Lambda-Guided Bounded Scale Adaptation for A Physics-Guided Virtual Mass-Airflow Estimator: A BMW M30B35 Simulation Study

Alaa A. Aldhalemi
Electrical Engineering Department,
University of Kufa, Iraq
alaaa.aldhalemi@uokufa.edu.iq
https://doi.org/10.56201/ijemt.vol.12.no10.2026.pg130.148

Abstract
The virtual mass-airflow (VMA) estimation for the BMW M30B35 engine was tested practically in simulation. The estimator combines a speed-density model with a (513-parameter) residual network. There is an online feedback loop for a single bounded scale parameter provided by the measurement of the injector voltage levels and lambda without a reference airflow sensor. The nominal mean absolute error (MAE) was reduced by 69.9% using residual learning and normalised root mean square error (NRMSE) for post-shift (scale adaptation after a simulated change in the plant) was reduced from 0.0152 to 0.0079. The four-stage supervisor did not significantly improve upon simpler policies on the original final test, and failed to satisfy the predeclared false-update criterion. Original results were then inspected and another fault test, exploratory, was conducted. For this purpose, it evaluates seven different feedback conditions with 60 run–seed pairs for each condition for 12 different load sequences and 5 frozen networks. Relative to validity-only adaptation (Holm-adjusted p = 0.0103), adaptation root-mean-square error (RMSE) was found to be 8.023 during a brief +15% lambda fault, versus 6.750 during acceptance during a similar error window in the later part of the hour, with no false updates, compared to 60 during the validity-only run. During this high of +15% there were 181 false updates in C2 and C3 each. Other recursive methods like random phase recurring fault did not give much benefit and recursive least squares (RLS) showed lower adaptation RMSE when the input was healthy feedback. This meant that only if the fault occurred after the window opened (and the feedback was accurate) did the benefit of later acceptance come. Before a practical, deployment, engine measurements, embedded timing tests and electrical compatibility checks need to be performed.

Keywords: *virtual mass-airflow estimator, physics-guided residual learning, lambda-guided calibration, excitation gating, adaptive calibration*

1. Introduction

Accurate fuel metering in a spark-ignition engine requires an estimate of the air entering the cylinders. Errors in this estimate affect the air–fuel ratio, particularly during throttle transients when exhaust-oxygen feedback is delayed (Stotsky et al., 2004; Wang et al., 2016). An upstream airflow meter measures air entering the intake, but manifold filling and emptying cause this flow to differ from cylinder airflow during tip-in and tip-out (L. Meng et al., 2019; Wang et al., 2016).

Speed-density estimation calculates cylinder airflow from engine speed, manifold pressure (MAP), intake temperature, displacement and a volumetric-efficiency (VE) map. The map requires calibration because VE varies with speed, load, temperature and engine condition (De Nicolao et al., 1996; de Nola et al., 2019; Stotsky et al., 2004).

Adaptive observers and self-learning VE maps can correct model error during operation (Gao et al., 2023; L. Meng et al., 2019; Monir et al., 2015; Stotsky et al., 2004). Neural virtual sensors offer another approach by learning airflow from data (Cardone et al., 2022; Fornaro et al., 2024; Savioli et al., 2024). In either case, online calibration depends on informative measurements and reliable feedback. An update may fit its calibration window well yet increase error on subsequent data. An adaptive soft sensor must therefore distinguish changes in the process from noise in the calibration signal (Kadlec et al., 2011; J. Li et al., 2023). A physical model with a learned correction can reduce dependence on training data (Q. Li et al., 2024; C. Meng et al., 2025). This study examines whether such an estimator can adapt without a reference airflow sensor and whether evaluating updates on a later window improves calibration reliability.

1.1 Related Work

Virtual sensors based on data and physics. Neural networks predict MAP and mass airflow (MAF), calibrate airflow and provide sensor redundancy (Alfieri et al., 2020; Cardone et al., 2022; Fornaro et al., 2024; Savioli et al., 2024). Their accuracy depends on the quality of the training labels and coverage of the operating range. Physics-informed learning incorporates physical knowledge through training data, model architecture or the learning objective (C. Meng et al., 2025; Seyyedi et al., 2024), as in VE estimation (Q. Li et al., 2024) and EGR-cooler temperature prediction (Kumar et al., 2025). A recent conceptual framework for diesel injection control combines physics-informed learning with recursive-least-squares correction of model factors and a deterministic fallback to static calibration. The framework remains conceptual, and its reported performance figures are projections rather than experimental measurements (Imafidon, 2026). Intake diagnostics compare measured and predicted MAF and MAP together with throttle position and engine speed to detect faults that remain within normal electrical signal limits (Komorska et al., 2017, 2018). The detection threshold depended on model uncertainty, and validation was restricted to steady operation (Komorska et al., 2017). Virtual sensors can also maintain estimates during sensor faults (Escobar-Jiménez et al., 2024). Calibration must therefore be tested with faulty as well as healthy feedback.

Gap and approach. The present estimator uses injector and lambda measurements to construct its calibration signal. A single scale parameter is updated only when the excitation criterion is met, the proposed change remains within its bounds and the candidate improves agreement on a later window. Simulated electrical channels represent the signals available to a retrofit installation. The contribution lies in this calibration procedure and its evaluation against simpler policies; residual learning itself is established. The results do not constitute evidence of general safety. Lambda-sensor dynamics and measurement validity remain relevant to calibration (Ricci et al., 2024). Oxygen sensors provide feedback on the air–fuel ratio relative to stoichiometric conditions, but coolant leaks, silicone sealants, burnt oil and some fuel additives can degrade their performance (Ahuchaogu & Ehibe, 2024).

1.2 Study Overview

The case study represents the BMW E34 535i equipped with the M30B35 engine and Motronic 1.3 engine control unit (ECU) (Bentley Publishers, 1998; BMW AG, 1991; Legg & Warren, 1997). The proposed retrofit reads electrical signals directly rather than through OBD-II. The proposed installation requires an additional MAP sensor and a continuous throttle-position sensor. Compatibility with the original ECU remains untested.

The study addresses four aspects of the calibration problem.

1. A compact residual airflow estimator that adapts one bounded scale parameter online, using a calibration signal reconstructed from injector and lambda measurements after validity checks and delay conditioning, without a reference airflow sensor.
2. A scalar excitation test, evaluated on eight steady-idle control trajectories.
3. Separate proposal and acceptance windows, compared with simpler policies and tested with corrupted lambda feedback.
4. A simulation with separate plant, estimator, fuel-delivery and calibration models.

Validation is restricted to simulation, and the estimated airflow does not control fuel injection or ignition. The proposed legacy voltage/frequency interface in Fig. 1, electrical compatibility and embedded operation require hardware validation.

2. Methodology

2.1 Virtual Test Bed

Plant and sensors. The M30B35 is a naturally aspirated six-cylinder engine with a displacement of 3430 cm^3 (Legg & Warren, 1997). Table 1 lists the engine and simulation parameters. Engine geometry follows the service literature. The VE surfaces, the manifold time constant (0.09 s), the injector flow (4.35 g s^{-1} at 300 kPa), the dead time (0.85 ms), and the noise and mismatch settings are simulation assumptions. The hidden plant is a mean-value model (Hendricks & Sorenson, 1990) with throttle and manifold dynamics, thermal states and pressure pulsation. The estimator uses a different VE table to introduce a controlled modelling error. The virtual ECU obtains injector pulse width from an imperfect fuel map. The fuel-delivery model has separate injector-scale, dead-time and wall-film parameters or states. Lambda is calculated from hidden airflow and delivered fuel, and is then delayed, filtered and corrupted. Consequently, the reconstructed calibration signal differs from the hidden plant airflow. The measured channels include range limits, bias, noise, quantization, lag and optional drift. Fault scenarios also include signal dropout and stuck readings.

Data. The full-mode dataset contains 132 runs and 720 000 samples recorded at 50 Hz, corresponding to four hours of simulation. Its 13 scenario groups cover nominal, boundary and corner conditions, unseen transients, four plant-mismatch families, calibration shifts and low-information operation. The training and validation procedure was followed by the complete runs and the threshold selection followed by the final testing (Table 2). All Run Identifiers, file checksums and separation checks between dataset splits were OK. Completely, the input scalers were applied during training runs only and every "update" and "inference" during online execution was completely without any reason.In full, an unknown input of the scalers that could not have been read out of the surrounding air and an unknown update of fake labels did not get into online inference or update decisions. The information a user can infer on the origin of the original simulation snapshot is limited because the source code is not stored and the

corresponding revision is not recorded. It will be documented their seed and software environment in the GitHub repository. The reported hash for the hash chain (for the latter re-run of the exploration) does not point back to the original source revision. Quick-mode runs were not included. In Stage A health flags were provided with the simulator; physical measurement-based fault detection was not assessed. The GitHub repository (under Data and code availability) includes configuration notes, calibration-data efficiency analyses, gate logs and event summaries.

Table 1. Engine and simulation parameters.

Parameter	**Value**	**Status**
Displacement, cylinders	3430 cm^3, 6	Service literature (Legg & Warren, 1997)
Speed and MAP range	500–6500 rpm; 10–110 kPa	Configured simulation domain
Stoichiometric AFR	14.7	Gasoline constant (Ahuchaogu & Ehibe, 2024)
Manifold / throttle time constant	0.09 s / 0.12 s	Simulation assumption
Injector flow, dead time	4.35 g s^{-1} at 300 kPa; 0.85 ms	Simulation assumption
Sampling rate	50 Hz	Configured

Table 2. Dataset split (132 runs; 50 Hz).

Split	**Runs**	**Samples**	**Duration**
Training	24	108 000	36 min
Validation	6	27 000	9 min
Threshold selection	8	96 000	32 min
Final test	94	489 000	163 min
Total	132	720 000	240 min

2.2 Estimators and Calibration Method

Estimators. Figure 1 shows the simulated airflow estimator and its calibration supervisor. The physics branch calculates airflow using the speed-density relation

$$\dot{m}_{physics} = \frac{P_{MAP} V_d N \eta_v(N, P_{MAP})}{120 R_{air} T_{IAT}}, \quad (1)$$

where manifold pressure is P_{MAP}, engine speed in rpm is N, and displacement is V_d. Intake-air temperature is T_{IAT}, the specific gas constant of air is R_{air}, and volumetric efficiency is η_v. The factor 120 converts rpm to intake strokes per second for a four-stroke engine. Pressure is in Pa, displacement in m^3 and temperature in K, and the output is converted from kg s^{-1} to g s^{-1}. VE is obtained by bounded bilinear interpolation on a speed–MAP table. The residual network adds a correction, $\phi_k = \dot{m}_{physics,k} + f_\theta(x_k)$, where k is the sample index, f_θ is the network with

weights θ, and x_k holds RPM, MAP, throttle position, intake temperature and coolant temperature. The network has 16–16–8–1 layers with ReLU units and 513 parameters. With float32 storage, the weights occupy 2052 bytes, increasing to 2092 bytes when the ten scaler coefficients are included. The pure-AI baseline has the same inputs and architecture but predicts airflow directly. Hidden airflow is used only for offline training and evaluation.

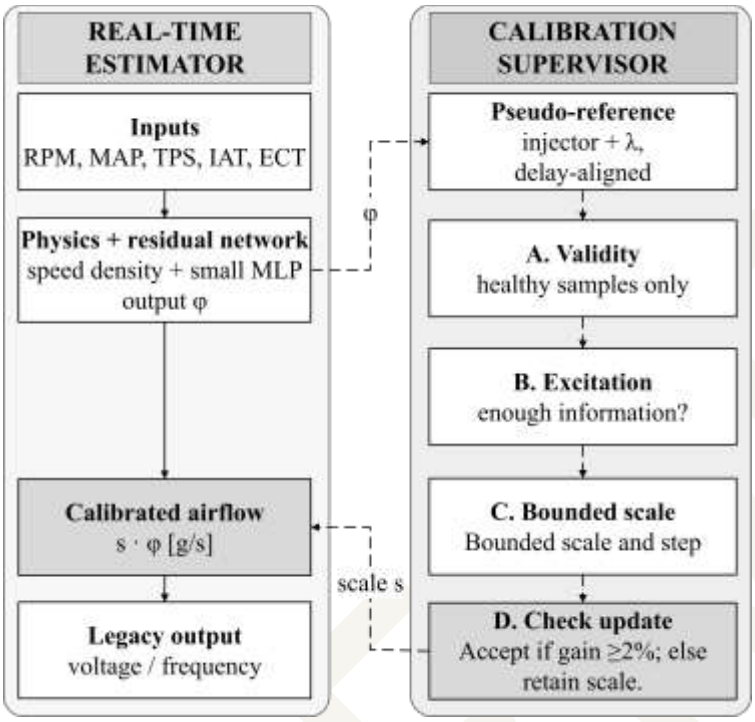


Fig. 1. Simulated airflow estimator (left) and calibration supervisor (right). Embedded timing and the proposed legacy output interface are unvalidated.

Both networks used scikit-learn MLPRegressor with Adam, squared-error loss, an L2 penalty of 10^{-4}, a learning rate of 0.005 and batches of 256. Inputs were standardised on training runs; targets were not standardised. Residual targets were clipped to $[-40, 40]$ g s^{-1}. Training ran for at most 400 epochs, stopping after 15 epochs without a validation MAE improvement greater than 10^{-6} g s^{-1}. The weights with the lowest validation MAE were restored. Both models used the same six validation runs and five seeds.

In this estimator, the physical guidance comes from the speed-density branch combined with a learned residual correction. The neural network itself uses squared-error training with L2 regularisation; no physics-based loss term or hard physical constraint is imposed.

Calibration signal. Fuel flow is reconstructed from the injector pulse width, dead time and pressure correction, and passed through a wall-film observer. Multiplying this estimate by the measured lambda λ_k and the stoichiometric ratio $AFR_{st} = 14.7$ provides the airflow pseudo-reference. Here $\hat{\dot{m}}_{fuel,k}$ is the reconstructed fuel flow:

$$\dot{m}_{air,k} = \hat{\dot{m}}_{fuel,k} AFR_{st} \lambda_k. \quad (2)$$

In the implementation, the current reconstructed fuel flow is multiplied by the measured lambda, delayed by 22 samples (0.44 s) and passed through a first-order low-pass filter with a 0.12 s time

constant (α = 1/7 at 50 Hz). The frozen estimator output undergoes the same delay and filtering before candidate fitting and acceptance. Lambda already includes the simulated exhaust and sensor dynamics, so this extra conditioning does not remove bias or transient misalignment. Calibration is disabled for the first 40 samples (0.80 s), allowing 22 samples for the delay and 18 for filter settling.

Four-stage supervisor. The online estimate is $\hat{\dot{m}} = s\phi_k$, where $\hat{\dot{m}}$ is the committed airflow, ϕ_k is the frozen network output and $s \in [0.80,1.20]$. Table 3 lists the fixed calibration settings.
– **Stage A, validity.** Simulator health flags exclude cold operation, fuel cut, enrichment, injector saturation, invalid lambda, misfire, dropout, stuck signals and observer warm-up. Measured coolant temperature must be at least 60 °C, and |dTPS/dt| must be at most 400 % s^{-1}. Invalid samples are skipped. Valid samples already stored in the proposal and acceptance buffers are retained across these gaps. An invalid sample while a proposal is being formed resets the 60-sample excitation buffer. The health-flag interface is assumed in the simulation; a physical fault detector was not implemented or tested.

– **Stage B, excitation.** Consider the approximate model $z_k = s\phi_k + v_k$, where z_k is the pseudo-reference airflow and v_k is zero-mean Gaussian noise. The sensitivity to the scale s is ϕ_k. Accumulating the squared sensitivity over a window W gives the scalar Fisher-information-inspired score

$$I_s(W) = \frac{1}{\sigma^2_{pseudo}} \sum_{k \in W} \phi_k^2 \geq I_{s,min}. \tag{3}$$

The excitation buffer holds $N_w = 60$ consecutive valid samples. With $\sigma^2_{pseudo} = 0.64$ and $I_{s,min} = 10^4$, the threshold corresponds to an RMS airflow of $\sqrt{I_{s,min} \frac{\sigma^2_{pseudo}}{N_w}} = 10.3 g s^{-1}$. The threshold was selected using the threshold-selection runs and remained fixed thereafter. Because filtering introduces correlation and the pseudo-reference contains bias, Eq. (3) serves as a fixed excitation score rather than a calibrated uncertainty measure.
– **Stage C, proposal.** The least-squares candidate is $s_{cand} = \sum_k \phi_k \dot{m}_k / \sum_k \phi_k^2$. A deadband of 0.002 suppresses small changes. The candidate step is limited to ± 0.02, the scale to [0.80, 1.20] and its rate to $0.01 s^{-1}$, using the time elapsed since the last commit. The candidate remains separate from the scale used for inference until the acceptance decision.
Each proposal and acceptance buffer must span at least 8 s and hold at least two valid samples. The excitation test uses the latest 60 consecutive valid samples when the proposal decision is made. C1 and C2 follow the same later-window schedule but accept every bounded candidate, skipping only the improvement test. The Stage D improvement-test column in Table 4 records this distinction; all online policies use only information available up to the decision time.

– **Stage D, acceptance.** The candidate is evaluated on a later window W_a that holds N_a valid samples and spans at least 8 s. The acceptance and proposal windows do not overlap. Its mean-squared error against the conditioned pseudo-reference is

$$L(s; W_a) = \frac{1}{N_a} \sum_{k \in W_a} \left[\dot{m}_{air,k} - s\phi_k\right]^2, \tag{4}$$

The pseudo-reference is defined by Eq. (2). Hidden airflow is used only for offline evaluation.

The candidate s_{cand} replaces the last-known-good scale s_{LKG} only if the relative error falls by at least 2%:

$$\rho = \frac{L(s_{LKG}; W_a) - L(s_{cand}; W_a)}{max[L(s_{LKG}; W_a), \epsilon]} \geq \rho_{min} = 0.02, \quad (5)$$

Here $\epsilon = 10^{-9}$ (g s^{-1})^2 prevents division by zero. If $\rho < 0.02$, the update is rejected and s_{LKG} is kept. Fig. 2 shows the separate proposal, acceptance and final-holdout windows. The holdout contributes no samples to candidate fitting or acceptance.

C5 uses scalar recursive least squares (RLS) with a forgetting factor of 0.985 and the same scale and rate limits. It updates without an excitation gate, a later-window acceptance test or rollback. Recursive least squares has also been proposed for correcting physics-model factors against sensor calibration drift (Imafidon, 2026).

False-update labels are assigned offline. An accepted update is false if its acceptance-window RMSE against identically delayed and filtered hidden airflow is more than 1.5 g s^{-1} above the RMSE of the previous scale. The false-update rate is the proportion of accepted updates that meet this degradation criterion. The label describes degradation within the acceptance window and does not measure failure in the final holdout. C5 has no discrete acceptance events, so it has no comparable label.

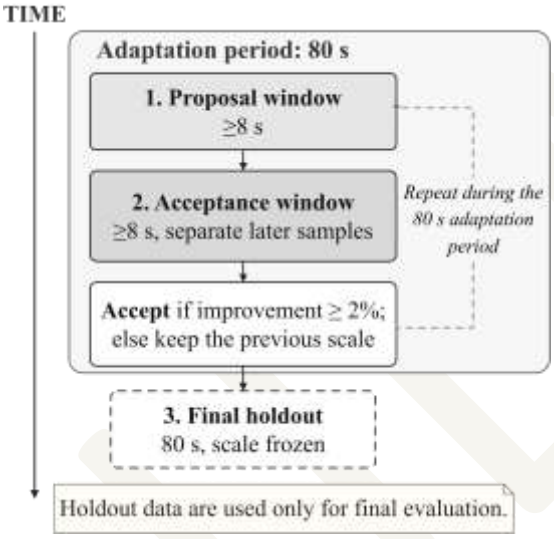


Fig. 2. Calibration and final evaluation windows.

Table 3. Calibration settings.

Parameter	Value	Parameter	Value
Initial scale / bounds	1.0; [0.80, 1.20]	Gate window / threshold	60 samples; 10 000
Per-update / rate limit	0.02; 0.01 s^{-1}	Pseudo-reference variance	0.64 (g s^{-1})^2
Deadband	0.002	Minimum improvement	2%
Proposal / acceptance	8 s / 8 s	Offline false-update limit	1.5 g s^{-1}

2.3 Experimental Design and Statistics

The original final-test evaluation used predeclared criteria. The high-load fault test was designed after inspecting those results and is reported separately as exploratory.

E1–E4 evaluate frozen estimators under nominal conditions, boundary and corner conditions, unseen transients and plant mismatch. E5 compares the six policies in Table 4 on 12 untouched calibration-shift runs and five network seeds, yielding 60 run–seed pairs for each policy. Each run contains an adaptation segment followed by a holdout segment evaluated with the final scale frozen. The secondary calibration-data efficiency experiment (E6) is reported in the GitHub repository.

Table 4. Calibration policies.

Mode	Description	Stage A	Excitation gate	Stage D improvement test	Online policy
C0	Frozen residual AI	–	–	–	Yes
C1	Bounded adaptation without validity checks	No	No	No	Yes
C2	Validity-only adaptation	Yes	No	No	Yes
C3	Four-stage supervisor	Yes	Yes	Yes	Yes
C4	True-airflow benchmark	Yes	No	Yes	No
C5	Scalar RLS	Yes	No	No	Yes

Policies marked "Online policy" do not use hidden airflow as feedback. Policies C2, C3 and C5 also assume the simulated health interface; C1 omits the validity check. C4 uses hidden airflow in both proposal and acceptance, so it serves only as a simulation benchmark.

The metrics are mean absolute error (MAE), root-mean-square error (RMSE) and normalised RMSE (NRMSE):

$$MAE = \frac{1}{n}\sum_{k} |\hat{y}_k - y_k|, \tag{6}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{k} (\hat{y}_k - y_k)^2}, \tag{7}$$

$$NRMSE = \frac{RMSE}{260 g s^{-1}}, \tag{8}$$

Here y_k is the reference airflow, $\hat{y}_k$ is the estimated airflow and n is the number of samples. NRMSE uses a fixed reporting span of 260 g s^{-1} to keep the normalisation identical across runs and policies. This denominator is not the standard deviation or a range estimated separately for each run. Signed bias and the 95th-percentile absolute error are also reported. Metrics were first computed for each run–seed pair and then averaged over seeds within each run.

Statistical inference used independent runs as the units of observation; time samples and network seeds were not treated as independent replicates. Paired differences were checked with the Shapiro–Wilk test, followed by a paired t-test or a Wilcoxon signed-rank test. The analysis uses 5000-resample bootstrap intervals, paired Cohen's d_z (the mean difference divided by its standard deviation) and Holm correction across six planned comparisons. The acceptance criteria were fixed before the final test was inspected: C3 NRMSE ≤ 0.12, at least 10% improvement over C0, no scale-bound violations and a false-update rate strictly lower than C1.

The six planned comparisons were recalculated after seed averaging and restricted to the 12

final-test calibration runs. The archived statistical_tests.csv pooled threshold-selection and final-test runs and was excluded from the p-values reported here. Nominal accuracy uses 5000-resample bootstraps over independent runs, with seed 0. The GitHub repository documents the corresponding correction to the E6 analysis.

Exploratory fault test. The protocol was written after inspection of the original final-test results and was not externally preregistered. The first execution files could not be recovered. All exploratory results reported here come from one re-run implemented from the written protocol. The configuration was locked before inspection of the output (SHA-256 prefix f18aa3b6206b3978); settings remained unchanged.

The test used twelve new 240 s high-load trajectories with seeds 2026100400–2026100411. All 12 trajectories were employed for each feedback condition as well as five frozen network seeds, resulting in 60 run–seed pairs for each policy. These trajectories adopted the same plant and sensor-noise models as before with severe VE mismatch added at 80 seconds. The commands were kept inside of the typical training parameters of 3400– 4400 RPM and 60–78% TPS.

The seven matched conditions included those with healthy feedback and had faults in the range of 15% with some faults being brief, others recurring, and some faults being persistent, in the range of 80–88 s, 80–160 s respectively. Almost every fault was arranged in blocks of 8 s on and 8 s off, either in synchrony with the proposal windows or with a random phase, sampled independently for each run. Dropout, stuck-signal and misfire events have been disabled to verify the lambda corruption only is the problem. Networks, controller thresholds and bounds were kept constant. The configurations and results can be found in the GitHub data repository provided in Data and code availability.

The primary exploration objective was RMSE of the committed estimation of the full 80 s adaptation period with noise obtained using unfiltered hidden airflow. The secondary outcomes were holdout NRMSE and original false update label. The average of 5 network seeds was taken in each run.

In each condition, two-sided exact sign-flip tests were done in each comparison for C3 versus C1, C2 and C5 where $\alpha = 0.05$ and Holm correction applied for 21 comparisons. This assumes that there is equality under the null in run-level difference. The average number of run-bootstrap resamples (seed 0) used to obtain the unadjusted 95% percentile interval is 5000. No negative or zero feedback was removed no matter the results. However, there has been no change in the acceptance criteria or endpoint for the original 12-run calibration experiment, the primary confirmatory analysis, as a result of the exploratory fault test.

3. Results

3.1 Estimator Accuracy

Nominal accuracy. On 10 untouched 90 s runs, residual AI reduced MAE from 1.453 to 0.438 g s^{-1}. The reductions relative to speed density and pure AI were 69.9% and 3.95%, respectively (Table 5). No separate paired test was predeclared for nominal conditions. Across all 94 matched test trajectories, residual AI had lower NRMSE than pure AI: mean difference -0.0097, $p_{Holm} = 4.6 \times 10^{-12}$ and $d_z = -0.43$ (Table 10).

Table 5. Nominal accuracy: means [95% bootstrap CIs], 10 runs.

Model	MAE (g s^{-1})	RMSE (g s^{-1})	NRMSE	P95 error (g s^{-1})
Physics-only	1.453 [1.014, 1.975]	1.895 [1.397, 2.455]	0.00729	3.820
Pure AI	0.456 [0.404, 0.509]	0.590 [0.529, 0.655]	0.00227	1.194
Residual AI	0.438 [0.386, 0.491]	0.573 [0.510, 0.637]	**0.00220**	**1.166**

Transients. The twelve unseen 120 s runs cover seven operating states (Table 6). Residual AI achieved an MAE of $0.319 \pm 0.022 gs^{-1}$, compared with 0.434 ± 0.072 for pure AI and 0.889 g s^{-1} for speed density. These correspond to MAE reductions of 26.6% relative to pure AI and 64.2% relative to speed density. Residual AI had the lowest MAE in six of the seven states. Pure AI had slightly lower MAE during acceleration (0.340 versus 0.345 g s^{-1}). Cross-correlation analysis detected no lag at the 20 ms sampling resolution. Fig. 3 shows one representative run.

Table 6. Transient MAE (g s^{-1}), averaged over 12 runs and five network seeds.

Estimator	Idle	Tip-in	Accel.	Cruise	High load	Tip-out	Decel.
Physics-only	0.372	0.308	0.899	0.936	1.497	1.097	0.643
Pure AI	0.300	0.169	**0.340**	0.398	0.699	0.756	0.410
Residual AI	**0.113**	**0.124**	0.345	**0.383**	**0.447**	**0.457**	**0.224**

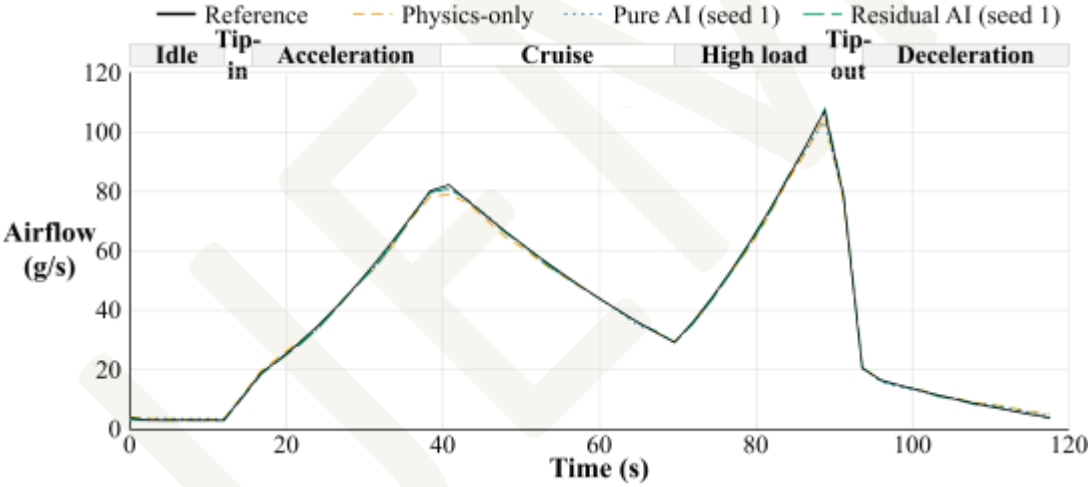


Fig. 3. Airflow estimates during an unseen transient run.

Boundary and corner conditions. Residual AI achieved the lowest MAE in five of the eleven conditions excluded from training, but speed density had lower pooled MAE in both groups (Table 7). Each group contains 10 runs and 30 000 samples, with the AI results averaged over five seeds. The seed standard deviation of corner MAE was 1.9 g s^{-1} for residual AI and 17.0 for pure AI. The physical baseline reduced variation between seeds, although the residual model remained unreliable outside its training coverage. Practical use would require a validated fallback to the physics estimate.

Table 7. Boundary and corner MAE (g s^{-1}); best values are bold.

Group	Condition	Physics	Pure AI	Residual AI
Boundary	Minimum-speed idle	0.407	1.036	**0.170**
	Redline	13.326	**3.602**	17.264
	Wide-open throttle	**1.064**	7.037	3.625
	Minimum MAP	1.809	1.726	**1.642**
	Extreme intake temperature	0.494	0.315	**0.290**
	Extreme coolant temperature	**0.719**	23.919	3.331
	Pooled	**2.970**	6.272	4.387
Corner	Lugging	1.164	0.555	**0.238**
	Engine braking	**1.534**	16.326	6.749
	Cold high load	**2.395**	19.255	4.316
	Hot wide-open throttle	**1.381**	19.370	5.890
	Redline, low MAP	5.229	38.910	**3.705**
	Pooled	**2.340**	18.883	4.180

Plant mismatch. Plant mismatch was evaluated with the estimators held fixed: 10 runs each covered nominal, low and high VE mismatch, and six each covered MAP-gain and manifold-time-constant error. Table 8 lists NRMSE for the nominal, VE-mismatch and MAP-gain conditions. At low VE mismatch (scale 1.06 and +150 rpm shift), residual AI improved on speed density but had slightly higher NRMSE than pure AI. At high mismatch (scale 0.90 and −300 rpm shift), speed density reached 0.01068, residual AI 0.01250 and pure AI 0.01267. Residual AI was then 17.1% worse than speed density.

Linear interpolation between the 6% and 10% VE-mismatch cases places the NRMSE crossover near 8.8% for both networks (8.6% using MAE). Both the VE scale and surface shape differed, so this is a descriptive estimate rather than a validated operating limit (Fig. 4). Residual AI had the lowest NRMSE under MAP-gain error. With a slower manifold, pure AI had slightly lower NRMSE, while residual AI had lower MAE. The plant-mismatch result files in the GitHub repository listed under Data and code availability provide that comparison.

Table 8. NRMSE under plant mismatch; best values are bold.

Estimator	Nominal	Low VE	High VE	MAP gain
Physics-only	0.00729	0.01234	**0.01068**	0.00578
Pure AI	0.00227	**0.00769**	0.01267	0.00599
Residual AI	**0.00220**	0.00786	0.01250	**0.00544**

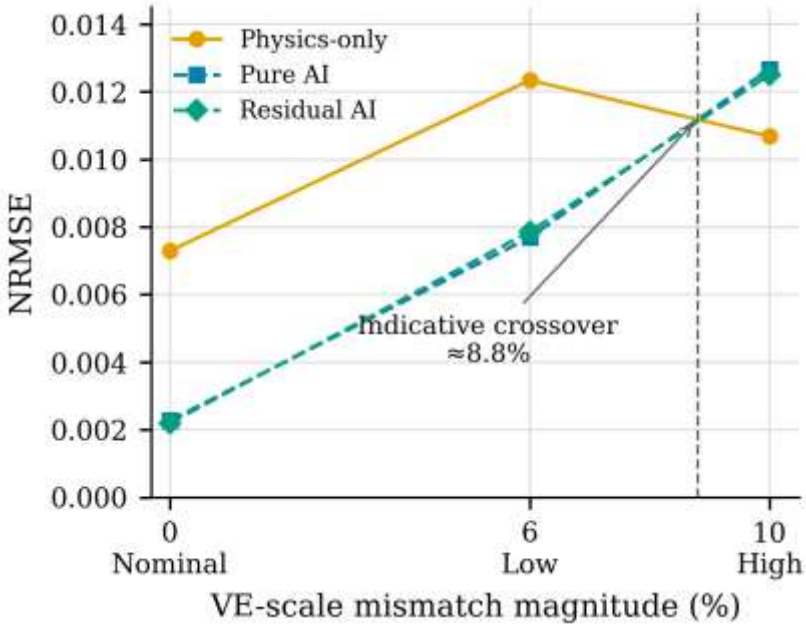


Fig. 4. NRMSE under VE mismatch.

3.2 Pseudo-Reference Quality

The conditioned pseudo-reference was compared with identically delayed and filtered hidden airflow using the coefficient of determination R^2_{pseudo} over the calibration-eligible samples, with the conditioned hidden airflow as the reference. Hidden airflow is excluded from the online calibration path. The pseudo-reference tracked the conditioned airflow closely: $R^2 = 0.9940 \pm 0.0010$ on eight threshold-selection runs and 0.9939 ± 0.0007 on 12 final-test runs, with individual values of 0.9922–0.9953. Its final-test bias was $+1.25 \pm 0.18 gs^{-1}$, about 3.5% of mean airflow, and arose from the imposed injector mismatch. The close agreement in trend therefore coexisted with a systematic calibration bias. A bounded, single-scale model limits adaptation, but later-window acceptance cannot identify a bias that both windows shares.

3.3 Excitation Gate Behaviour

The excitation gate blocked no proposal in the calibration-shift runs: the smallest proposal score was 13 397, above the 10 000 thresholds. The 15 rejected final-test updates failed the later-window improvement test. Fig. 5 shows scale changes, excitation and decisions in one shift run; the GitHub repository provides the detailed logs.

On eight steady-idle trajectories replayed with five frozen networks, no complete 60-sample window passed the excitation gate. The mean and maximum scores were 745 and 789, both below 8% of the threshold, at a mean hidden airflow of 2.76 g s^{-1}. In contrast, every tested window in final-test shift run 000 passed. Overlapping windows describe gate behaviour and are not independent statistical observations. The replay records are available in the GitHub repository.

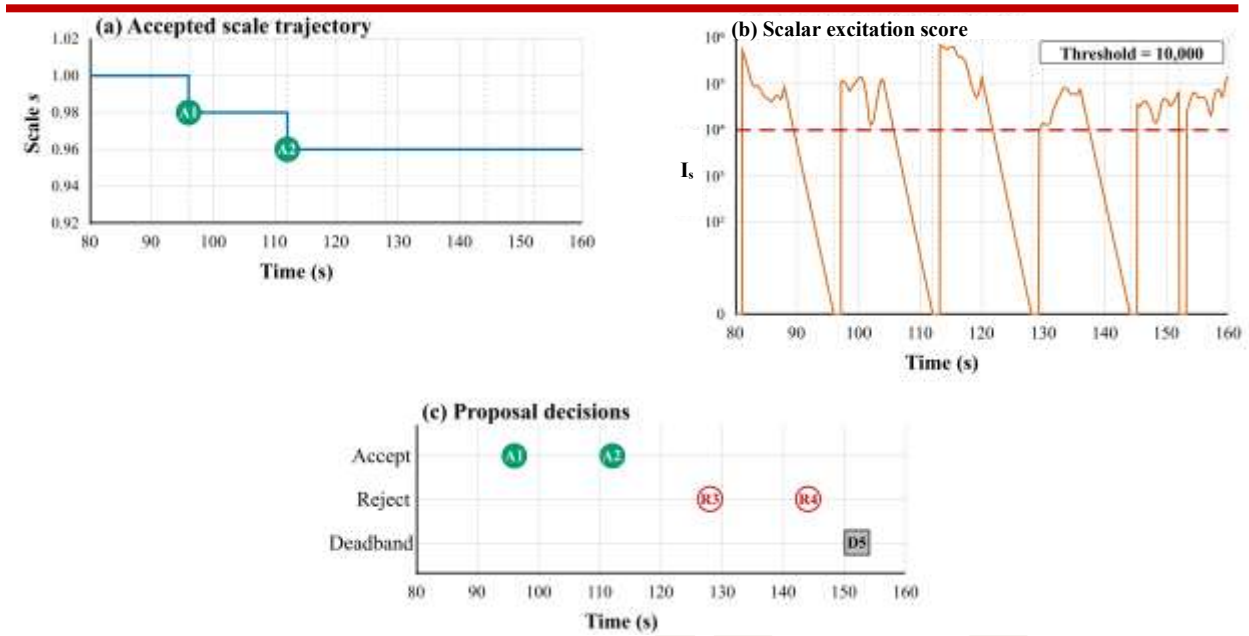


Fig. 5. Scale, the excitation score defined by Eq. (3) and update decisions during a shift run.

Steady idle was the only low-information condition tested. With one scale parameter, the excitation score is a sum of squared airflow, and therefore functions mainly as a minimum-airflow criterion. The replay shows that the gate rejects sustained idle and admits one shift trajectory. These results provide no evidence of protection against a biased pseudo-reference that still meets the excitation criterion. The single-parameter test also cannot establish identifiability for a model with several adjustable parameters.

3.4 Calibration-Policy Ablation

Table 9 compares post-shift holdout performance over the 12 final-test runs. C3, the four-stage supervisor, reduced mean NRMSE from 0.01522 for C0 to 0.00789 (48.2%) and MAE from 3.374 to 1.633 g s^{-1}. The paired t-test gave $t(11) = -8.269$ with 11 degrees of freedom, raw $p = 4.8 \times 10^{-6}$, $p_{Holm} = 2.4 \times 10^{-5}$ and $d_z = -2.39$. The committed scale remained within its bounds. C3 accepted 203 proposals and rejected 15.

Table 9. Post-shift holdout accuracy (mean ± SD; 12 runs).

Mode	NRMSE	MAE (g s^{-1})
C0	0.0152 ± 0.0023	3.374 ± 0.548
C1	0.0071 ± 0.0019	1.450 ± 0.433
C2	0.0079 ± 0.0036	1.629 ± 0.842
C3	0.0079 ± 0.0036	1.633 ± 0.844
C4	0.0048 ± 0.0042	0.972 ± 0.951
C5	0.0083 ± 0.0035	1.727 ± 0.833

Table 9 (continued). Final scale (mean ± SD) and events (60 run–seed pairs).

Mode	Final s	Accepted/rejected	Incomplete candidates	False acc.
C0	1.000 ± 0.000	0/0	0	–
C1	0.950 ± 0.011	238/0	43	0
C2	0.954 ± 0.018	218/0	38	0
C3	0.954 ± 0.018	**203/15**	34	**0**
C4	0.931 ± 0.022	220/0	30	0
C5	0.957 ± 0.018	–	–	–

C4 uses true airflow and is a simulation benchmark that cannot be deployed. C5 is continuous RLS and has no discrete acceptance or false-update labels. "Rejected" means the improvement test failed; "incomplete" means adaptation ended before acceptance finished. Both retain the previous scale. The false-update counts for C3 and C1 were 0/203 and 0/238 accepted updates. First acceptance occurred at 16.02 s, about 800 continuously valid samples; 55 of 60 pairs updated under C2–C4. The GitHub repository provides the sample-budget analysis and event totals. C3 did not differ significantly from C1, C2 or C5 in the planned paired tests (Table 10). The mean NRMSE difference relative to C2 was 2×10^{-5}; relative to RLS, mean NRMSE was 4.7% lower. These tests do not show that the policies are equivalent. C1 had the lowest deployable mean, largely because of stuck-sensor run 002. C1 accepted four corrupted-feedback updates on this run, whereas C2–C5 did not update. The resulting improvement for C1 does not establish the reliability of adaptation without validity checks (see the sensitivity analysis in the GitHub repository). With run 002 removed, the mean NRMSE of C3 was 0.00698 instead of 0.00789. The differences remained nonsignificant: C3–C1 $+2 \times 10^{-5}$ (Wilcoxon $p = 0.84$), C3–C2 $+2 \times 10^{-5}$ ($p = 0.81$), and C3–C5 -4.2×10^{-4} ($p = 0.58$).

Table 10. Paired NRMSE differences (first minus second); negative favours the first.

Comparison	Test	n	Mean diff.
Residual AI vs. physics	Wilcoxon	94	+0.00031
Residual AI vs. pure AI	Wilcoxon	94	−0.00972
C3 vs. C0	paired t	12	−0.00733
C3 vs. C1	Wilcoxon	12	+0.00078
C3 vs. C2	Wilcoxon	12	+0.00002
C3 vs. C5	paired t	12	−0.00039

Table 10 (continued). Holm-adjusted p-values and paired effect sizes.

Comparison	p_{Holm}	d_z
Residual AI vs. physics	1.000	$+0.06$
Residual AI vs. pure AI	4.6×10^{-12}	-0.43
C3 vs. C0	2.4×10^{-5}	-2.39
C3 vs. C1	1.000	$+0.30$
C3 vs. C2	1.000	$+0.09$
C3 vs. C5	1.000	-0.13

C3 met the predeclared accuracy, improvement and scale-bound criteria. It failed the requirement for a lower false-update rate than C1, because both rates were zero (Table 11). The overall predeclared outcome was therefore FAIL. No accepted update exceeded the specified hidden-error tolerance, so this label could not separate the policies. The result establishes neither a safety benefit nor the redundancy of the acceptance gate.

Table 11. Predeclared acceptance criteria.

Criterion	**Threshold**	**Outcome**
C3 final-test NRMSE	≤ 0.12	Pass (0.00789)
Improvement over C0	$\geq 10\%$	Pass (48.2%)
Scale-bound violations	0	Pass (0)
C3 false-update rate	lower than C1	**Fail** (0.000 vs. 0.000, tie)

3.5 Exploratory Adverse Calibration Test

This exploratory re-run followed inspection of the original final-test results and does not replace that evaluation. All 2520 policy replays completed successfully, and the frozen-input checks confirmed that engine inputs and hidden airflow were identical across lambda conditions. Only the measured lambda and its pseudo-reference changed. C0 and the benchmark C4 were not affected by lambda faults. Every adaptation sample passed Stage A, and every C3 proposal passed the excitation gate. The experiment thus evaluates the later-window acceptance test under conditions where the other gates do not restrict proposals. The RMS airflow of the frozen estimate was 84.2–88.0 g s^{-1}. At this high load, the existing 1.5 g s^{-1} false-update threshold is responsive: a scale change can alter window RMSE by at most $|\Delta s|$ RMS(ϕ). With a maximum step of 0.02, that threshold cannot be exceeded when RMS(ϕ) is at most 75 g s^{-1}.

Table 12 reports all seven feedback conditions from the exploratory re-run. During a brief +15% fault, the four-stage supervisor (C3) reduced adaptation RMSE by 1.273 g s^{-1} relative to validity-only adaptation (C2) (C3−C2 CI −1.292 to −1.257; Holm-adjusted p = 0.0103). It rejected all 60 first candidates and recorded zero false updates among 180 accepted updates, compared with 60 among 240 for C2. During aligned recurring faults, C3 rejected every candidate, kept the initial scale and reduced RMSE by 3.177 g s^{-1} (CI −3.239 to −3.115; adjusted p = 0.0103). Rejecting every candidate avoided harmful updates at the cost of leaving the actual plant shift uncorrected.

Table 12. Exploratory re-run: mean RMSE (g s^{-1}) and false/accepted updates.

Feedback condition	C2 RMSE	C3 RMSE	C5 RMSE	C2 false / accepted	C3 false / accepted
Healthy feedback	5.664	5.686	3.093	0 / 239	0 / 225
Brief +15%	8.023	6.750	5.874	60 / 240	0 / 180
Brief −15%	5.664	5.686	3.067	0 / 239	0 / 225
Recurring aligned +15%	11.686	8.509	10.951	173 / 240	0 / 0 (n/e)
Recurring random +15%	10.049	8.516	9.813	85 / 235	0 / 1
Persistent +15%	11.941	11.937	14.273	181 / 240	181 / 238
Persistent −15%	5.613	5.613	7.955	0 / 240	0 / 240

Means use 12 independent runs after averaging five network seeds, and event counts cover 60 run–seed pairs. n/e means the false-update rate cannot be estimated because no update was accepted (0/0). A rate calculated from a single accepted update (0/1) provides little evidence about reliability. C1 and C2 were identical because all samples passed the health mask. C5 has no comparable discrete false-update label. The C0–C5 holdout results, full event counts and 21 paired tests are available in the GitHub repository.

With random-phase recurring faults, C3 accepted only one update across 60 pairs. Its RMSE difference from C2 was −1.533 g s^{-1} (CI −2.408 to −0.486), but this was not significant after Holm correction ($p = 0.082$). Performance varied with the randomly selected fault phase. Under a persistent +15% bias, C3 and C2 recorded 181 false updates each, among 238 and 240 accepted updates, respectively. Their RMSE values were nearly identical (11.937 and 11.941 g s^{-1}). C3 had lower RMSE than RLS for aligned recurring faults and persistent biases, and also for random-phase recurring faults (adjusted $p = 0.043$). RLS had lower RMSE with healthy feedback and with both brief faults (each adjusted $p = 0.0103$). With healthy feedback, C3 was slightly worse than C2 (+0.021 g s^{-1}; adjusted $p = 0.043$). None of the deployable policies achieved the lowest error under every feedback condition. The exploratory results do not change the original FAIL outcome in Table 11.

5. Conclusion

Considering speed density, the nominal MAE was decreased 69.9% and transient MAE 64.2% by the residual airflow estimator in simulation. This benefit could only be encouraged within the range of supported operation, had no effect in boundary and corner circumstances, and operated in opposition in high volumetric efficiency mismatch.

The four-stage supervisor had only a marginal advantage over simpler policies for post-shift NRMSE, reducing it by 48.2%, and it could not meet the predeclared false-update criterion. Later-window acceptance minimized negative lambda checks for a short time or "aligned" recurring positive faults, while persistent bias defeated the check. The excitation gate has been found to reject all tested steady-ids windows and the RLS could get lower adaptation error for healthy feedback. It is important to understand that these results are not just an overall calibration improvement, but conditionality.

It has been decided that the study will be limited to simulation. For the implementation, it is necessary to measure the parameters of the engine; carry out extensive general fault testing; securely detect lambda drift and the error of the injector flow; embed engine parameters into the

timing parameter; and validate the ECU interface that is proposed. Results validate and support further experimental testing, but do not provide any means of determining safety or compatibility with the original ECU.

Data and code availability. The source code, simulation configurations, trained models, result files and supporting analyses are available at https://github.com/alaa198011/m30-virtual-maf-lambda-calibration. The repository includes reproducibility information, calibration-data efficiency analyses, gate replay records and calibration-event summaries.

IJEMT

References

Ahuchaogu, N., & Ehibe, P. (2024). Catalyst system and oxygen sensor performance in automobile exhaust emissions on injection engine. *International Journal of Engineering and Modern Technology, 10*(9), 19–29. https://doi.org/10.56201/ijemt.v10.no9.2024.pg19.29

Alfieri, V., Pedicini, C., & Possieri, C. (2020). Design of a neural virtual sensor for the air and charging system in a diesel engine. *IFAC-PapersOnLine, 53*(2), 14061–14066. https://doi.org/10.1016/j.ifacol.2020.12.937

Bentley Publishers. (1998). *BMW 5-series service manual: 1989–1995, 525i, 530i, 535i, 540i, including Touring*. Robert Bentley.

BMW AG. (1991). *1991 BMW 525i, 535i, M5 (E34) electrical troubleshooting manual: Vehicles produced 9/90 to 07/91*. BMW AG.

Cardone, M., Fornaro, E., Strano, S., Terzo, M., & Tordela, C. (2022). A neural network based approach for the intake air mass flow prediction in SI engines. In V. Niola, A. Gasparetto, G. Quaglia, & G. Carbone (Eds.), *Advances in Italian mechanism science: Proceedings of the 4th International Conference of IFToMM Italy* (pp. 866–873). Springer. https://doi.org/10.1007/978-3-031-10776-4_99

De Nicolao, G., Scattolini, R., & Siviero, C. (1996). Modelling the volumetric efficiency of IC engines: Parametric, non-parametric and neural techniques. *Control Engineering Practice, 4*(10), 1405–1415. https://doi.org/10.1016/0967-0661(96)00150-5

de Nola, F., Giardiello, G., Gimelli, A., Molteni, A., Muccillo, M., & Picariello, R. (2019). Volumetric efficiency estimation based on neural networks to reduce the experimental effort in engine base calibration. *Fuel, 244*, 31–39. https://doi.org/10.1016/j.fuel.2019.01.182

Escobar-Jiménez, R. F., Ocampo-Rodríguez, L. E., Osorio-Gordillo, G. L., & García-Beltrán, C. D. (2024). Sensor fault-tolerant control for an internal combustion engine. *International Journal of Adaptive Control and Signal Processing, 38*(2), 475–491. https://doi.org/10.1002/acs.3711

Fornaro, E., Cardone, M., Terzo, M., Strano, S., & Tordela, C. (2024). Experimentally validated neural networks for sensors redundancy purposes in spark ignition engines. *SAE International Journal of Engines, 17*(2), 203–220. https://doi.org/10.4271/03-17-02-0012

Gao, J., Liu, X., Sun, B., Zhao, W., Xing, H., Gao, B., & Chen, H. (2023). In-cylinder air mass flow estimation of gasoline engines based on map self-learning. *Control Engineering Practice, 140*, Article 105674. https://doi.org/10.1016/j.conengprac.2023.105674

Hendricks, E., & Sorenson, S. C. (1990). *Mean value modelling of spark ignition engines* (SAE Technical Paper 900616). SAE International. https://doi.org/10.4271/900616

Imafidon, O. (2026). A conceptual framework for intelligent fuel injection optimization in heavy-duty diesel engines: A machine learning-enhanced approach to real-time emissions control. *International Journal of Engineering and Modern Technology, 12*(4), 108–133. https://doi.org/10.56201/ijemt.vol.12.no4.2026.pg108.133

Kadlec, P., Grbić, R., & Gabrys, B. (2011). Review of adaptation mechanisms for data-driven soft sensors. *Computers & Chemical Engineering, 35*(1), 1–24. https://doi.org/10.1016/j.compchemeng.2010.07.034

Komorska, I., Wołczyński, Z., & Borczuch, A. (2017). Model-based analysis of sensor faults in SI engine. *Combustion Engines, 169*(2), 146–151. https://doi.org/10.19206/CE-140240

Komorska, I., Wołczyński, Z., & Borczuch, A. (2018). Fault diagnostics in air intake system of combustion engine using virtual sensors. *Diagnostyka, 19*(1), 25–32. https://doi.org/10.29354/diag/80972

Kumar, A., Kumar, R., Manojdharan, A., Challa, K., & Kramer, M. (2025). *A physics-guided neural network approach for EGR cooler temperature estimation: Addressing ECU model discrepancies* (SAE Technical Paper 2025-28-0393). SAE International. https://doi.org/10.4271/2025-28-0393

Legg, A. K., & Warren, L. (1997). *BMW 3- and 5-series service and repair manual* (Manual 1948). J H Haynes & Co.

Li, J., Zhou, Q., He, X., Chen, W., & Xu, H. (2023). Data-driven enabling technologies in soft sensors of modern internal combustion engines: Perspectives. *Energy, 272*, Article 127067. https://doi.org/10.1016/j.energy.2023.127067

Li, Q., Guo, F., Song, K., Xie, H., Zhou, S., & Sang, H. (2024). Physics-informed data-driven modeling for engine volumetric efficiency estimation. *IFAC-PapersOnLine, 58*(29), 403–408. https://doi.org/10.1016/j.ifacol.2024.11.178

Meng, C., Griesemer, S., Cao, D., Seo, S., & Liu, Y. (2025). When physics meets machine learning: A survey of physics-informed machine learning. *Machine Learning for Computational Science and Engineering, 1*, Article 20. https://doi.org/10.1007/s44379-025-00016-0

Meng, L., Luo, J., Yang, X., & Zeng, C. (2019). Intake air mass observer design based on extended Kalman filter for air-fuel ratio control on SI engine. *Energies, 12*(18), Article 3444. https://doi.org/10.3390/en12183444

Monir, V. H., Salehi, R., Salarieh, H., Alasty, A., & Vossoughi, G. (2015). Real-time estimation of the volumetric efficiency in spark ignition engines using an adaptive sliding-mode observer. *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering, 229*(14), 1925–1933. https://doi.org/10.1177/0954407015574805

Ricci, F., Avana, M., & Mariani, F. (2024). Enhancing lambda measurement in hydrogen-fueled SI engines through virtual sensor implementation. *Energies, 17*(16), Article 3932. https://doi.org/10.3390/en17163932

Savioli, T., Pampanini, M., Visani, G., Esposito, L., & Rinaldini, C. A. (2024). Engine mass flow estimation through neural network modeling in semi-transient conditions: A new calibration approach. *Fluids, 9*(10), Article 239. https://doi.org/10.3390/fluids9100239

Seyyedi, A., Bohlouli, M., & Nedaaee Oskoee, S. (2024). Machine learning and physics: A survey of integrated models. *ACM Computing Surveys, 56*(5), Article 115. https://doi.org/10.1145/3611383

Stotsky, A., Kolmanovsky, I., & Eriksson, S. (2004). Composite adaptive and input observer-based approaches to the cylinder flow estimation in spark ignition automotive engines. *International Journal of Adaptive Control and Signal Processing, 18*(2), 125–144. https://doi.org/10.1002/acs.785

Wang, Z., Zhu, Q., & Prucka, R. (2016). *A review of spark-ignition engine air charge estimation methods* (SAE Technical Paper 2016-01-0620). SAE International. https://doi.org/10.4271/2016-01-0620