

Article

A LightGBM-Regulated and Residual-Compensated Adaptive Extended Kalman Filter for State of Charge Estimation Under Unseen Driving Cycles

Mengqi Gao ^{1,*}, Gang Fang ¹, Li Cai ¹, Xiaojiang Zou ² and Nina Dai ^{1,*}

¹ School of Electronic and Information Engineering, Chongqing Sanxian University of Science and Technology, Chongqing 404000, China; fanggang@sanxian.edu.cn (G.F.); 20040001@sanxian.edu.cn (L.C.)

² Battery Division, Chongqing Hangying Automobile Manufacturing Co., Ltd., Chongqing 404100, China; 18311830071@163.com

* Correspondence: sawaiki@outlook.com (M.G.); a58238735@outlook.com (N.D.); Tel.: +86-1328-305-6477 (M.G.)

Abstract

Data-driven methods have been extensively used for state of charge (SOC) estimation because of their nonlinear regression capability. However, when the driving cycle differs from the training conditions, their estimation accuracy may deteriorate because of insufficient generalization ability. This problem becomes more pronounced under complex unseen dynamic conditions, while some data-driven methods also suffer from high computational complexity, which further limits their practical application. To overcome these issues, this study presents a LightGBM-regulated and residual-compensated adaptive extended Kalman filter method for SOC estimation. In the proposed method, battery dynamics are represented by a second-order equivalent circuit model, while SOC is recursively estimated using an adaptive extended Kalman filter (AEKF). LightGBM is used to predict covariance adjustment coefficients for adaptively regulating Q and R , and to compensate for the remaining SOC estimation error after AEKF processing. The proposed method is validated under multiple dynamic driving cycles and different temperature conditions. The experimental results of the main unseen-cycle tests show that the proposed method achieves good estimation accuracy and generalization performance under unseen driving cycles, with the maximum error (ME) and root mean square error (RMSE) not exceeding 2.26% and 1.28%, respectively.

Keywords: lithium-ion battery; state of charge estimation; LightGBM; adaptive extended Kalman filter; unseen driving cycles

Academic Editor: JongHoon Kim

Received: 26 May 2026

Revised: 28 June 2026

Accepted: 8 July 2026

Published: 13 July 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

As electric vehicles and energy storage systems continue to develop, lithium-ion batteries have become important power sources for electrified transportation and stationary energy storage applications [1,2]. To support safe and reliable battery operation, battery management systems (BMSs) are required to monitor and manage battery states in real time at the cell, module, and pack levels [3,4]. Among various battery state indicators, the state of charge (SOC) is a key parameter that reflects the remaining available capacity and plays an essential role in energy management and operational safety [5]. However, SOC is an internal battery state and cannot be directly measured by sensors during practical

operation. Therefore, SOC is commonly inferred from measurable variables, including current, voltage, and temperature [6]. Accurate SOC estimation from these signals remains a fundamental challenge in battery management research [7].

In early SOC estimation studies, coulomb counting and the open-circuit voltage (OCV) method were the two representative approaches [8]. Coulomb counting updates SOC through time-domain accumulation of battery current, but the estimation result is sensitive to the initial SOC setting. In addition, current sensor bias and measurement noise may be accumulated together with the integration process, which gradually causes SOC drift [9–11]. The OCV method determines SOC based on the empirical relationship between OCV and SOC, usually implemented by tabulated data or polynomial fitting. However, reliable OCV acquisition requires the battery to stay in a sufficiently rested condition, so this method is difficult to apply directly to electric vehicles or energy storage systems with continuously changing load profiles [12,13]. These conventional approaches still have limited accuracy when used for SOC estimation under dynamic operating conditions [14].

Model-based SOC estimation methods have been widely investigated to overcome the limitations of conventional approaches. According to the battery modeling mechanism, they can generally be divided into electrochemical models (EMs) and equivalent circuit models (ECMs). EMs describe the internal battery behavior from electrochemical reactions and mass-transfer processes, but their high model complexity and computational cost make them difficult to implement in online BMS applications [15]. ECMs use circuit elements, such as resistors and capacitors, to reproduce the terminal voltage response of batteries and are therefore more suitable for online SOC estimation when combined with filtering algorithms [16]. For example, Yang et al. [17] applied an enhanced equivalent circuit model together with an extended Kalman filter (EKF) to online SOC estimation of lithium-ion batteries. Song et al. [18] proposed a weighted multi-innovation Sage-Husa AEKF method, where historical innovation information was assigned different weights to improve noise-covariance updating and filtering stability. However, the performance of ECM-based methods is closely related to the accuracy of model parameters, which may change with temperature, SOC, and battery aging, thereby affecting SOC estimation accuracy [19].

With the accumulation of battery operating data, data-oriented SOC estimators have also been widely investigated. Unlike model-based methods that explicitly describe battery dynamics, these approaches infer SOC from measured signals such as current, voltage, and temperature through a learned nonlinear relationship. For instance, Fan et al. [20] designed a CNN-based estimator with a U-Net structure for SOC prediction. Saneep et al. [21] combined RF with PSO, where PSO was adopted to tune the RF hyperparameters. Xia et al. [22] presented an improved LSSVM model and verified its robustness in SOC estimation. These studies indicate that machine-learning models can capture useful information from battery operating profiles. However, their estimation performance is still strongly related to the representativeness and scale of the training samples, which may become a limitation when the test conditions are different from the training data.

To lessen the reliance of data-driven methods on extensive experimental datasets, Chianese et al. [23] employed a validated battery simulation model to produce training data for neural networks (NNs) in SOC estimation, thereby reducing the number of required experiments. The resulting SOC prediction accuracy was comparable to that obtained using neural networks trained with standard experimental data. However, the OCV–SOC relationship shows pronounced nonlinearity at low SOC levels, which may lead to increased prediction errors in this region under certain conditions. Moreover, the black-box nature of some data-driven methods limits their interpretability, making it difficult to explain the sources of estimation errors from a mechanistic perspective [24].

To improve model interpretability and introduce physical constraints into data-driven estimation, hybrid SOC estimation methods have been investigated in recent studies. Li et al. [25] proposed an SVM-assisted AEKF method and obtained good prediction accuracy under dynamic operating conditions. Tang et al. [26] developed an NN-ECM fusion method to improve SOC estimation under different temperatures and operating profiles. Zhu et al. [27] proposed an EKF-CNN-BiLSTM-Attention method, where EKF was used to obtain the initial SOC estimate and CNN-BiLSTM-Attention was further adopted for SOC prediction. Guo et al. [28] incorporated neural networks into the ECM to fit the modeling residuals of the classical ECM, thereby improving SOC estimation performance under complex operating conditions. These studies show that model-based structures combined with data-driven correction can preserve physical constraints and provide nonlinear compensation capability. However, many existing hybrid methods are still strongly affected by the representativeness and quality of the training data, and some of them have mainly been evaluated under identical or similar driving cycles. Their generalization ability and estimation stability under unseen driving cycles still need further investigation.

In addition, some data-driven algorithms have relatively high computational complexity, which may impose a considerable computational burden in practical online BMS applications [29,30]. Therefore, improving generalization ability under unseen driving cycles while maintaining SOC estimation accuracy, computational efficiency, and model compactness remains an important challenge.

Although the above methods can achieve relatively accurate SOC estimation, they still face potential limitations in generalization under unseen driving cycles and model complexity. In contrast, LightGBM is an efficient gradient boosting decision tree model with good feature interpretability and relatively low computational complexity. It has been shown to outperform algorithms such as XGBoost and stochastic gradient boosting (SGB) in terms of computational speed and memory consumption, making it well suited for regression tasks such as SOC estimation [31]. Motivated by these challenges, this study develops a LightGBM-regulated and residual-compensated AEKF method for SOC estimation under unseen dynamic operating conditions. In this method, the second-order equivalent circuit model provides the physical basis for battery dynamics, and the AEKF performs recursive SOC estimation. LightGBM is then employed to predict covariance adjustment coefficients from operating features, model-error features, and state-parameter features, so that the noise covariance matrices can be adjusted online. In addition, LightGBM is used to correct the remaining SOC error after AEKF estimation, enabling the proposed method to combine adaptive filtering with data-driven nonlinear compensation.

Recent hybrid SOC estimation studies have also combined online parameter identification, adaptive filtering, and data-driven error correction. For example, Ge et al. proposed a BWO-AEKF-CNN-LSTM fusion framework, in which online-identified model parameters were used in AEKF to obtain an initial SOC estimate, and a CNN-LSTM network was further adopted to compensate for the SOC estimation error [32]. Different from this post-compensation-oriented deep learning structure, the proposed method introduces LightGBM into both the internal covariance regulation of AEKF and the external residual compensation stage. Specifically, LightGBM predicts the adjustment coefficients of the process and measurement noise covariance matrices, allowing Q_k and R_k to be adaptively adjusted according to operating features, model-error features, and state-parameter features. Another LightGBM model is then used to compensate for the remaining SOC error after AEKF estimation. The proposed framework does not simply replace the model-based estimator with a data-driven model, nor does it only perform post-processing error correction; instead, it combines physical-model-based recursive estimation, adaptive co-

variance regulation, and nonlinear residual compensation in a unified framework. The main contributions are listed below.

- (1) An AEKF–LightGBM framework is developed for SOC estimation under unseen driving cycles. Using a second-order equivalent circuit model together with an AEKF, the proposed framework uses LightGBM to regulate the filtering process and compensate for estimation residuals, thereby retaining physical model constraints while improving adaptability to complex dynamic operating conditions.
- (2) A LightGBM-based module for covariance adjustment coefficient prediction and residual error compensation is designed. By using features related to operating excitation, terminal voltage residuals, prior SOC, and model parameters, the proposed module predicts covariance adjustment coefficients to adaptively regulate Q_k and R_k . In addition, the residual SOC estimation error of the AEKF is compensated by LightGBM to further improve estimation accuracy and stability.
- (3) The proposed method is validated across different driving cycles using the CALCE INR 18650-20R dataset. The model is trained under the DST cycle at different temperatures and tested under the FUDS, US06, and BJDST cycles. Under the same conditions, the proposed framework is compared with XGBoost, SVR, RF, EKF, GPR, and LSTM in terms of estimation accuracy, online computation time, and model storage size.

The subsequent sections are arranged as follows. Section 2 presents the proposed AEKF–LightGBM fusion estimation method. Section 3 describes the dataset construction and partitioning strategy. Section 4 reports and discusses the SOC estimation results. Finally, Section 5 provides the conclusions.

2. Proposed AEKF–LightGBM Fusion Framework

2.1. Overall Architecture

To address increased model mismatch, the limited adaptability of filters with fixed covariance settings, and complex nonlinear errors in SOC estimation under unseen driving cycles, this study develops a fusion estimation framework by combining model-driven and data-driven approaches. The proposed method is based on a second-order equivalent circuit model and an AEKF. The physical model represents the battery dynamic response, while LightGBM adaptively regulates the filter covariance and further compensates for the residual error. This design improves SOC estimation accuracy under unseen driving cycles while maintaining computational efficiency and model compactness. The workflow of the proposed method is shown in Figure 1.

2.2. Construction of the Second-Order Equivalent Circuit Model

In this study, the battery voltage dynamics are modeled using a second-order ECM. This model offers a compromise between computational simplicity and representation accuracy, because it avoids the increased complexity of higher-order ECMs while still capturing the main polarization behavior of the battery. Structurally, it contains an OCV source, one ohmic resistor, and two parallel RC branches.

Figure 2 shows the structure of the second-order ECM. In this circuit, U_{OC} is the OCV source, R_0 is the ohmic resistor, and the two RC branches, (R_1, C_1) and (R_2, C_2) , describe the battery polarization dynamics at different time scales. The terminal voltage is calculated as

$$U_t = U_{OC} - U_{p1} - U_{p2} - IR_0, \quad (1)$$

where U_t is the terminal voltage, I is the current, and U_{p1} and U_{p2} are the polarization voltages corresponding to the two RC branches, respectively.

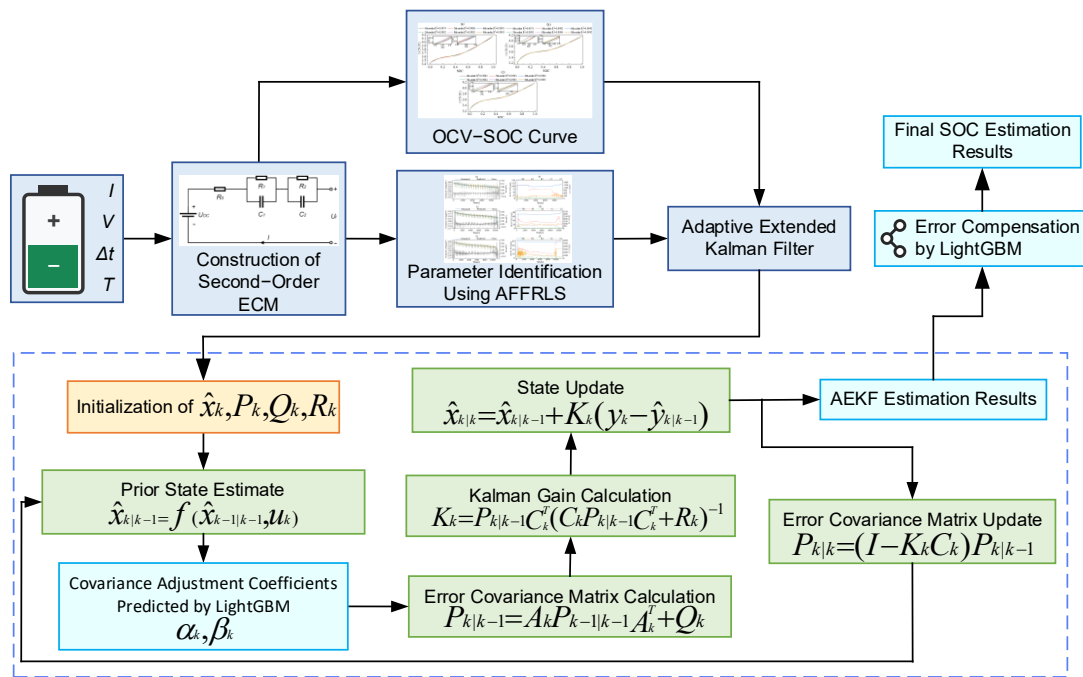


Figure 1. Overall framework of the proposed AEKF-LightGBM SOC estimation method.

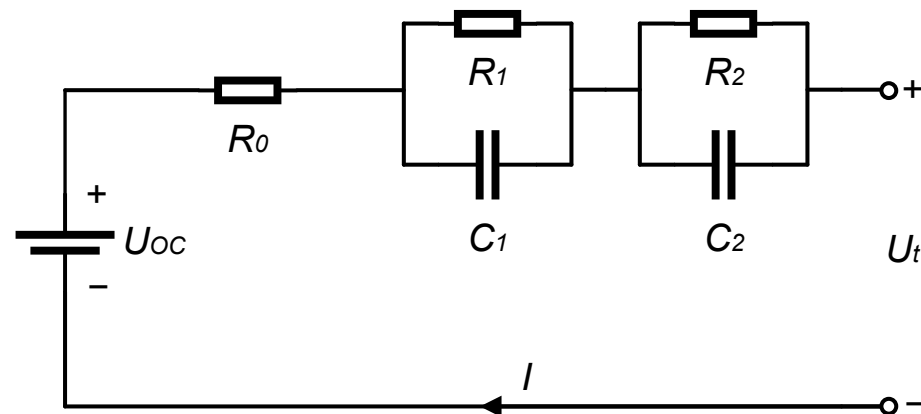


Figure 2. Second-order equivalent circuit model.

The model accuracy is mainly affected by five electrical parameters: R_0 , R_1 , R_2 , C_1 , and C_2 . These parameters are closely related to the ohmic and polarization characteristics of the battery and may vary with SOC, temperature, and battery chemistry. Zhang et al. systematically compared the temperature and SOC dependence of second-order RC parameters across LFP, NCM, and sodium-ion batteries, showing that parameter variations can affect the reliability of BMS modeling and algorithm migration [33]. Therefore, these parameters need to be identified before constructing the state-space model.

The OCV–SOC curve was first constructed from the experimental data, and a polynomial fitting method was then used to obtain the functional expression of OCV with respect to SOC. This function was used to construct the observation equation in the subsequent EKF process. Using the incremental-current OCV tests from the public CALCE battery dataset of the University of Maryland, 4th- to 9th-order polynomials were fitted to the OCV–SOC curves at 0 °C, 25 °C, and 45 °C, as shown in Figure 3. The fitting results show that the 5th-order polynomial achieves R^2 values greater than 0.999 at 25 °C and 45 °C and greater than 0.998 at 0 °C. Although higher-order polynomials can further increase R^2 , the improvement is limited while the model complexity increases substan-

tially. Therefore, the 5th-order polynomial is selected to represent the OCV–SOC curve in this paper, and the corresponding expressions are listed in Table 1.

To account for temperature-related parameter variations, AFFRLS-based parameter identification was conducted independently at 0 °C, 25 °C, and 45 °C. The DST data at each temperature were used as the identification data, from which R_0 , R_1 , R_2 , C_1 , and C_2 were obtained for the second-order ECM. Since online parameter identification is susceptible to measurement noise, physical plausibility constraints were imposed on the identified parameters. When an updated parameter value exceeded the predefined reasonable range at a given time step, the abnormal value was discarded and the valid estimate from the previous time step was retained, thereby preventing nonphysical parameters from being used in the subsequent EKF and LightGBM calculations. The parameter identification results are shown in Figure 4. In the subsequent SOC estimation under different driving cycles, the parameters identified online by AFFRLS for the corresponding driving cycle were used to construct the EKF state-space equations. It should be noted that the parameters identified from the test cycles in the following sections were used only for EKF model construction and were not involved in the training of the data-driven model.

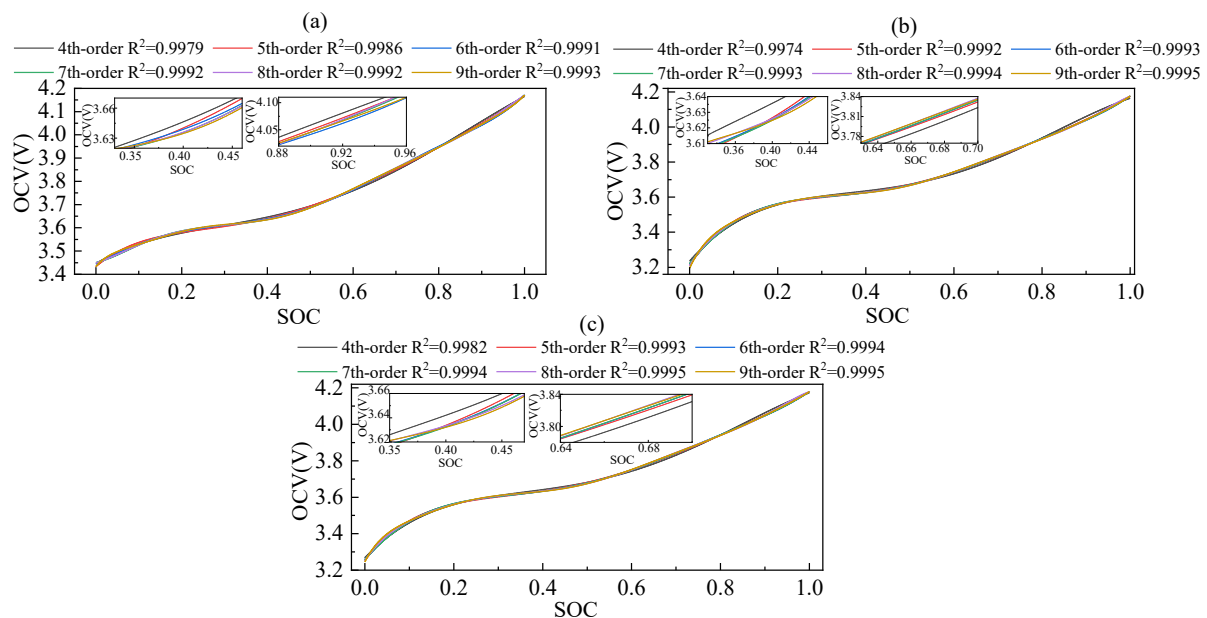


Figure 3. OCV–SOC curves fitted by 4th- to 9th-order polynomials at different temperatures: (a) 0 °C, (b) 25 °C, and (c) 45 °C.

Table 1. The 5th-order OCV–SOC polynomial expressions at different temperatures.

Temperature	5th-Order Polynomial
0 °C	$OCV = 3.9948 \cdot SOC^5 - 12.1362 \cdot SOC^4 + 13.7043 \cdot SOC^3 - 6.3756 \cdot SOC^2 + 1.5484 \cdot SOC + 3.4336$
25 °C	$OCV = 6.7774 \cdot SOC^5 - 21.5505 \cdot SOC^4 + 25.9825 \cdot SOC^3 - 13.8856 \cdot SOC^2 + 3.6383 \cdot SOC + 3.2122$
45 °C	$OCV = 5.3640 \cdot SOC^5 - 17.4299 \cdot SOC^4 + 21.4745 \cdot SOC^3 - 11.6294 \cdot SOC^2 + 3.1468 \cdot SOC + 3.2525$

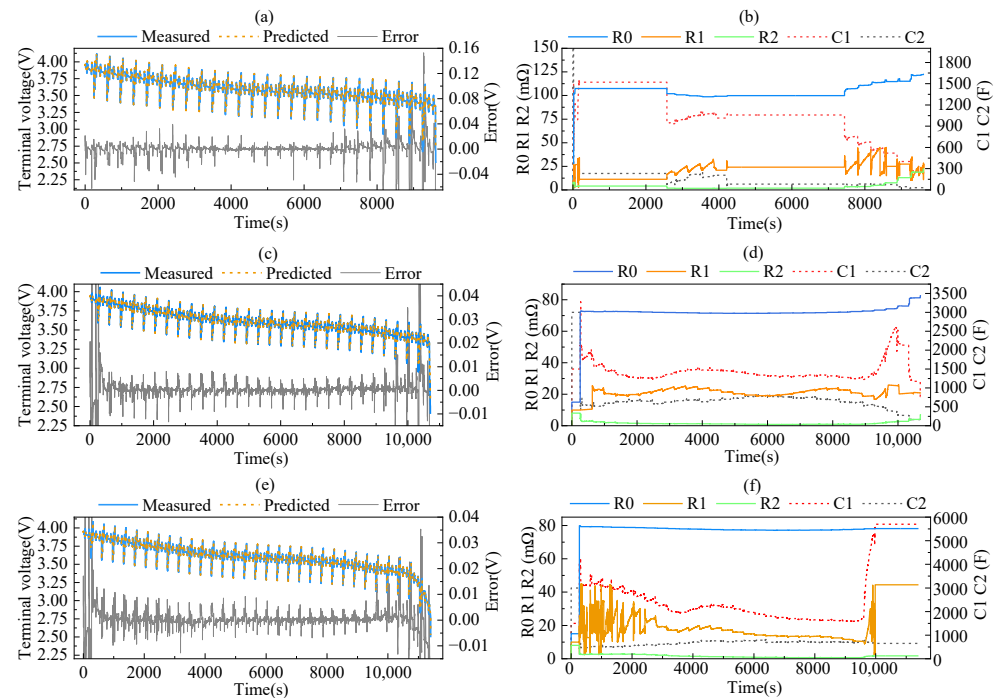


Figure 4. Predicted terminal voltage and AFFRLS-identified parameters of the second-order equivalent circuit model at different temperatures: (a,b) 0 °C, (c,d) 25 °C, and (e,f) 45 °C.

The identified ECM parameters show noticeable variations during the dynamic driving cycle. These variations are related to the changes in SOC, current excitation, temperature-dependent electrochemical behavior, and measurement noise, and therefore reflect the time-varying characteristics of the battery under dynamic operating conditions. In the second-order ECM, R_0 represents the ohmic resistance of the battery, including electrolyte resistance, electrode resistance, and contact resistance, and it mainly determines the instantaneous voltage drop caused by load current. R_1 and C_1 describe the fast polarization process, which is usually associated with electrochemical polarization and charge-transfer behavior. R_2 and C_2 describe the slow polarization process, which is mainly related to concentration polarization and diffusion effects. Therefore, the identified variations in these parameters have clear physical meanings rather than being only numerical fitting results.

Identification errors in different ECM parameters may influence SOC estimation through different paths. An error in R_0 directly changes the calculated ohmic voltage drop IR_0 , thereby affecting the terminal-voltage prediction and the voltage residual used for AEKF correction. Errors in R_1 and C_1 affect the fast polarization voltage response, which may lead to inaccurate voltage prediction during rapid current changes. Errors in R_2 and C_2 affect the slow polarization and relaxation behavior, which may introduce accumulated voltage prediction deviations during longer dynamic operation. These voltage prediction errors are further transmitted to the AEKF update process through the innovation term and Kalman gain, finally influencing the SOC estimation accuracy.

Previous studies have shown that AFFRLS can effectively improve the online identification accuracy of equivalent circuit model parameters under dynamic operating conditions, and AFFRLS combined with EKF has also been applied to lithium-ion battery SOC estimation [34,35]. In this study, the AFFRLS algorithm was implemented in Python 3.13.2 according to the recursive least-squares updating equations and the ECM parameter-conversion relationships described in this section and in previous AFFRLS-based studies. No external AFFRLS software package was used. In the implementation, the parameter vector of AFFRLS was initialized as $\theta_0 = 0$, and the covariance matrix was initialized as $P_0 =$

$10^6 I$. The adaptive forgetting factor was constrained by $\lambda_{\min} = 0.95$ and $\lambda_{\max} = 0.999$, with $\beta = 80.0$. The initial physical parameter values were set to $R_0 = 0.015 \Omega$, $R_1 = 0.010 \Omega$, $R_2 = 0.008 \Omega$, $C_1 = 1500F$, and $C_2 = 3000F$. When the updated parameters exceeded the predefined physical range or led to nonphysical values, the previous valid estimates were retained.

After the second-order ECM and parameter identification are completed, the state-space model for battery SOC estimation can be further established. In this study, SOC and the polarization voltages corresponding to the two RC branches are chosen as state variables, and the state vector is expressed as

$$x_k = [SOC_k, U_{p1,k}, U_{p2,k}]^T, \quad (2)$$

where the system input is the battery current I_k , and the observation output is the terminal voltage $U_{t,k}$.

The second-order equivalent circuit model can be discretized into the following state equations:

$$\begin{cases} SOC_{k+1} = SOC_k - \frac{\eta \Delta t}{Q_n} I_k, \\ U_{p1,k+1} = e^{-\Delta t/(R_1 C_1)} U_{p1,k} + R_1 (1 - e^{-\Delta t/(R_1 C_1)}) I_k, \\ U_{p2,k+1} = e^{-\Delta t/(R_2 C_2)} U_{p2,k} + R_2 (1 - e^{-\Delta t/(R_2 C_2)}) I_k. \end{cases} \quad (3)$$

The corresponding observation equation is given by:

$$U_{t,k} = U_{OC}(SOC_k) - U_{p1,k} - U_{p2,k} - I_k R_0, \quad (4)$$

where Δt denotes the sampling interval, η is the Coulombic efficiency, and Q_n is the nominal battery capacity.

2.3. LightGBM-Regulated Adaptive Extended Kalman Filter

After the second-order RC equivalent circuit model and its state-space equations are established in Section 2.2, the EKF can be further employed for recursive state estimation of the lithium-ion battery. The conventional EKF integrates model prediction and measurement update and can therefore achieve online SOC estimation to a certain extent. However, its filtering accuracy is highly affected by the settings of the process and measurement noise covariance matrices. Under unseen dynamic driving cycles, the degree of battery model mismatch, the excitation intensity of the driving cycle, and the measurement uncertainty usually vary with the operating state. Fixed covariance settings can hardly capture these variations adequately, thereby limiting the adaptability of the filter to complex operating conditions. To improve the robustness and state-tracking capability of the filter under unseen driving cycles, time-varying covariance matrices are introduced into the conventional EKF, forming the proposed improved AEKF.

In contrast to the conventional EKF with fixed Q and R , this study formulates the process and measurement noise covariance matrices as time-varying scaled versions of their baseline forms:

$$\begin{cases} Q_k = \alpha_k Q_{base}, \\ R_k = \beta_k R_{base}, \end{cases} \quad (5)$$

where Q_{base} and R_{base} are the baseline covariance matrices for process and measurement noise, respectively, and α_k and β_k are the corresponding adjustment coefficients at time step k . By introducing α_k and β_k , the filter can dynamically adjust the covariance matrices according to variations in model propagation uncertainty and measurement reliability under different driving cycles, thereby enhancing the adaptivity of the state estimation process. Specifically, α_k mainly reflects the uncertainty associated with model prediction under the current driving cycle, whereas β_k reflects the time-varying reliability of the observation information. When the operating condition changes sharply or the

model mismatch increases, a larger α_k reduces the filter's excessive dependence on model prediction. When the measurement noise increases or the observation information becomes less reliable, a larger β_k weakens the influence of the measurement update.

Based on the nonlinear state-space model established in Section 2.2 and the system state vector defined in Equation (2), the improved AEKF still follows the recursive prediction–update framework. In this framework, the battery current I_k is used as the system input, and the terminal voltage $U_{t,k}$ is used as the measurement output. Compared with the conventional EKF, the time-varying covariance matrices Q_k and R_k are introduced into the error covariance propagation and Kalman gain calculation to improve the adaptability of SOC estimation under unseen dynamic driving cycles.

In the prediction stage, the prior state estimate at the current time step is first calculated using the posterior state estimate from the previous time step and the current input:

$$\hat{x}_{k|k-1} = f(\hat{x}_{k-1|k-1}, u_k), \quad (6)$$

where $\hat{x}_{k|k-1}$ denotes the prior state estimate at time step k , $\hat{x}_{k-1|k-1}$ denotes the posterior state estimate at time step $k-1$, and u_k is the system input. To propagate the error covariance, the state equation is first-order linearized at the current operating point, yielding the state transition Jacobian matrix:

$$A_k = \frac{\partial f(x, u_k)}{\partial x} \Big|_{x=\hat{x}_{k-1|k-1}}. \quad (7)$$

Accordingly, the prior error covariance matrix can be updated as

$$P_{k|k-1} = A_k P_{k-1|k-1} A_k^T + Q_k, \quad (8)$$

where $P_{k|k-1}$ denotes the prior error covariance matrix at time step k , and $P_{k-1|k-1}$ denotes the posterior error covariance matrix at time step $k-1$. Compared with the conventional EKF, the time-varying Q_k introduced in this paper can more flexibly characterize the variation in model prediction uncertainty under different operating conditions.

After the state prediction is completed, the predicted output at the current time step is computed using the observation equation:

$$\hat{y}_{k|k-1} = h(\hat{x}_{k|k-1}, u_k), \quad (9)$$

where $\hat{y}_{k|k-1}$ denotes the predicted terminal voltage obtained from the prior state estimate. The observation equation is then linearized at the current prior state estimate, yielding the observation Jacobian matrix:

$$C_k = \frac{\partial h(x, u_k)}{\partial x} \Big|_{x=\hat{x}_{k|k-1}}. \quad (10)$$

The Kalman gain at time step k can then be expressed as

$$K_k = P_{k|k-1} C_k^T (C_k P_{k|k-1} C_k^T + R_k)^{-1}, \quad (11)$$

where K_k denotes the Kalman gain at time step k , and R_k denotes the measurement noise covariance matrix at time step k . Since R_k reflects the filter's assessment of measurement reliability, using a time-varying R_k enables a more appropriate measurement update under varying measurement noise levels and operating conditions.

In the update stage, the prior state estimate is corrected according to the voltage difference between the measured and predicted terminal voltages at the current time step. The innovation term is expressed as

$$v_k = y_k - \hat{y}_{k|k-1}, \quad (12)$$

where y_k denotes the measured terminal voltage at time step k . The posterior state estimate is then updated as

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + K_k (y_k - \hat{y}_{k|k-1}). \quad (13)$$

The corresponding posterior error covariance matrix is updated as

$$P_{k|k} = (I - K_k C_k) P_{k|k-1}, \quad (14)$$

where I is the identity matrix. Through the above recursive process, the filter integrates model prediction and measurement information at each time step to estimate the battery SOC and the polarization voltages of the two RC branches online. The first element of the updated state vector is taken as the base SOC estimate obtained by the improved AEKF at the current time step, denoted by $SOC_{AEKF,k}$.

Compared with the standard EKF, the key feature of the improved AEKF developed in this paper lies in the introduction of the covariance adjustment coefficients α_k and β_k . By extending the fixed process and measurement noise covariance matrices to time-varying forms, the filter can dynamically adjust the weighting between state propagation and measurement update according to variations in model mismatch and measurement reliability under different operating conditions. This improves the robustness and adaptability of SOC estimation under unseen dynamic driving cycles. It should be noted that α_k and β_k are not manually fixed in advance but are closely related to the current operating state, model mismatch features, and system dynamic conditions. Their construction and prediction processes are further described in Section 2.4.

2.4. Construction and Prediction of Covariance Adjustment Coefficients Based on LightGBM

In Section 2.3, the covariance matrices of process and measurement noise were expressed in time-varying forms by introducing the covariance adjustment coefficients α_k and β_k , namely $Q_k = \alpha_k Q_{\text{base}}$ and $R_k = \beta_k R_{\text{base}}$. This formulation enhances the adaptability of the improved AEKF to variations in model mismatch and measurement uncertainty under unseen dynamic driving cycles. However, α_k and β_k do not have directly measurable ground-truth values, and their values are closely related to the current operating condition, model prediction error, measurement residual, and battery dynamic parameters. If fixed coefficients are determined only by empirical settings, it is difficult to adequately capture the dynamic variation in the filter's confidence in model prediction and measurement information under different driving cycles. Therefore, this paper uses LightGBM to establish a covariance adjustment coefficient prediction model, which learns the mapping between current operating features and the optimal covariance adjustment strategy.

LightGBM is an efficient gradient boosting decision tree model with strong nonlinear mapping and feature interaction modeling capabilities. It is well suited for tabular feature data composed of current, voltage, model residuals, and identified parameters. In this study, the covariance adjustment coefficients and residual compensation values need to be predicted from multiple numerical features, including operating features, model-error features, and state-parameter features. This task requires a regression model that can capture nonlinear feature interactions while maintaining fast inference and compact storage for online BMS implementation. LightGBM was selected because it provides a suitable balance among nonlinear fitting capability, computational efficiency, and model compactness. Compared with MLP/ANN models, tree-based boosting models are generally more suitable for tabular regression tasks with limited experimental data and usually require less hyperparameter tuning. Compared with XGBoost, LightGBM uses histogram-based decision tree learning and leaf-wise tree growth with depth constraints, which can improve computational efficiency and reduce memory consumption while preserving nonlinear fitting capability. Compared with CatBoost, LightGBM is more straightforward for the numerical battery features used in this study, since categorical-feature handling is not a major requirement in the present SOC estimation task. LightGBM is suitable for covariance regulation and residual compensation in online BMS-oriented SOC estimation,

where estimation accuracy, inference speed, and model size need to be considered simultaneously. Therefore, LightGBM is used in this study as the regression model for covariance adjustment coefficient prediction and SOC residual error compensation. The workflow for constructing and predicting the covariance adjustment coefficients is shown in Figure 5.

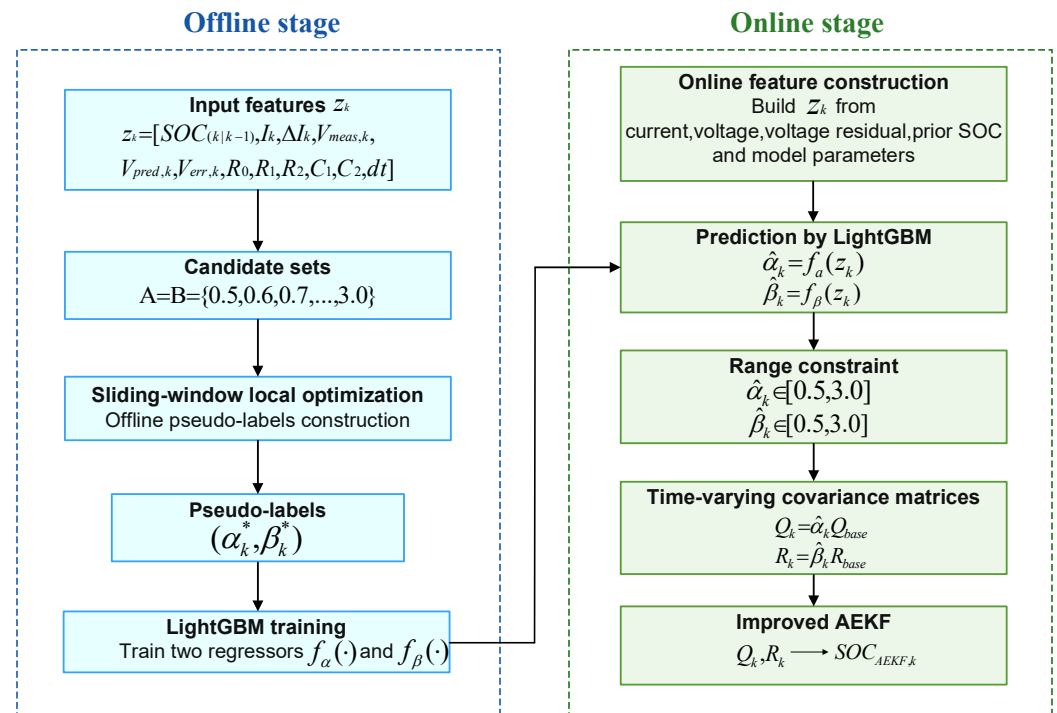


Figure 5. Workflow for constructing and predicting covariance adjustment coefficients using LightGBM. The superscript * denotes the locally optimized pseudo-label values obtained in the offline stage.

To enable LightGBM to effectively characterize variations in model uncertainty and measurement uncertainty under the current driving condition, an input feature vector z_k is constructed in this paper. This feature vector mainly consists of three types of information. The first type is operating-excitation features, including the current I_k , current variation ΔI_k , and sampling interval Δt , which are used to reflect the intensity of load variation. The second type is model-mismatch features, including the measured terminal voltage $V_{meas,k}$, predicted terminal voltage $V_{pred,k}$, and voltage prediction error $V_{err,k}$, where

$$\begin{cases} V_{meas,k} = U_{t,k}^{meas}, \\ V_{pred,k} = \hat{y}_{k|k-1}, \\ V_{err,k} = V_{meas,k} - V_{pred,k}. \end{cases} \quad (15)$$

These features reflect the deviation between the second-order ECM and the actual terminal voltage response. The third type is state and parameter features, including the prior SOC estimate $SOC_{k|k-1}$ and the identified parameters R_0 , R_1 , R_2 , C_1 , and C_2 , which characterize the current battery state and model dynamics. Accordingly, the feature vector constructed in this paper can be expressed as

$$z_k = [SOC_{k|k-1}, I_k, \Delta I_k, V_{meas,k}, V_{pred,k}, V_{err,k}, R_0, R_1, R_2, C_1, C_2, \Delta t], \quad (16)$$

where $\Delta I_k = I_k - I_{k-1}$. The above features describe the current operating state from the perspectives of operating excitation, model mismatch, and system dynamic parameters, thereby providing the input basis for predicting the covariance adjustment coefficients.

Although $V_{\text{pred},k}$ and $V_{\text{err},k}$ are mathematically related through $V_{\text{err},k} = V_{\text{meas},k} - V_{\text{pred},k}$, they provide different information for the covariance adjustment model. $V_{\text{pred},k}$ represents the terminal-voltage response predicted by the second-order ECM under the current state and identified parameters, reflecting the model-predicted operating point of the battery. In contrast, $V_{\text{err},k}$ directly characterizes the mismatch between the measured and predicted terminal voltages, which is closely related to model uncertainty and measurement residuals. Therefore, $V_{\text{pred},k}$ provides absolute model-response information, whereas $V_{\text{err},k}$ provides residual mismatch information. In addition, LightGBM is a tree-based model and is less sensitive to feature collinearity than linear regression models, because redundant features can be automatically selected or ignored according to their split gains during tree construction. Therefore, both features are retained to provide complementary information for covariance adjustment coefficient prediction.

Since α_k and β_k do not have directly available supervised labels, an offline local optimization method is adopted in this paper to construct pseudo-labels. Specifically, during the training stage, the training driving-cycle data are divided into several local segments using a sliding window. For each window, the optimal α and β are selected from predefined candidate coefficient sets by minimizing the SOC estimation error obtained by the improved AEKF within that window. In this paper, the candidate search ranges of both α and β are set to $[0.5, 3.0]$, with a search step of 0.1, namely

$$\mathcal{A} = \mathcal{B} = \{0.5, 0.6, 0.7, \dots, 3.0\}, \quad (17)$$

where $\mathcal{A}$ and $\mathcal{B}$ denote the candidate sets of α and β , respectively. Since this search process is performed only during the offline pseudo-label construction stage, it does not increase the computational burden of online SOC estimation. This process can be expressed as

$$(\alpha_k^*, \beta_k^*) = \arg \min_{\alpha \in \mathcal{A}, \beta \in \mathcal{B}} \sum_{i=k}^{k+L-1} (SOC_i^{\text{est}}(\alpha, \beta) - SOC_i^{\text{true}})^2, \quad (18)$$

where L is the sliding-window length, k denotes the starting sample point of the sliding window, and $\mathcal{A}$ and $\mathcal{B}$ are the candidate sets of α and β , respectively. $SOC_i^{\text{est}}(\alpha, \beta)$ denotes the SOC estimate at the i -th sample point obtained by the improved AEKF under the given α and β , and SOC_i^{true} denotes the corresponding reference SOC. Through the above offline local optimization process, the optimal covariance adjustment coefficients (α_k^*, β_k^*) corresponding to each window can be obtained and used as pseudo-labels for supervised learning in LightGBM.

After the pseudo-labels are obtained, two LightGBM regression models are established to predict the adjustment coefficients for process noise covariance and measurement noise covariance, respectively. These mapping relationships can be expressed as

$$\begin{cases} \hat{\alpha}_k = f_{\alpha}(z_k), \\ \hat{\beta}_k = f_{\beta}(z_k), \end{cases} \quad (19)$$

where $f_{\alpha}(\cdot)$ and $f_{\beta}(\cdot)$ denote the trained LightGBM regression models, respectively. During model training, the feature vector z_k is used as the input, and the offline-optimized α_k^* and β_k^* are used as labels, enabling LightGBM to learn the nonlinear mapping relationship between the current operating state and the optimal covariance adjustment strategy.

In the online estimation stage, the feature vector z_k is first constructed using the current, measured terminal voltage, predicted terminal voltage, voltage residual, prior SOC estimate, and identified model parameters at the current time step. The constructed feature vector is then input to the trained LightGBM models to obtain the predicted covariance adjustment coefficients $\hat{\alpha}_k$ and $\hat{\beta}_k$. To prevent abnormal predictions from destabilizing the filter, the predicted covariance adjustment coefficients are constrained within the same positive interval as the offline candidate sets:

$$\begin{cases} \hat{\alpha}_k \in [0.5, 3.0], \\ \hat{\beta}_k \in [0.5, 3.0]. \end{cases} \quad (20)$$

This range is consistent with the candidate sets used in the offline pseudo-label construction stage, which prevents LightGBM from outputting abnormal coefficients outside the training-label distribution during online prediction and thereby improves the stability of the covariance adjustment process. It should be noted that $\hat{\alpha}_k$ and $\hat{\beta}_k$ output by LightGBM as a regression model are continuous values and are not limited to the discrete values in the offline candidate sets. Therefore, they are only constrained within the same positive interval as the offline candidate sets, rather than being forcibly mapped to a specific discrete candidate value.

Subsequently, the time-varying Q and R matrices are constructed based on the predicted covariance adjustment coefficients:

$$\begin{cases} Q_k = \hat{\alpha}_k Q_{\text{base}}, \\ R_k = \hat{\beta}_k R_{\text{base}}. \end{cases} \quad (21)$$

The constructed Q_k and R_k are then substituted into the recursive process of the improved AEKF described in Section 2.3 to complete the state prediction, Kalman gain calculation, and state update at the current time step.

To improve the reproducibility of the proposed method, the pseudo-label construction process for the covariance adjustment coefficients is configured as follows. In the offline stage, a sliding window of length $L = 100$ is used to divide the DST training driving-cycle data into local segments, and the sliding step is set to $s = 1$. For the window $[k, k + L - 1]$, α and β are searched over the candidate sets $\mathcal{A} = \mathcal{B} = \{0.5, 0.6, \dots, 3.0\}$ with a search step of 0.1, and the optimal coefficient pair (α_k^*, β_k^*) is obtained by minimizing the sum of squared SOC estimation errors within this window. The selection of L affects the balance between pseudo-label stability and local dynamic adaptability. A smaller L may make the optimized pseudo-labels more sensitive to short-term current fluctuations, voltage noise, and transient model errors. For example, a relatively small window length, such as $L = 50$, may emphasize short-term local variations but may also lead to more fluctuating pseudo-labels. In contrast, an excessively large L , such as $L = 150$ or $L = 200$, may smooth the local dynamic characteristics of the driving cycle and weaken the response of the covariance adjustment coefficients to rapid operating changes. Therefore, $L = 100$ is selected as a compromise between local adaptability and pseudo-label stability. The coefficient interval $[0.5, 3.0]$ allows both downward and upward scaling of the base covariance matrices while avoiding excessively small or large covariance values that may destabilize the AEKF, and the step size of 0.1 provides a balance between pseudo-label resolution and offline grid-search cost. This optimal coefficient pair is used as the pseudo-label of the starting sample k of the window to train the two LightGBM regression models. The baseline process and measurement noise covariance matrices are set as $Q_{\text{base}} = \text{diag}([10^{-7}, 10^{-6}, 10^{-6}])$ and $R_{\text{base}} = [2 \times 10^{-4}]$, respectively. In the online estimation stage, $\hat{\alpha}_k$ and $\hat{\beta}_k$ are predicted by LightGBM, and $Q_k = \hat{\alpha}_k Q_{\text{base}}$ and $R_k = \hat{\beta}_k R_{\text{base}}$ are then constructed. The above pseudo-label construction is performed only under the DST training driving cycle; the FUDS, US06, and BJDST driving cycles are not involved in this process or in the training of the LightGBM models.

Through the above method, the covariance adjustment coefficients no longer rely on fixed empirical settings but are adaptively predicted by LightGBM according to the current operating features, model-mismatch features, and state-parameter features. This mechanism enables the filter to dynamically adjust the confidence assigned to model prediction and measurement information under complex dynamic driving cycles, thereby enhancing the robustness and generalization ability of SOC estimation under unseen driving cycles. Meanwhile, LightGBM has low computational complexity, which helps meet the lightweight requirement of the model while maintaining its adaptive capability and

provides a more stable base SOC estimate for the subsequent residual compensation module.

2.5. Residual Error Compensation Based on LightGBM

Although the improved AEKF described in Sections 2.3 and 2.4 can enhance the adaptability of the filter to variations in model mismatch and measurement uncertainty through time-varying covariance adjustment, the second-order ECM remains an approximate model. In addition, parameter identification errors, the nonlinear characteristics of the OCV–SOC relationship, and polarization effects under complex dynamic driving cycles may still affect the SOC estimation results. Therefore, the SOC estimates output by the AEKF may still contain certain systematic residuals. To further improve the final SOC estimation accuracy, LightGBM is further introduced on the basis of the improved AEKF to learn and compensate for the residual errors after AEKF estimation.

In the proposed framework, three LightGBM models are trained separately, namely the covariance adjustment coefficient prediction model for α_k , the covariance adjustment coefficient prediction model for β_k , and the residual compensation model. This separate-training strategy is adopted because the three models serve different roles and have different supervised learning targets. Specifically, the prediction of α_k is embedded in the AEKF recursion to regulate the process noise covariance matrix Q_k , thereby adjusting the confidence assigned to the state prediction process. The prediction of β_k is used to regulate the measurement noise covariance matrix R_k , thereby adjusting the confidence assigned to the terminal-voltage measurement update. In contrast, the residual compensation model is applied after the AEKF recursion to correct the remaining nonlinear SOC estimation error. Therefore, the α_k and β_k prediction models act inside the filtering process, whereas the residual compensation model acts after the filtering output. Separate training helps maintain the interpretability of the covariance regulation and residual correction modules, reduces the coupling among different learning objectives, and avoids introducing additional model complexity. Although a joint learning framework may further integrate covariance regulation and residual compensation, it may also increase the coupling among different learning objectives and make the individual contribution of each module less distinguishable. Therefore, separate training was adopted in this study to keep the roles of covariance regulation and residual compensation clear and interpretable.

The base SOC estimate obtained by the improved AEKF is denoted as $SOC_{AEKF,k}$, and the reference SOC is denoted as $SOC_{true,k}$. The residual SOC error at time step k can then be defined as

$$e_k = SOC_{true,k} - SOC_{AEKF,k}, \quad (22)$$

where e_k denotes the residual between the AEKF estimate and the reference SOC. The LightGBM error compensation model aims to capture the nonlinear relationship between the input features and this residual error, thereby obtaining the SOC residual compensation value for the current time step. The prediction process can be expressed as

$$\Delta SOC_k = f_{res}(z_k^{res}), \quad (23)$$

where $f_{res}(\cdot)$ denotes the trained LightGBM residual compensation model, z_k^{res} is the input feature vector in the error compensation stage, and ΔSOC_k is the SOC compensation value predicted by LightGBM.

To enable LightGBM to effectively characterize the relationship between the AEKF residual error and the current operating state, the input feature vector in the error compensation stage is constructed as

$$z_k^{\text{res}} = [\text{SOC}_{\text{AEKF},k}, I_k, \Delta I_k, V_{\text{meas},k}, V_{\text{pred},k}, V_{\text{err},k}, R_0, R_1, R_2, C_1, C_2, \Delta t], \quad (24)$$

where $\text{SOC}_{\text{AEKF},k}$ is the SOC estimate output by the improved AEKF; I_k and ΔI_k denote the battery current and current variation, respectively; $V_{\text{meas},k}$, $V_{\text{pred},k}$, and $V_{\text{err},k}$ denote the measured terminal voltage, predicted terminal voltage, and voltage residual, respectively; and R_0 , R_1 , R_2 , C_1 , and C_2 are the second-order ECM parameters identified under the corresponding driving cycle. These features capture the sources of residual error from the perspectives of the AEKF estimation state, operating-excitation intensity, terminal-voltage deviation, and model parameter characteristics. The same feature-retention strategy is also used in the residual compensation model. Although $V_{\text{pred},k}$ and $V_{\text{err},k}$ are correlated, $V_{\text{pred},k}$ describes the model-predicted voltage level, whereas $V_{\text{err},k}$ emphasizes the remaining voltage mismatch after model prediction. Retaining both features helps the residual compensation model distinguish between the absolute voltage operating point and the local model deviation.

In the offline training stage, z_k^{res} is used as the input, and e_k is used as the supervised label to train the LightGBM residual compensation model. Through its gradient boosting decision tree structure, the model learns the nonlinear relationship between the AEKF residual error and the input features. Compared with complex deep learning models, LightGBM has high training efficiency, fast inference speed, and low model complexity, making it suitable for residual correction in online SOC estimation.

In the online estimation stage, the SOC estimate at the current time step is first obtained by the improved AEKF. The feature vector z_k^{res} is then constructed using this base SOC estimate, the battery current, measured terminal voltage, predicted terminal voltage, voltage residual, and identified model parameters. Subsequently, z_k^{res} is input to the trained LightGBM residual compensation model to obtain the SOC compensation value ΔSOC_k . The final SOC estimate can be expressed as

$$\text{SOC}_{\text{final},k} = \text{SOC}_{\text{AEKF},k} + \Delta \text{SOC}_k, \quad (25)$$

where $\text{SOC}_{\text{final},k}$ denotes the final SOC estimate after LightGBM-based residual compensation. Through the above error compensation process, LightGBM further performs data-driven correction of complex nonlinear residual errors while retaining the physical model constraints and recursive estimation capability of the improved AEKF, thereby improving the accuracy and generalization ability of SOC estimation under unseen dynamic driving cycles.

3. Dataset and Data Preparation

This study uses the publicly available INR 18650-20R lithium-ion battery dataset provided by the Center for Advanced Life Cycle Engineering (CALCE), University of Maryland [36]. The INR 18650-20R cell was manufactured by Samsung SDI Co., Ltd. (Yongin-si, Republic of Korea). The battery has a rated capacity of 2000 mAh. This dataset contains incremental-current OCV tests and several typical dynamic driving-cycle tests. Because of its public availability and multiple types of experimental records, this dataset has been widely used in lithium-ion battery modeling and state estimation studies.

The data used in this paper include the incremental-current OCV tests and four dynamic driving-cycle tests, namely the Dynamic Stress Test (DST), Federal Urban Driving Schedule (FUDS), US06 driving cycle, and Beijing Dynamic Stress Test (BJDST). The incremental-current OCV test data were used in Section 2 to construct the OCV–SOC curves at different temperatures, whereas the dynamic driving-cycle data were used for training and testing the proposed AEKF–LightGBM method. Since the original dynamic driving-cycle data include charging, discharging, and rest processes before the formal dynamic segment to set the battery to an initial SOC of 80%, the raw data were preprocessed, and

only the data from the beginning to the end of the dynamic driving-cycle segment were retained for the experiments.

Regarding dataset usage and partitioning, for the SOC estimation experiments under different unseen dynamic driving cycles in Section 4.2.1, data at 0 °C, 25 °C, and 45 °C were selected for experimental analysis because temperature has a significant influence on battery model parameters and SOC estimation results. To examine whether the proposed method can generalize to unseen driving cycles, the complete DST driving cycle at each temperature was used as the training set for the LightGBM covariance adjustment coefficient prediction model and the LightGBM residual compensation model, while the FUDS, US06, and BJDST dynamic driving cycles were used as unseen test sets. Since the test driving-cycle data were not involved in the training of the LightGBM covariance adjustment model, the residual compensation model, or the pseudo-label construction process for the covariance adjustment coefficients and were used only for AFFRLS-AEKF online recursive estimation during testing, this partitioning strategy can more effectively evaluate the SOC estimation performance of the proposed method under unseen driving cycles. Compared with random sample partitioning, partitioning the data by driving cycle can reduce overfitting caused by overlapping data distributions and is more consistent with the research objective of SOC estimation under unseen driving cycles.

For the comparative analyses in Sections 4.2.2 and 4.2.3, the model training was performed using the 25 °C DST data, whereas the 25 °C US06 data were used for testing.

For the additional initial-SOC analysis in Section 4.2.4, the 25 °C data with an initial SOC of 50% were used, in which DST was used for training and FUDS, US06, and BJDST were used as unseen test cycles. For the robustness analysis in Section 4.2.5, the 25 °C US06 test data were used. Voltage noise, constant voltage sensor bias, and proportional perturbations of the identified ECM parameters were introduced separately, while the DST cycle at 25 °C was used as the training condition. For the additional cross-dataset validation in Section 4.2.6, the 25 °C data from the TU Berlin Lithium-Ion Battery Drive Cycle Dataset were used, where the FTP-72 cycle of cell NMC068 was used for training and the HWFET and CSHVC cycles of the same cell were used as unseen test cycles. For the ablation study in Section 4.2.7, the 25 °C US06 test condition was used to compare EKF, LightGBM-regulated AEKF, EKF with LightGBM residual compensation, and the complete proposed framework.

4. Results and Discussion

To evaluate the SOC estimation performance of the proposed AEKF-LightGBM method under unseen dynamic driving cycles, experiments were conducted from multiple aspects. First, SOC estimation under different unseen driving cycles and temperatures was examined. Second, different data-driven models were compared within the proposed fusion framework. Third, the proposed method was compared with typical SOC estimation methods. Fourth, experiments under a lower initial SOC condition were conducted. Fifth, robustness under voltage noise, voltage sensor bias, and ECM parameter uncertainty were evaluated. Sixth, an additional validation was performed on the TU Berlin dataset. Finally, an ablation study was conducted to quantify the contribution of covariance regulation and residual compensation.

4.1. Evaluation Metrics

Three error metrics are used to evaluate the proposed SOC estimation method: maximum error (ME), mean absolute error (MAE), and root mean square error (RMSE). The SOC estimation error at sample i is defined as

$$e_i = SOC_{est,i} - SOC_{true,i} \quad (26)$$

where $SOC_{est,i}$ and $SOC_{true,i}$ denote the estimated SOC and reference SOC at sample i , respectively. ME, MAE, and RMSE are then defined as

$$\begin{cases} ME = \max_{1 \leq i \leq N} |e_i|, \\ MAE = \frac{1}{N} \sum_{i=1}^N |e_i|, \\ RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N e_i^2}. \end{cases} \quad (27)$$

where N is the number of test samples. ME indicates the largest local deviation in SOC estimation and reflects the worst-case estimation error under dynamic operating conditions. MAE describes the average absolute estimation error over the whole test cycle and is used to evaluate the overall estimation accuracy. RMSE gives greater weight to large errors and is therefore more sensitive to error fluctuation and occasional large deviations. By jointly using ME, MAE, and RMSE, the SOC estimation performance can be evaluated from the perspectives of local maximum deviation, average estimation accuracy, and overall error stability.

4.2. Experimental Results

4.2.1. SOC Estimation Results Under Different Unseen Dynamic Driving Cycles

Following the data partitioning strategy in Section 3, the proposed AEKF-LightGBM method was tested under three temperatures: 0 °C, 25 °C, and 45 °C. For each temperature, the DST driving-cycle data served as the training set, whereas the FUDS, US06, and BJDST data were used as unseen test sets.

Figure 6 displays the SOC estimation and error curves of the proposed AEKF-LightGBM method under different temperatures and unseen dynamic driving cycles. Table 2 presents the corresponding ME, MAE, and RMSE metrics under each test condition.

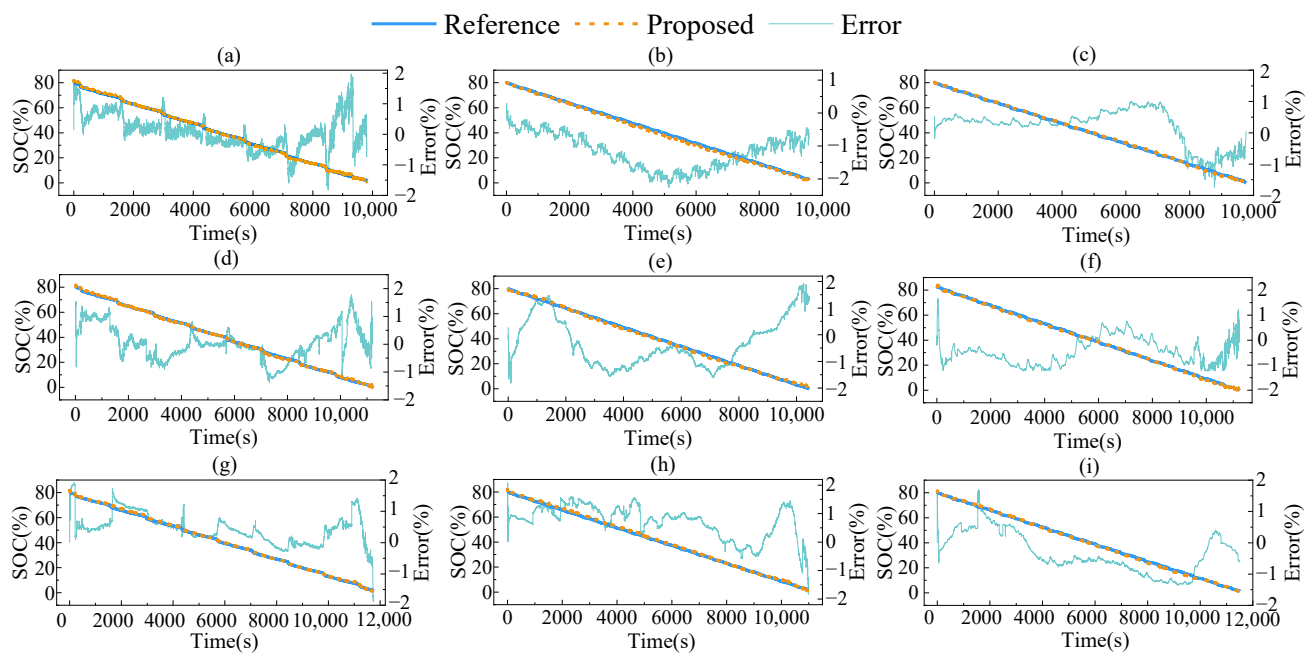


Figure 6. SOC estimation results of the proposed AEKF-LightGBM method under different temperatures and unseen driving cycles: (a–c) 0 °C under FUDS, US06, and BJDST; (d–f) 25 °C under FUDS, US06, and BJDST; and (g–i) 45 °C under FUDS, US06, and BJDST.

Table 2. Error metrics of the proposed AEKF–LightGBM method under different unseen driving cycles and temperatures.

Conditions		ME (%)	MAE (%)	RMSE (%)
FUDS	0 °C	1.97	0.43	0.56
	25 °C	1.78	0.44	0.58
	45 °C	1.90	0.52	0.65
US06	0 °C	2.26	1.17	1.28
	25 °C	1.91	0.83	0.94
	45 °C	2.08	0.82	0.92
BJDST	0 °C	1.75	0.54	0.60
	25 °C	1.65	0.59	0.69
	45 °C	1.69	0.63	0.72

4.2.2. Comparison of Different Data-Driven Models

To compare the effects of different data-driven models within the proposed fusion framework, Extreme Gradient Boosting (XGBoost), Support Vector Regression (SVR), Random Forest (RF), LightGBM were separately adopted as the data-driven module under the same second-order ECM, AEKF structure, and input features. For each data-driven model, it was used for both covariance adjustment coefficient prediction and AEKF residual error compensation. Specifically, the input features were first used to predict $\hat{\alpha}_k$ and $\hat{\beta}_k$, which were then used to construct the time-varying covariance matrices Q_k and R_k . The residual error of the AEKF output was subsequently compensated. For this comparison, the 25 °C DST data were used for model training, and the 25 °C US06 data were used for testing to assess the overall SOC estimation performance of different data-driven models under a complex unseen dynamic driving cycle.

Figure 7 shows the SOC estimation curves and error obtained using different data-driven models. Table 3 presents the corresponding ME, MAE, and RMSE metrics for each model.

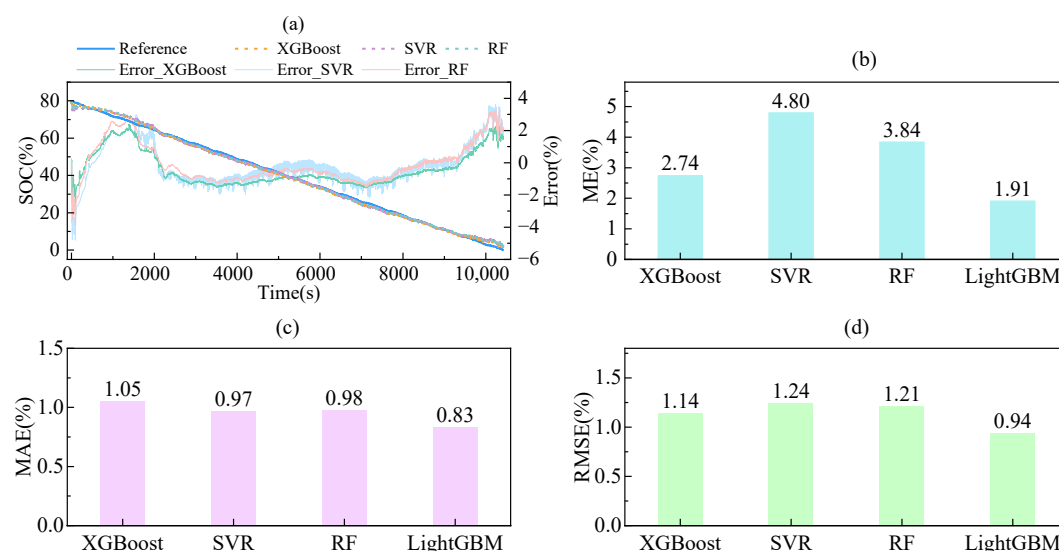
**Figure 7.** SOC estimation results of different data-driven models in the proposed framework under the unseen US06 driving cycle at 25 °C: (a) SOC estimation and error curves; (b) ME; (c) MAE; and (d) RMSE.

Table 3. Error metrics of different data-driven models in the proposed framework under the unseen US06 driving cycle at 25 °C.

Data-Driven Models	ME (%)	MAE (%)	RMSE (%)
XGBoost	2.74	1.05	1.14
SVR	4.80	0.97	1.24
RF	3.84	0.98	1.21
LightGBM	1.91	0.83	0.94

4.2.3. Comparison of Different SOC Estimation Methods

To further evaluate the SOC estimation capability of the proposed AEKF–LightGBM method, three comparison methods were selected, including the standalone Extended Kalman Filter (EKF), Gaussian Process Regression (GPR), and Long Short-Term Memory (LSTM). In this experiment, the 25 °C DST data served as the training dataset, while the 25 °C US06 data were used as the test dataset. Among the compared methods, EKF performs recursive estimation only with the second-order ECM and fixed covariance matrices. GPR and LSTM were trained and evaluated under the same data partitioning strategy.

The main parameters of LightGBM were set as follows: $n_estimators = 100$, $learning_rate = 0.1$, $max_depth = 6$, and $num_leaves = 31$, while the remaining parameters were kept at their default values. For the GPR model, the key parameters included the kernel function, α , and $n_restarts_optimizer$. The kernel was defined as $ConstantKernel \times (RationalQuadratic + RBF) + WhiteKernel$, with $\alpha = 1 \times 10^{-5}$ and $n_restarts_optimizer = 2$. For the LSTM model, the key parameters included the sequence length, hidden size, and number of LSTM layers, which were set to 100, 256, and 1, respectively.

Figure 8 shows the SOC estimation curves and error curves of different SOC estimation methods under the US06 driving cycle at 25 °C. Table 4 presents the corresponding ME, MAE, and RMSE metrics for each method.

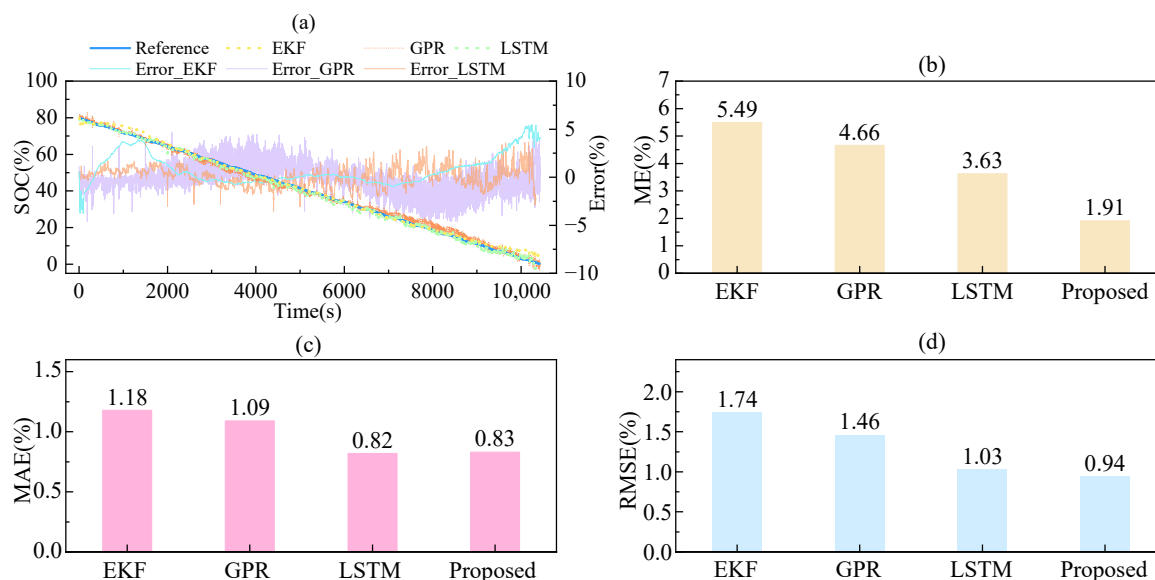
**Figure 8.** SOC estimation results of different SOC estimation methods under the unseen US06 driving cycle at 25 °C: (a) SOC estimation and error curves; (b) ME; (c) MAE; and (d) RMSE.

Table 4. Error metrics of different SOC estimation methods under the unseen US06 driving cycle at 25 °C.

Methods	ME (%)	MAE (%)	RMSE (%)
EKF	5.49	1.18	1.74
GPR	4.66	1.09	1.46
LSTM	3.63	0.82	1.03
Proposed	1.91	0.83	0.94

4.2.4. Comparison Under Different Initial SOC Conditions

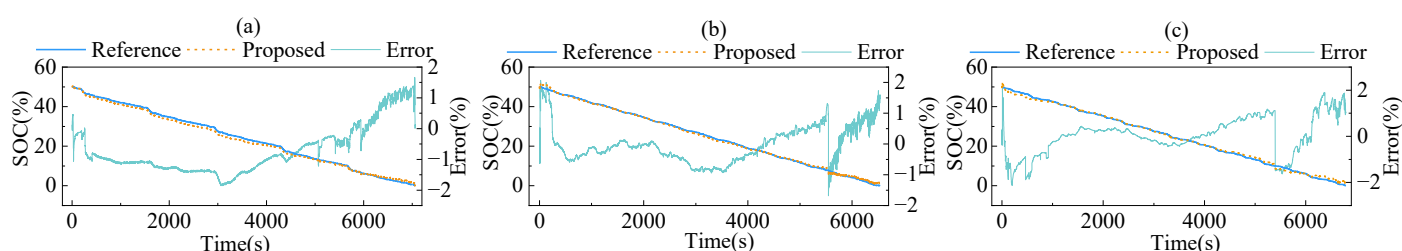
To examine the stability of the proposed method under different initial SOC conditions, additional experiments were carried out on the CALCE INR 18650-20R dataset with a lower initial SOC. In practical applications, the initial SOC of a battery may differ depending on the operating task and previous usage history. Therefore, it is necessary to verify whether the proposed method can still provide stable SOC estimation when the initial SOC changes.

In this section, the experiments were conducted at 25 °C using the data with an initial SOC of 50%. Following the same data partitioning strategy as the 80% initial SOC experiments, the DST driving cycle was used for model training, while the FUDS, US06, and BJDST driving cycles were used as unseen test cycles. The same feature construction, covariance adjustment coefficient prediction, and residual compensation processes described in Section 2 were adopted.

Figure 9 shows that the proposed method follows the reference SOC profiles under the FUDS, US06, and BJDST cycles when the initial SOC is set to 50%. The corresponding error results are summarized in Table 5. The ME, MAE, and RMSE values are all kept within reasonable limits for the three unseen cycles, which suggests that the proposed method retains stable SOC estimation capability under the lower initial SOC condition.

Table 5. Error metrics of the proposed method under the 50% initial SOC condition at 25 °C.

Conditions	ME (%)	MAE (%)	RMSE (%)
FUDS	1.86	0.95	1.05
US06	2.07	0.47	0.60
BJDST	2.13	0.61	0.79

**Figure 9.** SOC estimation results of the proposed method under the 50% initial SOC condition at 25 °C: (a) FUDS, (b) US06, and (c) BJDST.

4.2.5. Robustness Analysis Under Measurement Noise, Sensor Bias, and Parameter Uncertainty

To further examine the noise robustness of the proposed method, voltage-noise tests were conducted under the US06 driving cycle at 25 °C. In practical BMS applications, the measured terminal voltage may be affected by sensor noise, sampling errors, and external disturbances. Therefore, zero-mean Gaussian noise with standard deviations of 20 mV, 30 mV, and 40 mV was added to the measured terminal voltage of the US06 test data to sim-

ulate different levels of voltage measurement noise. The DST driving cycle at 25 °C was used as the training condition, and the trained LightGBM models were kept unchanged during the noise tests.

Figure 10 shows the SOC estimation curves and error curves of the proposed method under different voltage noise levels. It can be observed that, although the voltage noise causes additional fluctuations in the estimation error, the estimated SOC curves can still generally track the reference SOC trajectories under the three noise levels. As the voltage noise level increases from 20 mV to 40 mV, the fluctuation range of the SOC estimation error increases gradually, indicating that the measurement noise has a certain influence on the filtering update and residual compensation processes.

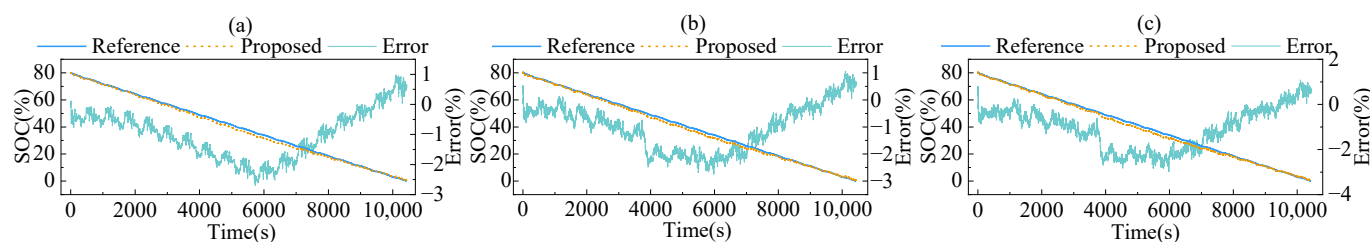


Figure 10. SOC estimation results of the proposed method under different voltage noise levels at 25 °C: (a) 20 mV, (b) 30 mV, and (c) 40 mV.

The corresponding error metrics are listed in Table 6. Under voltage noise levels of 20 mV, 30 mV, and 40 mV, the MAE values are 1.06%, 1.10%, and 1.13%, respectively, and the RMSE values are 1.26%, 1.32%, and 1.39%, respectively. The ME values increase from 2.69% to 2.98% as the noise level increases. Compared with the noise-free US06 result in Table 2, the estimation errors increase under noisy voltage measurements. However, even under the relatively large voltage noise level of 40 mV, the MAE and RMSE remain close to 1%, and the ME remains below 3%.

Table 6. Error metrics of the proposed method under different voltage noise levels.

Voltage Noise Level	ME (%)	MAE (%)	RMSE (%)
20 mV	2.69	1.06	1.26
30 mV	2.74	1.10	1.32
40 mV	2.98	1.13	1.39

In addition to random measurement noise, voltage sensors in practical BMS applications may also suffer from systematic bias. To evaluate the influence of voltage sensor bias, constant voltage offsets of +20 mV, +30 mV, −20 mV, and −30 mV were added to the measured terminal voltage of the 25 °C US06 test data. Different from zero-mean Gaussian noise, these offsets represent persistent measurement deviations during online SOC estimation. The corresponding SOC estimation and error curves are shown in Figure 11, and the quantitative results are listed in Table 7.

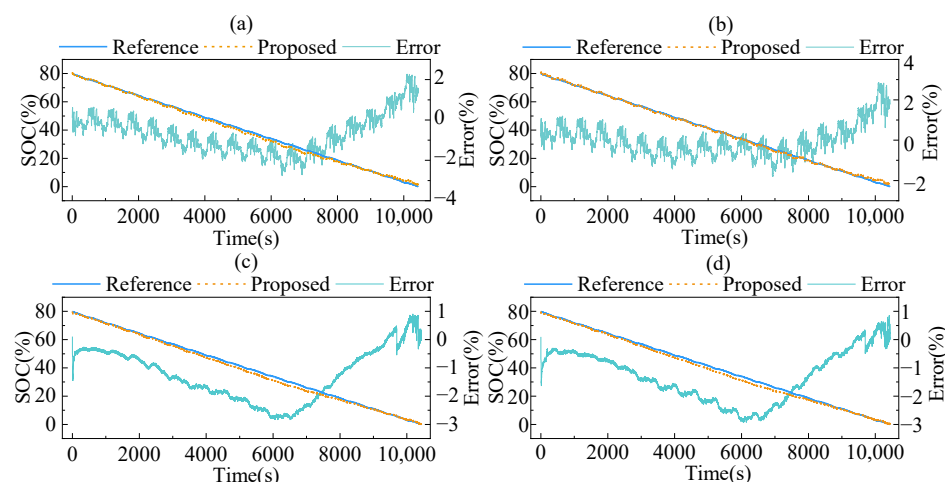


Figure 11. SOC estimation and error curves of the proposed method under different voltage sensor bias levels: (a) +20 mV, (b) +30 mV, (c) −20 mV, and (d) −30 mV.

Table 7. SOC estimation errors of the proposed method under different voltage sensor bias levels.

Voltage Sensor Bias Level	ME (%)	MAE (%)	RMSE (%)
+20 mV	2.76	0.97	1.15
+30 mV	2.84	1.08	1.21
−20 mV	2.83	1.29	1.54
−30 mV	2.93	1.32	1.56

As shown in Table 7, the ME values under different voltage bias levels remain below 2.93%, while the MAE and RMSE values remain within 1.32% and 1.56%, respectively. Compared with the positive bias cases, the negative voltage bias produces slightly larger MAE and RMSE values. Overall, the proposed method still maintains acceptable SOC estimation accuracy under the tested voltage sensor bias conditions, indicating its robustness against systematic voltage measurement deviations.

Furthermore, the influence of ECM parameter uncertainty was evaluated using the 25 °C US06 test data. Random proportional perturbations were added to the identified parameters R_0 , R_1 , R_2 , C_1 , and C_2 , with perturbation levels set to $\pm 5\%$, $\pm 10\%$, and $\pm 15\%$. During this test, the measured voltage signal and the trained LightGBM models were kept unchanged so that only the influence of parameter uncertainty was considered. The SOC estimation and error curves are shown in Figure 12, and the quantitative results are listed in Table 8.

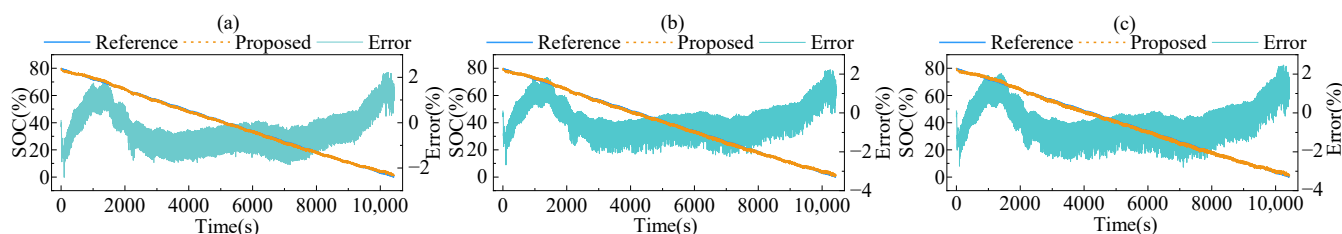


Figure 12. SOC estimation and error curves of the proposed method under different ECM parameter uncertainty levels: (a) $\pm 5\%$, (b) $\pm 10\%$, and (c) $\pm 15\%$.

Table 8. SOC estimation errors of the proposed method under different ECM parameter uncertainty levels.

Parameter Uncertainty Level	ME (%)	MAE (%)	RMSE (%)
±5%	2.41	0.97	1.14
±10%	2.67	1.06	1.23
±15%	2.79	1.24	1.37

As shown in Table 8, the ME increases from 2.41% to 2.79% as the perturbation level increases from ±5% to ±15%, while the MAE and RMSE remain within 1.37%. These results indicate that although parameter uncertainty increases the estimation error, the proposed method can still maintain acceptable SOC estimation accuracy under the tested parameter perturbation conditions.

4.2.6. Validation on a Different Public Dataset

To further evaluate the applicability of the proposed method on a dataset different from CALCE, an additional validation was conducted using the TU Berlin Lithium-Ion Battery Drive Cycle Dataset [37]. This dataset contains data from three brand-new LG INR21700 M50LT lithium-ion cells with a nominal capacity of 4.93 Ah. The cells were tested under 12 drive cycles and five ambient temperatures, including 45 °C, 35 °C, 25 °C, 15 °C, and 5 °C. In this subsection, the 25 °C data were selected for validation. The data in the 2_preprocessed folder were used, where the drive-cycle profiles had been resampled to a strict sampling frequency of 10 Hz, corresponding to a fixed sampling interval of 0.1 s. The used data columns include time, terminal voltage, current, cell surface temperature, and reference SOC calculated by coulomb counting.

According to the dataset description, the 12 drive cycles are assigned to three different cells rather than being measured from the same cell. To reduce the influence of cell-to-cell differences, the experiments in this subsection were conducted using the drive cycles from the same cell, NMC068. Specifically, the FTP-72 cycle at 25 °C was used as the training cycle, while the HWFET and CSHVC cycles at 25 °C were used as unseen test cycles. The HWFET cycle represents a highway driving condition with relatively smoother current variations, whereas the CSHVC cycle corresponds to a city-suburban heavy-vehicle driving condition with stronger dynamic fluctuations.

The SOC estimation and error curves are shown in Figure 13, and the corresponding quantitative results are listed in Table 9. Under the HWFET cycle, the proposed method achieves an ME of 1.96%, an MAE of 0.61%, and an RMSE of 0.75%. Under the CSHVC cycle, the ME, MAE, and RMSE are 2.43%, 0.91%, and 1.12%, respectively. Compared with HWFET, the CSHVC cycle produces slightly larger estimation errors, which is consistent with its stronger dynamic characteristics. The SOC estimation curves under both test cycles remain close to the reference SOC, indicating that the proposed method can still maintain good estimation accuracy when applied to the TU Berlin dataset.

Table 9. SOC estimation errors of the proposed method on the TU Berlin dataset under different unseen test cycles at 25 °C.

Test Cycle	ME (%)	MAE (%)	RMSE (%)
HWFET	1.96	0.61	0.75
CSHVC	2.43	0.91	1.12

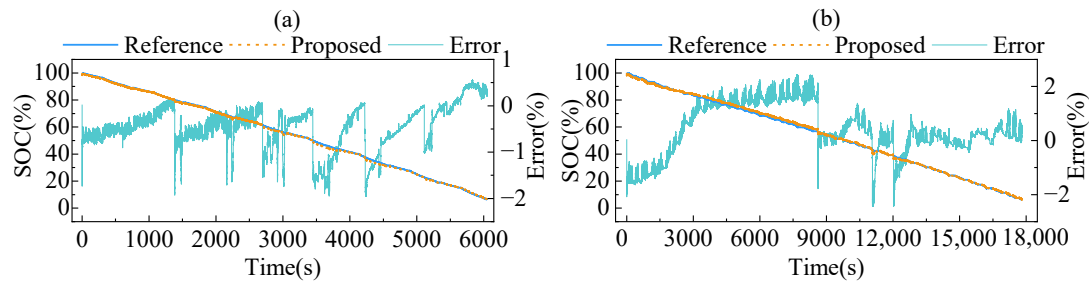


Figure 13. SOC estimation and error curves of the proposed method on the TU Berlin dataset under different unseen test cycles at 25 °C: (a) HWFET and (b) CSHVC.

4.2.7. Ablation Study of the Proposed Framework

To quantify the contribution of each module in the proposed framework, an ablation study was conducted under the 25 °C US06 test condition. Four variants were compared: EKF, LightGBM-regulated AEKF, EKF + LightGBM residual compensation, and the proposed method. The EKF uses fixed process and measurement noise covariance matrices without LightGBM-based covariance regulation or residual compensation. The LightGBM-regulated AEKF only introduces LightGBM-based covariance adjustment into the filtering process, while the residual compensation module is not used. The EKF + LightGBM residual compensation only applies LightGBM to compensate for the SOC residual after EKF estimation, while the covariance matrices remain fixed. The proposed method integrates both LightGBM-based covariance regulation and LightGBM-based residual compensation.

The SOC estimation and error curves are shown in Figure 14, and the quantitative results are listed in Table 10. Compared with the EKF baseline, the LightGBM-regulated AEKF reduces the estimation errors, indicating that covariance regulation affects the filter adaptation under the tested dynamic condition. The EKF + LightGBM residual compensation variant also reduces the errors compared with EKF, suggesting that residual compensation can correct part of the remaining SOC estimation deviation. When both modules are used together, the ME, MAE, and RMSE are further reduced under the 25 °C US06 condition. These results indicate that the two modules contribute differently to the final SOC estimation result.

Table 10. Ablation study results of different variants under the 25 °C US06 test condition.

Methods	ME (%)	MAE (%)	RMSE (%)
EKF	5.49	1.18	1.74
LightGBM-regulated AEKF	4.07	1.11	1.45
EKF + LightGBM residual compensation	2.27	1.08	1.23
Proposed	1.91	0.83	0.94

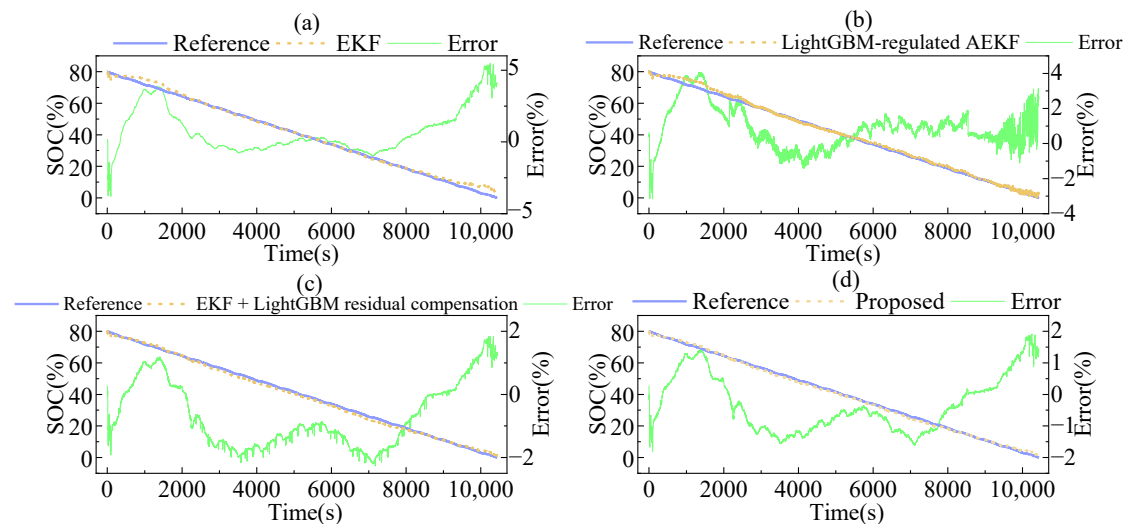


Figure 14. SOC estimation and error curves of different ablation variants under the 25 °C US06 test condition: (a) EKF, (b) LightGBM-regulated AEKF, (c) EKF + LightGBM residual compensation, and (d) proposed method.

4.3. Discussion

4.3.1. Analysis of Results

The results in Section 4.2 show that the proposed AEKF-LightGBM method provides good SOC estimation performance at different temperatures and under multiple unseen dynamic driving cycles. In the main unseen-cycle experiments in Section 4.2.1, the ME and RMSE of the proposed method do not exceed 2.26% and 1.28%, respectively, under all temperature and driving-cycle test conditions, indicating that the method maintains good cross-cycle generalization ability beyond the training driving cycle.

Across different unseen dynamic driving cycles, the overall SOC estimation errors remain relatively small for FUDS and BJDST, whereas larger errors appear under the US06 cycle. This is mainly because the US06 cycle involves more drastic current variations and more complex load fluctuations, which intensify the mismatch between the second-order ECM and the actual battery dynamic response, thereby increasing the difficulty of AEKF state estimation and LightGBM residual compensation. From the perspective of temperature effects, the estimation results are generally more stable at 25 °C, while the errors increase in some cases at 0 °C and 45 °C. This is related to enhanced polarization characteristics, increased model parameter fluctuations, and larger terminal-voltage prediction errors at 0 °C and 45 °C. In addition, local error fluctuations occur at the initial stage and near the end of some tests. The former is mainly associated with the incomplete convergence of the initial AEKF state, while the latter is related to the stronger nonlinearity of the OCV–SOC relationship in the low-SOC region and the accumulation of recursive estimation errors. Overall, the proposed method maintains low RMSE values under different temperatures and unseen driving cycles, indicating good cross-cycle estimation stability.

In the comparison of different data-driven models, LightGBM achieves the best performance in terms of ME, MAE, and RMSE, with values of 1.91%, 0.83%, and 0.94%, respectively, outperforming XGBoost, SVR, and RF. This indicates that, within the same fusion framework, using LightGBM as the data-driven module can achieve lower overall estimation error and smaller peak error. Compared with SVR and RF, LightGBM has stronger capability in modeling nonlinear relationships and feature interactions in tabular feature data. Compared with XGBoost, LightGBM shows higher training efficiency and good fitting capability when handling relatively complex feature combinations. Therefore, using LightGBM to predict $\hat{\alpha}_k$ and $\hat{\beta}_k$ and further compensate for the AEKF residual

error can more effectively improve SOC estimation performance under complex unseen driving cycles.

In the comparison of different SOC estimation methods, the RMSE values obtained by the standalone EKF, GPR, LSTM, and the proposed method under the 25 °C US06 driving cycle are 1.74%, 1.46%, 1.03%, and 0.94%, respectively, and the corresponding ME values are 5.49%, 4.66%, 3.63%, and 1.91%, respectively. These results show that the proposed AEKF-LightGBM method gives the lowest ME and RMSE, indicating its advantage in suppressing maximum local errors and reducing overall error fluctuations.

It should be noted that, under the US06 driving cycle, the MAE of LSTM is 0.82%, which is slightly lower than that of the proposed method at 0.83%. This indicates that the proposed method does not outperform LSTM in terms of average absolute error under this specific test condition. However, the difference is very small, and MAE mainly reflects the average error magnitude, while it does not fully characterize local peak errors or error fluctuation stability. Therefore, relying only on the MAE comparison under a single unseen driving cycle may not be sufficient to comprehensively evaluate the generalization performance of the two methods.

To further examine the comparison with LSTM under different unseen driving cycles, additional LSTM experiments were conducted under the FUDS and BJDST driving cycles at 25 °C using the same DST driving cycle as the training condition. The corresponding SOC estimation curves and error metrics are shown in Figure 15 and Table 11, respectively. The results show that the LSTM error level varies across different unseen driving cycles, indicating that its performance is affected not only by the intensity of current excitation, but also by the similarity between the training and testing temporal patterns. Although LSTM achieves a slightly lower MAE under US06, it produces higher MAE, ME, and RMSE values than the proposed method under both FUDS and BJDST when compared with the corresponding results in Table 2. In particular, the large ME under BJDST suggests that LSTM may still suffer from local error amplification under certain unseen dynamic segments. In contrast, the proposed method combines AEKF-based physical constraints, adaptive covariance regulation, and LightGBM-based residual compensation, which helps improve estimation stability and suppress local peak errors across different unseen driving cycles.

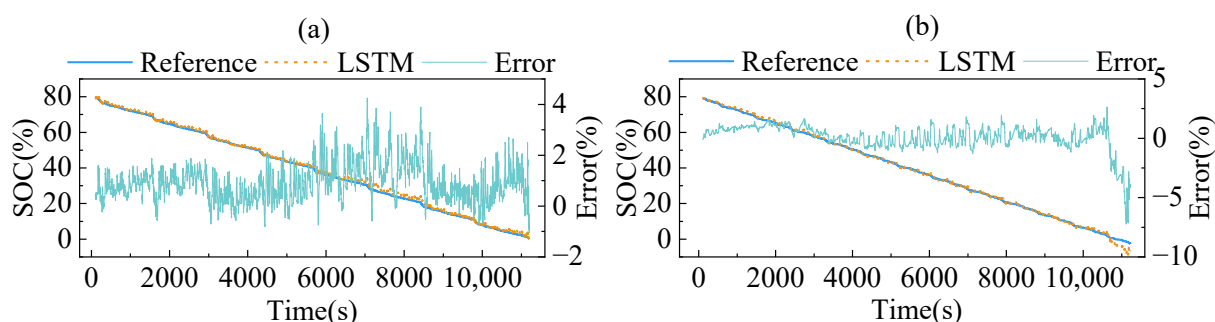


Figure 15. SOC estimation and error curves of LSTM under additional unseen driving cycles at 25 °C: (a) FUDS and (b) BJDST.

Table 11. Error metrics of LSTM under additional unseen driving cycles at 25 °C.

Conditions	ME (%)	MAE (%)	RMSE (%)
FUDS	4.25	0.95	1.17
BJDST	7.23	0.74	1.08

The standalone EKF has a clear physical model basis, but its fixed covariance settings make it difficult to adapt to variations in model mismatch and measurement uncertainty

caused by drastic load changes under the US06 cycle, resulting in relatively large errors. Although GPR and LSTM have strong nonlinear mapping capabilities, their generalization ability may still be limited by the coverage of the training driving cycle under the cross-cycle setting of DST training and US06 testing. In contrast, the proposed method is based on the physical recursive estimation of the AEKF and uses LightGBM for adaptive covariance adjustment and residual error compensation. This combination allows the proposed method to retain the physical constraints of model-driven estimation while exploiting the nonlinear correction capability of data-driven modeling, leading to more stable SOC estimation results under unseen dynamic driving cycles.

The results under the 50% initial SOC condition further indicate that the proposed method is not limited to the original 80% initial SOC setting. Changing the initial SOC condition shifts the SOC operating range and may affect the OCV-SOC relationship, the identified ECM parameters, and the residual error distribution. Therefore, some variation in the estimation error is expected when the operating SOC range changes. However, as shown in Figure 9 and Table 5, the proposed method can still maintain stable SOC tracking under FUDS, US06, and BJDST at the 50% initial SOC condition. This suggests that the combination of AFFRLS-based parameter identification, AEKF-based recursive estimation, LightGBM-based covariance adjustment, and residual compensation can reduce the dependence of the estimation process on a single initial SOC condition. Nevertheless, only two initial SOC levels, namely 80% and 50%, are considered in this study. Further validation under more initial SOC levels and wider SOC operating ranges is still needed.

The robustness analysis shows that measurement noise, voltage sensor bias, and ECM parameter uncertainty all affect SOC estimation to different degrees. Under these disturbances, the estimation errors increase compared with the undisturbed case, but no severe divergence appears in the SOC estimation curves. This suggests that the proposed framework has a certain tolerance to measurement and parameter uncertainties within the tested disturbance ranges.

The TU Berlin dataset validation further shows that the proposed method can be applied to a dataset different from CALCE. The different error levels between HWFET and CSHVC indicate that the estimation performance is still affected by the dynamic characteristics of the test cycle. In particular, stronger current fluctuations and more frequent transient changes tend to increase model mismatch and residual errors, which explains the slightly larger error observed under CSHVC.

The ablation study provides a more direct comparison of the two main modules in the proposed framework. Compared with the EKF baseline, the LightGBM-regulated AEKF reduces the estimation error, indicating that covariance regulation has a positive effect on filter adaptation under the tested condition. The EKF + LightGBM residual compensation variant also reduces the error, showing that residual compensation can correct part of the remaining SOC estimation deviation. When the two modules are used together, the error is further reduced under the 25 °C US06 condition.

Overall, the performance improvement of the proposed AEKF-LightGBM method is mainly related to two aspects. First, the covariance adjustment coefficients predicted by LightGBM allow the AEKF to update Q_k and R_k according to operating features, voltage residuals, prior SOC, and identified model parameters. This helps reduce the limitation of fixed-covariance settings when the driving cycle differs from the training condition. Second, the residual compensation module learns part of the remaining nonlinear error after AEKF estimation and provides an additional correction for the deviations caused by ECM approximation, parameter identification uncertainty, and polarization dynamics. Therefore, the reduction in ME, MAE, and RMSE observed in the experiments is consistent with the roles of covariance regulation and residual compensation in the proposed framework.

4.3.2. Computational Complexity and Embedded Implementation Feasibility

All experiments were conducted in a Python environment on a computer equipped with an Intel(R) i7-11800H @ 2.30 GHz CPU and an NVIDIA GeForce RTX 3060 GPU. The implementation of the proposed method was based on custom Python scripts developed by the authors in Python 3.13.2, using NumPy 2.3.4, Pandas 2.3.3, Matplotlib 3.10.7, LightGBM 4.6.0 and Scikit-learn 1.7.2. The AFFRLS and AEKF algorithms were coded by the authors according to the recursive equations and ECM parameter-conversion relationships described in Section 2, and no external AFFRLS or AEKF software package was used. In the online SOC estimation task under the unseen US06 driving cycle at 25 °C, a total of 10,354 samples were processed. Compared with GPR and LSTM, the proposed method achieved the shortest online SOC estimation time, approximately 6 s, whereas the online estimation times of GPR and LSTM were approximately 126 s and 91 s, respectively. After training, the total size of the three LightGBM models, $f_a(\cdot)$, $f_\beta(\cdot)$, and $f_{res}(\cdot)$, was approximately 663 kB, while the model sizes of GPR and LSTM were 35,617 kB and 1064 kB, respectively. This indicates that the proposed method maintains generalization ability while achieving lower complexity in terms of online estimation time and model storage size.

The real-time feasibility of the proposed method mainly depends on the recursive AEKF calculation and the online inference of the LightGBM models. For the AEKF part, the state vector contains SOC and two polarization voltages, resulting in a low-dimensional state-space model. Therefore, the online computation mainly includes state prediction, terminal-voltage prediction, Jacobian calculation, Kalman gain calculation, and covariance update with small-size matrices. This recursive structure has relatively low computational complexity and is suitable for real-time SOC estimation in BMS applications.

For the LightGBM part, the training process is completed offline, and only model inference is required during online SOC estimation. During inference, each decision tree only performs a sequence of threshold comparisons according to the current input features and then outputs the corresponding leaf value. Therefore, the online calculation of LightGBM mainly consists of comparison, branch judgment, and addition operations, without backpropagation or large-scale matrix multiplication. Compared with deep learning models such as LSTM or CNN-LSTM, this inference process is more lightweight and easier to implement on embedded controllers.

From the perspective of memory consumption, the main storage requirement of LightGBM comes from the tree structure, including feature indices, split thresholds, and leaf values. These parameters can be exported after offline training and stored in the Flash memory of the BMS controller. During online operation, only the current feature vector, AEKF state variables, covariance matrices, and temporary inference variables need to be stored in RAM. Therefore, the proposed method has practical potential for embedded BMS implementation. Nevertheless, further code-level optimization, fixed-point or mixed-precision implementation, and hardware-in-the-loop validation are still needed before deployment on resource-constrained BMS hardware platforms.

4.3.3. Limitations

Although the proposed method provides accurate SOC estimation under the tested conditions, some limitations remain. First, the present validation is mainly conducted using data from a healthy INR 18650-20R cell, and the influence of battery aging and SOH variation is not explicitly considered. During long-term operation, battery aging may lead to capacity degradation, increased internal resistance, changes in polarization characteristics, and variation in the OCV-SOC relationship. These changes may affect the AFFRLS-based parameter identification process, the AEKF recursive estimation, and the residual error distribution learned by LightGBM. Although the online identification of ECM pa-

rameters can update part of the model dynamics, the nominal capacity Q_n , OCV-SOC relationship, and residual compensation model may still need to be updated or recalibrated when significant aging occurs. Therefore, future work should further validate the proposed framework under different SOH conditions and incorporate SOH-related features or aging-aware model adaptation strategies to improve SOC estimation robustness over the battery lifetime.

Second, the main validation is based on the CALCE INR 18650-20R dataset; an additional validation using the TU Berlin LG INR21700 M50LT dataset was also included in this study. The current validation is still limited to a small number of public datasets, and the applicability of the proposed method to broader battery chemistries and operating conditions has not yet been systematically verified. Different lithium-ion battery chemistries, such as LFP, may exhibit different OCV-SOC characteristics, voltage plateaus, polarization behavior, temperature sensitivity, and aging mechanisms. These differences may influence the accuracy of the ECM, the stability of AFFRLS parameter identification, and the feature distribution learned by LightGBM. The current results cannot be directly generalized to other battery chemistries without further validation. When applying the proposed framework to other battery types, the OCV-SOC relationship, ECM parameters, covariance adjustment model, and residual compensation model should be recalibrated or retrained using data from the corresponding battery chemistry. Future work will further investigate the applicability of the proposed method to different lithium-ion battery chemistries, cell formats, and pack-level operating conditions.

It should also be noted that temperature is not used as a direct input feature of LightGBM in the proposed framework. This design is related to the experimental setting of the CALCE dataset used in this study. The tests at 0 °C, 25 °C, and 45 °C are conducted under discrete and approximately constant temperature conditions. Therefore, within each driving-cycle test, temperature acts mainly as a condition-level variable rather than a time-varying input signal. Directly adding such a nearly constant temperature value to the LightGBM feature vector may provide limited additional information for sample-wise covariance adjustment and residual compensation. In contrast, the OCV-SOC relationship and the ECM parameters identified under the corresponding temperature condition can reflect the temperature-dependent battery behavior more directly. Specifically, the parameters R_0 , R_1 , R_2 , C_1 , and C_2 are related to the ohmic resistance, polarization dynamics, and terminal-voltage response of the battery, and they vary with the operating state and identified model dynamics under each temperature condition. Therefore, these parameters provide more informative temperature-related representations than a constant temperature label in the present discrete-temperature experimental setting. As shown in Table 2, the proposed method maintains acceptable estimation accuracy under both 0 °C and 45 °C across the tested unseen driving cycles, suggesting that the temperature-related information contained in the OCV-SOC relationship and identified ECM parameters can support the LightGBM-based covariance adjustment and residual compensation processes under the three discrete temperature conditions considered in this study. Nevertheless, this conclusion should not be extended to all thermal scenarios. Under continuous temperature variations, rapid thermal transients, or wider temperature ranges, temperature may become a time-varying state that directly affects battery dynamics, and introducing temperature or thermal-state-related features may be necessary in future work.

In this study, the SOC estimation performance was mainly evaluated using ME, MAE, and RMSE, which describe the maximum local deviation, average estimation error, and overall error fluctuation, respectively. These metrics provide a direct and commonly used quantitative basis for comparing different SOC estimation methods under the tested operating conditions. In future work, confidence intervals and repeated-run statistical anal-

ysis can be further introduced to provide a more comprehensive evaluation of the statistical reliability and stability of the proposed method.

5. Conclusions

This study focuses on SOC estimation for lithium-ion batteries under unseen dynamic driving cycles, where model mismatch, limited covariance adaptability, and insufficient data-driven generalization may degrade estimation accuracy. To address these issues, an AEKF-LightGBM method is developed. In this method, a second-order equivalent circuit model and AEKF are used to provide the model-based estimation framework, while LightGBM predicts covariance adjustment coefficients to adaptively tune the process and measurement noise covariance matrices. In addition, LightGBM compensates for the remaining AEKF estimation errors, thereby improving SOC estimation accuracy and robustness under unseen dynamic driving cycles.

The experimental results demonstrate that the proposed AEKF-LightGBM method follows the reference SOC well under the unseen FUDS, US06, and BJDST cycles at 0 °C, 25 °C, and 45 °C, with ME and RMSE remaining below 2.26% and 1.28%, respectively. For the online estimation task using 10,354 samples from the unseen US06 cycle at 25 °C, the proposed method obtains lower ME and RMSE than EKF, GPR, and LSTM. Its total online estimation time is about 6 s, equivalent to an average estimation time of about 0.58 ms for each sample, and the combined size of the three LightGBM models is about 663 kB. These results verify that the proposed method achieves accurate SOC estimation, efficient online computation, and lightweight model storage. Additional validation on the TU Berlin dataset and the ablation study further show that the method can be applied to a public dataset different from CALCE and that both covariance regulation and residual compensation contribute to the reduction in SOC estimation errors under the tested conditions. However, this study still has several limitations. The proposed method is mainly validated using data from a healthy INR 18650-20R cell, and the effects of battery aging and SOH variation are not fully considered. Although voltage-noise tests were conducted under the US06 cycle at 25 °C, the added Gaussian noise is still a simplified representation of real BMS measurement disturbances. The proposed method has not yet been validated on other battery chemistries, such as LFP. In future work, aging-aware parameter updating, SOH-related feature incorporation, validation under real-world measurement disturbances, and cross-chemistry testing will be further investigated to improve the reliability and generalization ability of the proposed method in practical BMS applications.

Author Contributions: Conceptualization, M.G.; methodology, M.G. and L.C.; software, M.G.; validation, M.G.; formal analysis, M.G. and L.C.; investigation, M.G.; resources, X.Z.; data curation, X.Z.; writing—original draft preparation, M.G.; writing—review and editing, G.F., L.C. and N.D.; visualization, M.G.; supervision, G.F., L.C. and N.D.; project administration, G.F., L.C. and N.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Chongqing Natural Science Foundation Innovation and Development Joint Fund, grant numbers CSTB2025NSCQ-LZX0152 and CSTB2025NSCQ-WZQLHJMSX0009.

Data Availability Statement: The CALCE battery dataset used in the main experiments is publicly available from the CALCE Battery Research Group at the University of Maryland: <https://calce.umd.edu/battery-data> (accessed on 15 May 2026). The TU Berlin Lithium-Ion Battery Drive Cycle Dataset used for the additional validation is publicly available from the Technische Universität Berlin repository: <https://depositonce.tu-berlin.de/items/7f68932b-4d43-4f49-a5d8-914b00039f87> (accessed on 25 June 2026), <https://doi.org/10.14279/depositonce-21133>.

Conflicts of Interest: Xiaojang Zou was employed by Chongqing Hangying Automobile Manufacturing Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Zhang, Y.; Lai, X.; Zhang, X.; Fan, Y.; Cheng, E.; Zheng, Y.; Tang, X.; Tang, B.; Zhu, Z. Fusion of Stress and Electrical Signals for High-Accuracy Joint Estimation of SOC and SOH in Lithium-Ion Batteries. *Energy* **2025**, *331*, 137063. <https://doi.org/10.1016/j.energy.2025.137063>.
2. Xiong, R.; Li, Z.; Li, H.; Wang, J.; Liu, G. A Novel Method for State of Charge Estimation of Lithium-Ion Batteries at Low-Temperatures. *Appl. Energy* **2025**, *377*, 124514. <https://doi.org/10.1016/j.apenergy.2024.124514>.
3. Xu, Q.; Wang, X.; Ye, H.; Guo, J.; Gong, L.; Tan, P.; Pan, T. An Expansion Force-Based Method for Accurate State of Charge Estimation across the Entire Lifecycle of Lithium-Ion Batteries. *J. Energy Storage* **2025**, *137*, 118705. <https://doi.org/10.1016/j.est.2025.118705>.
4. Wang, C.; Yang, M.; Wang, X.; Xiong, Z.; Qian, F.; Deng, C.; Yu, C.; Zhang, Z.; Guo, X. A Review of Battery SOC Estimation Based on Equivalent Circuit Models. *J. Energy Storage* **2025**, *110*, 115346. <https://doi.org/10.1016/j.est.2025.115346>.
5. Kuang, P.; Zhou, F.; Xu, S.; Li, K.; Xu, X. State-of-Charge Estimation Hybrid Method for Lithium-Ion Batteries Using BiGRU and AM Co-Modified Seq2Seq Network and H-Infinity Filter. *Energy* **2024**, *300*, 131602. <https://doi.org/10.1016/j.energy.2024.131602>.
6. Çimen, H.; Uçar, K.; Arabacı, H. Enhancing Generalization Performance of CNN-Based State-of-Charge Estimation for Lithium-Ion Batteries. *J. Energy Storage* **2025**, *136*, 118597. <https://doi.org/10.1016/j.est.2025.118597>.
7. Zou, M.; Wang, J.; Yan, D.; Li, Y.; Tang, X. Novel State of Charge Estimation Method of Containerized Lithium-Ion Battery Energy Storage System Based on Deep Learning. *J. Power Sources* **2024**, *624*, 235609. <https://doi.org/10.1016/j.jpowsour.2024.235609>.
8. How, D.N.; Hannan, M.; Lipu, M.H.; Ker, P.J. State of Charge Estimation for Lithium-Ion Batteries Using Model-Based and Data-Driven Methods: A Review. *IEEE Access* **2019**, *7*, 136116–136136. <https://doi.org/10.1109/ACCESS.2019.2942213>.
9. Mohammadi, F. Lithium-Ion Battery State-of-Charge Estimation Based on an Improved Coulomb-Counting Algorithm and Uncertainty Evaluation. *J. Energy Storage* **2022**, *48*, 104061. <https://doi.org/10.1016/j.est.2022.104061>.
10. Zhu, Y.; Xiong, Y.; Xiao, J.; Yi, T.; Li, C.; Sun, Y. An Improved Coulomb Counting Method Based on Non-Destructive Charge and Discharge Differentiation for the SOC Estimation of NCM Lithium-Ion Battery. *J. Energy Storage* **2023**, *73*, 108917. <https://doi.org/10.1016/j.est.2023.108917>.
11. Movassagh, K.; Raihan, A.; Balasingam, B.; Pattipati, K. A Critical Look at Coulomb Counting Approach for State of Charge Estimation in Batteries. *Energies* **2021**, *14*, 4074. <https://doi.org/10.3390/en14144074>.
12. Lin, C.; Yu, Q.; Xiong, R.; Wang, L.Y. A Study on the Impact of Open Circuit Voltage Tests on State of Charge Estimation for Lithium-Ion Batteries. *Appl. Energy* **2017**, *205*, 892–902. <https://doi.org/10.1016/j.apenergy.2017.08.124>.
13. Pattipati, B.; Balasingam, B.; Avvari, G.V.; Pattipati, K.R.; Bar-Shalom, Y. Open Circuit Voltage Characterization of Lithium-Ion Batteries. *J. Power Sources* **2014**, *269*, 317–333. <https://doi.org/10.1016/j.jpowsour.2014.06.152>.
14. Fan, Y.; Huang, Z.; Li, H.; Kaleem, M.B.; Zhang, R.; Liu, W. State of Charge Estimation for Lithium-Ion Batteries with Enhanced Open-Circuit Voltage Model. *Measurement* **2025**, *251*, 117124. <https://doi.org/10.1016/j.measurement.2025.117124>.
15. Guo, F.; Couto, L.D.; Mulder, G.; Trad, K.; Hu, G.; Capron, O.; Haghverdi, K. A Systematic Review of Electrochemical Model-Based Lithium-Ion Battery State Estimation in Battery Management Systems. *J. Energy Storage* **2024**, *101*, 113850. <https://doi.org/10.1016/j.est.2024.113850>.
16. Koseoglou, M.; Tsioumas, E.; Panagiotidis, I.; Papagiannis, D.; Jabbour, N.; Mademlis, C. A Lithium-Ion Battery Equivalent Circuit Model Based on a Hybrid Parametrization Approach. *J. Energy Storage* **2023**, *73*, 109051. <https://doi.org/10.1016/j.est.2023.109051>.
17. Yang, L.; He, M.; Ren, Y.; Gao, B.; Qi, H. State of Charge Estimation for Lithium-Ion Batteries via a Diffusion-Enhanced Equivalent Circuit Model and an Improved Extended Kalman Filter. *Energy* **2026**, *345*, 140181. <https://doi.org/10.1016/j.energy.2026.140181>.
18. Song, W.; Liu, R.; Jin, X.; Guo, W. SOC Estimation for Lithium-Ion Batteries Based on Weighted Multi-Innovation Sage–Husa Adaptive EKF. *Energies* **2025**, *18*, 4364. <https://doi.org/10.3390/en18164364>.

19. Shu, X.; Li, Y.; Wei, K.; Yang, W.; Yang, B.; Zhang, M. Research on the Output Characteristics and SOC Estimation Method of Lithium-Ion Batteries over a Wide Range of Operating Temperature Conditions. *Energy* **2025**, *317*, 134726. <https://doi.org/10.1016/j.energy.2025.134726>.
20. Fan, X.; Zhang, W.; Zhang, C.; Chen, A.; An, F. SOC Estimation of Li-Ion Battery Using Convolutional Neural Network with U-Net Architecture. *Energy* **2022**, *256*, 124612. <https://doi.org/10.1016/j.energy.2022.124612>.
21. Saneep, K.; Sundareswaran, K.; Nayak, P.S.R.; Puthusserry, G.V. State of Charge Estimation of Lithium-Ion Batteries Using PSO Optimized Random Forest Algorithm and Performance Analysis. *J. Energy Storage* **2025**, *114*, 115879. <https://doi.org/10.1016/j.est.2025.115879>.
22. Xia, B.; Ye, M.; Wang, Q.; Lian, G.; Li, Y.; Zhang, B.; Zhang, Y. Robustness Estimation for State-of-Charge of a Lithium-Ion Battery Based on Feature Fusion. *Energy Rep.* **2024**, *12*, 4016–4025. <https://doi.org/10.1016/j.egy.2024.09.072>.
23. Chianese, G.; Iannucci, L.; Veneri, O.; Capasso, C. Real-Time Estimation of Battery SoC through Neural Networks Trained with Model-Based Datasets: Experimental Implementation and Performance Comparison. *Appl. Energy* **2025**, *389*, 125783. <https://doi.org/10.1016/j.apenergy.2025.125783>.
24. Isa, A.A.H.; Sulthan, S.M.; Dani, M.N.; Tan, S.J. Enhanced Lithium-Ion Battery State-of-Charge Estimation via Akima–Savitzky–Golay OCV-SOC Mapping Reconstruction and Bayesian-Optimized Adaptive Extended Kalman Filter. *Energies* **2025**, *18*, 6192. <https://doi.org/10.3390/en18236192>.
25. Li, S.; Yang, S.; Jiang, Z.; Jiang, W.; Zhu, Z.; Ma, Y. An Online State of Charge Estimation Method: Support Vector Machine Battery Model Fusing a Filtering Algorithm. *Electr. Power Syst. Res.* **2026**, *253*, 112530. <https://doi.org/10.1016/j.epsr.2025.112530>.
26. Tang, A.; Huang, Y.; Liu, S.; Yu, Q.; Shen, W.; Xiong, R. A Novel Lithium-Ion Battery State of Charge Estimation Method Based on the Fusion of Neural Network and Equivalent Circuit Models. *Appl. Energy* **2023**, *348*, 121578. <https://doi.org/10.1016/j.apenergy.2023.121578>.
27. Zhu, M.; Wang, Y.; Wang, D. SOC Prediction of Li-Ion Battery Based on EKF and CNN-BiLSTM-Attention. *ACS Omega* **2025**, *10*, 57157–57168. <https://doi.org/10.1021/acsomega.5c06451>.
28. Guo, Z.; Li, Y.; Yan, Z.; Chow, M.-Y. A Neural-Network-Embedded Equivalent Circuit Model for Lithium-Ion Battery State Estimation. *arXiv* **2024**, arXiv:2407.20262. <https://doi.org/10.48550/arXiv.2407.20262>.
29. Sesidhar, D.; Badachi, C.; Green, R.C., II. A Review on Data-Driven SOC Estimation with Li-Ion Batteries: Implementation Methods & Future Aspirations. *J. Energy Storage* **2023**, *72*, 108420. <https://doi.org/10.1016/j.est.2023.108420>.
30. Zhao, F.; Guo, Y.; Chen, B. A Review of Lithium-Ion Battery State of Charge Estimation Methods Based on Machine Learning. *World Electr. Veh. J.* **2024**, *15*, 131. <https://doi.org/10.3390/wevj15040131>.
31. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Proceedings of the Advances in Neural Information Processing Systems*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
32. Ge, Z.; Yu, Z.; Lei, Z.; Lang, L.; Bai, X.; Xu, G.; Ma, M.; Zhang, H.; Yan, Y.; Xu, Y. A Novel Hybrid Framework for State of Charge Estimation of Lithium-Ion Batteries: Fusing Online Parameter Identification with Adaptive Filter and Deep Learning Models. *J. Energy Storage* **2026**, *172*, 122921. <https://doi.org/10.1016/j.est.2026.122921>.
33. Zhang, H.; Liu, T.; Zhao, Y.; Hu, J.; Zuo, Q.; Jiang, L. From Lithium to Sodium: A Comparative Study of Temperature and SOC Dependence in Second-Order RC Parameters for Battery Management Migration. *Energy* **2026**, *355*, 141122. <https://doi.org/10.1016/j.energy.2026.141122>.
34. Sun, X.; Ji, J.; Ren, B.; Xie, C.; Yan, D. Adaptive Forgetting Factor Recursive Least Square Algorithm for Online Identification of Equivalent Circuit Model Parameters of a Lithium-Ion Battery. *Energies* **2019**, *12*, 2242. <https://doi.org/10.3390/en12122242>.
35. Li, M.; Zhang, Y.; Hu, Z.; Zhang, Y.; Zhang, J. A Battery SOC Estimation Method Based on AFFRLS-EKF. *Sensors* **2021**, *21*, 5698. <https://doi.org/10.3390/s21175698>.
36. Center for Advanced Life Cycle Engineering. INR 18650-20R Battery Data. Available online: <https://calce.umd.edu/battery-data> (accessed on 22 March 2026).
37. Yao, J. *Lithium-Ion Battery Drive Cycle Dataset*; Technische Universität Berlin: Berlin, Germany, 2024.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.