2023-01-0374

Modeling of Transient Gasoline Engine Emissions using Data-Driven Modeling Techniques

Ganesh Sundaram, Tobias Gehra, Jonas Ulmen, Mirjan Heubaum, Daniel Görges, and Michael Günthner
RPTU Kaiserslautern-Landau

Abstract

In recent years, the automotive industry has shifted from purely combustion engine-driven vehicles towards hybridization due to the introduction of CO_2 emission legislation. Hybrid powertrains also represent an important pillar and starting point in the journey towards zero-emission and full electrification. Fulfilling the most recent emission standards requires efficient control strategies for the engine, capable of real-time operation. Model accuracy is one of the main parameters which directly influence the performance of such control strategies. Specific methodologies developed in the past, such as physically- or phenomenologically-based approaches, have already facilitated the modeling of the combustion engine. Even though these models can accurately predict emissions in steady state conditions, their performance during transient engine operation is time-consuming and still not sufficiently reliable. The major contribution of the current work is to clarify and apply the recent advancements in data-driven modeling techniques, especially in time series forecasting with feedforward neural networks (FFNNs) and long short-term memory networks (LSTMs), to address the limitations mentioned above and to compare the different approaches.

The quantity and quality of data are significant challenges for data-driven modeling. This paper studies the modeling of gasoline engine emissions using FFNNs and LSTMs. The data quantity and quality requirements are studied based on a portable emission measurement system (PEMS), measuring at 1 Hz, and additional analyses on an engine test bench with a HiL setup, providing the possibility of increasing the measurement frequency with more sophisticated devices by a factor of five. Subsequently, the training and validation of the FFNNs and LSTMs are outlined, and finally, the model accuracy is discussed.

Introduction

Reducing the emissions of both greenhouse gases and pollutants are the dominating topics of current drivetrain developments for road vehicles. Both in the United States [1] and the European Union [2], [3] as well as in most other developed countries of the western world, ambitious targets for the reduction of CO_2 emissions in the transport sector have been set. Alongside, emission standards for pollutants are also tightened to ultra-low levels, altogether aiming at zero-impact emissions from road vehicles.

For combustion engine-driven vehicles, massive efforts have been made in the last two decades to improve both efficiency and pollutant emissions based on the findings from research activities worldwide. Part of this research was identifying vehicle operating conditions leading to increased emissions. One of these areas is low-speed vehicle operation, including when the vehicle starts from a standstill. The studies helped to develop appropriate operational strategies to reduce both pollutant and CO_2 emissions, e.g., by running the engine at an optimum load point and by substituting the combustion engine at low vehicle speeds, e.g., through the use of an electric motor for propulsion. These insights led to the development of Hybrid Electric Vehicles, commonly known as HEVs. Hybrid drive systems offer numerous opportunities to optimize engine operation and to downsize the combustion engine.

Engine load point optimization is a crucial step in improving engine efficiency and reducing CO_2 emissions. Based on the characteristic maps of fuel economy for the corresponding engine, which can be obtained from testbed measurements, strategies for optimal operation can be developed. Whenever it is identified that the current operating mode is deviating from the optimal region, various strategies can be applied to move the engine operation to a more efficient condition. Such deviations from the optimal operating point include two directions, thus both high-load conditions, such as sudden acceleration, and low-load conditions, such as low-speed driving. In addition to non-optimal fuel economy, both situations may also cause an increase in pollutant emissions. As an example, strong accelerations entail highly transient operating point changes of the engine, which may favor effects such as incomplete combustion or high nitrous oxide (NO_x) emissions. On the other side, extensive low-load operation typically results in a decrease in exhaust gas temperature, which might fall below the conversion limit temperature of the exhaust gas after-treatment system (EGA), resulting in a deterioration of tailpipe emissions. Both situations must be avoided to improve real-world pollutant emissions and comply with current and future emission standards.

In order to assess the emission level and fuel economy of vehicles in realistic operation, various driving cycles have been developed for the assessment on roller dynamometers, serving as a standard metric for type approval. In Europe, the Worldwide Harmonized Light Duty Vehicle Test Procedure (WLTP) and the associated Worldwide Harmonized Light Duty Vehicle Test Cycle (WLTC) was introduced in the year 2017, aiming at more realistic driving conditions than the previously used New European Driving Cycle (NEDC). As a further step, European emission legislation today also requires assessing pollutant emissions in real-world operations on the road corresponding to the Real Driving Emissions legislation (RDE). These developments further increase the requirements for operational strategies to control the entire drivetrain. In this context, the performance of the engine and drivetrain control system depends heavily on how well the emission dynamics of the engine are understood and transferred into a model to be used in the operational strategy. However, modeling engine emissions is subject to various challenges, as will be discussed in the upcoming sections.

In recent years, various projects have been carried out in Machine

Learning-based modeling of engine processes for application in the powertrain control unit. Most of these projects have focused on partial aspects of engine control requirements (e.g., lambda control [4] or EGR control [5]) or the realization of an efficiency-optimized hybrid operating strategy, but without consideration of pollutant emissions [6], [7], [8]. Other works address modeling for engine emissions and fuel consumption using machine learning methods, but without the objective of an application to engine control [9]. In contrast, the present work aims at making the transient pollutant formation in real road operation describable and applicable as a control variable for engine control.

Problem Definition

The vital contribution of the operational strategy in a hybrid electric vehicle is the efficient load distribution corresponding to the torque requested by the driver. The electric motor generally shifts the combustion engine to its optimal (or at least a more efficient) operating point. Usually, this optimum is understood as the most efficient engine operation. However, considering the high requirements concerning pollutant emissions, this choice can no longer be taken without considering the impact on tailpipe emission quality. This approach implies that the operational strategy should already have an emission model included by which it can predict how the emission parameters propagate from the current state. Modeling emissions involves various challenges. As a first classification, exhaust emissions can be divided into two major categories: during steady-state and during transient operation. Emission models must include both of these operating modes. To optimize tailpipe emissions, future operating strategies will require real-time prediction of pollutant formation and the various factors influencing exhaust gas after-treatment. For this purpose, it is also beneficial to apply additional sensors in the exhaust system or to analyze the existing raw data with modern evaluation methods to provide new information to the higher-level models or diagnostics, as presented, for example, in [10].

Steady-State Operation

Whenever the vehicle is moving at a constant speed, and a consistent load, the Engine Control Units (ECU) have ample time to set the parameters influencing emissions, such as ignition timing, valve timing, and lambda (the air-fuel ratio) to optimum values to make the combustion process as complete as possible. This measure has a positive effect on raw emissions. As an additional advantage, the steady-state operation of an internal combustion engine is easier to model when compared to the transient conditions, as the engine behavior during this condition can be described physically and phenomenologically. This has been a proven standard in recent years. However, it is time-consuming due to the difficulty of obtaining the parameters required for customizing the existing emissions for the respective application. Generally, the combustion quality inside the engine, which decisively influences the emission level, can be described and determined through a heat release analysis. For this purpose, transient measurements of pressure at three points inside the engine are required – the pressure in the intake duct, inside the engine cylinder, and in the exhaust manifold – which is why this approach is sometimes also referred to as "three-pressure analysis" (TPA). The heat release curve obtained for a particular engine operating point provides ample information on the combustion process. In further steps, a combustion model may be derived and optimized from these known combustion characteristics, providing the required data for the emission models.

Transient Operation

The operation of an internal combustion engine in a transient or highly dynamic condition differs fundamentally from the steady-state or quasi-steady-state conditions. The essential fact in the steady-state operation that the operating points remain the same for a considerable time is violated here. The ECUs cannot set optimum parameters for these situations as the operating points change continuously and rapidly. Additionally, the inertia of the systems also plays a role here, especially in the form of phenomenons such as boost pressure buildup. Even modern engines struggle to support these stepless calibration requirements to avoid such non-optimal conditions, resulting in emission peaks in transient operation. Unfortunately, the physically/phenomenologically describable processes mentioned with steady-state emission modeling overlap for every operational point and increase the uncertainty when used to model the transients. The pressure analysis method fails as the intermediate states will be passed in short intervals from point A to point B during an operation. A possible solution to the problem is building steady-state sectional areas within a single transient operation; however, this subdivided approach would cost considerable computational effort. Adding to the existing overheads, it is also fundamentally impossible to measure, analyze and optimize all operating points during transient conditions. Due to this, these situations are better handled with interpolation techniques rather than an arbitrary change of operating point in transient operation.

Operation Strategy Development

More knowledge regarding the emission dynamics and operating environment parameters thoroughly aids the development of operational strategies for HEVs. With new intelligent onboard systems, acquiring information such as driver behavioral patterns, route prediction, and diagnostics is much easier. As mentioned in the previous section, obtaining steady-state emission models is possible. The major challenge exists in modeling engine emissions during transient operating conditions. Developing an ideal operational strategy that assures the lowest potential emissions during real-world vehicle operation on the road is only achievable when these critical pieces of information are combined.

This work mainly focuses on exploring the possibilities of understanding the transient emission modeling process and building an emission model for a particular engine. The newly developed models aim to be compatible with a controller deciding on the operational strategy which takes action to avoid excessive pollutant emissions in critical driving maneuvers. The entire setup should ideally be real-time capable and respect the driver's demands. As pollutant formation is an extraordinarily complex process - even though it can be described relatively well in steady-state conditions - its dynamics in transients are challenging to capture mathematically. Hence, the possibilities of data-driven modeling will be explored in depth. The data-driven model should ideally help in the real-time detection and prediction of pollutant formation, which will be used inside the operational strategy. The modeling approaches will be explained in detail in the upcoming sections.

State of the Art

Emission Models

In recent works, various emission models have been presented, which describe the formation and prediction of the emission components. These models are mainly divided into two subcategories: physical and data-driven models.

Physical modeling techniques comprehend how each gaseous component reacts during combustion individually and identifies which combustion models can represent them. These combustion models calculate the substantial, non-measurable quantities, such as the partitioning and the state quantities of the burned and unburned zone. Such models have been developed by Pischinger et al. [11] and Merker et al. [12] for example. They divide combustion models into three categories: thermodynamic (0-dimensional), phenomenological (quasi-dimensional), and multi-dimensional models. These models differ in their complexity. The model selection depends upon the emission species that are of interest. For example, in Esposito et al. [13], the carbon monoxide model used is based on Cai et al. [14] and employs the reduced fuel oxidation mechanism utilizing a 0D-1D combustion model. The data-driven emission models are explained in more detail below.

Emission Measurement Systems

This section briefly introduces the underlying measurement and test bench systems. Additionally, their capabilities and limitations are also explored. These devices can be classified into portable and stationary measuring devices.

PEMS: The portable emission measurement systems (PEMS) are specifically designed for operation on the complete vehicle in real road traffic and allow an analysis according to the legally required standards. For this purpose, they are usually mounted on the vehicle's trailer coupling, and the exhaust gas is discharged directly at the rear silencer. In addition to gaseous and particle emissions analyzers, these devices have an exhaust flow meter to measure the exhaust gas mass flow and a GPS to track the route, especially the altitude profile. The measurement device requires a host computer recording the vehicle's OBD values. In terms of design, the devices available on the market may differ. However, the features mentioned above are standard and can be found on all PEMS devices as a standard feature. Due to the physical limitations, these devices are usually designed for a minimum acquisition rate of 1 Hz and are limited for high dynamics emission measurements which require higher frequency [15].

Stationary Measurement System: The major limitations of PEMS are their form factor and frequency. These factors do not constrain stationary measuring devices used in laboratories or test facilities. As they are not required to be mobile, they can be built larger and heavier if needed. This flexibility allows the integration of more sensitive and sophisticated analyzers. Additionally, probe mounting points could be varied, allowing the ability to decide the measurement location. This capability becomes handier when unique measurements like raw emissions (directly after leaving the combustion chamber) have to be made. Furthermore, the measurement system is modular and can be replaced depending on the requirement. More advanced measurement systems provide higher data collection frequency.

Experimental Setup

Test Vehicle and Test Bench

As the primary goal of this work is to develop a model for transient operating conditions, and the physics-based modeling techniques were already identified, the data-driven modeling technique becomes the go-to solution. The general character of a transient operation in a vehicle is being short in time intervals. This nature demands high-frequency data collection to catch the internal dynamics entirely in the model. A BMW 530e test vehicle and its equivalent engine on the test bench have been set up to provide the possibility of interdisciplinary tests and measurements comprising real-world operation and HiL tests. The first objective of the test bench setup was to derive real-world realistic load cycles and their corresponding RDE cycles. These profiles will later be assigned as the test bench's base profiles. The test vehicle was equipped with a PEMS to capture emission data for various speed-torque profiles. One such profile is shown in Fig. 1. For all of these evaluations, AVL Concerto and Matlab have been used as efficient tools for the automation of evaluation procedures and for the visualization of test results.

As a starting point for emission modeling and test bench setup, the PEMS data provides comprehensive insight and reference data, particularly on the emission-critical operating ranges of the combustion engine. However, these measurements could not be directly used for data-driven modeling because of their lower data point frequency. Additionally, the PEMS measurements are taken from the exhaust system's tailpipe after the exhaust gas treatment system, whereas in this work, more importance is placed on providing an emission model for the output from the combustion chamber, which is usually designated as raw emission. Hence, measurements taken in front of the exhaust gas after-treatment system were considered. To further improve comparability between the test vehicle and the test bench, the same BMW B48 type engine was installed on the test bench. The engine parameters are detailed in Table 1.

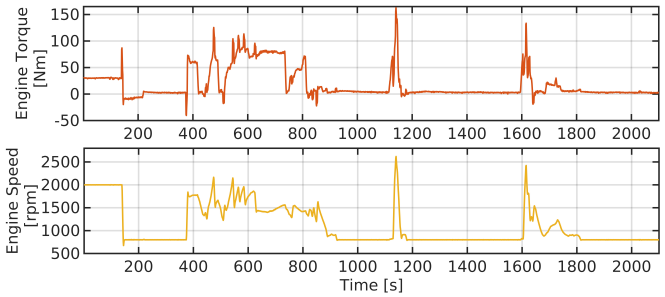

Figure 1: Speed-torque profile acquired during real driving tests.

Table 1: Technical details of the BMW B48 test engine.

Engine Parameters	Value	Unit
Maximum Power	135	kW
Maximum Torque	290	Nm
Displacement	1.998	dm^3
Compression Ratio	11.0	-
Bore	82.0	mm
Stroke	94.6	mm

As data-driven modeling ideally requires data containing all the dominant dynamics inside the system, conducting such complex experiments on an actual test vehicle could be challenging. Additionally, a measurement system with a higher frequency might have mounting space constraints on the vehicle. These limitations could be eliminated with the test bench. Various sophisticated measurement systems can be used for emission measurement, and complex driving profiles can be performed at the test bench as it is operated in a controlled environment. The current test bench setup is highly dynamic and automated, with an electric machine as the load unit and the engine control unit completely open, allowing access to all desired parameters while ensuring a fault-free combustion engine operation. The engine is also equipped with both in-cylinder and low-pressure (intake/exhaust) pressure indication, ensuring component safety and providing data for further investigations.

The raw emissions are sampled directly downstream of the turbocharger (Fig. 2), as it can be assumed that the exhaust gas is well mixed at this point. Due to the twin-scroll system, this is also the first possible sampling point at which the influence of all cylinders can be observed. The partial flow is transported to the emission-measuring device through a heated line. The measuring device analyzes the exhaust gas according to the principle of Fourier transform infrared spectroscopy. In order to analyze THC and O_2, additional modules are also integrated in the same system. The measurement system allows precise emission analysis at a frequency of 5 Hz - five times the resolution compared with on-road tests with PEMS. Additionally, the fast response feature in the measurement system improves the dynamic behavior.

Fig. 3 shows a schematic of the complete test bench set up with the engine, the electric machine used as the load, and the central measurement equipment. Standard equipment, such as temperature and pressure measuring points, are omitted to reduce the figure complexity.

Modeling Methodology

In this section, we utilize data-driven techniques to gain insights into the transient emission behavior of an HEV. Firstly, we highlight the

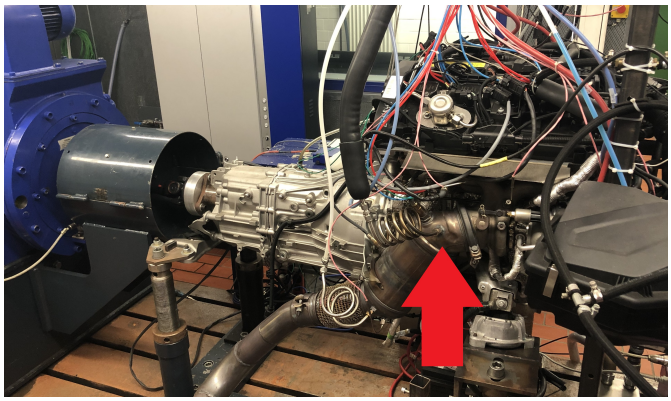

Figure 2: Test bench setup - raw emissions are sampled between the turbocharger and catalytic converter - red arrow.

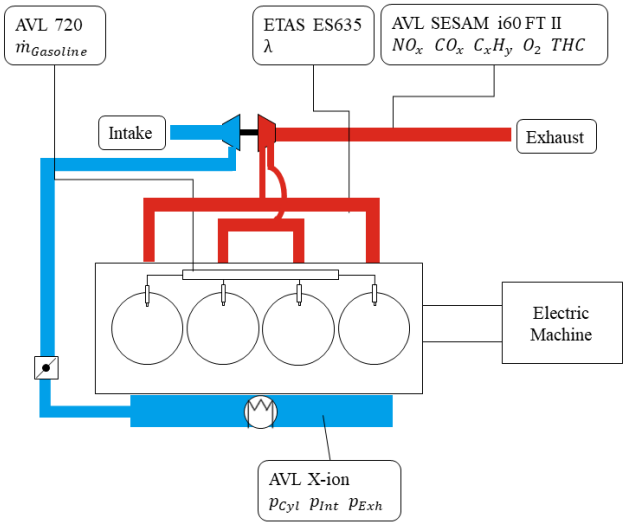

Figure 3: Schematic representation of the test bench setup.

most important interrelationships between the different measurements, and secondly, different methods are used to derive a data-driven model of the emission system.

Data Generation and Acquisition/ Experiment Design

As a data-driven model's quality depends on the data quality used for its training, it is crucial to obtain measurements that yield maximum information [16] from the system. The model quality is defined as how well the intrinsic dynamics of the system are captured and how explicitly they are differentiated. Hence, generating and recording data for all the system's dynamics is ideal. Therefore, it is essential to design trajectories that persistently excite the system at hand.

The initial focus in the dataset generation was put on the replication of the profiles initially done on the test vehicle BMW 530e during the PEMS analysis. Specific real-world driving profiles are used for this purpose which had load-speed profiles of various 90-minute RDE-compliant drives, with a roughly equal ratio between city, interurban, and highway driving. Distinctive driving profiles, which generally are outside the legally compliant accelerations, were also conducted. These unique longitudinal dynamics help to understand dynamic load-changing scenarios. In addition to the available OBD signals, the values determined by the PEMS (emissions, GPS, and temperatures) were also recorded. However, it was concluded that sticking to those profiles only captures particular dynamics rather than the full emission dynamics of the engine. Hence the techniques from classical system identification are also investigated. Literature refers to trajectory models that are spline-variants [17] or variants of periodic excitation [16, 18, 19], which could better touch the most of the dynamics spectrum. Among them, the popular model is the Fourier series, which is given for the measurement j as follows:

$$x_j(t) = \sum_{k=1}^{K} \left(\frac{a_k^{(j)}}{\omega_f^{(j)} k} \sin(\omega_f^{(j)} kt) - \frac{b_k^{(j)}}{\omega_f^{(j)} k} \cos(\omega_f^{(j)} kt) \right) + c_{j,0} \quad (1)$$

The fundamental frequency $\omega_f^{(j)}$ of measurement j can be chosen to be equal for all measurements ($=: \omega_f$), to ensure periodicity. $a_k^{(j)}$ and $b_k^{(j)}$ are the amplitudes for the $1 \dots K$ harmonic terms. $c_{j,0}$ is the offset and the trajectory evolves along time t.

As a rule of thumb, the conventional system identification literature suggests creating a trajectory with a sufficient number of sinusoids [20]. The parameters $a_k^{(j)}$, $b_k^{(j)}$, and $c_{j,0}$ should be determined to maximize the information content of the measurement time series. The book written by Khalil and Dombre [21] describes various criteria that can be chosen for selecting these parameters. However, only some of those criteria can be used for formulating black-box models. Finally, it was decided to produce multiple trajectories covering most of the measurement space, featuring many sine and cosine terms. This can be achieved using the following optimization problem:

$$\begin{aligned} &\underset{\mathbf{v}}{\operatorname{argmax}} \big(H(\mathbf{v}) + \Lambda(\mathbf{v}) \big) \\ &\text{s.t.} \quad \mathbf{x_{lb}} \leq \mathbf{x(t)} \leq \mathbf{x_{ub}} \quad \forall t \in [0, T]. \end{aligned} \quad (2)$$

The vector $\mathbf{v}$ summarizes the Fourier coefficients $[\mathbf{a}, \mathbf{b}, \mathbf{c_0}]^T$. The function $H(\mathbf{v})$ is a measurement of the dissimilarity of the Fourier coefficients, e.g., minimizing the correlation between them. In contrast, the function $\Lambda(\mathbf{v})$ consists of a weighted sum of the 1-norm and 2-norm of $\mathbf{v}$ to ensure that the coefficients become non-zero. The constraints $\mathbf{x_{lb}}$ and $\mathbf{x_{ub}}$ are the bounds for possible measurements. With different starting values for $\mathbf{v}$, presumably rich trajectories can be obtained.

Fig. 4 shows the final trajectories obtained for both engine rpm and throttle position. They feature 50 Fourier coefficients to parameterize a curve for both throttle pedal position and reference rpm. For the 30 minutes of data, the duration of one cycle is set to five minutes and repeated five times. The repetitions extend the data and mitigate the noise effect since each trajectory part is performed multiple times with a fundamental frequency set to $\omega_f = 2\pi/600$ s.

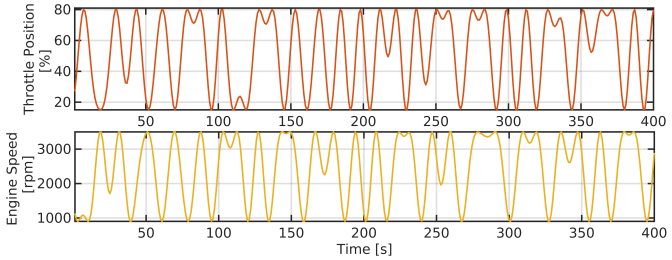

Figure 4: Fourier series trajectory.

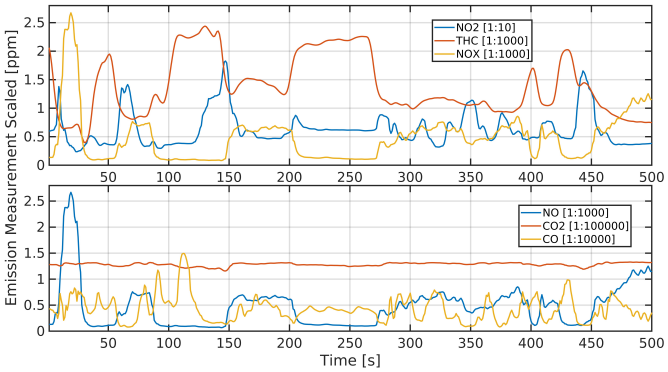

Figure 5: Measured emission during the RDE profile (scaled to fit).

The objective function term $H(\mathbf{v})$ was set to $5 \cdot \text{var}(\mathbf{v})$, and $\Lambda(\mathbf{v}) = 0.1\,\|\mathbf{v}\|_1 + 0.002\,\|\mathbf{v}\|_2$. The throttle pedal position is kept between 15% and 80% with rpm between 800 min^{-1} and 3500 min^{-1}. A sample measurement of emission recorded for 500 seconds with those trajectories can be seen in Fig. 5.

Data Processing

With the current configuration of the measurement and automation system from AVL PUMA, more than 80 parameters are recorded and collected in the dataset. However, not all of these measured parameters can directly be used for data-based modeling as this would exponentially increase the complexity and computational requirement. Moreover, some parameters might not even bring significant contributions to the model. These parameters, therefore, have to be filtered out, so various data preprocessing steps are done before the final data is used for modeling.

Correlation Analysis: At first, the significant emission outputs, which are detectable with the existing measurement technology, are identified. These included CO, CO_2, NO, NO_2, THC, and NO_x and have to be the output from the data-driven model. Other parameters contributing to changes in these emission values must be determined here. Correlation matrices are used for this analysis. A correlation matrix calculates the relationship between each column in the dataset and maps its intensity of it from a negative one to a positive one. A negative correlation value shows how deeply each parameter is inversely related, and a positive correlation value shows how much they are proportional to each other. For the dataset $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N]^T$, where $\mathbf{X}_j$ are vectors consisting of the time series data of measurement j, we can build the correlation matrix as

$$\mathbf{P} = \begin{bmatrix} 1 & \rho_{12} & \ldots & \rho_{1n} \\ \rho_{21} & 1 & \ldots & \rho_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1} & \rho_{n2} & \ldots & 1 \end{bmatrix}, \; \rho_{ij} = \frac{\text{Cov}(\mathbf{X}_i, \mathbf{X}_j)}{\sqrt{\text{Var}(\mathbf{X}_i)\text{Var}(\mathbf{X}_j)}}. \quad (3)$$

Shifting Data: Previous experience shows that some parameters have a direct temporal influence on emissions. This can be observed, for example, with increased load and thus increased temperature of nitrogen oxides. Due to the spatial separation of the exhaust gas measuring system from the object under investigation, a temporal distortion in the correlation must be assumed. The correlation analysis showed an improvement in the correlation values with time-lagged input. A detailed investigation revealed that the gas travel time from the combustion engine to the measuring device takes several seconds and is, therefore, not instantaneous. Consequently, the entire dataset is time-shifted to improve its correlation. Later in the result analysis, it is also observed that this shifting improved the modeling accuracy. The next step was to quantify the time shift required. A separate measurement program based on defined torque ramps, as shown in Fig. 6, is used for the estimation purpose. The results are filtered with positive load sequences and torque values as input. The choice of torque as input avoids the uncertainties that the engine speed might bring about.

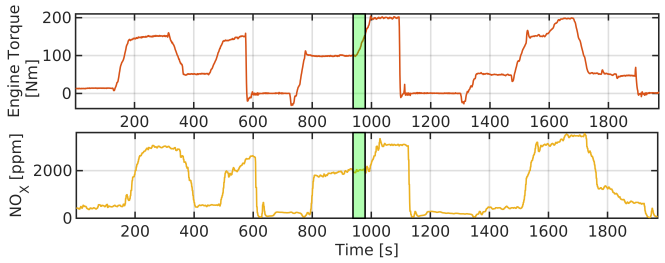

Figure 6: Ramp program to validate gas travel time - band cursor indicates signal delay.

The output measurements are then advanced in every discrete step, and the corresponding correlation value is calculated and plotted. One such sample could be seen for NO_x in Fig. 7. The final shift value (τ) is estimated with

$$\tau_{\text{estimated}} = \underset{\tau}{\text{argmax}}\,(Corr(\tau)) \quad (4)$$

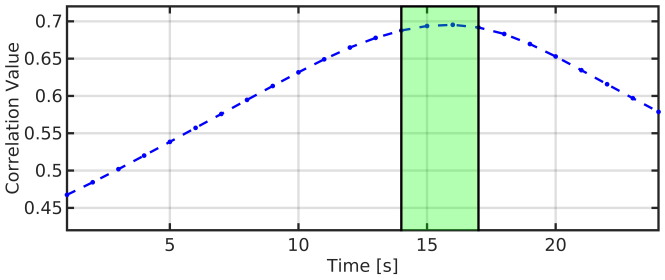

Figure 7: Time-shifted correlation analysis on NO_x.

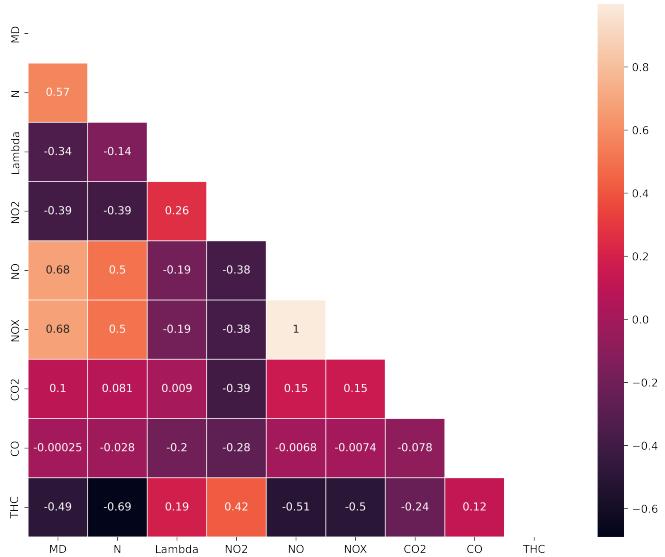

Figure 8: Correlation heat map obtained after data shifting.

Finally, a time-shifted correlation heat map, shown in Fig. 8 is obtained. This excerpt shows only a part of the parameters considered for modeling in this work.

Data-Driven Modeling

Most modern control and estimation techniques require mathematical models to obtain future predictions. A time-discrete dynamical system

in state space representation is denoted by

$$x_{k+1} = f\left(x_k, u_k\right) \tag{5}$$

which is fundamentally a one-step prediction model. If the function $f(x_k, u_k)$ is applied recursively

$$x_{k+T} = f^{(T)}\left(f^{(T-1)}\left(\dots f^{(1)}\left(x_k, u_k\right)\dots\right)\right) \tag{6}$$

we obtain a T-step prediction of our system. The function $f(x_k, u_k)$ might be derived analytically (white box), using data-driven techniques (black box), or by a mixture of both (grey box). In this paper, we focus on black-box techniques to find a model of the form

$$\mathbf{y} + \varepsilon = F_\beta\left(\mathbf{x}\right), \tag{7}$$

where the desired model $F_\beta(\cdot)$ maps the input vector $\mathbf{x}$ to the output vector $\mathbf{y}$ as accurate as possible. This can be achieved by varying the parameters β so that the error ε is minimized, which can be stated as the regression problem

$$\underset{\beta}{\operatorname{argmin}} = \left\|F_\beta\left(\mathbf{x}\right) - \mathbf{y}\right\|_2^2, \tag{8}$$

that can be approached through the use of conventional machine learning as well as deep learning. Note that both (9) and (6) can be modeled in the form of (7).

Nonlinear AutoRegression with eXogenous (NARX) models

Autoregressive functions try to model a predictive function using p past input values

$$x_{k+1} = f\left(x_k, x_{k-1}, \dots, x_{k-p}, u_k, u_{k-1}, \dots, u_{k-p}\right), \tag{9}$$

where the arguments are the exogenous inputs u_i and the internal states x_j. The function $f(\cdot)$ can be modeled using the following approaches.

1) Random Forest Regression: Originating from a root node, decision tree learning looks for optimal top-down data partitioning into subsets (branches). The most distal nodes (leaves) are the outputs of these models. Random Forest Algorithms generate an ensemble of decision trees to allow for more representation power and prevent overfitting, which might happen when polling only from a single decision tree [22].

2) XGBoost Regression (XGB): The XGB algorithm recently came into the limelight with its efficient implementation and capability to combine various weak learning models to create a model with better accuracy. It belongs to the gradient-boosting family of algorithms that use tree ensemble base learners to generate different branches and various learners. Initially, each learner is equally treated. However, in each iteration, the weak classifier observations are weighed higher than the ones from the robust classifiers. In general, machine learning algorithms use random guessing, especially in the initial stages of learning. Tree ensemble-based algorithms consider that weak learners are much better than random guessing. XGBoost not only adapts this technique but also ensures that the weak learners are less commonly used as the iteration progresses.

3) Support Vector Regression (SVR): In ordinary regressions, the main objective is to reduce the error sum in the deviation between the predicted and actual value $\mathbf{W}$. Additionally, penalty parameters are introduced as Lasso, Ridge, and ElasticNet techniques to reduce the number of feature parameters used for the modeling. However, an SVR's main objective is shifted to reducing the regression model's coefficient values. Generally, the *l2*-norm of the coefficient vector is used for this purpose. A bound for the absolute error ε (epsilon) is set prior and lets the objective function find hyperplanes in higher dimensions to fit the data. The problem can be stated as

$$\begin{aligned} &\min\left(\frac{1}{2}\left\|\mathbf{W}\right\|^2 + C\sum_{i=1}^{n}\left|\varepsilon_i\right|\right) \\ &\text{s.t.} \quad \left|y_i - w_i x_i\right| \le \epsilon + \left|\varepsilon_i\right| \ , \end{aligned} \tag{10}$$

where w_i is the coefficient, and x_i is the predictor. An error value deviation margin ε is also added to relax the constraint and avoid non-solvable situations.

Multi-Layer Perceptrons

Artificial neural networks are sets of connected computing nodes called artificial neurons. Typically, these nodes are clustered into layers. The most simplistic artificial neural network architecture is called multi-layer perceptron and is built by multiple layers, including an input layer, hidden layers, and an output layer [23]. All units from adjacent layers are connected. This is described by N chained, nonlinear functions

$$\mathbf{y} = f_{\beta_N}^{(N)}\left(f_{\beta_{N-1}}^{(N-1)}\left(\dots f_{\beta_1}^{(1)}\left(\mathbf{x}\right)\dots\right)\right), \tag{11}$$

where β_i consists of the weights $\mathbf{W}_i^T$ and biases $\mathbf{b}_i$ of each - so-called - a layer of the network. Every layer performs a transformation

$$f_{\beta_i}^{(i)} = g(\mathbf{W}_i^T\,\mathbf{h}_{i-1} + \mathbf{b}_i), \tag{12}$$

where $g(\cdot)$ is a nonlinear activation function and $\mathbf{h}_{i-1}$ is the output of the previous layer [24].

Long Short-Term Memory Neural Networks

The application of feedforward neural networks, including MLPs, is limited to static classification and regression tasks. In contrast, Recurrent Neural Networks (RNNs) have an internal state, making them a dynamic system. There can be feedback connections between higher and lower layers and self-feedback of neurons, allowing data from earlier processing steps to influence the current step. In that way, a memory of time series events is generated. However, RNNs can only learn short time series because they face the problem of vanishing or exploding errors [25]. Long Short-Term Memory Neural Networks (LSTMs) extend RNNs by gating units that switch specific flows of information on and off [24], leading to the ability to learn up to more than 1,000 steps [25]. Featuring these properties, LSTMs are amongst the most commonly used elements for time series predictions.

Results and Discussion

In the evaluation stage, the initial focus was given to the correlation matrix (Fig. 8) and its reliability. For this purpose, a plot with the input torque (in Nm) and the engine speed (in rpm) against each pollutant is obtained to verify the assumed correlations. Such a plot obtained for NO_x can be seen in Fig. 9. The positive correlation (value = 0.68) can be qualitatively confirmed from the figure as an increase in torque leads to a rise in the NO_x concentration as well. In the choice of input parameters, the primary goal was to utilize easily measurable variables, which are also simple to predetermine in the subsequent optimization of the operating strategy and yet have a good correlation with the output variables. Thus, the engine speed and air-fuel ratio were selected in addition to the torque. These parameters allow a good comparison of the following modeling approaches, whereby the best approach can be extended in complexity.

All models are trained and validated on a dataset consisting of several RDE measurements, followed by a Fourier series profile obtained to solve the optimization problem (2). The test dataset is another 15-minute RDE profile with highly dynamic transients, shown in Figure 10. It has to be mentioned that this data has not been part of the training data. The test data is very dynamic regarding the speed signal with various up- and downshifts and the torque demand. Lambda is usually very close to 1 for the engine under study and typically ranges from 0.95 to 1.05. The training data is first normalized with the min-max scalar technique and split for training and validation. Ideal hyperparameters are essential for model accuracy, but finding them can only be done with heuristics. A parameter sweep technique was implemented

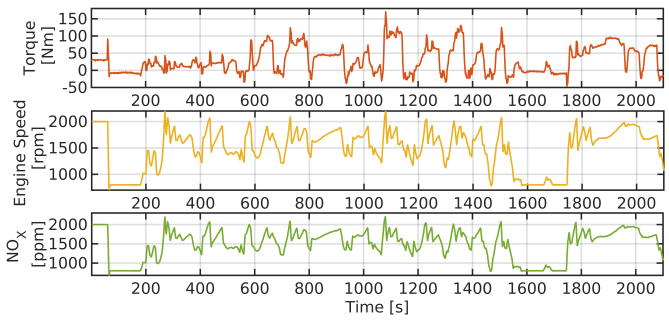

Figure 9: Measured RDE profile with NO_x.

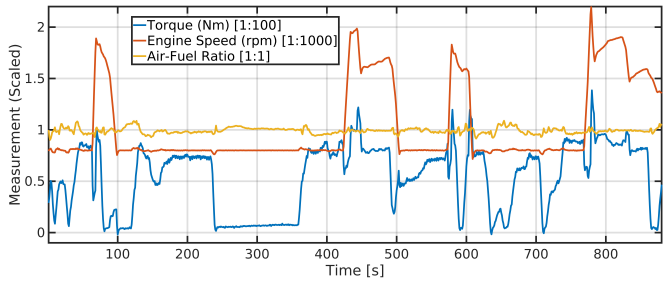

Figure 10: Torque, engine speed, and air-fuel ratio measured for the 15-minute profile.

on the university's high-performance computer for this purpose. All following graphs and error measures refer to the test data set.

Table 2 summarizes the chosen models and the corresponding hyperparameter choices. The NARX models create a unique model for each pollutant, whereas the neural network models predict multiple outputs for the corresponding input matrix. The inputs to the neural network models are only the exogenous input time series. However, the NARX models require not only the exogenous inputs but also the past internal states. These previous state requirements are satisfied with the ground truth data. The RMSE (Root Mean Squared Error) between the related model predictions and the test data time series is used as a comparison metric. Each NARX approach consists of multiple submodels of pollutants, so their average RMSE is chosen as the final RMSE. Figures 11 and 12 show the prediction performance of the feedforward neural network, LSTM network, and the NARX models against the actual emission output.

The LSTM model performed better than the FFNN model within the neural networks. The RMSE value is approximately 10 percent lower. This capability comes from the model characteristics itself. The FFNN model receives specific inputs at a single time step and tries to model the output regardless of the previous time steps. Hence, in steady-state conditions, FFNN models can be beneficial. However, this practice becomes a restriction during transient engine operation and emission prediction, as the torque and speed gradients significantly influence the engine's emission behavior. In these situations, LSTMs perform better than FFNNs as they consider the effects of earlier step data for the current step. Looking at the different emission components, there are differences in the performance of the models. For example, FFN and LSTM models estimate the NO emissions reasonably well, whereas, for THC, the FFN model performs significantly worse than the LSTM model.

The NARX models generally approximate the emissions well and capture the dynamics of all emission types. Comparing the NARX models with the neural networks concerning the RMSE value, it is noticeable that the values of the neural networks are 5 to 7 times higher. Among the NARX models, XGBoost and Random Forest regressions exaggerate the dynamics at some spikes. Qualitatively, the support vector regression approximates the dynamics best. The NARX models are more sample efficient and can produce better results on the small training data set. In this study, it is assumed that the ground truth of the internal states $x_k, x_{k-1}, \ldots, x_{k-p}$ are accessible to produce a one-step prediction x_{k+1}. In practical scenarios, for example model-based control applications, the individual pollutants would need to be measured online, which is likely non-feasible. Additionally, it is found that the FFNN and LSTM models suffer from the size of the training data set, even though the results are comparable even for vastly different hyperparameters. For further studies, the NARX models have the highest potential to be considered accurate enough to model the transient emission behavior of engines for advanced control applications.

Conclusion

In this paper, data-driven modeling of emissions from gasoline engines under transient operation was presented with both conventional machine learning and deep learning methods. The focus was primarily on the methodology and comparison of the different approaches. The real-world driving profiles were collected from a test vehicle and then repeated on a transient engine test bench. The data collected from various such profiles were used for training the models. It was concluded that the non-linear auto-regression models have a significant advantage and perform better than the deep learning models, improving the RMSE by [almost] a factor of ten. Not surprisingly, stateful models considering the influence of past time steps, like NARX and LSTMs, are superior to feedforward networks when predicting transient emissions. Despite the small dataset and the use of just three fundamental input parameters — torque, speed, and lambda — the models could deliver good qualitative performance and accurate quantitative results except for a few situations. These results provide an excellent intuition regarding the creation of emission formation models which can be used to optimize the operating strategy. In this respect, an increase in the dataset size and the choice of additional input parameters should lead to better results in the prediction quality.

Outlook

Regarding measurement technology, two approaches in the present investigations offer optimization potential for future data-based emissions modeling, especially in transient operation. The current 5 Hz measurement system - even though it can approximate the primary dynamic processes - results in specific dynamics inevitably getting smoothed out and never reaching the actual emission peaks in the measurement dataset. If models are built on these datasets, the smoothing of these spikes might reduce the efficiency of the operating strategy with respect to avoiding emissions.

Furthermore, the spatial separation of the emission measurement device and the measurement location could be improved in the current setup. As discussed earlier, the gas transit time is necessary to correct the measurement data to compensate for the time discrepancy. It is challenging to find a robust methodological solution to obtain this value. Additionally, further mixing of the measurement gas or subsequent reactions in the heated line to the measurement system cannot be excluded. Therefore, in further investigations, a laser-based measurement system will be used to analyze the exhaust gas in the process (in situ) without a significant time delay and with an elevated measurement frequency of up to 1 kHz. For the reasons mentioned above, this should improve the data-based methods and provide new insights into the transient emission formation in the present experimental setup.

Another potential area of further study is the search for suitable input parameters which could significantly improve the quality of the models. In addition to the selected parameters, other parameters can be measured directly or entered as maps. It would also be possible to use variables originating from physical modeling and extend the black box model to a grey box model. Values characterizing the conditions in the combustion chamber during transient operation, such as pressure gradients or the residual gas content, would be interesting for this purpose.

Table 2: Modeling methods and respective root mean squared errors. For each method, the best configuration of hyperparameters was determined by a parameter sweep. The hyperparameter ranges used during the parameter sweep, the best configurations, and the corresponding root means square errors are included. The final model configuration chosen is highlighted by underlining the corresponding parameter value.

Model	**RMSE (Best)**	**Varied Parameters (best configuration is underlined)**
Non-linear Auto-regression		
• Random Forest	0.0029	Estimators: [20, 30, <u>50</u>, 100], Regression Order: [2, 3, <u>6</u>]
• XGBRegressor	0.0023	Estimators: [20, 30, <u>50</u>, 100], Regression Order: [2, 3, <u>6</u>]
• SVR	0.0021	Kernal (gamma) : [0.001, 0.01, <u>0.05</u>], Regularization (C): [0.1, <u>1</u>, 100], Training loss penalty factor (epsilon): [0.001, 0.01, <u>0.03</u>]
Deep Learning Models		
• Feedfoward-NN	0.0168	Fully connected layers: [4, 10, <u>16</u>]; Nodes: [200, <u>300</u>, 400]; Circulating learning rate scheduling; All resulting RMSE were in similiar range.
• LSTM-NN	0.0152	LSTM layers [1, 2, <u>4</u>, 8]; Nodes: [4, 8, <u>16</u>, 32]; Circulating learning rate scheduling; All resulting RMSE were in similiar range.

Figure 11: Emission prediction results with LSTM and FFNN. All pollutants are drawn on a unitless and normalized scale between 0 and 1.

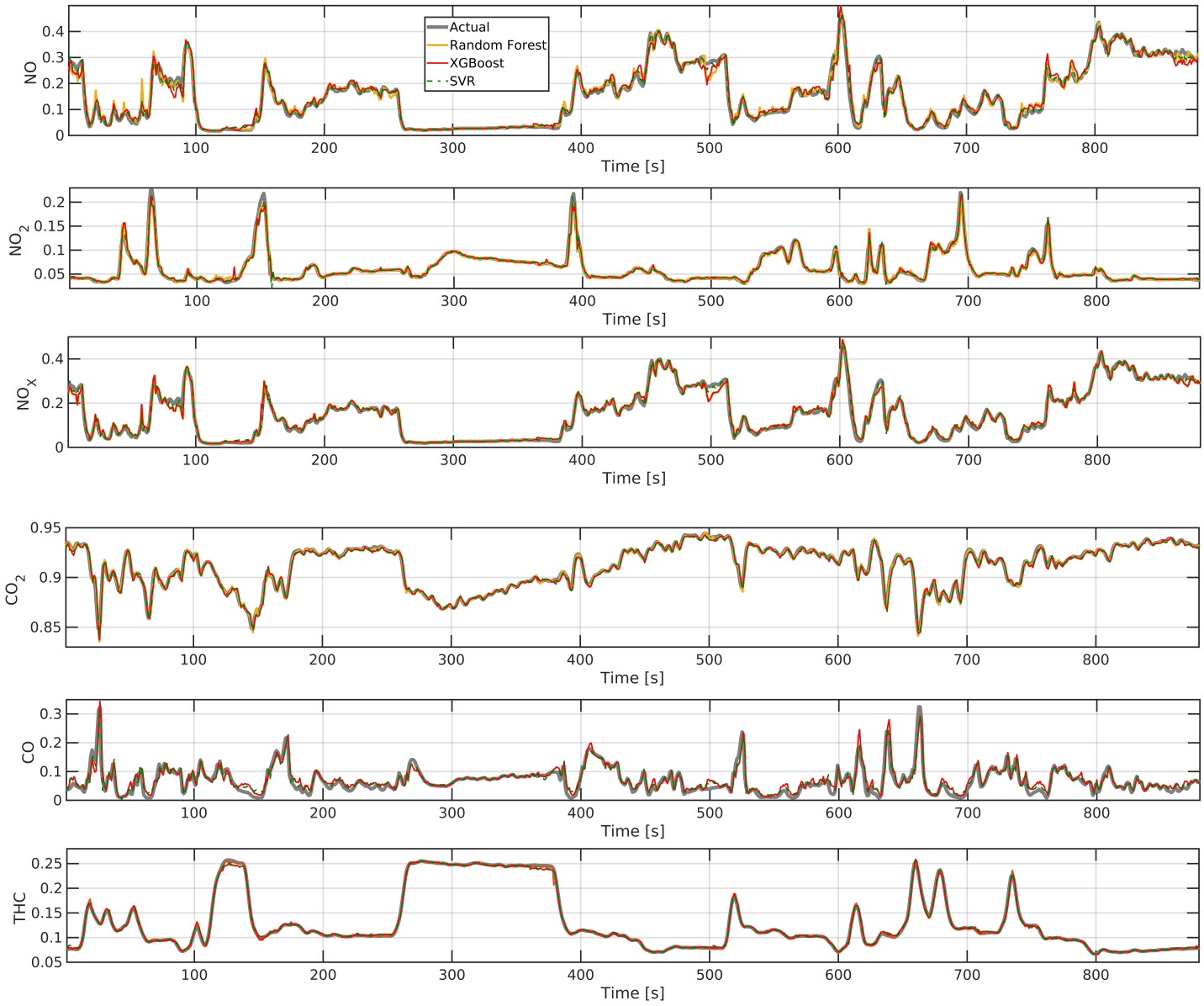

Figure 12: Emission prediction results with Random Forest, XGBoost, and SVR. All pollutants are drawn on a unitless and normalized scale between 0 and 1.

References

1. L. Schipper, C. Saenger, and A. Sudardshan, "Transport and carbon emissions in the united states: the long view," *Energies*, vol. 4, no. 4, pp. 563–581, 2011.

2. J. Delbeke, P. Vis, G. Klaassen, J. Lefevere, and M. Damien, *EU Climate Policy Explained*. Routledge London, 2015.

3. Y. Liu and C. Cirillo, "Evaluating policies to reduce greenhouse gas emissions from private transportation," *Transportation Research Part D: Transport and Environment*, vol. 44, pp. 219–233, 2016.

4. P. K. Wong, H. C. Wong, C. M. Vong, Z. Xie, and S. Huang, "Model predictive engine air-ratio control using online sequential extreme learning machine," *Neural Computing and Applications*, vol. 27, no. 1, pp. 79–92, 2016.

5. S. Lee, H. Choi, and K. Min, "Reduction of engine emissions via a real-time engine combustion control with an egr rate estimation model," *International Journal of Automotive Technology*, vol. 18, no. 4, pp. 571–578, 2017.

6. R. Liessner, A. M. Dietermann, and B. Bäker, "Safe deep reinforcement learning hybrid electric vehicle energy management," in *International conference on agents and artificial intelligence*, pp. 161–181, Springer, 2019.

7. M. Schudeleit, W. Gu, F. Küçükay, and M. Zhang, "Fuel and co2 savings in real driving using machine learning hev operating strategy," in *17. Internationales Stuttgarter Symposium*, pp. 659–677, Springer, 2017.

8. A. Taghavipour, N. L. Azad, and J. McPhee, "Real-time predictive control strategy for a plug-in hybrid electric powertrain," *Mechatronics*, vol. 29, pp. 13–27, 2015.

9. A. S. Silitonga, M. H. Hassan, H. C. Ong, and F. Kusumo, "Analysis of the performance, emission and combustion characteristics of a turbocharged diesel engine fuelled with jatropha curcas biodiesel-diesel blends using kernel-based extreme learning machine," *Environmental Science and Pollution Research*, vol. 24, no. 32, pp. 25383–25405, 2017.

10. S. R. Dadam, M. Van Nieuwstadt, A. Lehmen, V. K. Ravi, V. Kumar, and R. Bhat, "A unique application of gasoline particulate filter pressure sensing diagnostics," *SAE International Journal of Passenger Cars-Mechanical Systems*, vol. 14, no. 06-14-02-0007, pp. 105–116, 2021.

11. R. Pischinger, M. Klell, and T. Sams, *Thermodynamik der Verbrennungskraftmaschine*. Springer-Verlag, 2009.

12. G. P. Merker and R. Teichmann, *Grundlagen Verbrennungsmotoren: Funktionsweise und alternative Antriebssysteme Verbrennung, Messtechnik und Simulation*. Springer-Verlag, 2019.

13. S. Esposito, L. Diekhoff, and S. Pischinger, "Prediction of gaseous pollutant emissions from a spark-ignition direct-injection engine with gas-exchange simulation," *International Journal of Engine Research*, vol. 22, no. 12, pp. 3533–3547, 2021.

14. L. Cai, A. Ramalingam, H. Minwegen, K. A. Heufer, and H. Pitsch, "Impact of exhaust gas recirculation on ignition delay times of gasoline fuel: An experimental and modeling study," *Proceedings of the Combustion Institute*, vol. 37, no. 1, pp. 639–647, 2019.

15. H. Tschöke, "Hintergrund und Motivation," in *Real Driving Emissions (RDE)*, pp. 1–15, Springer, 2019.

16. L. Ljung, *System Identification*, pp. 163–173. Birkhäuser Boston, 1998.

17. W. Rackl, R. Lampariello, and G. Hirzinger, "Robot excitation trajectories for dynamic parameter estimation using optimized b-splines," in *2012 IEEE International Conference on Robotics and Automation*, pp. 2042–2047, 2012.

18. J. Swevers, W. Verdonck, and J. De Schutter, "Dynamic model identification for industrial robots," *IEEE Control Systems Magazine*, vol. 27, no. 5, pp. 58–71, 2007.

19. J. Swevers, C. Ganseman, J. De Schutter, and H. Van Brussel, "Experimental robot identification using optimised periodic trajectories," *Mechanical Systems and Signal Processing*, vol. 10, no. 5, pp. 561–577, 1996.

20. P. Annus, R. Land, M. Min, and J. Ojarand, "Simple signals for system identification," *Fourier Transform-Signal Processing. In-Tech: Rijeka*, pp. 257–276, 2012.

21. W. Khalil and E. Dombre, *Modeling, Identification and Control of Robots*. Butterworth-Heinemann, 2004.

22. S. L. Brunton and J. N. Kutz, *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*. USA: Cambridge University Press, 1st ed., 2019.

23. F. Murtagh, "Multilayer perceptrons for classification and regression," *Neurocomputing*, vol. 2, no. 5, pp. 183–197, 1991.

24. I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

25. R. C. Staudemeyer and E. R. Morris, "Understanding LSTM - a tutorial into long short-term memory recurrent neural networks," *ArXiv*, vol. abs/1909.09586, 2019.

Contact Information

Tobias Gehra, M.Sc.
RPTU Kaiserslautern-Landau, Germany
Institute of Vehicle Propulsion Systems (LAF)
tobias.gehra@mv.uni-kl.de

Ganesh Sundaram, M.Sc.
RPTU Kaiserslautern-Landau, Germany
Institute of Electromobility (LEM)
sundaram@eit.uni-kl.de

Acknowledgments

The work presented in this paper was carried out within the framework of the project ML-MoRE (Maschinelles Lernen für die Modellierung und Regelung der Emissionen von Hybridfahrzeugen in Realfahrzyklen). The German Federal Ministry of Education and Research (BMBF) funded the project under grant 01-S20007A. The responsibility for the content of this publication lies with the authors. The authors would like to express their gratitude for the funding.

SPONSORED BY THE

Federal Ministry
of Education
and Research

Furthermore, the authors would like to thank the project partners KST Motorenversuch and RA Consulting for their support throughout the project.