Reconstructing Missing NOx Emissions in Heavy-Duty Diesel Vehicle OBD Data：A Machine learning approach

Author names: Zeping Cao [a], Yanan Wang [a], Xiaoyang Zhao [a], Jiawei Yin [a], Zhenyu Jia [a], Yanjie Zhan [b], Yan Liu [a] *, Qijun Zhang [a] **, Hongjun Mao [a]

Author Address: a Tianjin Key Laboratory of Urban Transport Emission Research & State Environmental Protection Key Laboratory of Urban Ambient Air Particulate Matter Pollution Prevention and Control, College of Environmental Science and Engineering, Nankai University, Tianjin, 300071, China

b Tianjin Youmei Environment Technology, Ltd., Tianjin, 300380, China

Corresponding Author's e-mail address: liuyanwork@nankai.edu.cn (Y.Liu), zhangqijun@nankai.edu.cn (Q.Zhang).

ABSTRACT. On-Board Diagnostic (OBD) systems enable real-time monitoring of heavy-duty diesel vehicle operations and NOx emissions. However, existing OBD systems have inherent limitations, including systematic missing values. To address this issue, this study develops a data-driven approach. The method utilizes OBD-recorded upstream and downstream NOx emissions data to build machine learning models for reconstructing real-world OBD data. The modeling results indicate that machine learning models perform well in predicting upstream emissions, achieving an R^2 above 0.9 on the test set. However, its performance in predicting downstream emissions was highly variable, with R^2 ranging from 0.05 to 0.98, and showed a

positive correlation with fuel-based emission factors. A case study was conducted on a selected vehicle. The total NOx emission associated with missing data for this specific vehicle was estimated at 15,741.3 g, whereas the recorded emission from available data was 6,157.3 g. Missing data were then imputed for an additional 31 vehicles, revealing that normal emitters showed significantly higher emission associated with missing data. The proposed approach is highly compatible with existing big data platforms and can be easily extended to other vehicles. This will improve the platform's representation of real-world emission, enabling policymakers to implement more targeted pollution mitigation strategies.

Environmental Implication. On-Board Diagnostics (OBD) systems are widely deployed for real-time monitoring of NOx emissions from HDDVs. However, systematic data errors and missing values hinder their effectiveness in accurately assessing real-world emissions. This study develops a machine learning-based approach to reconstruct missing NOx emissions, enabling more reliable emission estimations. By improving data integrity, this method enhances the precision of identifying high-emission events and vehicles, facilitating more effective regulatory enforcement. Given that NOx is a major contributor to air pollution and adverse health effects, accurately quantifying its emissions is crucial for developing targeted mitigation strategies and reducing environmental harm.

KEYWORDS: On-board diagnostics, NOx emission, Heavy-duty diesel vehicle, Machine learning, data imputation.

1. Introduction.

With the ongoing advancements in battery and motor technologies, vehicle electrification has increasingly emerged as a predominant trend in the development of the motor vehicle sector in China, leading to a substantial reduction in emissions from mobile sources[1-3]. Among various vehicle types, heavy-duty vehicles (HDVs) are characterized by their substantial weight and long transport distances. These vehicles predominantly rely on diesel power and, in the short term, face significant challenges in achieving large-scale electrification[4]. Although heavy-duty diesel vehicles (HDDVs) constitute a relatively small proportion of the total number of motor vehicles, their contribution to emissions from mobile sources is highly significant[5-7]. According to the 2023 China Mobile Source Management Annual Report, their NOx emissions account for 76.1% of the total emissions from mobile sources, while particulate matter emissions represent 50.7%[8]. Therefore, controlling emissions from HDDVs has become a critical pathway in reducing emissions from mobile sources.

In order to control emissions from HDDVs, the Chinese government has formulated and implemented the China VI emission standards[9]. These standards introduced the World Harmonized Stationary Cycle (WHSC) and World Harmonized Transient Cycle (WHTC) as testing cycles, replacing the previous European Stationary Cycle (ESC) and European Transient Cycle (ETC)[10]. Building upon the China V emission standards, the China VI standards further tighten emission limits, reducing the NOx and particulate matter (PM) limits in standard test cycles by 77%

and 67%, respectively[11]. In addition, the standards introduce requirements for World Harmonized Not-to-Exceed (WNTE) and Portable Emission Measurement System (PEMS), and for the first time, include the requirement for a remote emissions management in national standards, aiming to enable real-time monitoring and management of NOx emissions from HDDVs[12]. Specifically, China VI-a mandates that newly sold HDVs be equipped with an On-Board Diagnostics (OBD) system that includes a remote emissions management terminal, while China VI-b stipulates that vehicles should be connected to the network and upload data throughout their entire lifecycle[13].

OBD is an automated remote monitoring device that collects vehicle operation and NOx emission data transmitted via the CAN bus through the OBD interface[14]. Under the coordination of the microcontroller unit (MCU), it integrates the location information obtained by the positioning module and transmits the data remotely to the monitoring platform via the network module, enabling efficient data monitoring and transmission[15]. In contrast to PEMS testing and chassis dynamometer testing, which are expensive, cumbersome, and only reflect NOx emissions of certain vehicles within limited temporal and spatial scopes[16-18], the OBD system provides an opportunity to capture the operational status and NOx emissions of HDDVs across their entire lifecycle, thereby enabling precise and scientific dynamic monitoring and management of NOx emissions[19]. With the implementation of regulations, a large number of HDDVs have been equipped with OBD systems through retrofit and original equipment manufacturer (OEM)-performed installation methods[20]. In order

to validate the accuracy of OBD data, a few works have been accomplished. Cheng et al. conducted a 20-day experiment by integrating OBD and PEMS devices, and reported a strong Pearson correlation between the NOx emission concentrations measured by the two systems[21]. Zhang et al. further confirmed that OBD data on fuel consumption, engine speed, and vehicle speed were reliable and aligned with PEMS measurements[22]. Based on the reliable conclusions drawn from OBD data, researchers have conducted substantial work in policy recommendations, high-emission vehicle screening, and inventory preparation, with much of this work built upon data cleaning[23-28]. Zhang et al. proposed a standard data processing procedure, which includes handling timestamp errors, missing or invalid NOx data, and speed errors[22]. Zhao et al. adopted the data processing methods from Zhang et al., analyzing the large-scale application of OBD data in Beijing, China[23]. Lv et al. recommended addressing out-of-bound errors, repetition errors, invalid data errors, and timestamp errors[25]. Based on the processed data, they constructed a real-time emission inventory for HDVs in Tangshan, China. Similarly, Wang et al. filtered valid China VI OBD data, calculated the average emission factors (EFs) of vehicles, and estimated the emission reduction benefits of implementing the China VI standard[26]. The data cleaning–analysis workflow has become a critical component of the mainstream technical pathway in OBD-related research. However, the NOx emission control monitoring system in HDDVs is deactivated when the engine coolant temperature drops below 70°C[9], inevitably resulting in biased data loss, which compromises the completeness and rigor of related research efforts.

Unlike the systematic loss of NOx sensor data, OBD's other parameters demonstrate exceptionally high completeness[25], particularly for OEM-performed China VI HDDVs. These data include transient vehicle operating conditions and environmental parameters beyond NOx emissions (Table S1). When the NOx sensor functions properly, data mining and analysis enable the construction of a mapping relationship between these parameters and the vehicle's transient NOx emissions. Extending this mapping to datasets with missing NOx data allows for a comprehensive view of real-world emissions. The development of individualized emission models serves as an effective technical approach to capturing such mapping relationships. With advancements in artificial intelligence, machine learning offers a feasible and efficient solution to this challenge[29-31] and has been extensively applied in emission modeling[32, 33]. However, current research primarily relies on PEMS data, and the applicability of machine learning models to OBD data, particularly data downloaded from data platforms, requires further investigation.

In order to address the aforementioned issues, this study utilized OBD data from 32 HDDVs downloaded from the Tianjin Ecology and Environment Bureau On-Board Diagnostic Platform. Based on this data, two distinct models were developed for each vehicle to estimate NOx emissions at the SCR inlet and outlet. Based on these models, missing data were categorized into three types according to SCR inlet temperature, and different strategies were applied using the two types of models to perform data imputation. Finally, the emissions of HDDVs were reconstructed based on the imputed results. The contributions of this study to the existing literature are as follows:

First, it examines the applicability of machine learning models to OBD data. Second, it proposes a historical data-driven imputation method for predicting the NOx emissions corresponding to missing data rows in OBD datasets, addressing data gaps caused by temporary sensor malfunctions and engine coolant temperatures below 70°C. The proposed method aims to reconstruct the full picture of real-world HDDVs emissions as accurately as possible. Moreover, since the reconstructed data include cold-start emissions, this approach could assist policymakers in better understanding real-world cold-start emissions from vehicle fleets, thereby enabling the development of more targeted and effective policies.

2. Material and Method.

2.1. Data acquisition and processing.

The data used in this study were sourced from the Tianjin Ecology and Environment Bureau On-Board Diagnostic Platform, covering OBD records from 32 HDDVs. These vehicles share the same model, with data from 20 vehicles covering a one-month period, 6 vehicles covering two months, and another 6 vehicles covering one year. Table S2 provides a detailed listing of the vehicle ID, vehicle model, and engine model.

In the raw OBD data, three main types of data errors are identified: timestamp errors, out-of-bound errors, and constant-value errors. Timestamp errors can be further classified into two types: (1) temporal duplication, which includes both cases where the timestamp and data values are exactly repeated, and where the timestamp is repeated but the data values differ; (2) temporal gaps. Out-of-bound errors refer to

cases where the value of one or more specific data variables exceeds the normal measurement range of the sensor, with the range criteria provided in Table S1. Constant-value errors occur when a sensor value remains constant over an unreasonable period. In this study, a threshold of 15 was selected to identify such occurrences. The variables assessed for this criterion include Fuel Rate, Intake Flow Rate, SCR Upstream NOx Concentration, SCR Upstream Temperature, SCR Downstream NOx Concentration, and SCR Downstream Temperature. Figure S1 illustrates specific examples of these data errors.

This study categorizes the data into three levels based on the extent of preprocessing applied. Raw Data refers to the original dataset before any processing has been applied. Preprocessed Data (PD) encompasses data that has been processed for timestamp errors and out-of-bound errors, excluding NOx sensor data. The procedure consists of two steps. (1) For temporal duplication errors, rows with identical data values are discarded. In cases where the data values are not repeated, and timestamp breaks coincide with the duplicated timestamps, the duplicated rows are sequentially filled into the breaks. Otherwise, the repeated sequences are resampled by averaging. (2) Out-of-bound errors are addressed using linear interpolation based on adjacent data points, as other variables, apart from the NOx sensor data, rarely exhibit continuous errors (less than 1% for most vehicle brands)[25].

Cleaned Data (CD) is built upon the PD dataset, which undergoes additional out-of-bound error handling for NOx sensor data and constant value error correction. To

ensure data reliability, all rows containing errors are removed from the dataset. The focus of this study's data imputation is on PD dataset because it retains the majority of operational information. In contrast, CD is used to establish the relationship between operational conditions and upstream/downstream NOx emissions. Table 1 presents the processing methods implemented for different data types.

Table 1. Data processing implementation for different data types

Error Names	Raw Data	Preprocessed Data	Cleaned Data
Time stamp error	Unprocessed	Identical data – Delete	Identical data – Delete
		Not identical – Timestamp fix/ resample	Not identical – Timestamp fix/ resample
Out-of-bound error	Unprocessed	Outliers-prone columns (Excluding NOx data) – Resample	Outliers-prone columns (Including NOx data) –Resample
Constant-value error	Unprocessed	Unprocessed	Constant-prone columns – Delete

2.2. Machine Learning Model Construction.

Two types of models were developed in this study. Engine-out Emission Model (EEM) establishes the relationship between features and NOx emissions at the upstream of the SCR system, while Pipe-out Emission Model (PEM) directly establishes the relationship between features and NOx emissions at the downstream of the SCR system. For EEM, the selected raw features include: vehicle speed (v), engine net output torque (T_{out}), friction torque (T_f), engine speed (RPM), engine fuel flow (F_{fuel}), and intake air flow (MAF). To enhance the model's R^2, time-series features

were introduced for each raw feature. For instance, for the vehicle speed (v), additional input features were created by incorporating the speed at the previous time step (v_{t-1}) and the speed at the next time step (v_{t+1}). It is important to note that these features intentionally exclude temperature-related data from the raw data, such as engine coolant temperature (ECT), SCR inlet temperature ($T_{SCR\text{-}in}$), and SCR outlet temperature ($T_{SCR-out}$), to avoid the influence of missing data at lower temperatures.

The target for EEM is the instantaneous upstream NOx emission mass ($NO_{x,up,t}^{mass}$), calculated as follows:

$$NO_{x,up,t}^{mass} = \frac{0.001 \times 46}{3600 \times 22.4} \times \frac{1}{\rho_e} \times NO_{x,up\text{-}stream,t} \times \left(F_{fuel,t} \times \rho_{fuel} + MAF_t\right)$$

Where ρ_e is the air density, equal to 1.29 kg/m³. $NO_{x,up\text{-}stream,t}$ represents the output value of the NOx upstream sensor at time t (in ppm), $F_{fuel,t}$ is the engine fuel flow rate at time t, and ρ_{fuel} is the fuel density, taken as 0.85 kg/L in this study. MAF_t is the intake air flow rate at time t.

For PEM, the selected raw features include v, T_{out}, T_f, RPM, F_{fuel}, MAF, ECT, $T_{SCR\text{-}in}$, $T_{SCR-out}$. Similar to EEM, temporal features before and after each raw feature are also introduced. The target for PEM is the instantaneous downstream NOx mass emission ($NO_{x,down,t}^{mass}$), and the calculation formula is as follows:

$$NO_{x,down,t}^{mass} = \frac{0.001 \times 46}{3600 \times 22.4} \times \frac{1}{\rho_e} \times NO_{x,down\text{-}stream,t} \times \left(F_{fuel,t} \times \rho_{fuel} + MAF_t\right)$$

Where $NO_{x,down,t}^{mass}$ is the output value of the downstream NOx sensor at time t (ppm). Table 2 presents the input parameters for the two different models and their corresponding labels.

To ensure the accuracy and reliability of the machine learning model predictions,

while accounting for the influence of environmental temperature and vehicle conditions on NOx emissions, the PD and CD dataset for each vehicle are segmented into multiple monthly sub-datasets. Machine learning models are then established for each month based on the CD dataset. The LightGBM model is selected for its efficient training speed and outstanding performance, particularly in handling large-scale data and high-dimensional features[34]. Bayesian optimization is employed to fine-tune the model hyperparameters in a data-efficient and automated manner[35]. This process enables the model to achieve optimal performance with minimal manual intervention. The overall workflow is structured as follows: First, the original data is split into training and validation sets in a 9:1 ratio, in chronological order. The training set is used for adjusting the model parameters, and the validation set is used to evaluate the model's generalization ability. Five-fold cross-validation is performed on the training set, with the negative root mean square error (NRMSE) as the objective function. Bayesian optimization process starts by randomly sampling 30 initial hyperparameter configurations to establish a prior distribution of the objective function. A surrogate model, typically a Gaussian Process, is constructed to approximate the true objective function. Based on this surrogate, an acquisition function—such as expected improvement—is used to guide the selection of the next hyperparameter set to evaluate, effectively balancing the trade-off between exploration and exploitation. This iterative optimization continues for 120 steps, progressively refining the surrogate model and converging toward the global optimum. The best-performing hyperparameter combination identified during this process is

subsequently used to train the final model on the full training set. The corresponding MSE and R^2 are calculated on the test and validation set.

Table 2. Features and Labels of Machine Learning Models

Model Type	Features	Unit	Labels
EEM	v, v_{t-1}, v_{t+1}	Km/h	$NO_{x,up,t}^{mass}(g)$
	T_{out}, $T_{out,t-1}$, $T_{out,t+1}$,	%	
	T_f, $T_{f,t-1}$, $T_{f,t+1}$	%	
	RPM, RPM_{t-1}, RPM_{t+1}	RPM	
	F_{fuel}, $F_{fuel,t-1}$, $F_{fuel,t+1}$	L/h	
	MAF, MAF_{t-1}, MAF_{t+1}	Kg/h	
PEM	v, v_{t-1}, v_{t+1}	Km/h	$NO_{x,down,t}^{mass}(g)$
	T_{out}, $T_{out,t-1}$, $T_{out,t+1}$,	%	
	T_f, $T_{f,t-1}$, $T_{f,t+1}$	%	
	RPM, RPM_{t-1}, RPM_{t+1}	RPM	
	F_{fuel}, $F_{fuel,t-1}$, $F_{fuel,t+1}$	L/h	
	MAF, MAF_{t-1}, MAF_{t+1}	Kg/h	
	ECT, ECT_{t-1}, ECT_{t+1}	℃	
	T_{SCR-in}, $T_{SCR-in,t-1}$, $T_{SCR-in,t+1}$	℃	
	$T_{SCR-out}$, $T_{SCR-out,t-1}$, $T_{SCR-out,t+1}$	℃	

2.3. NOx Emission Prediction and Imputation.

2.3.1. Missing Data Classification.

In the PD dataset, features that may be used (v, T_{out}, T_f, RPM, F_{fuel}, MAF, ECT, T_{SCR-in}, and $T_{SCR-out}$) are processed according to the method described earlier to create time-series data features for the previous and subsequent seconds. Then, by comparing the timestamps of PD and CD dataset, rows of data corresponding to timestamps that exist in PD but are missing in CD are extracted as the dataset to be imputed. The missing data is divided into multiple parts according to the number of

243 months they span, and imputation is performed on a monthly basis. Based on the
244 missing data records' SCR inlet temperature, the data to be imputed is classified into
245 three types: Extreme-temperature missing data, Low-temperature missing data,
246 Normal-operation missing data.

247 *2.3.2 Imputation for* Extreme-temperature *Missing Data.*

248 For Extreme-temperature missing data, this study assumes that the SCR system is
249 non-operational, and the upstream NOx emissions correspond to the actual emissions.
250 In fact, when the SCR inlet temperature is below 100°C, the commonly used catalysts
251 in the current HDDV market (such as iron-based, copper-based, and vanadium-based
252 catalysts) have almost no catalytic activity[36, 37], this phenomenon is well
253 recognized in the field. Of course, the catalytic materials in the SCR system have a
254 porous structure, which can lead to some adsorption of NOx. However, the accuracy
255 of the predictions can be improved through simple correction. By extracting the
256 corresponding features from the missing dataset and constructing other time-series
257 dependent features (as described earlier, Table 2), and coupling this with the EEM, the
258 real-world emissions can be estimated by multiplying with the estimated material
259 adsorption capacity. Additionally, material adsorption capacity A_{ads} can be estimated
260 using extreme-temperature operation data from specific vehicles (SCR inlet
261 temperature below 100°C), as given by the following equation:

$$A_{ads} = \frac{\sum_1^T \mathrm{NO}_{\mathrm{x,down,t}}^{\mathrm{mass}}}{\sum_1^T \mathrm{NO}_{\mathrm{x,up,t}}^{\mathrm{mass}}}$$

262 where T represents the total duration of the data segment.

2.3.3 Imputation for Low-temperature Missing Data.

For Low-temperature missing data, when the SCR inlet temperature falls within the 100-200°C range, the engine coolant temperature is typically low (below 70°C), which causes the NOx sensor to remain in the off state within this temperature range. This is particularly true when the inlet temperature is below 150°C, resulting in a minimal, or even nonexistent, amount of valid data. The absence of low-temperature data hampers the accuracy of the monthly-established PEM in estimating the missing values. To resolve this issue, this study examined data from various vehicles and found that, in some cases, NOx sensors do not shut off properly when the engine coolant temperature is lower than 70 ℃. These vehicles can be used to establish benchmark models for estimating the emissions of other vehicles. Based on these models, we have completed the imputation of Low-temperature missing data. Detailed process is as follows:

(1) Establishing the benchmark model. First, we used SQL statements to filter suitable vehicles on the data platform. Subsequently, all data points with inlet temperatures between 100°C and 200°C were filtered. Using the method described in Section 2.2, PEMs were constructed. Notably, the time-series features were extracted prior to data filtering to maximize model performance and minimize errors associated with these features. (2) Predicting NOx emissions for each vehicle using the benchmark model. For any given vehicle, we first calculated its fuel-based EF when the engine coolant temperature was above 70°C. Then, we identified the model in the baseline set with the closest EF and coupled it with the missing data in the SCR inlet

temperature range of 100–200°C. This coupling yielded instantaneous predictions of NOx emissions.

2.3.4 Imputation for Normal-operation Missing Data.

For Normal-operation missing data, the performance of the PEM is sometimes suboptimal due to the low output values of the NOx sensor, which are easily affected by noise, making it difficult to effectively capture the patterns of NOx emissions. In contrast, the EEM is less affected by sensor noise and exhibits relatively stable SCR conversion efficiency in the temperature range above 200°C, with minimal variation. Therefore, we first evaluate the accuracy of the monthly trained PEMs. If the model achieves $R^2 > 0.6$, it is directly used to fill the missing data. Otherwise, the estimation from the EEM, combined with the estimated SCR conversion efficiency η_{SCR}, is employed to complete the missing values. The estimation process for η_{SCR} is as follows:

(1) Arrange T_{SCR-in} in ascending order, denoted as $T_{SCR-sorted}$, and record the minimum temperature as $t_{SCR-in,\ min}$. (2) Starting from $t_{SCR-in,\ min}$, create bins every 5°C, and sequentially assign the data that meet the criteria to the corresponding bins. (3) For bin i, estimate the SCR conversion efficiency $\eta_{SCR,\,i}$ using the following formula:

$$\eta_{SCR,\,i} = \frac{1}{T}\left(\frac{\sum_{t=1}^{T} NO_{x,\text{up-stream},t,i} - \sum_{t=1}^{T} NO_{x,\text{down-stream},t,i}}{\sum_{t=1}^{T} NO_{x,\text{up-stream},t,i}}\right) \times 100\%$$

2.4. Overview and Validation of the Methodology

To enhance clarity, the overall process of the methodology proposed in this study is illustrated in the Figure 1. The methodology consists of three main steps: data

306 acquisition and cleaning, model construction, and data imputation. The process of
307 data processing is relatively fixed and can be directly applied to the OBD data of
308 other China VI vehicles. The modeling step is flexible and can be adapted to other
309 models based on modeling costs. For instance, to ensure the accuracy of imputation,
310 this study uses all CD data for the corresponding month as training data for the
311 machine learning model. In practice, however, only a subset of the data (e.g., selecting
312 data from a single day each month) can be used to train the model and still achieve a
313 high level of accuracy. The data imputation step is robust and data-driven, relying
314 solely on historical information, which makes it cost-effective.

315 Overall, the methodology proposed in this study is versatile and suitable for the
316 imputation of erroneous and missing values in OBD data. Moreover, it is highly
317 compatible with the underlying computational frameworks used by current OBD big
318 data management platforms, making it well-suited for large-scale deployment. The
319 imputed data can largely reflect the overall situation in the real world.

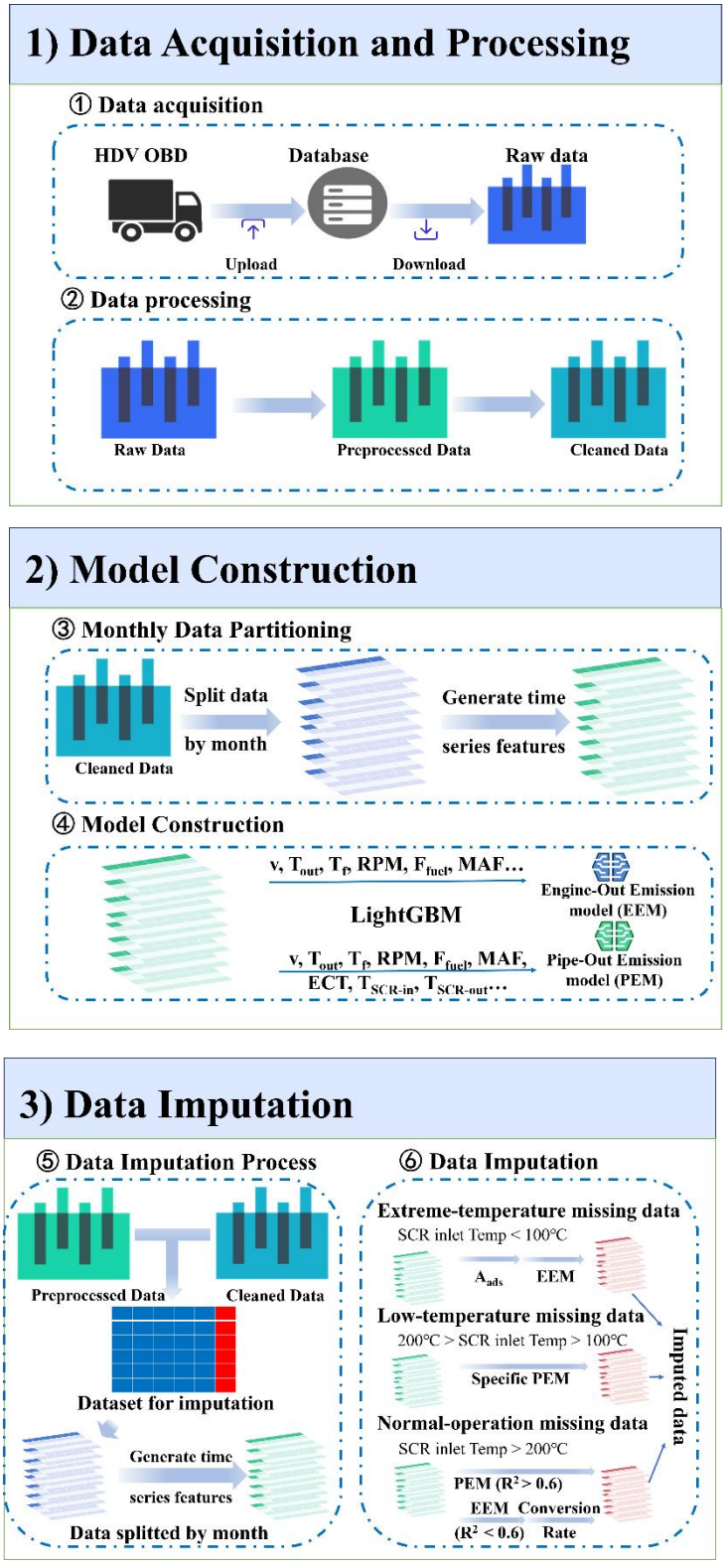


Figure 1. Overall Process of OBD Data Imputation.

3. Result and Discussion.

3.1. Data Processing and Analysis.

We processed the data of all vehicles using the above-mentioned workflow. Figure S2 shows the proportions of analyzable data, constant-value errors, timestamp errors, and exceeding boundaries errors in the data of all vehicles. The figure indicates that over 60% of the data from each vehicle is analyzable. Out-of-bounds errors are the primary cause of data discrepancies, followed by timestamp errors, with constant-value errors being the least frequent. Figure 2 includes box plots that illustrate the proportions of out-of-bounds errors across different vehicles and variables. From Figure 2, it is evident that the majority of out-of-bounds errors are caused by the vehicle's NOx sensor being turned off. In contrast, other parameters generally exhibit high reliability. Constant-value errors are rare and tend to occur mainly within the NOx sensors. Timestamp errors have a relatively minor impact on data integrity, as they affect all variables uniformly, preventing any bias or inconsistencies. Overall, the remaining data provides sufficient real-world vehicle operational information, which is adequate for reconstructing missing data.

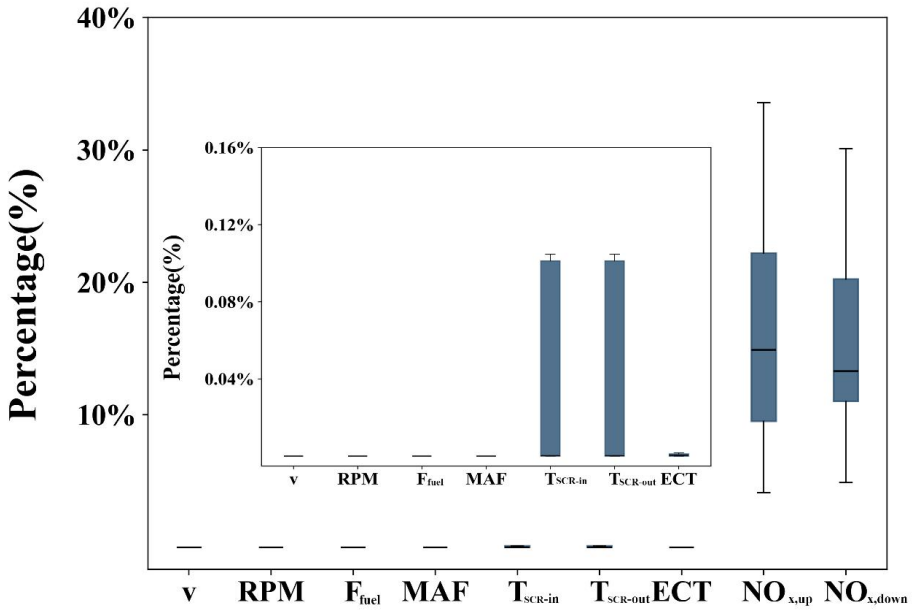


Figure 2. Proportion of Out-of-Bounds Errors for Different Vehicles and Columns.

Figure 3 shows the fuel-based EFs for vehicles based on CD data. Among the 32 vehicles, 26 had EFs below both limits, while 6 vehicles exceeded both thresholds. Specifically, vehicle #27 slightly exceeded the PEMS in-use limit, while vehicles #28 to #32 far exceeded both limits, with EFs 9.2 to 20.3 times the Diesel Engine Limit and 6.1 to 13.6 times the PEMS in-use limit. Among the 32 vehicles, 26 had EFs below both limits, while 6 vehicles exceeded both thresholds. Specifically, vehicle #27 slightly exceeded the PEMS in-use limit, while vehicles #28 to #32 far exceeded both limits, with EFs 9.2 to 20.3 times the Diesel Engine Limit and 6.1 to 13.6 times the PEMS in-use limit. Figure 4 further illustrates the monthly EFs for the six vehicles with one year of data. As shown in Figure 4a, the EFs for vehicles #4, #9, and #16–18 remained relatively stable. And in Figure 4b, vehicle #31 experienced a sharp increase in emissions around July 2022, jumping from 4 g/kg-fuel to 54 g/kg-fuel. To investigate the cause of the sudden spike in emissions for vehicle #31, we identified the anomaly period (September 1–3, 2022; Figure 4c) and subsequently used ArcGIS

Pro to locate areas where the vehicle remained stationary for over an hour during this timeframe (Figure S3). Notably, several automotive repair facilities are located in the vicinity of this area. It is possible that vehicle #31 experienced some issues during the maintenance process. To investigate further, we examined the EFs corresponding to the SCR inlet, which represents the untreated gases, before and after the anomaly, as shown in Figure 4d. From the figure, it is evident that the upstream EF experienced only minimal changes before and after the spike. This suggests that the issue was not related to the engine itself, but rather to a malfunction in the SCR system. This study suggests that relevant authorities could adopt a similar approach to identify high-emission vehicles, locate the malfunctioning components, and preliminarily diagnose the causes of the high emissions. This would enable precise fault detection and management of high-emission vehicles, ultimately addressing the root causes of NOx emissions from HDDVs.

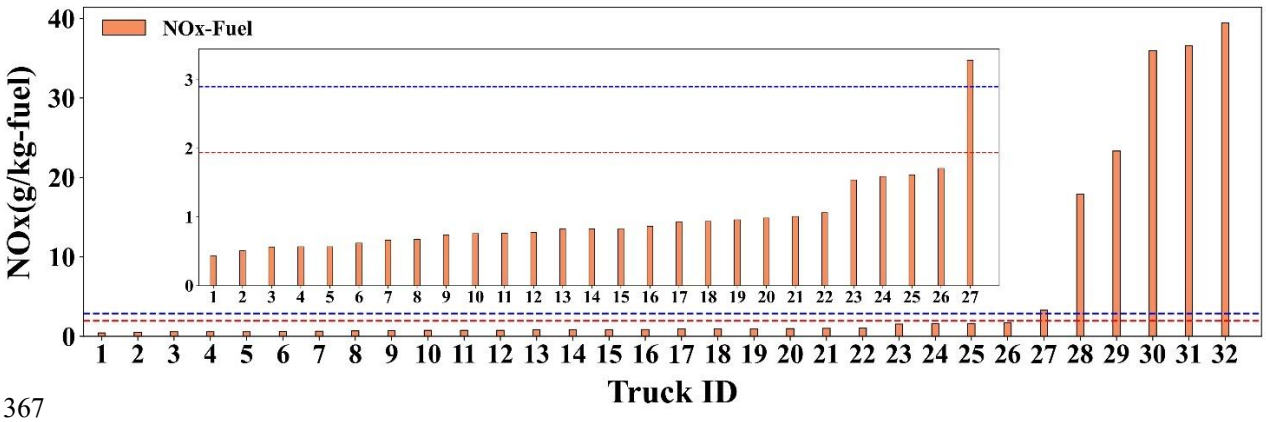


Figure 3. Fuel-based EFs for Different Vehicles Based on Cleaned Data. The red line represents the Diesel Engine Limits (1.94 g/kg-fuel), and the blue line represents the PEMS in-use limits (2.9 g/kg-fuel). These limits are derived and calculated from power-based limits[23, 38].

3.2. Model Construction.

This study developed Engine-out and Pipe-out LightGBM machine learning models for each of the 26 vehicles with 1–2 months of data. For the remaining 6 vehicles with one year of data, monthly machine learning models were constructed. Figure 5a and 5b present the R^2 values of the EEMs and PEMs for the 26 vehicles on the validation and test sets, while Figure 5c and 5d illustrate the monthly R^2 values of the EEMs and PEMs for vehicle #31. The EEMs exhibit very high accuracy, while the PEMs generally show lower. Furthermore, there is a certain correlation between the fuel-based EFs and the R^2 of the PEMs; as the former increases, the latter also increases. However, this effect is not observed in the EEMs. The test-validation R^2 values for #31 further illustrate this point. Before September 2022, the R^2 values of the PEMs on the test set ranged from 0 to 0.5, corresponding to the lower emission levels shown in previous figures. After September 2022, the R^2 values on the PEM's test set increased to above 0.9, corresponding to the higher emission levels presented earlier.

This phenomenon may stem from multiple factors. Under one possible explanation, NOx sensor drift may introduce noise into the training data. When emissions are relatively high, this noise has little impact, allowing the machine learning model to effectively capture the main relationships within the data. However, when emissions are lower, the noise can severely degrade model performance. Another possibility is that the dataset's representation of the SCR system (via ECT, $T_{SCR\text{-}in}$, and $T_{SCR\text{-}out}$) may not fully capture its actual state. SCR systems often contain multiple catalysts, which function more actively in low-emission vehicles, making the

system more complex. In high-emission vehicles, catalyst performance is weaker, allowing the inlet and outlet temperatures of the SCR system to serve as a more effective proxy for its state, thereby yielding higher R^2 values.

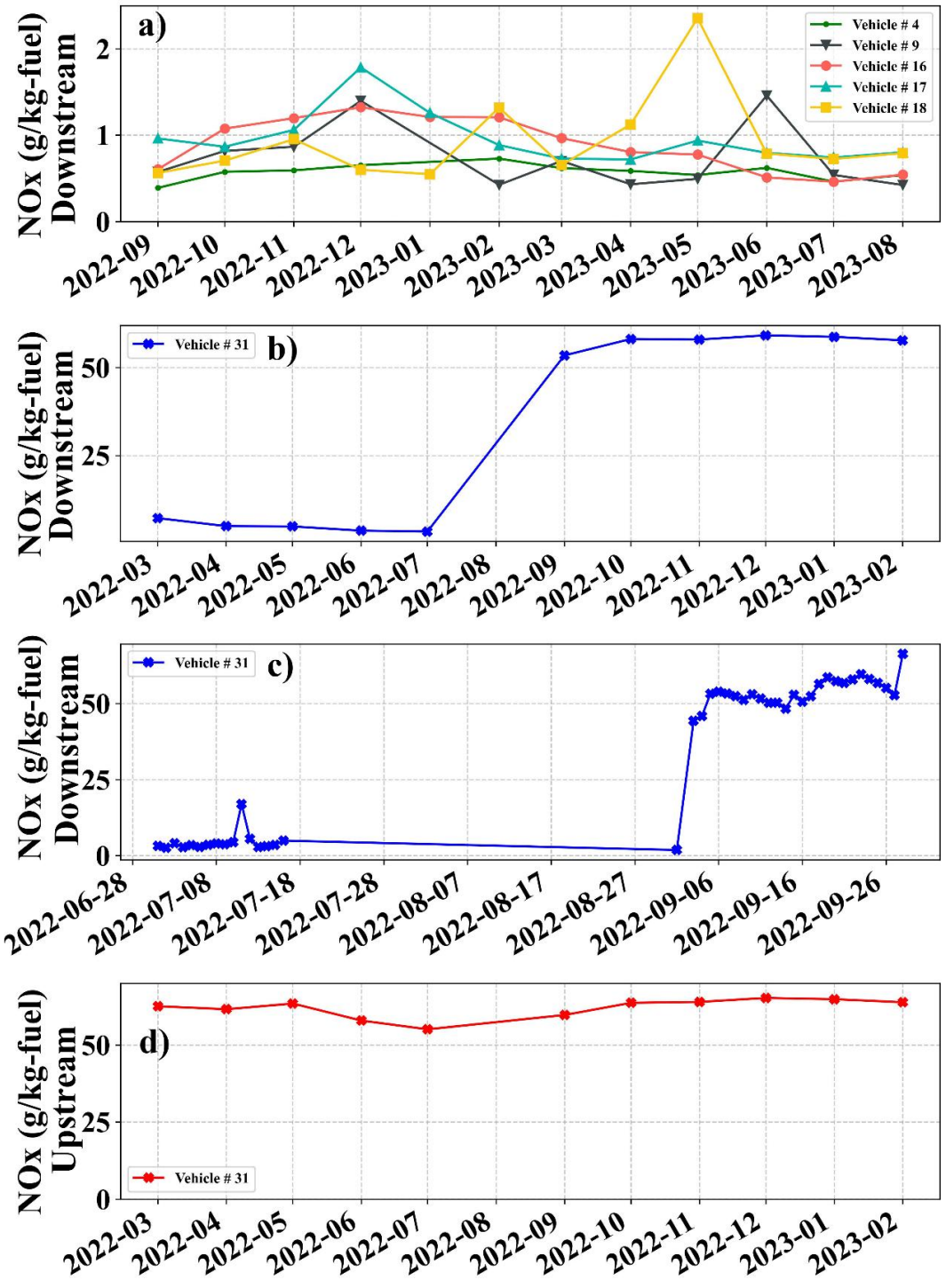
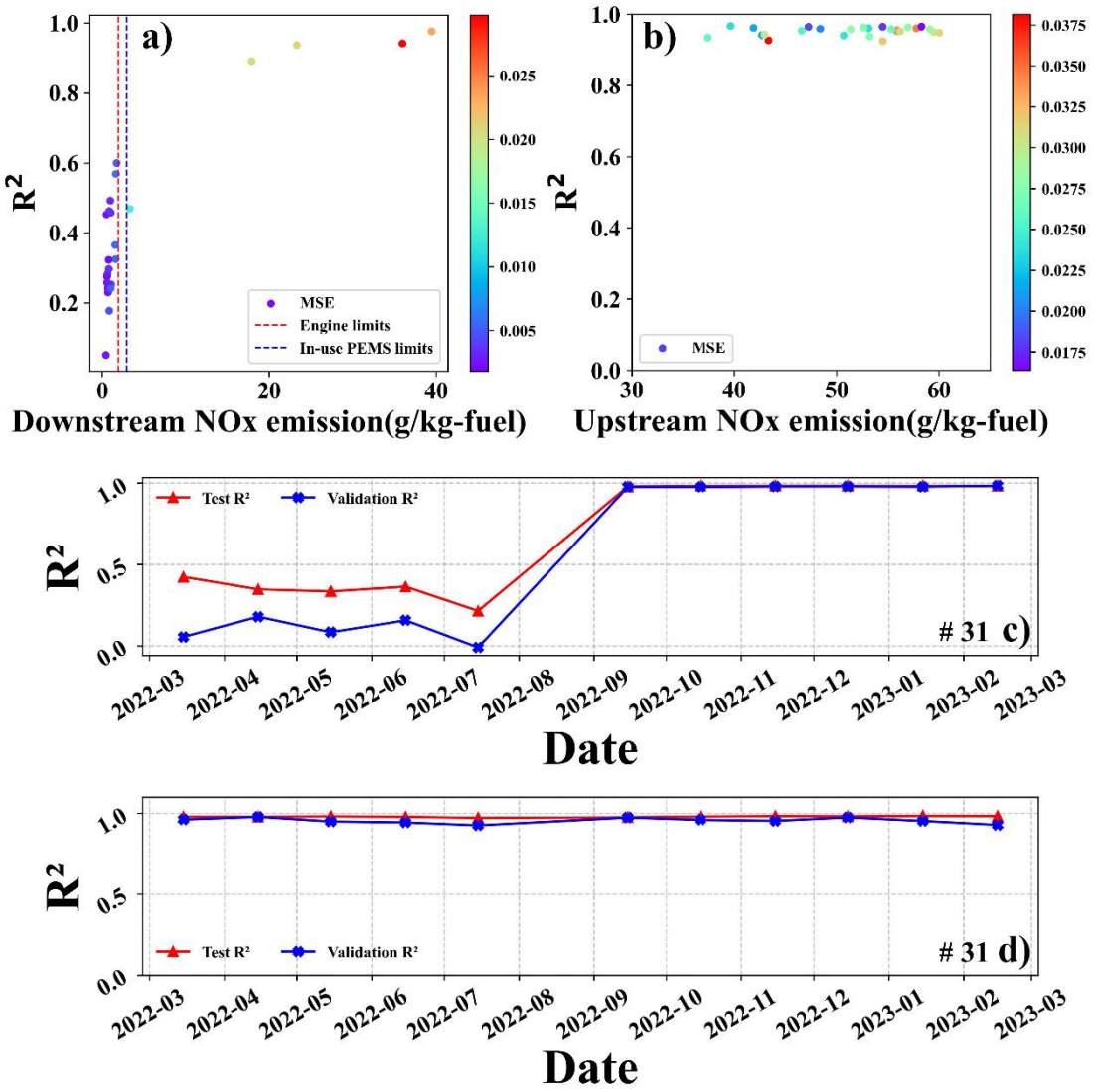


Figure 4. Monthly/Daily Fuel-based EFs for Vehicles with One Year of Data. (a) Monthly EFs for vehicles #4, #9, #16, #17, and #18. (b) Monthly EFs for vehicle #31. (c) Daily EFs for vehicle #31

400 from July to September 2022. (d) EFs for vehicle #31 based on upstream NOx emissions.

401 To maximize the accuracy of data imputation, this study constructed models
402 using all available data for each month and estimated NOx emissions for missing
403 entries on a monthly basis. However, alternative strategies can be applied in practical
404 scenarios. To explore this, an EEM was built for each month using a randomly
405 selected single day of data from #17. Figure S4 illustrates the predictive performance
406 of these models on other days, showing that their performance within the same month
407 remained largely stable. Some models even maintained high R^2 values across the
408 entire year. This suggests that, in practical applications, it is feasible to build models
409 using only a subset of the data, significantly reducing computational costs. This
410 finding supports large-scale data imputation efforts for OBM platform data.



411

412 **Figure 5.** Performance of EEMs and PEMs. The red line represents engine limits, while the blue

413 line represents in-use PEMS limits. (a) R^2 of PEMs on the test set for 26 vehicles with one to two

414 months of data. (b) R^2 of EEMs on the test set for the same 26 vehicles. (c) R^2 of PEMs on the test

415 set for #31. (d) R^2 of EEMs on the test set for #31.

416 **3.3 Data Imputation.**

417 To investigate the patterns of data loss. Figure 6 illustrates the frequency of

418 missing data across all vehicles and the proportion of cumulative missing duration

419 relative to the total duration. It can be observed that short-duration data gaps (under

420 500 seconds) account for over 80% of occurrences and approximately 85% of the

total missing duration, making them the dominant type of data loss. In contrast, data gaps exceeding 500 seconds exhibit lower frequency and proportion, with almost no occurrences exceeding 3,000 seconds. Shorter data gaps are typically associated with poor network or GPS signals, whereas longer gaps often correspond to cold starts and prolonged cold operations, during which NOx emissions are significantly elevated. The systematic absence of these data often results in an underestimation of emission inventories or EFs derived from processed OBD data. Thus, data imputation is essential to accurately capture the full scope of real-world HDDV emissions.

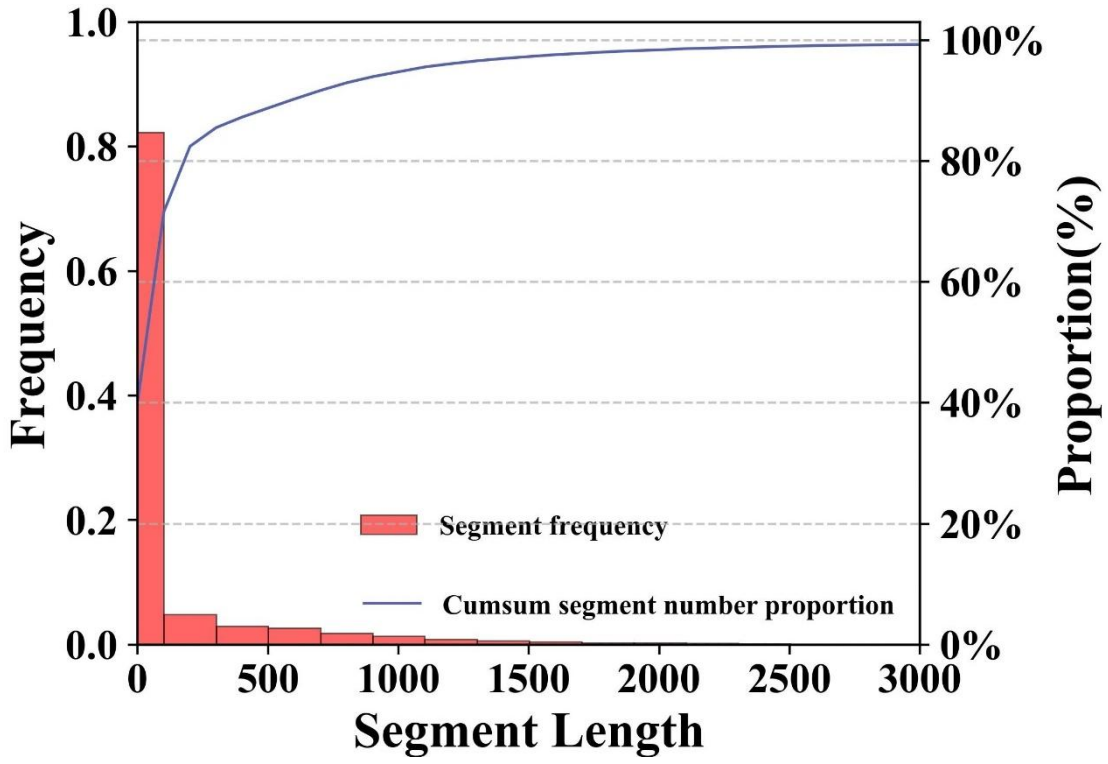


Figure 6. Frequency and duration proportion of missing data rows relative to the total missing data.

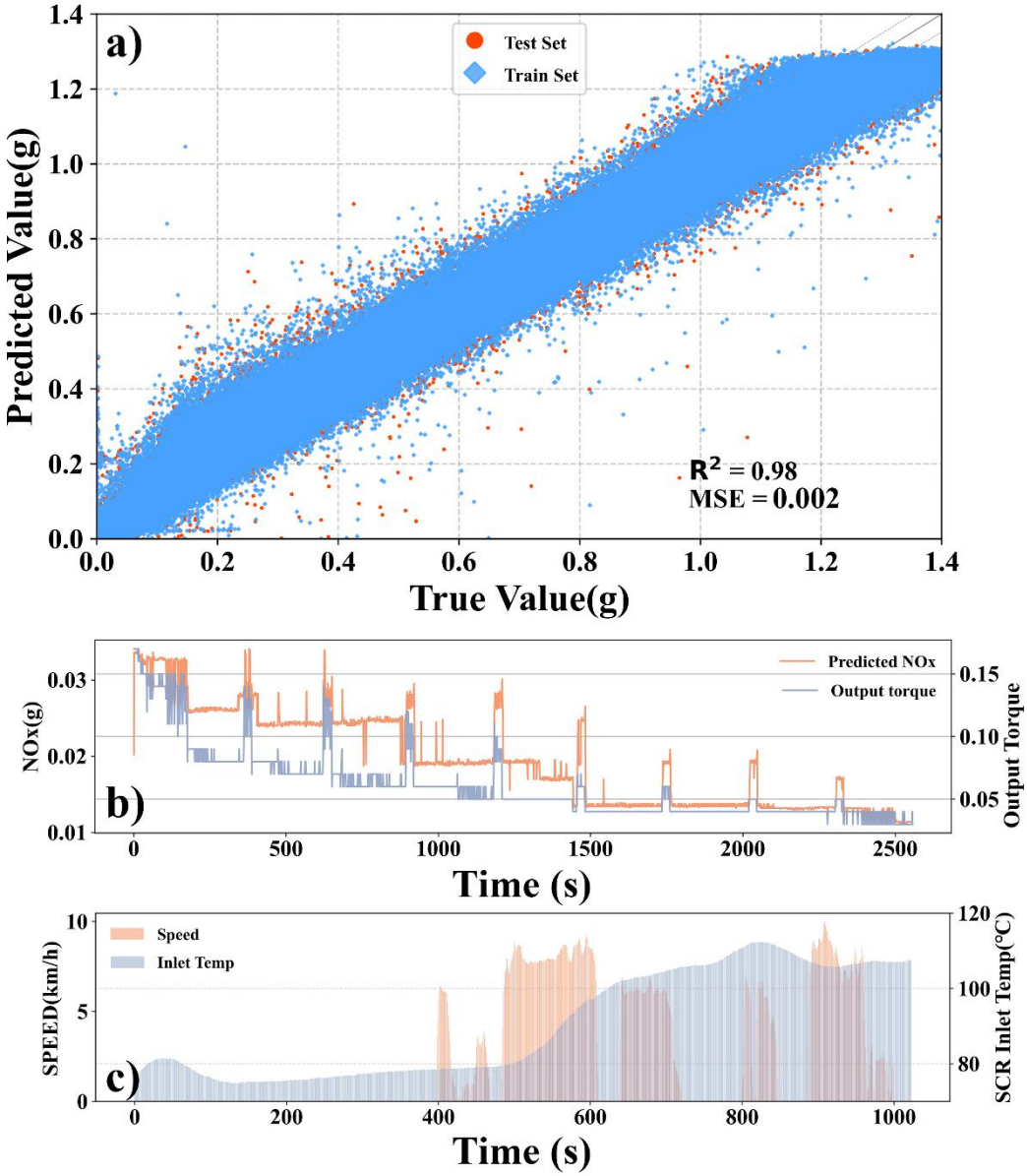
Taking #17 as an example, this study illustrates the complete process of data imputation. First, we proceed with Extreme-temperature data imputation. However, before starting, we will investigate the adsorption performance of the SCR system used in this study. Since vehicle #17 lacked sufficient low-temperature records, we

estimated A_{ads} using data from other vehicles. Through data analysis, we found that vehicle #31 recorded data where the engine coolant temperature was below 70°C. Based on this data, we investigated the adsorption performance of the catalyst. We first isolated the data where no emission factor anomaly occurred, specifically the data before and

including July 2022, where the SCR inlet temperature was below 100°C. We then calculated the upstream NOx emissions $NO_{x,up,t}^{mass}$, downstream NOx emissions $NO_{x,down,t}^{mass}$, and their adsorption coefficient A_{ads}. The calculated value of A_{ads} is 0.95. Using these two values, we computed their moving averages (window sizes: 1, 5, 10, 20), and finally calculated the corresponding R^2 for each moving window. The results are shown in Figure S5. From the figure, it can be observed that as the window length increases, the R^2 between $NO_{x,up,t}^{mass}$ and $NO_{x,down,t}^{mass}$ gradually increases, rising from 0.88 for the raw data to 0.99 at the end. Additionally, the relationship between $NO_{x,up,t}^{mass}$ and $NO_{x,down,t}^{mass}$ becomes increasingly close to $NO_{x,down,t}^{mass} = A_{ads} \times 0.95$. Considering the generally close-to-0.95 R^2 values for EEMs, this estimation scheme is feasible.

Figure 7a presents a scatter plot comparing the true and predicted values for the training and test sets of #17's EEM in October,2022, while Figure S6 provides results for other months. The model achieves an R^2 of 0.98 and an MSE of 0.002 on the training set, demonstrating exceptional generalization capability. The near-perfect overlap between the training and test set scatter points further confirms the absence of overfitting. Figure 7b) illustrates a missing segment of approximately 2500 seconds in October, including transient net output torque and the results of data imputation. In

this segment, the vehicle speed remains constant at 0, and the SCR inlet temperature is below 100°C. From the figure, it can be observed that the transient NOx emissions trend generally follows the trend of the net output torque, although small fluctuations occur due to the influence of other parameters. Overall, the predictions from the EEM are robust and reliable. It is worth mentioning the corresponding driving behavior for this segment. The vehicle starts in a completely cooled state but appears to have only turned on the engine without moving. Alternatively, in another similar scenario, the vehicle continues to operate at a very low speed in a low-temperature state, as shown in Figure 7c). This type of driving behavior results in the vehicle running persistently under low-temperature conditions, preventing the SCR system from reaching its operational temperature, thus leading to higher emissions. Additionally, the engine coolant temperature in this data segment is often below 70°C, and the performance of the NOx sensor is limited, making it difficult for the OBD system to accurately reflect the impact of poor driving behavior on NOx emissions. Using the EEMs established for each vehicle on a monthly basis, this study completed the data imputation for all instances where the SCR inlet temperature is below 100°C. After imputation and calculation, the total estimated emissions for Extreme-temperature missing data amount to 6188.4 g.



475

476 **Figure 7.** Performance of the EEM and relevant parameters and imputation results for certain
477 segments. a) Performance of the October model for #17. b) Predicted NOx and corresponding net
478 output torque for a specific segment. c) Vehicle speed and SCR inlet temperature corresponding to
479 the other specific segment.

480 For the Low-temperature data imputation, we first constructed baseline models
481 using the filtered data. By examining the emissions of vehicles with the same model
482 in the data platform, we identified five eligible vehicles with fuel-based EFs above

70°C of 1.33 g/kg-fuel, 1.48 g/kg-fuel, 1.66 g/kg-fuel, 1.66 g/kg-fuel, and 39.32g/kg-fuel, respectively. Machine learning models were developed for each vehicle, forming a baseline prediction model matrix. Their performance on the training and test sets is shown in Figure S7. The model's R^2 ranges from 0.61 to 0.82, indicating that the machine learning model effectively captures the emission patterns within this temperature range. Among these models, the one with an EF of 1.33 g/kg-fuel was closest to that of #17, and was therefore used to predict the missing data. After imputation and calculation, the total estimated emissions for Low-temperature missing data amount to 8866.4 g.

Finally, we proceed with Normal-operation data imputation. The R^2 values of the PEMs for vehicle #17 were all below 0.6; therefore, the EEM was employed in conjunction with the conversion efficiency approach. Figure 8 shows the monthly estimated SCR conversion efficiency for vehicle #17. It can be observed that the SCR conversion efficiency fluctuates around the overall conversion rate for each month. In the temperature range of 200–350°C at the SCR inlet, the fluctuations are relatively small, and the rate of change in conversion efficiency with temperature is also low. However, when the inlet temperature exceeds 350°C, the fluctuations become more significant. This is due to the different frequency distributions of the inlet temperature, as shown in Figure S8b. The data above 350°C is sparse, and the limited data volume is insufficient to offset the shift in NOx sensor readings, leading to greater fluctuations. Fortunately, the data requiring imputation, where the SCR inlet temperature exceeds 350°C, is very limited and can almost be ignored as seen in

Figure S8c. Therefore, the estimated SCR conversion efficiency can generally meet the high-precision estimation requirements. After imputation and calculation, the total estimated emissions for Normal-operation missing data amount to 686.4 g.

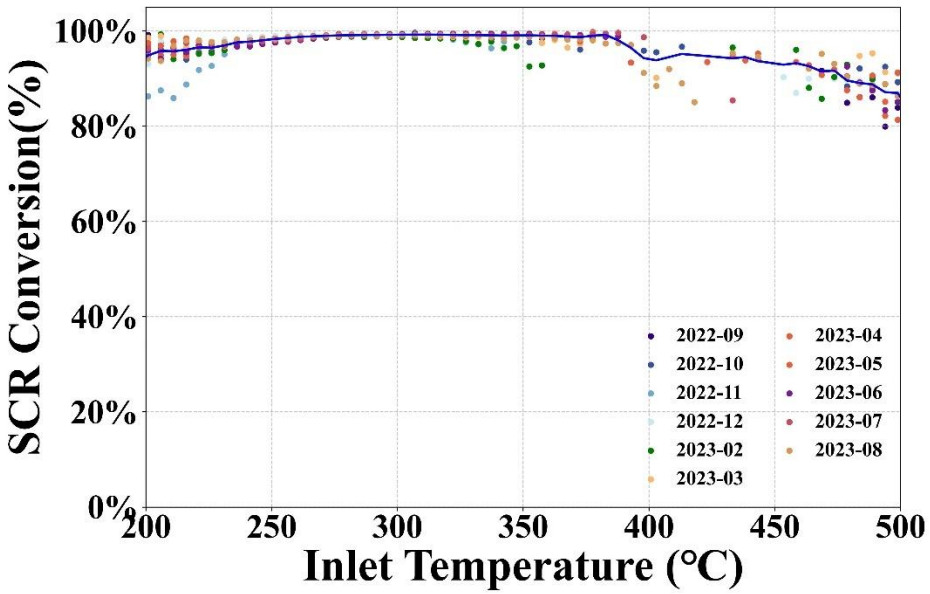


Figure 8. SCR Conversion Efficiency Estimation for Vehicle #17. The blue line represents the conversion efficiency estimated using all the data.

Through the entire process, the data imputation for #17 was successfully completed. To better illustrate the imputation results, Figure 9 presents NOx emissions histograms binned by SCR inlet temperature after different types of data imputation. Figure 9a shows the emissions from CD, which is regularly used by other studies. As observed, NOx emissions below an SCR inlet temperature of 200°C are nearly absent, indicating that the OBD system failed to record these data. After successive imputations in Figures 9b, 9c, and 9d, these missing emissions were recovered. The results reveal that when the NOx sensor in the OBD system fails, a substantial amount of NOx emissions actually occurs. For #17, these missing emissions are 2.5 times higher than the recorded emissions, which is 15,741.3g versus

6,157.3g. While this does not necessarily imply that all HDDVs —or even other vehicles of this model—exhibit similarly high unrecorded emissions, it highlights a critical issue: the current OBD system has certain limitations. In specific scenarios, such as port logistics zones, or industrial parks, HDDVs may frequently operate under start-stop conditions, keeping engine coolant temperatures persistently below 70°C. This results in prolonged SCR inactivity and substantial NOx emissions, leading to severe localized pollution. If high-emission vehicle identification or emission inventory development relies solely on available OBD data, policymakers and enforcement agencies may misjudge the actual emissions situation, potentially compromising the effectiveness of mitigation measures. To address this issue, we recommend adopting the data imputation approach proposed in this study for large-scale OBD data enhancement, which could significantly improve the accuracy of emission assessments and regulatory decisions.

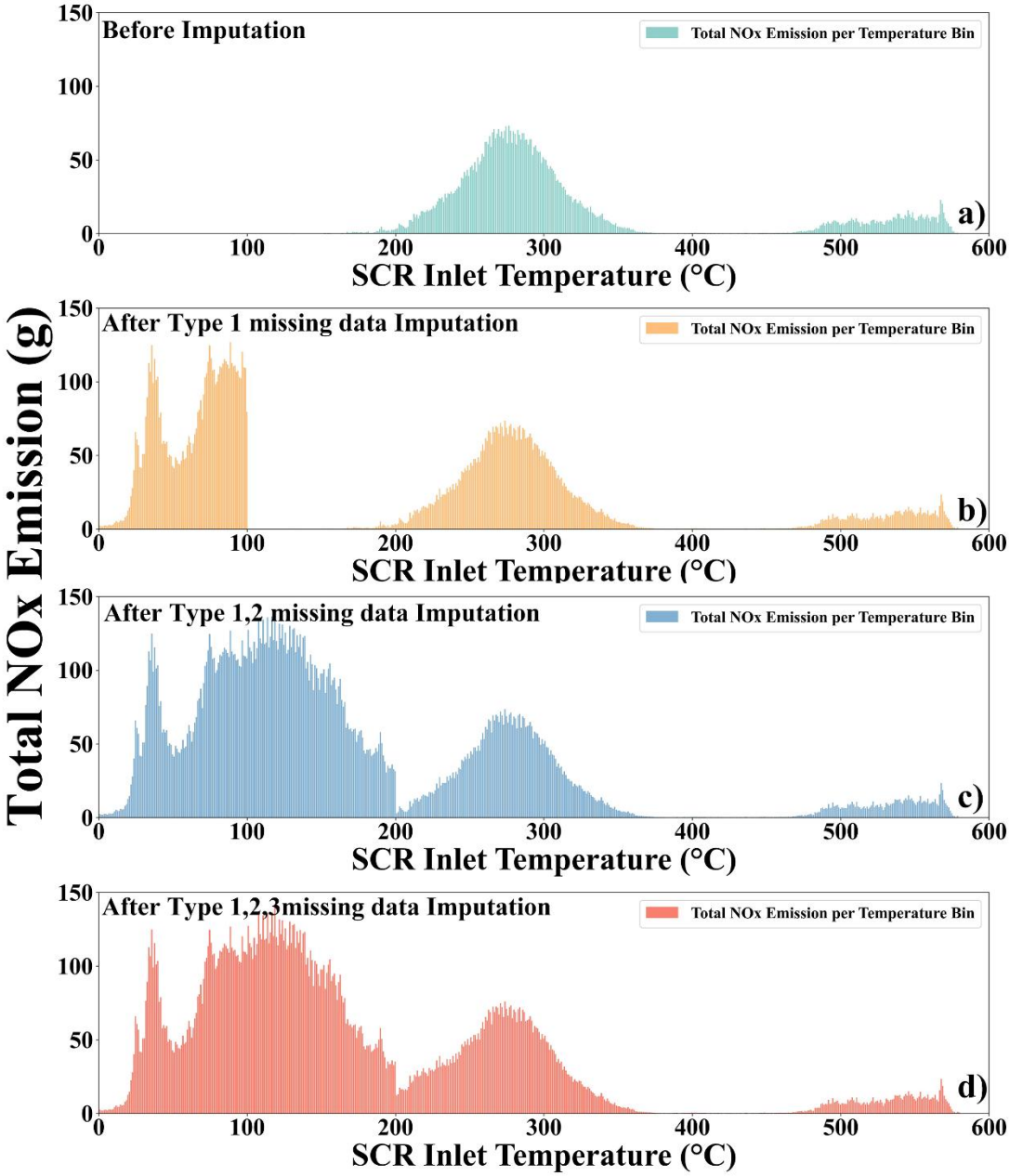
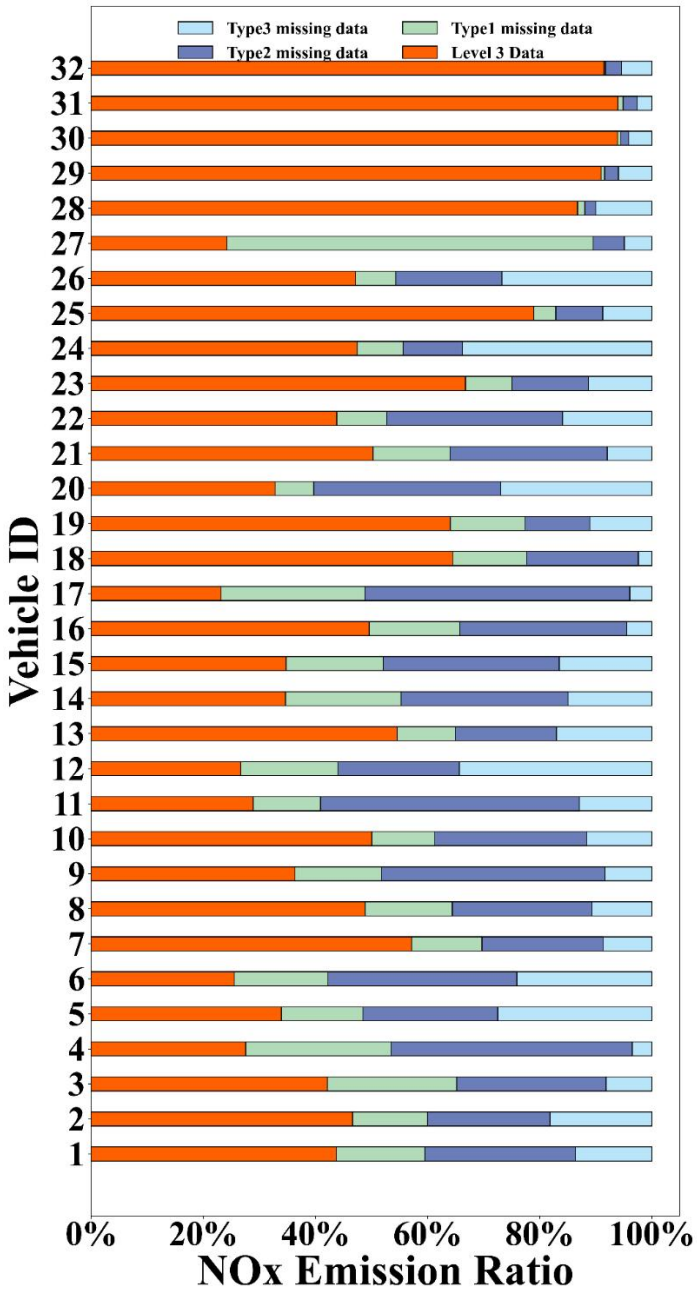


Figure 9. NOx emissions from Cleaned Data and data progressively imputed through sequential processing. (a) Emissions from Cleaned Data, (b) emissions after Extreme-temperature imputation, (c) emissions after Extreme-temperature and Low-temperature data imputation, and (d) emissions after Extreme-temperature, Low-temperature, and Normal-operation data imputation.

To verify the generalizability of our method, we applied the data imputation process to each vehicle. Figure 10 illustrates the proportion of emissions corresponding to CD, Extreme-temperature data, Low-temperature data, and Normal-

operation data relative to the total emissions after imputation. Figure S9 illustrate the vehicles' fuel-based EF before and after imputation. It appears that for most vehicles, the emissions associated with missing data exceed those recorded in CD. However, for vehicles #28–32, which exhibit extremely high emissions, CD account for the majority of emissions in the imputed dataset. This may be attributed to their exceptionally low SCR conversion efficiency, leading to near-engine-out NOx emissions even under normal operation. As a result, their low-temperature NOx emissions do not significantly exceed those during regular operation, as observed in normal vehicles. Moreover, since missing data typically occur during low-power operating, the emissions from missing segments are relatively lower than those in CD. For lower-emission vehicles, the situation is entirely different. Under normal driving conditions, these vehicles maintain a consistently high SCR conversion efficiency. However, when data loss occurs, a substantial portion corresponds to low-temperature operating conditions, which result in emissions far exceeding those during regular operation. Comparing with Figure S2, we find that for normal-emission vehicles, the higher the proportion of missing data, the greater the emissions associated with these missing segments. However, this effect is not absolute, as factors such as SCR temperature and vehicle operating conditions vary across missing data, introducing multifaceted influences. We recommend that regulatory authorities conduct basic analyses of missing data, such as the proportion of missing data duration, the duration of low-temperature operation, and engine net output torque under low-temperature conditions. Utilizing these indicators for a multifaceted and dynamic approach to

564 high-emission vehicle management could yield significant environmental benefits.



565

566 **Figure 10.** Proportion of total NOx emissions contributed by different types of missing data for all

567 vehicles.

568 **4. Conclusion.**

569 The main contributions of this study are threefold. First, it investigates the

570 applicability of machine learning models to OBD data. Second, it establishes a data

571 imputation method that effectively utilizes well-recorded data to reconstruct real-
572 world NOx emission. Third, it applies this method to impute missing data for 32
573 selected vehicles, demonstrating its generalizability. The key findings are as follows:

574 (1) For OBD data, EEMs generally exhibit high R^2 values, indicating considerable
575 accuracy, with R^2 exceeding 0.9. In contrast, PEMs show lower and more variable R^2
576 values, ranging from 0.05 to 0.98.

577 (2) Fuel-based EF is largely uncorrelated with the R^2 of EEMs but shows some
578 correlation with the R^2 of PEMs. Specifically, higher fuel-based EF values tend to
579 correspond to higher R^2 values in PEMs.

580 (3) The missing data imputation for #17 was completed for Extreme-temperature,
581 Low-temperature, and Normal-operation missing data. The emissions associated with
582 missing data amount to 15,741.3 g, while the recorded CD account for 6,157.3 g. The
583 emissions from missing data are 2.5 times higher than those from recorded data for
584 #17.

585 (4) The imputation results for 32 vehicles indicate that the proportion of emissions
586 attributed to missing data is relatively low for extreme emitters. In contrast, normal
587 emitters exhibit significantly higher emissions associated with missing data, and this
588 proportion is related to the extent of missing data.

589 Currently, OBD serves as a crucial monitoring tool for measuring and
590 characterizing emissions from HDDVs. However, the widely deployed OBD systems
591 in China have a critical limitation: they fail to capture low-temperature emissions,
592 which are non-negligible. This study takes a step forward in exploring the

applicability of machine learning to OBD data and, on this basis, establishes a method for imputing missing data. The proposed method is generalizable and can be applied for large-scale data imputation. Admittedly, the method proposed in this study could be costly—developing EEM and PEM on a monthly basis for each vehicle would require considerable resources. However, by simplifying the data requirements and model complexity (e.g., through data sampling and the use of simpler machine learning algorithms), or by developing generalized PEM and EEM models, the computational burden can be significantly reduced, making large-scale application feasible. In addition, most current big data platforms are capable of supporting machine learning tasks, which would further facilitate the integration and practical implementation of the proposed approach. Nevertheless, this study has certain limitations. First, the data imputation was applied exploratorily to a single vehicle model with 32 trucks, raising concerns about its generalizability. For instance, other models may employ different types of SCR systems, rendering the current temperature-based categorization insufficient. Second, the proposed method requires further validation. To address these issues, future research will focus on expanding the method to a broader range of vehicles and conducting on-road PEMS-OBD combined experiments to verify its applicability.

ASSOCIATED CONTENT

Supplementary Material. Detailed information on variables and data recorded by OBD devices (Table S1), detailed vehicle information (Table S2), examples of time series errors, exceeding boundaries errors, and constant value errors (Figure S1),

proportion of analyzable data, timestamp errors, constant value errors, and exceeding boundaries errors for all vehicles (Figure S2), location where vehicle #31 was parked for more than 1 hour (Figure S3), predictive performance of EEMs using randomly selected days from vehicle #17 (Figure S4), scatter plot of sliding averages between upstream and downstream SCR NOx emissions for various window lengths: a) window=1, b) window=5, c) window=10, d) window=20 (Figure S5), performance of EEMs for vehicle #4 in other months (Figure S6), performance of the benchmark model. All figures (a)-(e) display scatter plots of true and predicted values on the test set, with R^2 representing the model's R^2 on the test set. EF refers to the fuel-based emission factor, which represents the vehicle's emission factor when the engine coolant temperature is above 70°C (Figure S7), distribution frequency of SCR inlet temperature for all vehicles before and after processing, and for filtered data: a) distribution frequency of the raw data, b) distribution frequency of the processed data, c) distribution frequency of the filtered data (Figure S8), Vehicles' fuel-based EF before and after imputation (Figure S9).

Notes

The authors declare no competing financial interest

ACKNOWLEDGMENT

This work was supported by the National Key Research and Development Program of China (2024YFC3712302, 2024YFE0112100, 2023YFC3705405), the National Natural Science Foundation of China (42177084), the Fundamental Research Funds for the Central Universities of China (63241318, 63241322,63243126) and Natural

Science Foundation of Tianjin City (22JCYBJC01330).

ABBREVIATIONS

HDVs, heavy-duty vehicles; HDDVs, heavy-duty diesel vehicles; OBD, On-Board Diagnostic; MCUs, microcontroller units; EFs, emission factors; WHSC, World Harmonized Stationary Cycle; ESC, Stationary Cycle; ETC, European Transient Cycle; PM, particulate matter; WNTE, World Harmonized Not-to-Exceed; PEMS, Portable Emission Measurement System; MCU, microcontroller unit; NRMSE, negative root mean square error; LightGBM, Light Gradient Boosting Machine; PD, Preprocessed Data; CD, Cleaned Data; EEM, Engine-Out Emission model; PEM, Pipe-Out Emission model.

REFERENCES

[1] W. Ke, S. Zhang, Y. Wu, B. Zhao, S. Wang, J. Hao, Assessing the future vehicle fleet electrification: the impacts on regional and urban air quality, Environmental science & technology, 51 (2017) 1007-1016.

[2] C. McCaffery, H. Zhu, T. Tang, C. Li, G. Karavalakis, S. Cao, A. Oshinuga, A. Burnette, K.C. Johnson, T.D. Durbin, Real-world NOx emissions from heavy-duty diesel, natural gas, and diesel hybrid electric vehicles of different vocations on California roadways, Science of the Total Environment, 784 (2021) 147224.

[3] E. Mulholland, J. Miller, Y. Bernard, K. Lee, F. Rodríguez, The role of NOx emission reductions in Euro 7/VII vehicle emission standards to reduce adverse health impacts in the EU27 through 2050, Transportation Engineering, 9 (2022) 100133.

[4] S.J. Davis, N.S. Lewis, M. Shaner, S. Aggarwal, D. Arent, I.L. Azevedo, S.M. Benson, T. Bradley, J. Brouwer, Y.-M. Chiang, Net-zero emissions energy systems, Science, 360 (2018) eaas9793.

[5] C. Wen, J. Lang, Y. Zhou, X. Fan, Z. Bian, D. Chen, J. Tian, P. Wang, Emission and influences of non-road mobile sources on air quality in China, 2000–2019, Environmental Pollution, 324 (2023) 121404.

[6] C. Song, C. Ma, Y. Zhang, T. Wang, L. Wu, P. Wang, Y. Liu, Q. Li, J. Zhang, Q. Dai, Heavy-duty diesel vehicles dominate vehicle emissions in a tunnel study in northern China, Science of the Total Environment, 637 (2018) 431-442.

[7] S. Sun, L. Sun, G. Liu, C. Zou, Y. Wang, L. Wu, H. Mao, Developing a vehicle emission inventory with high temporal-spatial resolution in Tianjin, China, Science of the Total Environment, 776 (2021) 145873.

[8] M.o. Ecology, E.o.t.P.s.R.o. China, China mobile source environmental management annual report,

in, Ministry of Ecology and Environment of the People's Republic of China ···, 2023.
[9] MEE, SAMR, Limits and Measurement Methods for Emissions from Diesel Fuelled Heavy-Duty Vehicles (CHINA VI), GB 17691–2018, (2018).
[10] L. Gang, Y. Ying, Z. Minghui, Z. Xin, J. Liang, Key technical contents of the China VI emission standards for diesel fuelled heavy-duty vehicles, Johnson Matthey Technology Review, 63 (2019) 21-31.
[11] H. Cui, F. Posada, Z. Lv, Z. Shao, L. Yang, H. Liu, Cost-benefit assessment of the China VI emission standard for new heavy-duty vehicles, The International Council on Clean Air Transportation: Washington DC, USA, (2018) 13.
[12] P.J.P. UPDATE, CHINA'S STAGE VI EMISSION STANDARD FOR HEAVY-DUTY VEHICLES (FINAL RULE), (2018).
[13] G. Xu, W. Shan, Y. Yu, Y. Shan, X. Wu, Y. Wu, S. Zhang, L. He, S. Shuai, H. Pang, Advances in emission control of diesel vehicles in China, Journal of Environmental Sciences, 123 (2023) 15-29.
[14] Z. Chen, Q. Liu, H. Liu, T. Wang, Recent advances in SCR systems of heavy-duty diesel vehicles—low-temperature NOx reduction technology and combination of SCR with remote OBD, Atmosphere, 15 (2024) 997.
[15] S.-F. Wang, Y.-T. Chen, C.-W. Yang, W.-X. Chang, Y.-H. Liang, C.-C. Chen, High-performance vehicle diagnostic information collection device for vehicle big data, in: 2019 International Conference on Image and Video Processing, and Artificial Intelligence, SPIE, 2019, pp. 717-720.
[16] Y. Jiang, Y. Tan, J. Yang, G. Karavalakis, K.C. Johnson, S. Yoon, J. Herner, T.D. Durbin, Understanding elevated real-world NOx emissions: heavy-duty diesel engine certification testing versus in-use vehicle testing, Fuel, 307 (2022) 121771.
[17] X. Li, Y. Ai, Y. Ge, J. Qi, Q. Feng, J. Hu, W.C. Porter, Y. Miao, H. Mao, T. Jin, Integrated effects of SCR, velocity, and Air-fuel Ratio on gaseous pollutants and CO2 emissions from China V and VI heavy-duty diesel vehicles, Science of the Total Environment, 811 (2022) 152311.
[18] A.K. Agarwal, N.N. Mustafi, Real-world automotive emissions: Monitoring methodologies, and control measures, Renewable and Sustainable Energy Reviews, 137 (2021) 110624.
[19] Y. Tan, P. Henderick, S. Yoon, J. Herner, T. Montes, K. Boriboonsomsin, K. Johnson, G. Scora, D. Sandez, T.D. Durbin, On-board sensor-based NO x emissions from heavy-duty diesel vehicles, Environmental science & technology, 53 (2019) 5504-5511.
[20] H. Wang, Q. Liu, B. Bai, J. Wang, H. Xiao, H. Liu, J. Liang, Z. Lin, D. He, H. Yin, Exploring heavy-duty truck operational characteristics through On-Board Diagnostics (OBD) data, Research in Transportation Business & Management, 57 (2024) 101204.
[21] Y. Cheng, L. He, W. He, P. Zhao, P. Wang, J. Zhao, K. Zhang, S. Zhang, Evaluating on-board sensing-based nitrogen oxides (NOX) emissions from a heavy-duty diesel truck in China, Atmospheric Environment, 216 (2019) 116908.
[22] S. Zhang, P. Zhao, L. He, Y. Yang, B. Liu, W. He, Y. Cheng, Y. Liu, S. Liu, Q. Hu, On-board monitoring (OBM) for heavy-duty vehicle emissions in China: Regulations, early-stage evaluation and policy recommendations, Science of The Total Environment, 731 (2020) 139045.
[23] P. Zhao, X. Wu, S. Zhang, L. He, Y. Yang, Q. Hu, C. Huang, B. Yu, Y. Wu, Regulatory Insights for On-Board Monitoring of Vehicular NOx Emission Compliance, Environmental Science & Technology, 58 (2024) 7968-7976.
[24] F. Deng, Z. Lv, L. Qi, X. Wang, M. Shi, H. Liu, A big data approach to improving the vehicle emission inventory in China, Nature communications, 11 (2020) 2801.

[25] Z. Lv, Y. Zhang, Z. Ji, F. Deng, M. Shi, Q. Li, M. He, L. Xiao, Y. Huang, H. Liu, A real-time NOx emission inventory from heavy-duty vehicles based on on-board diagnostics big data with acceptable quality in China, Journal of Cleaner Production, 422 (2023) 138592.
[26] J. Wang, R. Wang, H. Yin, Y. Wang, H. Wang, C. He, J. Liang, D. He, H. Yin, K. He, Assessing heavy-duty vehicles (HDVs) on-road NOx emission in China from on-board diagnostics (OBD) remote report data, Science of The Total Environment, 846 (2022) 157209.
[27] Z. Yang, K. Han, L. Liao, J. Wu, Using Multi-Source Data to Identify High-Emitting Heavy-Duty Diesel Vehicles, arXiv preprint arXiv:2404.10243, (2024).
[28] X. Zhang, J. Li, H. Liu, Y. Li, T. Li, K. Sun, T. Wang, A fuel-consumption based window method for PEMS NOx emission calculation of heavy-duty diesel vehicles: Method description and case demonstration, Journal of Environmental Management, 325 (2023) 116446.
[29] N. Wei, Z. Jia, Z. Men, C. Ren, Y. Zhang, J. Peng, L. Wu, T. Wang, Q. Zhang, H. Mao, Machine learning predicts emissions of brake wear PM2. 5: model construction and interpretation, Environmental Science & Technology Letters, 9 (2022) 352-358.
[30] Y. Jia, X. Hu, W. Kang, X. Dong, Unveiling Microbial Nitrogen Metabolism in Rivers using a Machine Learning Approach, Environmental Science & Technology, 58 (2024) 6605-6615.
[31] Z. Jia, J. Yin, Z. Cao, N. Wei, Z. Jiang, Y. Zhang, L. Wu, Q. Zhang, H. Mao, Large-scale deployment of intelligent transportation to help achieve low-carbon and clean sustainable transportation, Science of the total environment, 949 (2024) 174724.
[32] G.H. Shahariar, T.A. Bodisco, N. Surawski, M.M.R. Komol, M. Sajjad, T. Chu-Van, Z. Ristovski, R.J. Brown, Real-driving CO2, NOx and fuel consumption estimation using machine learning approaches, Next Energy, 1 (2023) 100060.
[33] N. Wei, Q. Zhang, Y. Zhang, J. Jin, J. Chang, Z. Yang, C. Ma, Z. Jia, C. Ren, L. Wu, Super-learner model realizes the transient prediction of CO2 and NOx of diesel trucks: Model development, evaluation and interpretation, Environment International, 158 (2022) 106977.
[34] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, Lightgbm: A highly efficient gradient boosting decision tree, Advances in neural information processing systems, 30 (2017).
[35] P.I. Frazier, Bayesian optimization, in: Recent advances in optimization and modeling of contemporary problems, Informs, 2018, pp. 255-278.
[36] B. Guan, R. Zhan, H. Lin, Z. Huang, Review of state of the art technologies of selective catalytic reduction of NOx from diesel engine exhaust, Applied Thermal Engineering, 66 (2014) 395-414.
[37] Y. Inomata, S. Hata, M. Mino, E. Kiyonaga, K. Morita, K. Hikino, K. Yoshida, H. Kubota, T. Toyao, K.-i. Shimizu, Bulk vanadium oxide versus conventional V2O5/TiO2: NH3–SCR catalysts working at a low temperature below 150° C, Acs Catalysis, 9 (2019) 9327-9331.