyses of low salinity water flooding in sandstones. Fuel 2018;234:674–86.
[27] Esmaeili S, Sarma H, Harding T, et al. A data-driven model for predicting the effect of temperature on oil-water relative permeability. Fuel 2019;236:264–77.
[28] Masnadi MS, El-Houjeiri HM, Schunack D, et al. Global carbon intensity of crude oil production. Science 2018;361(6405):851–3.
[29] Orellana A, Laurenzi IJ, MacLean HL, et al. Statistically enhanced model of in situ oil sands extraction operations: an evaluation of variability in greenhouse gas emissions. Environ Sci Technol 2018;52(3):947–54.
[30] Meng F, Wijesinghe A, Colvin J, et al. Conversion of exhaust gases from dual-fuel (natural gas-diesel) engine under Ni-Co-Cu/ZSM-5 catalysts. SAE Technical Papers. 2017..
[31] Steinmann ZJN, Venkatesh A, Hauck M, et al. How To address data gaps in life cycle inventories: a case study on estimating CO2 emissions from coal-fired electricity plants on a global scale. Environ Sci Technol 2014;48(9):5282–9.
[32] AWS Amazon Machine Learning - Splitting the Data into Training and Evaluation Data. 2019, https://docs.aws.amazon.com/machine-learning/latest/dg/splitting-the-data-into-training-and-evaluation-data.html.
[33] Todeschini, R. Tutorial 5, Useful and Unuseful Summaries of Regression Models. http://www.moleculardescriptors.eu/tutorials/tutorials.htm.
[34] OPGEE: the Oil Production Greenhouse Gas Emissions Estimator. 2018, https://eao-stanford-edu.ezproxy.lib.uh.edu/research/opgee-oil-production-greenhouse-gas-emissions-estimator.
[35] Cooney G, Jamieson M, Marriott J, et al. Updating the US Life cycle ghg petroleum baseline to 2014 with projections to 2040 using open-source engineering-based models. Environ Sci Technol 2017;51(2):977–87.
[36] El-Houjeiri HM, Brandt AR, Duffy JE. Open-source LCA tool for estimating greenhouse gas emissions from crude oil production using field characteristics. Environ Sci Technol 2013;47(11):5998–6006.
[37] Brandt AR, Masnadi MS, Englander JG, et al. Climate-wise choices in a world of oil abundance. Environ Res Lett 2018;13, (4):9.
[38] Masnadi MS, El-Houjeiri HM, Schunack D, et al. Well-to-refinery emissions and net-energy analysis of China's crude-oil supply. Nat Energy 2018;3(3):220–6.
[39] Wang JF, O'Donnell J, Brandt AR. Potential solar energy use in the global petroleum sector. Energy 2017;118:884–92.
[40] PRELIM: the Petroleum Refinery Life Cycle Inventory Model. 2017, https://www.ucalgary.ca/lcaost/prelim.