Research Article
doi.org/10.1002/batt.202500456

Chemistry Europe
European Chemical Societies Publishing

www.batteries-supercaps.org

Benchmarking the Transferability of Real-Time State of Charge Algorithms to Sodium-Ion Cells Using an Open-Source Diagnostics Framework

Katharina Lilith Quade,* Elias Hempen, Hanna van den Berg, Dominik Jöst, Franziska Berger, Florian Ringbeck, and Dirk Uwe Sauer

Despite the abundance of battery state estimation algorithms in the BMS literature, their applicability to emerging cell chemistries remains uncertain, as evaluating their performance across diverse use cases is complex, resource-intensive, and time-consuming. In this work, we introduce an evaluation framework designed to assess the performance of diagnostic algorithms across various applications and external conditions. Our framework relies on simulations of different operational scenarios grouped into categories and evaluates algorithm performance using statistical metrics that represent accuracy, bias, and precision. While our framework can be used for various state estimators and applications, we demonstrate its functionality by benchmarking three common and real-time capable State of Charge (SOC) algorithms across three cell types, including a sodium-ion battery. Within our case study, we show that the tested model-based algorithms offer excellent transferability for the specific sodium-ion battery considering different operation conditions. Additionally, we demonstrate that the characteristic OCV(SOC) relationship of the sodium-ion cell allows for the use of lower-quality and more affordable sensors, as the cell is less sensitive to measurement inaccuracies. Overall, our open-source framework supports the systematic assessment of diagnostic algorithm transferability and provides a foundation for informed decision-making when selecting algorithms for a specific application and cell type.

1. Introduction

Batteries have long powered consumer electronics, stationary storage, and vehicles.[1–4] As electrification advances, their role in other sectors is also rapidly evolving, bringing increased interest in alternative cell chemistries that meet the requirements of sustainability, low cost, and energy density. With similar electrochemical characteristics to lithium-based systems, sodium-ion batteries (SIBs) offer potential as drop-in replacements for certain applications.[5,6] The growing relevance of SIBs is reflected in recent industrial announcements and production efforts.[5,7–9] Nonetheless, as a relatively new battery technology with distinct storage mechanisms, the suitability of SIBs for energy transition applications remains to be fully demonstrated, partly to the potential need for adjustments in real-time diagnostic algorithms. Real-time diagnostic algorithms, integrated into a battery management system (BMS), are essential for ensuring efficient, long-lasting, and safe operation by estimating internal states such as the State of Charge (SOC). Existing diagnostic methods, however, have been primarily developed and optimized for lithium-ion batteries (LIBs), leaving the performance of state estimation in SIBs uncertain and not yet well understood. Ultimately, this uncertainty may discourage battery system designers from adopting SIBs.[9,10]

In the literature, a wide array of algorithms for estimating the SOC exist. They range from conventional coulomb counting[11,12]

K. L. Quade, E. Hempen, H. van den Berg, D. Jöst, F. Berger, F. Ringbeck, D. U. Sauer
Chair for Electrochemical Energy Conversion and Storage Systems
Institute for Power Electronics and Electrical Drives (ISEA)
RWTH Aachen University
Campus-Boulevard 89, 52074 Aachen, Germany
E-mail: katharina.quade@isea.rwth-aachen.de

K. L. Quade, E. Hempen, H. van den Berg, D. Jöst, F. Berger, F. Ringbeck, D. U. Sauer
Center for Ageing, Reliability and Lifetime Prediction of Electrochemical and Power Electronic Systems (CARL)
RWTH Aachen University
Campus-Boulevard 89, 52074 Aachen, Germany

K. L. Quade, E. Hempen, D. Jöst, F. Berger, F. Ringbeck, D. U. Sauer
Jülich Aachen Research Alliance
JARA Energy
Templergraben 55, 52056 Aachen, Germany

D. U. Sauer
Institute for Power Generation and Storage Systems (PGS)
E.ON Energy Research Center (E.ON ERC)
RWTH Aachen University
Mathieustrasse 10, 52074 Aachen, Germany

D. U. Sauer
Helmholtz Institute Münster: Ionics in Energy Storage (HI MS)
IMD-4 Forschungszentrum Jülich
52425 Jülich, Germany

Supporting information for this article is available on the WWW under https://doi.org/10.1002/batt.202500456

and model-based techniques[13–16] to data-based methods,[17–22] each with its own advantages and limitations.[23–25] However, selecting the most suitable algorithm for a particular application and battery cell presents a challenge due to the limited evaluation of algorithms found in the literature that consider a diversity of operational scenarios, making direct performance comparisons difficult.[26] Often, SOC algorithms are evaluated for only one specific load profile, such as the worldwide harmonized light vehicles test procedure (WLTP), at a single temperature,[12,13,15,19] assuming ideal BMS measurements, and considering one initial SOC[12–15] using a single statistical metric such as root mean square error (RMSE).[17,27] Yet, in reality, batteries experience various operation scenarios, influencing the algorithm's overall performance.[28] Without a comprehensive evaluation of BMS algorithms, diagnostic algorithms might fail to perform reliably in specific situations, even if they perform well on average.[29]

To address these challenges, researchers have already developed methods for evaluating diagnostic algorithms. For instance, varying load profiles,[28,30] temperatures,[31] and various aging states[30] were considered for SOC algorithm assessment. Simulations were performed under the influence of disturbances such as noise, offset, or faulty initialization.[16,30,32] Klee Barillas et al.[30] and Campestrini et al.[26] evaluated various diagnostic algorithms using a point-based system. The points achieved in different categories were graphically compared for all the algorithms investigated in a radar chart, allowing an intuitive visual comparison of the algorithms based on individual criteria. However, the evaluation of these algorithms either overlooks algorithm performance across diverse operating scenarios or relies on a scale with no direct physical meaning. Additionally, existing evaluation methods focus solely on measuring algorithm accuracy, neglecting assessments of bias, and precision. This can contribute to misleading results and potentially cause inadequate algorithm selection, as demonstrated in the literature.[27]

In this work, we introduce an open-source multicriteria simulation-based evaluation framework designed to assess the performance of real-time diagnostic algorithms across a wide range of application scenarios, and external conditions. As there already exists a large number of real-time SOC estimation methods in the literature, this work does not focus on developing new algorithms but rather on presenting a framework to systematically assess their performance. While this simulation-based approach is well-suited for identifying critical scenarios and determining optimal operational and environmental parameters, the framework is ideally used as a preparatory step for final hardware-based validation. By narrowing the scope to the most relevant test cases, it helps reduce the required resources, time, and cost of subsequent physical testing while ensuring that hardware validation efforts are focused where they are most impactful. To illustrate our workflow, we further present a case study, showing how existing algorithms for LIBs can be assessed for their transferability to sodium-based systems in an automated and comparable manner. As the literature reports substantial performance differences depending on the algorithm type, our study aims to approximate the performance and sensitivity of algorithms before committing to long-term testing.[33] By applying our framework, we gain insight into the complex interplay of BMS properties, cell chemistry, and SOC algorithm performance. The open-source nature allows researchers to systematically benchmark various diagnostic algorithms, fostering transparency and comparability in the field.

2. Experimental Section

Our simulation-based evaluation framework, depicted in **Figure 1**, automates the assessment of diagnostic algorithm performance through a top-down data generation and bottom-up evaluation process. To generate data for our framework, we must first identify critical performance categories covering environmental, operational, and BMS-specific influences (Section 3.1) and derive specific scenarios, for example, 40 °C ambient temperature or 80% State of Health (SOH). The framework presented in this work is agnostic to the algorithm type, meaning that only external factors such as operating conditions and sensor disturbances are varied, while the algorithms' estimated states and parameters are not interfered with. Scenario data for each category, including battery and BMS time-series data, is generated using a model-based simulation

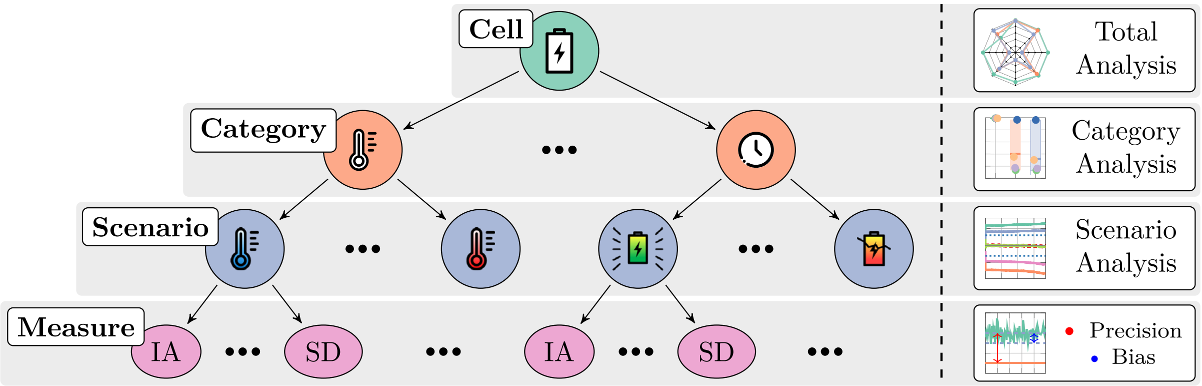

Figure 1. Evaluation framework consisting of top-down data generation and bottom-up evaluation process. We evaluate several metrics for every scenario, averaging the results for every category and cell.

toolchain that supports both battery and BMS simulations. The simulation toolchain, adapted from Berger et al.[29] and detailed in Section 2.1, is provided with this publication to enable readers to perform BMS simulations.[34]

For each simulated scenario, we compare the reference values to the estimated ones and calculate a set of statistical metrics, described in Section 2.2, including measures of accuracy, bias, and precision. These metrics are then aggregated and analyzed across all scenarios within each category. After evaluating each category individually, a final summary of the results is provided, including average values per category and metric and aggregated data across all categories to assess overall algorithm performance. The following sections describe the data generation and evaluation processes in more detail.

2.1. Data Generation Toolchain

Evaluating algorithms under diverse operating conditions with actual battery specimens is both time-consuming and resource-intensive. As a remedy, we employ a model-based simulation toolchain, originally developed in MATLAB/Simulink, for our algorithm performance benchmarking that is acting as a digital twin. This approach, depicted in **Figure 2**, simulates the entire BMS signal path, from the load profile to the diagnostic algorithms.

To simulate battery behavior in our toolchain, we use a high precision model based on the measurement and fitting procedure by Bihn et al.[35] The model provides reference values, which we use to compare against the outputs of the diagnostic algorithms. To ensure the accuracy of our reference values and a sufficient comparability to real-world battery behavior we calibrate the parameters of the high-precision model to result in an average voltage-response error of $\leq$5mV at all temperatures compared to the physical battery measurements. We demonstrate this process exemplary in Supporting Information Section Parameterization of ECMs for the SIB and further show the accuracy compared with R1RC models. To account for different hardware designs, we emulate the BMS hardware and introduce measurement noise, sampling effects, and offsets to the reference data based on components commonly used in modern systems.[36] The BMS hardware emulation method has been previously benchmarked to real-world BMS connected to actual sensors in our previous publication and generates sufficiently accurate results for this work.[29] Since the emulation is primarily hardware-dependent and the hardware parameters in this work are intended to remain adjustable, we refer to a previous study[37] for a detailed quantification of emulation errors under typical hardware configurations. The processed data serves as input for the diagnostic algorithms. The detailed steps of the BMS simulation are as follows: 1) Load profile, where the simulation of a load profile that is scaled to the specific battery is the starting point of the toolchain. We select a variety of relevant profiles, including edge cases. The profiles serve as an input for the subsequent high precision battery model. 2) High precision battery model: utilizing a high-order equivalent circuit model (ECM), we simulate battery characteristics such as voltage, current, and temperature within a battery simulation toolchain. The model structure, defined in an eXtensible Markup Language (XML) file, contains the parameter values dependent on the SOC and temperature. While these parameters are on the cell level in this work, Barbers et al. demonstrated how this model can be used to generate virtual battery packs considering electrochemical, thermal, and aging aspects.[38] Barbers et al. demonstrated that a normally distributed parameter range in a high precision battery model can adequately capture cell-to-cell variation in both initial and aging parameters. In this work, however, we fix these values for each cell, as our focus is on providing an initial demonstration

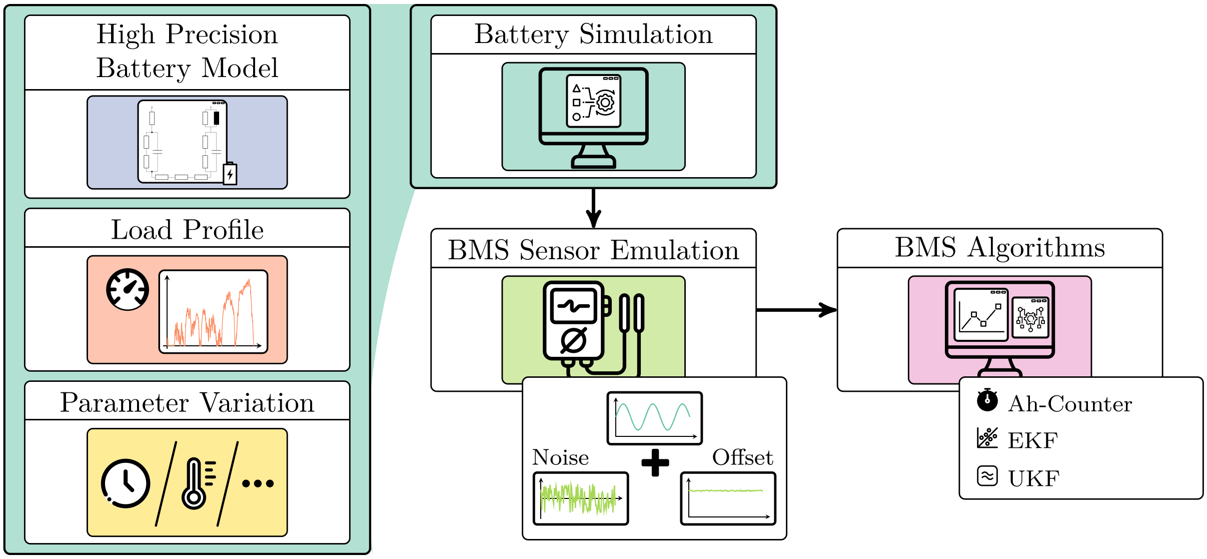

Figure 2. A high-precision battery model takes a load profile and specific environmental conditions as inputs, simulating realistic battery behavior. The outputs are then passed through a BMS sensor emulation stage, introducing noise and offsets to mimic real-world sensors. Finally, using different, real-time BMS algorithms we estimate the SOC.

of the framework's capabilities.[38] Regarding the aging parameters, we adopt a simplified approach that linearly scales the OCV and overpotentials with aging as a first-order approximation, following common practice in the literature[39] (see Supporting Information Section Setup of Robustness Categories). For each defined parameter, the voltage is calculated and the current state is updated. More information about the battery simulation can be found in ref. [38,40]. For the lithium-based systems, we use the battery models described in ref. [41]. For the sodium-based system, we build upon our previous publication about a commercial SIB,[6] where we performed DC pulse and electrochemical impedance spectroscopy (EIS) measurements to model the electrical behavior of the SIB. Further information about the fitting results can be found in the Supporting Information Section Parameterization of ECMs. The ECM of the SIB is openly shared as part of the open-source framework.[34] 3) Sensor emulation: The battery measurements are affected by noise, quantization, and perturbations, replicating real-world BMS hardware effects. Quantization effects are modeled based on the resolution of a BMS analog-digital converter, with separate quantization steps for temperature, voltage (0.15 mV), and current (0.3 mA), complemented by configurable measurement noise and gain/offset parameters. The sensor system introduces for each simulation a sensor offset of 0.025% of the nominal value, a (linear) gain error of 10^{-4}, and white noise with a variance of 0.15% relative to the nominal value. These values are based on the commercial sensor data sheets used in relevant applications.[36,42] For all conducted simulations, we use a fixed sampling time of 0.01s, consistent with the findings of Berger et al.[29] It is important to emphasize that the primary objective of this work is to showcase the benchmarking toolchain and its flexibility, rather than to develop an holistic fault model for each sensor error category. The toolchain is deliberately designed to remain flexible, enabling users to introduce more complex and dynamic error models such as temperature drift, stochastic faults, or time-varying noise. Future work can leverage this tool to explore detailed fault models for each category and systematically analyze their cumulative impact on SOC estimation. 4) BMS algorithms: The output from the BMS hardware emulation is further processed to simulate real-time diagnostic algorithms. As a demonstration of our evaluation framework, we benchmark four common real-time capable SOC algorithms, i.e., Ampere-hour counter with full charge detection (AhC & FC), extended Kalman filter with first-order Thévenin model (EKF R1RC), unscented Kalman filter with first-order Thévenin model (UKF R1RC), and adaptive UKF R1RC (AUKF R1RC) on two lithium-based and one sodium-based cells. Although our framework is compatible with any algorithm type, we focus on real-time capable SOC algorithms in this work, as they are currently used in most applications due to their lower computational complexity.[43–48] The process and results of the R1RC fitting for the SIBs is described in more detail in the Supporting Information Section Parameterization of ECMs. The R1RC models, and algorithm initialization procedures are predefined and not the focus in this work, as they are well-documented in existing literature. Instead, we focus on evaluating the operational behavior of these algorithms using predefined models to assess their performance and transferability. To ensure transparency and reproducibility, the algorithms and parameters used are given in Supporting Information Section Real-time SOC Algorithms. However, it is worth noting that we use R1RC models for the filter-based algorithms,[49] offering a computationally efficient alternative to more complex models discussed in the literature, which in turn might generate more accurate estimations.[50]

To demonstrate the functionality and reliability of our framework, we conducted a systematic hyperparameter sensitivity analysis for the model-based algorithms, which is described in detail in the Supporting Information Section Hyperparameter Sensitivity. This analysis illustrates how the framework can be used to explore a range of parameter combinations, identify parameters that minimize the RMSE, and ensure consistent and fair comparisons between algorithms.

2.2. Evaluation Criteria

To address the limitations of current evaluation practices, we introduce a set of statistical metrics for assessing diagnostic algorithm performance. Within the existing literature, numerous qualitative metrics are used for the statistical evaluation of algorithm accuracy.[27,51,52] Accuracy measures the difference between the estimated and reference values, with maximum accuracy indicating full congruence. It combines bias (systematic error) and precision (random error). When accuracy is low, distinguishing between bias and precision helps identify the source of inaccuracy. The concept of evaluation criteria is illustrated in Supporting Information Figure S5. In our work, we select statistical metrics for each criterion, i.e. accuracy, bias, and precision, as shown in **Table 1**.

Including the Index of Agreement (IA) as an accuracy metric provides a normalized measure ranging from 0 to 1. As errors increase, the IA decreases, but the metric tends to be less sensitive to large errors because it also depends on the variance of the observed data. This means that large, but short errors do not drastically reduce the IA.[53] The IA complements maximum error ($e_{\max}$) and RMSE, which are well-established in the literature, offering a widely recognized basis for comparison alongside a relative measure of accuracy. The RMSE increases sharply with error magnitude because it squares individual differences. Larger errors, especially outliers, have a disproportionately strong impact, making RMSE more sensitive than the IA. Bias is quantified using the (ME), while precision is assessed with the standard deviation (σ). In Table 1, the target value denotes the metric value that should be achieved with complete freedom from errors.

3. Case Study

Our framework is designed to be flexible and adaptable to various battery chemistries, operating conditions, and performance metrics. In this work, we present a case study to demonstrate its capabilities; however, this case study cannot cover the full range of possible applications. Therefore, the Supporting Information

Table 1. Statistical metrics for the evaluation of diagnostic algorithms with the indication of the target value and range of the metric. The estimated value is denoted as y, while the reference value is denoted as w. e indicates the deviation of the estimated value y from the corresponding reference value w.

Criterion	Metric	Definition	Range	Target value
Accuracy	Maximum error	$e_{\max} = \begin{cases} +\max_{i \in 1,\ldots,N} \lvert e_i \rvert, & \lvert \max e_i \rvert \geq \lvert \min e_i \rvert \\ -\max_{i \in 1,\ldots,N} \lvert e_i \rvert, & \text{otherwise} \end{cases}$	$(-\infty, +\infty)$	0
Accuracy	Root mean square error	$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i - w_i)^2}$	$[0, +\infty)$	0
Accuracy	Index of agreement	$\mathrm{IA} = 1 - \frac{\sum_{i=1}^{N} \lvert y_i - w_i \rvert}{\sum_{i=1}^{N} (\lvert y_i - \overline{w} \rvert + \lvert w_i - \overline{w} \rvert)}$	$[0, 1]$	1
Bias	Mean error	$\mathrm{ME} = \overline{e} = \frac{1}{N}\sum_{i=1}^{N} (y_i - w_i) = \overline{y} - \overline{w}$	$(-\infty, +\infty)$	0
Precision	Standard deviation	$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (ei - \overline{e})^2}$	$[0, +\infty)$	0

Section Framework Usability and Extensibility provides a step-by-step guide for executing the framework using an LFP/C cell and explains how to design customized test scenarios.

3.1. Simulation-Based Design of Experiment

To systematically analyze the performance of SOC estimation algorithms, it is essential to first define the application context and identify relevant usage scenarios. We focus on battery electric vehicles (BEVs) as a representative case study, as they impose distinct requirements on BMS functions due to their diverse and dynamic operating conditions. Based on this context, we define seven evaluation categories to assess factors influencing SOC estimation accuracy, encompassing environmental, operational, and BMS-specific conditions.[37,54] A *Base scenario* serves as a reference, testing each estimator without disturbances, as in prior studies.[26] We then vary one feature per category while keeping others fixed: load profiles (*Operation profile*), initial SOC (*Initial SOC*), ambient temperature (*Temperature*), aging (*Aging*), sensor errors (*Current sensor* and *Voltage sensor*), and faulty initialization (*Faulty initialization*). These categories are based on the analyses from Berger et al.[29] Sensor variations are based on realistic deviations from state-of-the-art BMS references, considering offset errors, variance, and gain effects.[26,36] Each cell undergoes a series of simulations (see Supporting Information Table S1 and Section Set-up of Robustness Categories). Figure 2A summarizes the considered BEV scenarios for our case study (**Figure 3**).

Figure 2B illustrates the profiles used in the *Operation profile* category, representing different driving states such as charging, driving, and resting to assess algorithm performance under varying operational dynamics and durations. To ensure comparability across different battery cells, all profiles are normalized and presented in (C-rate)-equivalents. For consistency in evaluation, we include the WLTP cycle and the standardized *DC0* profile, specifically developed for BEVs. These two profiles are relatively short and offer well-defined test conditions. In contrast, the remaining profiles reflect longer real-world usage patterns, derived from recorded driving data to ensure a representative assessment of algorithm robustness under practical conditions.

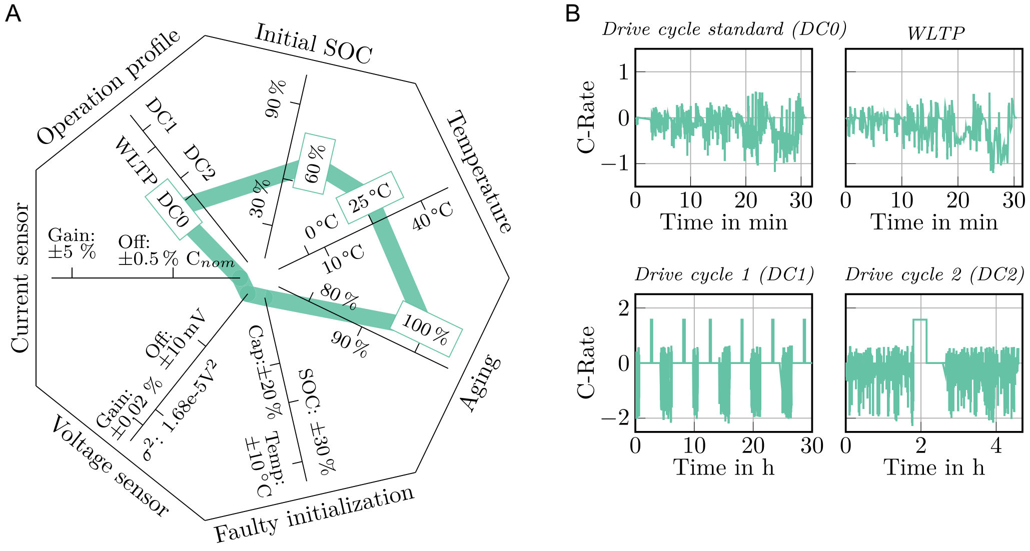

Figure 3. Design of experiment as a base for algorithm benchmark. A) Set-up of scenarios within each category, with the green line indicating the base scenario. Within each category, only one feature is varied. B) Drive cycles employed in the category *Operation profile*. The drive cycles have distinct driving dynamics and durations. They are composed of different vehicle states, such as charge, rest, and drive phases. The WLTP is included as it is a widely used profile in the literature.

3.2. Results

We apply our framework to three cells, two LIBs and one SIB, and compare four real-time capable SOC algorithms. The selected cells, each with distinct electrochemical properties, are as follows

$$\text{LiNiMnCoO}_2\,(\text{NMC})\ \text{vs.}\ \text{Li}_4\text{Ti}_5\text{O}_{12}\,(\text{LTO})\ \text{cell} \tag{1}$$

$$\text{LiNiCoAlO}_2\,(\text{NCA})\ \text{vs. graphite} + \text{silicon}\,(\text{SiC})\ \text{cell} \tag{2}$$

$$\text{NaMn}_{1/3}\text{Fe}_{1/3}\text{Ni}_{1/3}\text{O}_2\ (\text{NaMO}_2)\ \text{vs. hard carbon}\ (\text{HC})\ \text{cell} \tag{3}$$

The selection contrasts established lithium-ion chemistries that were previously analyzed in the literature[55,56] with an early commercial sodium-ion cell previously analyzed by Laufen et al.[6] Supporting Information Figure S7 shows the respective open-circuit voltage (OCV) (SOC) curves for all three cells. The $NaMO_2$/HC cell exhibits the steepest OCV curve among all tested configurations and shows a notably linear voltage behavior at high SOCs, in contrast to the other cells. Since model-based algorithms for state estimation rely on the voltage response to infer internal states, the shape and steepness of the OCV(SOC) relationship play a critical role in their performance and interpretability. Further properties of the cells are provided in the Supporting Information Section Battery Cell Properties.

Figure 4 presents the evaluation results of the three SOC algorithms across all cells. The left panel summarizes algorithm performance across the evaluation categories using the average IA value, chosen for its direct comparability in the radar chart. The scale is limited to the range from 0.6 to 1 to enhance visibility of performance differences between algorithms. All cells show good performance across all algorithms in the base scenario, with accuracy values above 0.93. However, the performance of each algorithm varies significantly with changes in operating conditions and cell chemistry. The right panel provides a detailed view of algorithm performance across the individual scenarios within the *Temperature* category as an example. For the EKF R1RC, additional plots to the right of each panel illustrate the deviation between estimated and reference OCV, along with the corresponding SOC difference. Similar analyses can be conducted for any other category and algorithm of interest.

Overall, for the NMC/LTO cell (Figure 4A), the AhC & FC performs well across most categories, with the exception of *Faulty initialization* and *Aging*, as expected. This can be attributed to the fact that the AhC & FC does not recalibrate unless a full charge or relaxation occurs, neither of which is included in the standard profile *DC0*. Since the profiles for all cell chemistries are C-rate equivalent, the results for the AhC & FC are the same as for the other two cells. In contrast, both model-based methods generally show reduced performance in the *Temperature* and *Voltage sensor* categories for the NMC/LTO cell. For the latter, both filters perform poorly, likely due to the cell's particularly narrow OCV range (see Figure S7A, Supporting Information). As a result, voltage measurement errors have a larger relative impact, since they represent a greater portion of the cell's overall OCV window compared to other chemistries.

In the *Temperature* category, performance deteriorates especially at low temperatures, where OCV estimation errors can exceed 10 mV. For instance, in the case of the EKF R1RC, this leads to substantial SOC estimation errors of around 4%. The UKF R1RC exhibits smaller deviations, resulting in an IA of ≈0.8. These inaccuracies are likely caused by increased overpotentials at low temperatures, which amplify modeling inaccuracies in the R1RC model. The AUKF R1RC achieves more accurate SOC estimations at 0 °C compared to the EKF R1RC; however, due to higher errors in the 10 °C case, its average IA value in this category is the lowest among the three filter-based methods.

For the NCA/SiC cell (Figure 4B), the model-based algorithms show significantly higher sensitivity to external factors compared to the base scenario, resulting in a much more heterogeneous radar chart. In the *Temperature* category, the OCV estimation errors of the EKF R1RC lie between 1 and 6 mV, resulting in SOC estimation errors below 1%. Although the voltage estimation error increases at lower temperatures, the SOC error remains negligible for the EKF R1RC. While the UKF R1RC is generally better suited for nonlinear systems, its performance degrades at low temperatures due to increased sensitivity to R1RC modeling inaccuracies and measurement noise. The AUKF R1RC achieves better results at low temperatures than the UKF R1RC, while maintaining similar accuracy at higher temperatures. With improved filter initialization, particularly under challenging conditions, these estimation errors could potentially be reduced. According to our framework, the EKF R1RC appears more robust in aging scenarios. Both filters use a fixed design without parameter adaptation to aging effects, whereas the UKF R1RC suffers from increased sensitivity to modeling inaccuracies caused by higher impedance and pronounced OCV nonlinearities.[57] This leads to larger SOC estimation errors for the UKF R1RC, while the EKF R1RC remains more stable under these fixed modeling conditions. Again, the AUKF R1RC outperforms the UKF R1RC in the *Aging* category. In contrast, all three filter-based algorithms exhibit a weaker, yet still noticeable, sensitivity to real-world BMS voltage measurement errors compared with the NMC/LTO cell, likely due to the steeper OCV curve, which helps mitigate the impact of voltage inaccuracies. As visible in the radar chart, the EKF R1RC performs worse in the *Faulty initialization* category compared to the UKF R1RC and AUKF R1RC. In this case, the filter corrects an initial SOC misestimation only slowly, failing to fully converge within the scenario duration.

Interestingly, the model-based algorithms applied to the SIB (Figure 4C) exhibit notably lower sensitivity to both external and internal influences and maintain a high SOC estimation accuracy even in the presence of accumulated voltage sensor errors. Regarding temperature sensitivity, despite relatively high OCV estimation errors, even higher than those observed for the NMC/LTO cell, the EKF R1RC achieves SOC estimation errors below 1%, whereas the NMC/LTO cell experiences a maximum of 4% error. This behavior can be attributed to the wide and linear voltage range of the $NaMO_2$/HC cell. For the remaining categories, the model-based algorithms continue to perform well, maintaining consistently high accuracy. The specific values of the remaining metrics, namely RMSE, e_{max}, ME, and σ, for the SIB

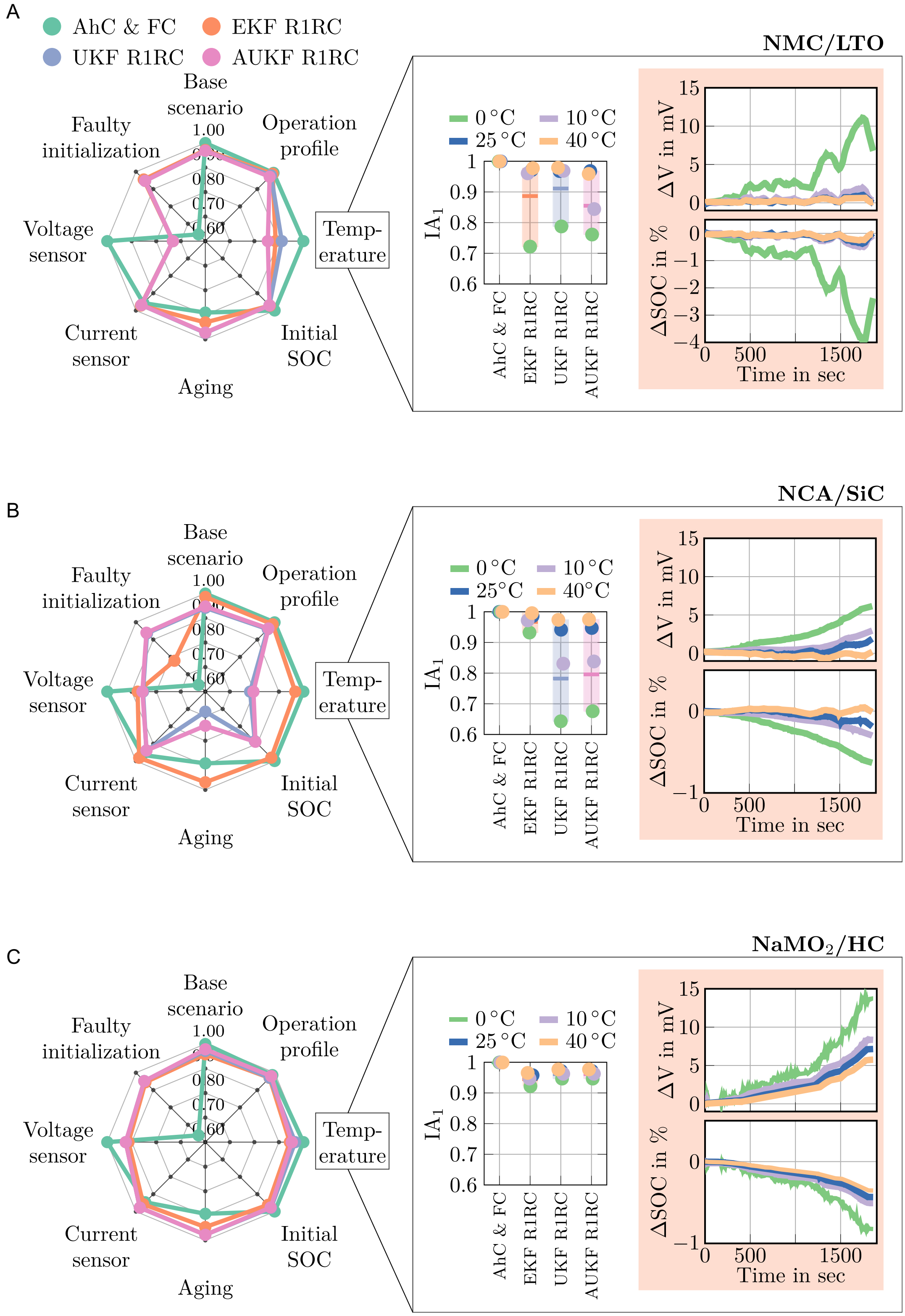

Figure 4. Summarized results using the IA as a measure of accuracy across all three cell chemistries. The IA provides an initial indication within the framework of which categories may exhibit irregularities. The right panel focuses on the *Temperature* category, first presenting an overview of the individual scenarios and algorithms, followed by the deviation of estimated and reference voltage and algorithm performance for each scenario for the EKF R1RC. A) NMC/LTO. B) NCA/SiC. C) NaMO$_2$/HC.

are shown in Supporting Information Figure S8 highlighting the strong performance of all algorithms for the SIB even under challenging conditions. When it comes to the model-based algorithms, slight performance losses are observed in the presence of voltage offsets and an initial capacity error of 20%, which leads to increased bias and, consequently, less accuracy. Overall, the AUKF R1RC consistently outperforms the EKF R1RC in terms of accuracy, bias, and precision across all evaluated scenarios. According to the IA, the performance of the AUKF R1RC is similarly high compared with the UKF R1RC.

In addition to the IA-based radar chart presented in Figure 4C for the SIB, Supporting Information Figure 4C shows radar charts of additional metrics, including RMSE, ME, and σ. We extend our comparison to all cell chemistries using the average metric values shown in Figure S10, Supporting Information. For the NCA/SiC cell, we observe that accuracy errors in all categories for the UKF R1RC and AUKF R1RC are caused by a combination of high bias and random error. Within the *Voltage sensor* category, the EKF R1RC, the UKF R1RC, and the AUKF R1RC show high bias, although the AUKF R1RC exhibits the lowest bias across all three cell chemistries. Particularly in the *Faulty initialization* category, we find that the RMSE is largely driven by systematic error, as indicated by high ME values across all cells in the AhC & FC. **Table 2** summarizes the best and worst algorithms for each cell based on the different averaged metrics. It is evident that, according to the measures of accuracy, all metrics favor the model-based algorithms, with the UKF R1RC performing best for the NMC/LTO and $NaMO_2$/HC cells according to the metric IA. However, when considering bias and precision, different algorithms rank as the best or worst. For the two lithium-based cells, the AhC & FC achieves the highest precision, as it is largely unaffected by environmental conditions and low random error occurs. This highlights that while model-based algorithms offer superior accuracy, direct approaches like the AhC & FC may be preferable if consistent performance and low random error are prioritized over accuracy.

3.3. Implications

To further analyze the sensitivity of all algorithms to hardware-related disturbances, **Figure 5**A,B present the accuracy metrics for the *Voltage sensor* and *Current sensor* scenarios, respectively, across all three cells. For the *Voltage sensor* category, the AhC & FC behaves as expected and remains unaffected by voltage deviations. While the AhC & FC includes a recalibration mechanism that could, in principle, respond to incorrect voltage readings, such

Table 2. Comparison of the performance evaluation results for the NMC/LTO, the NCA/SiC, and the $NaMO_2$/HC cells.

(a) Best and poorest average IA values across all robustness categories.

Battery	Best performance	$\overline{IA}$	Poorest performance	$\overline{IA}$
NMC/LTO	UKF R1RC	0.92	AhC & FC	0.91
NCA/SiC	EKF R1RC	0.94	UKF R1RC	0.87
$NaMO_2$/HC	UKF R1RC	0.97	AhC & FC	0.92

(b) Best and poorest average RMSE values across all robustness categories.

Battery	Best performance	$\overline{RMSE}$	Poorest performance	$\overline{RMSE}$
NMC/LTO	UKF R1RC	0.33	AhC & FC	1.41
NCA/SiC	EKF R1RC	0.32	AhC & FC	1.39
$NaMO_2$/HC	AUKF R1RC	0.20	AhC & FC	1.48

(c) Best and poorest average e_{max} values across all robustness categories.

Battery	Best performance	$\overline{e_{max}}$	Poorest performance	$\overline{e_{max}}$
NMC/LTO	UKF R1RC	0.59	AhC & FC	1.57
NCA/SiC	EKF R1RC	0.59	UKF R1RC	1.87
$NaMO_2$/HC	UKF R1RC	0.46	AhC & FC	1.7

(d) Best and poorest average ME values across all robustness categories.

Battery	Best performance	$\overline{ME}$	Poorest performance	$\overline{ME}$
NMC/LTO	EKF R1RC	0.28	AhC & FC	1.37
NCA/SiC	EKF R1RC	0.47	AhC & FC	1.38
$NaMO_2$/HC	AUKF R1RC	0.16	AhC & FC	1.42

(e) Best and poorest average SD values across all robustness categories.

Battery	Best performance	$\overline{SD}$	Poorest performance	$\overline{SD}$
NMC/LTO	AhC & FC	0.1	EKF R1RC	0.22
NCA/SiC	AhC & FC	0.09	UKF R1RC	0.55
$NaMO_2$/HC	UKF R1RC	0.12	EKF R1RC	0.19

recalibrations are not triggered during the standard *DC0* profile. In contrast, model-based algorithms are clearly sensitive to voltage inaccuracies. Among the tested disturbances, an offset of 10 mV has the most significant impact on estimation accuracy. Such an offset falls within the typical range of voltage measurement deviations in commercial BMSs[58] and results in a notable drop in performance for the model-based estimators. The NMC/LTO cell, which operates within a narrow voltage window and features a relatively flat OCV(SOC) curve (see Figure S7A, Supporting Information), is particularly sensitive to voltage offsets. These errors propagate into the OCV reference and thereby strongly affect SOC estimation. In contrast, the NCA/SiC cell shows reduced sensitivity, while the $NaMO_2$/HC cell, characterized by the steepest OCV curve with the broadest voltage window among the three, exhibits the lowest susceptibility to voltage measurement errors.

Inaccuracies in current sensors affect both model-based estimators and the AhC & FC. Because the test profiles are C-rate-equivalent across all cell chemistries, the impact of current sensor errors on the AhC & FC remains consistent regardless of the cell type. Similar to the *Voltage sensor* category, offset errors in the *Current sensor* category have the greatest impact, particularly on the AhC & FC. These offsets lead to linear integration drift over time, resulting in significant SOC deviations. In contrast, model-based estimators are less affected due to their closed-loop structure and the ability to partially compensate for erroneous current values through model dynamics. Across all chemistries, model-based algorithms consistently show lower variability around the reference case than the AhC & FC, highlighting their higher robustness against current sensor disturbances. When comparing realistic sensor inaccuracies across all scenarios, voltage sensor errors consistently cause greater estimation deviations than current sensor errors. In this regard, the $NaMO_2$/HC cell exhibits the highest robustness, benefiting from its steep OCV(SOC) slope, which mitigates the impact of voltage deviations. In contrast, the sensitivity to current sensor errors shows less dependence on cell chemistry and is more strongly influenced by filter initialization and the confidence placed in measurements and models.

Based on the previous observations, the UKF R1RC demonstrates the most accurate performance in estimating the SOC of the selected SIB according to the IA metric. Although our analysis is based on a finite set of simulated load profiles, the framework enables a targeted evaluation of how specific profile segments influence estimation accuracy, supporting the derivation of generalizable conclusions beyond the individual cases. To illustrate this in more detail, we break down the estimator behavior for the *DC1* profile into three operation phases, i.e., drive, rest, and charge. **Figure 6**A presents different metrics for accuracy, bias, and precision of the estimators during each of these phases. For the AhC & FC, the largest RMSE occurs during the drive phase, driven primarily by a high ME, indicating error accumulation due to the absence of voltage-based

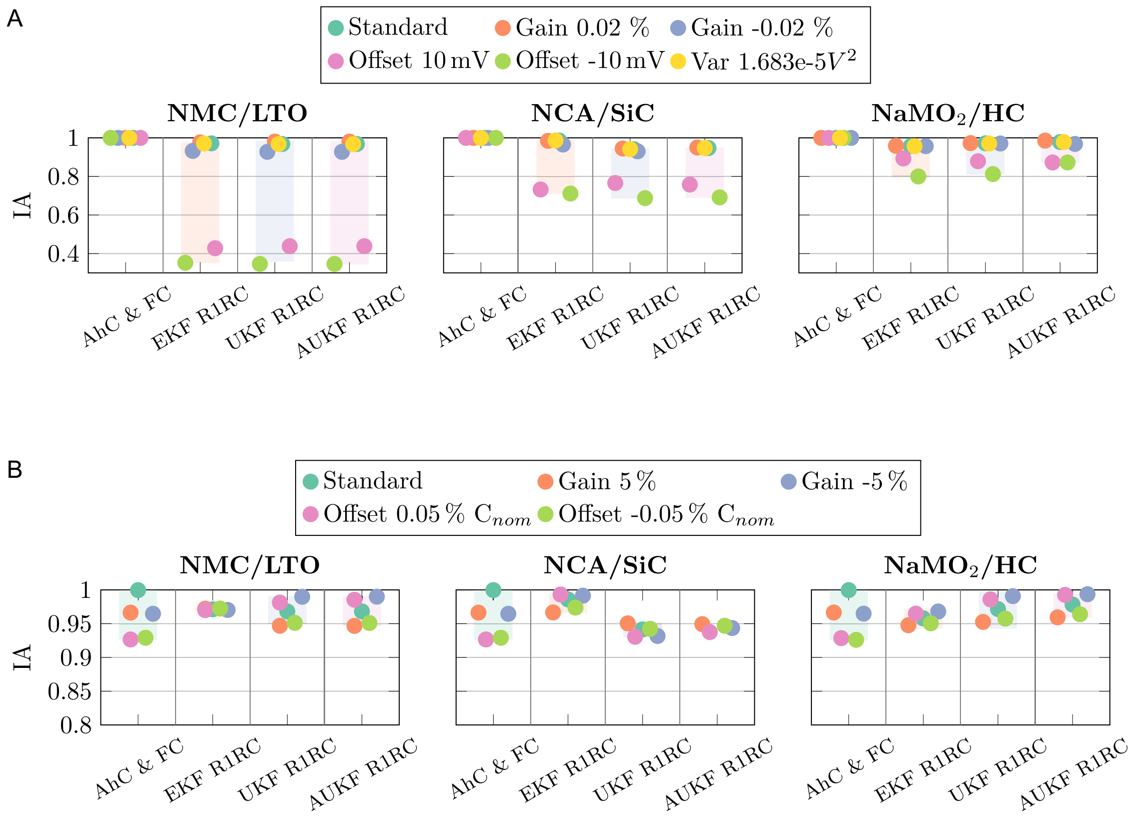

Figure 5. Comparison of algorithm performance considering IA across the three cell chemistries. A) Voltage sensor. B) Current sensor.

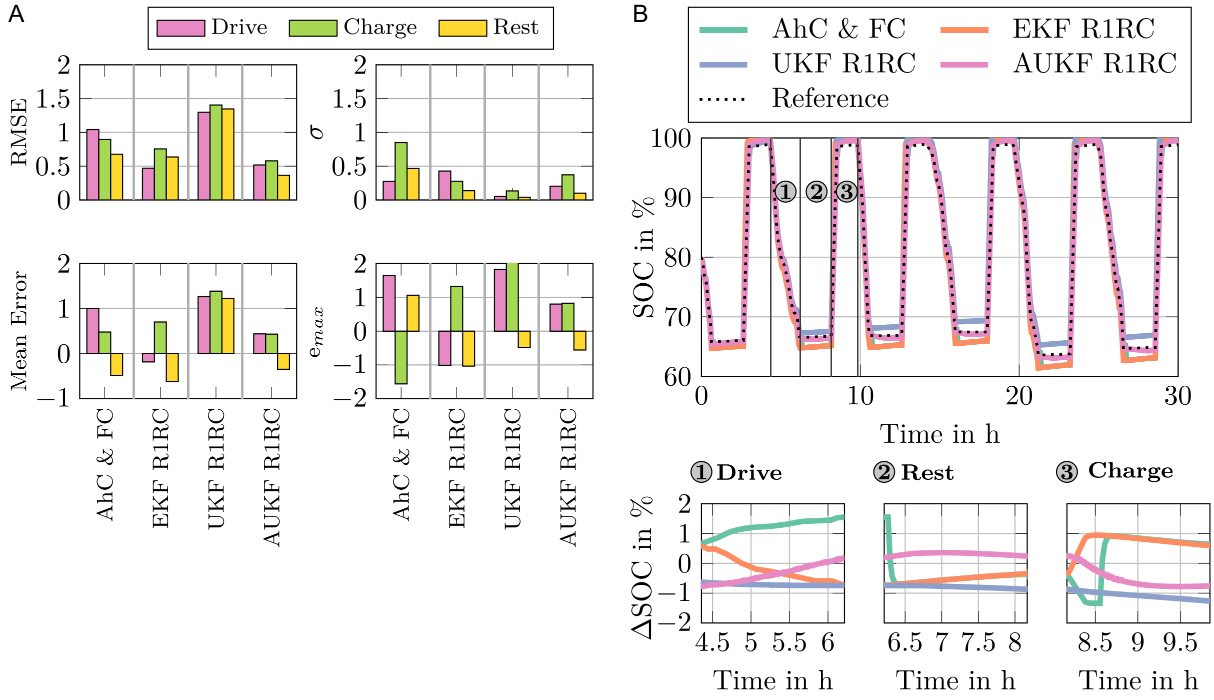

Figure 6. Detailed algorithm performance analysis for the Drive Cycle 1 profile consisting of charge and discharge phases. A) Algorithm performance for different operation phase according to the metrics of accuracy, bias and precision. B) Estimated and reference SOC as time series (top). Zoomed views of specific operation phases for detailed performance analysis (bottom).

correction during this segment. The lowest precision for the AhC & FC is seen in the charge phase, which undergoes abrupt corrections during voltage-based recalibrations. In contrast, the model-based filters show the lowest accuracy during the charge phase. Across all phases and metrics, the AUKF R1RC achieves the most accurate and robust performance, highlighting its robustness for our SIB under the profiles conditions. Figure 6B further illustrates this behavior over time for the *DC1* profile. The top plot shows the evolution of SOC estimates from all algorithms compared with the reference. We see that the AhC & FC increases its error during drive periods and realigns during OCV-based corrections, which are triggered during rest and charge phases. This leads to the step-like behavior and sharp transitions in the SOC estimation. The numbers correspond to zoomed views of the drive, rest, and charge segments, respectively: During driving, the AhC & FC estimate gradually accumulates error, while during rest and charging, jumps occur due to recalibration. The EKF R1RC shows smoother transitions and moderate drift during drive, while the AUKF R1RC maintains minimal error accumulation and low sensitivity to phase changes. Taken together, these results highlight the stability of the estimators for our SIB across all operational phases. The AhC & FC benefits from OCV based corrections, but its reliance on intermittent recalibration leads to phase-dependent error accumulation and corrections that compromise its precision. The EKF R1RC performs more consistently but still shows slight deviations during changing conditions. The UKF R1RC shows a bias for the profile under testing, while the AUKF R1RC offers the most balanced and robust SOC estimation for this specific profile.

3.4. Limitations of Evaluation Framework

The developed evaluation framework provides a structured and reproducible way to assess algorithm performance under an array of conditions. By isolating individual influencing factors, such as temperature, aging, or sensor accuracy, the framework allows to attribute performance clearly to specific conditions. While this controlled setup supports transparency and comparability, it does not fully reflect real-world applications, where multiple factors often interact simultaneously, such as low temperatures coinciding with sensor inaccuracies or the interplay between aging and voltage measurement errors. Moreover, while the use of simulation models ensures scalability and efficiency, it inherently involves abstractions of actual battery behavior. As a result, while the framework excels in controlled benchmarking and comparison of estimation methods, its direct transferability to real-world deployments must be considered with care and should rather serve as a quantitative sensitivity study. In future work, we aim to extend the framework by integrating scenarios with interacting influences and coupling it with hardware-in-the-loop testing environments to evaluate algorithm performance under more realistic and dynamic conditions. At the same time, we are expanding the framework to include additional relevant state estimations beyond SOC, further enhancing its practical value for BMS development.

4. Conclusion

To ensure optimal performance in a given application, it is essential to evaluate and compare different algorithms for each cell candidate under relevant operating conditions. Moreover, for new cell chemistries, this involves examining whether established approaches still apply or require adaptation. To support these tasks, we introduce a multicriteria evaluation framework capable of systematically assessing the operational behavior of real-time diagnostic algorithms. The framework, that is based on simulations, enables this analysis without requiring direct conceptual knowledge of algorithms. Instead, we focus on evaluating the operational behavior of existing algorithms with predefined initial parametrization.

Within our case study, we employ the proposed framework to two lithium-based and one sodium-based battery intended for use in a BEV application. For each cell, we identify the best performing SOC algorithm across relevant application scenarios. Considering industry-relevant hardware, we demonstrate that voltage measurement errors and R1RC model inaccuracies exacerbated by low temperatures are the most critical factors affecting accurate SOC estimation. However, the SIB is least affected by these influences due to its broad and steep OCV(SOC) curve, which reduces the impact of voltage errors. Ultimately, we demonstrate that real-time capable SOC algorithms can be transferred to the investigated SIB without being hindered by sensor limitations. The possibility of using lower-accuracy sensors without substantially compromising estimation quality may help reduce system costs and support the use of sodium-ion cells in cost-sensitive applications that have previously been constrained by high costs of cells and system hardware.

The open-source nature of our framework encourages further development and collaboration, enabling researchers and developers to expand the evaluation scope and explore more complex, overlapping conditions. As cell design and choice also greatly influences pack topology, future work should examine the effect of cell connections and other pack properties on pack-level SOC estimation. Overall, our simulation-based evaluation provides a valuable framework for identifying strengths and weaknesses in algorithm design and employment before committing to long-term real-world testing, ultimately supporting efficient BMS development.

Acknowledgements

This work was financially supported by the German Federal Ministry of Research, Technology and Space (BMFTR) within the "SIB:DE FORSCHUNG" project (03XP0627I).

Open Access funding enabled and organized by Projekt DEAL.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are openly available in BMS Simulation as MIL at https://git.rwth-aachen.de/isea/bms-simulation-as-mil, reference number 109806.

Keywords: algorithm benchmark · battery management system · diagnostics · electrochemistry · sodium

[1] A. Bills, S. Sripad, W. L. Fredericks, M. Singh, V. Viswanathan, *ACS Energy Lett.* **2020**, *5*, 663.

[2] J. Kersey, N. D. Popovich, A. A. Phadke, *Nat. Energy* **2022**, *7*, 664.

[3] A. Masias, J. Marcicki, W. A. Paxton, *ACS Energy Lett.* **2021**, *6*, 621.

[4] C. M. Tan, Y. Yang, K. J. M. Kumar, D. D. Mishra, T.-Y. Liu, *Sci. Rep.* **2024**, *14*, 10126.

[5] X. Cai, Y. Yue, Z. Yi, J. Liu, Y. Sheng, Y. Lu, *Nano Energy* **2024**, *129*, 110052.

[6] H. Laufen, S. Klick, H. Ditler, K. L. Quade, A. Mikitisin, A. Blömeke, M. Schütte, D. Wasylowski, M. Sonnet, L. Henrich, A. Schwedt, G. Stahl, F. Ringbeck, J. Mayer, D. U. Sauer, *Cell Rep. Phys. Sci.* **2024**, *5*, 101945.

[7] A. Rudola, R. Sayers, C. J. Wright, J. Barker, *Nat. Energy* **2023**, *8*, 215.

[8] K. Bischof, V. Marangon, M. Kasper, A. A. Regalado, M. Wohlfahrt-Mehrens, M. Hölzle, D. Bresser, T. Waldmann, *J. Power Sources Adv.* **2024**, *27*, 100148.

[9] L. Zhao, T. Zhang, W. Li, T. Li, L. Zhang, X. Zhang, Z. Wang, *Engineering* **2023**, *24*, 172.

[10] Y.-F. Feng, J.-N. Shen, Z.-F. Ma, Y.-J. He, *J. Energy Storage* **2021**, *43*,103233.

[11] Y. Zhang, W. Song, S. Lin, Z. Feng, *J. Power Sources* **2014**, *248*, 1028.

[12] K. Ng, C. Moo, Y.-p. Chen, Y. Hsieh, *Appl. Energy* **2009**, *86*, 1506.

[13] C. Lin, H. Mu, R. Xiong, W. Shen, *Appl. Energy* **2016**, *166*, 76.

[14] S. Yang, S. Zhou, Y. Hua, X. Zhou, X. Liu, Y. Pan, H. Ling, B. Wu, *Sci. Rep.* **2021**, *11*, 5805.

[15] Q. Wang, J. Wang, P. Zhao, J. Kang, F. Yan, C. Du, *Electrochim. Acta* **2017**, *228*, 146.

[16] C. Campestrini, T. Heil, S. Kosch, A. Jossen, *J. Energy Storage* **2016**, *8*, 142.

[17] J. Zhu, Y. Wang, Y. Huang, R. Bhushan Gopaluni, Y. Cao, M. Heere, M. J. Mühlbauer, L. Mereacre, H. Dai, X. Liu, A. Senyshyn, X. Wei, M. Knapp, H. Ehrenberg, in *Nat. Commun.* **2022**, *13*, 2261.

[18] M. Dubarry, N. Costa, D. Matthews, in *Nat. Commun*. **2023**, *14*, 3138.

[19] D. Roman, S. Saxena, V. Robu, M. Pecht, D. Flynn, in *Nat. Mach. Intell.* **2021**, *3*, 447.

[20] S. Khaleghi, D. Karimi, S. H. Beheshti, M. S. Hosen, H. Behi, M. Berecibar, J. Van Mierlo, *Appl. Energy* **2021**, *282*, 116159.

[21] M. A. Hannan, M. S. H. Lipu, A. Hussain, P. J. Ker, T. M. I. Mahlia, M. Mansor, A. Ayob, M. H. Saad, Z. Y. Dong, *Sci. Rep.* **2020**, *10*, 4687.

[22] W. Li, M. Rentemeister, J. Badeda, D. Jöst, D. Schulte, D. U. Sauer, *J. Energy Storage* **2020**, *30*, 101557.

[23] N. Ghaeminezhad, Q. Ouyang, J. Wei, Y. Xue, Z. Wang, *J. Energy Storage* **2023**, *72*, 108707.

[24] W. Waag, C. Fleischer, D. U. Sauer, *J. Power Sources* **2014**, *258*, 321.

[25] K. L. Quade, D. Jöst, D. U. Sauer, W. Li, *Batteries Supercaps 6*, e202300152.

[26] C. Campestrini, M. F. Horsche, I. Zilberman, T. Heil, T. Zimmermann, A. Jossen, *J. Energy Storage* **2016**, *7*, 38.

[27] M. W. Liemohn, A. D. Shane, A. R. Azari, A. K. Petersen, B. M. Swiger, A. Mukhopadhyay, *J. Atmos. Sol. Terr. Phys.* **2021**, *218*, 105624.

[28] S. Tewiele, *Development of representative driving and load cycles based on real-world driving data of battery electric vehicles* **2020**, https://doi.org/10.17185/duepublico/72728.

[29] F. Berger, D. Joest, E. Barbers, K. Quade, Z. Wu, D. U. Sauer, P. Dechent, *eTransportation* **2024**, *22*, 100355.

[30] J. Klee Barillas, J. Li, C. Günther, M. A. Danzer, *Appl. Energy* **2015**, *155*, 455.

[31] R. Guo, C. Hu, W. Shen, *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 15131.

[32] M. Saeed, S. Lu, Z. Song, X. Hu, *IEEE Trans. Power Electron.* **2024**, *39*, 8813.

[33] M. A. Awadallah, B. Venkatesh, *J. Energy Storage* **2016**, *6*, 95.

[34] ISEA / BMS Simulation as MIL, https://git.rwth-aachen.de/isea/bms-simulation-as-mil (accessed: June 2025).

[35] S. Bihn, J. Rinner, H. Witzenhausen, F. Krause, F. Ringbeck, D. U. Sauer, *Batteries* **2024**, *10*,314.

[36] S. Waldhoer, S. Bockrath, M. Wenger, R. Schwarz, V. R. H. Lorentz, *Renewable Energy Energy Manage,* **2020**, 1.

[37] M. Lelie, T. Braun, M. Knips, H. Nordmann, F. Ringbeck, H. Zappen, D. U. Sauer, *Appl. Sci.* **2018**, *8*, 534.

[38] E. Barbers, F. E. Hust, F. E. A. Hildenbrand, F. Frie, K. L. Quade, S. Bihn, D. U. Sauer, P. Dechent, *J. Energy Storage* **2024**, *84*, 110851.

[39] J. Schmalstieg, S. Käbitz, M. Ecker, D. U. Sauer, *J. Power Sources* **2014**, *257*, 325.

[40] F. Hust, C. Leroi, P. Sabet, H. Witzenhausen, F. Frie, S. Zappulla, C. Abele, A. Hase, C. Abele, E. Barbers, *Isea Framework* **2023**, https://git.rwth-aachen.de/isea/framework.

[41] S. Bihn, in *Automatic Parameterisation of Electrical Equivalent Circuit Models for Virtual Battery Cell Design*, RWTH Aachen University **2024**, https://doi.org/10.18154/RWTH-2024-10636.

[42] IVT-s Current Sensor Datasheet, https://www.isabellenhuette.com/solutions/products/ivt-s (accessed: June 2025).

[43] R. Singh, S. Das, S. Samanta, in *2024 Third Int. Conf. on Power, Control and Computing Technologies (ICPC2T)*, IEEE, Raipur, India, January **2024**, pp. 151–156, https://doi.org/10.1109/ICPC2T60072.2024.10474776.

[44] A. Wadi, M. Abdel-Hafez, A. A. Hussein, *Energies* **2022**, *15*, 3717.

[45] N. Wassiliadis, J. Adermann, A. Frericks, M. Pak, C. Reiter, B. Lohmann, M. Lienkamp, *J. Energy Storage* **2018**, *19*, 73.

[46] C. Campestrini, *Practical Feasibility of Kalman Filters for the State Estimation of Lithium-ion Batteries* **2018**, https://mediatum.ub.tum.de/1362581.

[47] Z. Chen, N. Ahmed, S. Julier, C. Heckman (Preprint), arXiv:1912.08601, v1, Submitted: Dec. **2019**, https://doi.org/10.48550/arXiv.1912.08601.

[48] Z. Fan, D. Shen, Y. Bao, K. Pham, E. Blasch, G. Chen, in *2024 27th Int. Conf. on Information Fusion (FUSION)*, IEEE, Venice, Italy, July **2024**, pp. 1–8, https://doi.org/10.23919/FUSION59988.2024.10706523.

[49] Getting Started with impedance.py impedance.py 1.7.1 Documentation, https://impedancepy.readthedocs.io/en/latest/getting-started.html (accessed: June 2025).

[50] R. Guo, W. Shen, *IEEE Trans. Ind. Electron.* **2023**, *70*, 10123.

[51] J. M. G. Sopeña, V. Pakrashi, B. Ghosh, *Sustainable Energy Technol. Assess.* **2023**, *57*, 103246.

[52] K. A. Maupin, L. P. Swiler, N. W. Porter, *J. Verif. Valid. Uncert. Quantif.* **2019**, *3*, 031002.

[53] C. J. Willmott, *Phys. Geogr.* **1981**, *2*, 184.

[54] A. Geslin, L. Xu, D. Ganapathi, K. Moy, W. C. Chueh, S. Onori, *Nat. Energy* **2024**, *10*, 172.

[55] L. K. Willenberg, P. Dechent, G. Fuchs, D. U. Sauer, E. Figgemeier, *Sustainability* **2020**, *12*, 557.

[56] T. Bank, S. Klamor, D. U. Sauer, *J. Energy Storage*, **2020**, *30*, 101465.

[57] L. Wang, X. Zhao, Z. Deng, L. Yang, *J. Energy Storage* **2023**, *57*, 106275.

[58] LTC6804-1 Datasheet and Product Info | Analog Devices, https://www.analog.com/en/products/ltc6804-1.html (accessed: June 2025).

Manuscript received: June 13, 2025
Revised manuscript received: August 16, 2025
Version of record online: