Annals of Operations Research
https://doi.org/10.1007/s10479-026-07342-3

ORIGINAL RESEARCH

Approximating end-of-life electric vehicle battery flows: surrogate modeling of a discrete event and agent-based simulation model

Sandra Huster[1,2] · **Mona Faraji-Niri**[2,3] · **Andreas Rudi**[1] · **Frank Schultmann**[1]

Received: 11 July 2025 / Accepted: 26 June 2026

Abstract
Decision support tools are essential for planning end-of-life pathways of electric vehicle (EV) batteries, including recycling, repurposing, and reconditioning. As large volumes of EV batteries reach end-of-life in the coming decades, estimating material flows under different scenarios becomes increasingly important. Detailed simulation models, such as those based on discrete event and agent-based modeling, can capture the complex dynamics of battery life cycles. However, their high computational cost limits their applicability in decision-making contexts that require extensive scenario analysis. To overcome this challenge, machine learning-based surrogate models are evaluated for approximating the outputs of a simulation model that estimates future end-of-life quantities of EV batteries. The simulation model uses static input configurations and produces time-dependent outputs, for which no historical time series exist. Two surrogate modeling strategies are compared: an all-at-once approach that predicts all time steps simultaneously, and a recursive approach that predicts them sequentially. Several machine learning algorithms are evaluated, including neural networks, Gaussian process regression, random forests, and XGBoost. Results show that neither approach consistently outperforms the other, and model performance varies strongly across key performance indicators. While surrogate models can substantially accelerate analysis, predictive accuracy is limited for outputs with sparse or highly variable data. The results underline the importance of selecting surrogate modeling approaches based on the specific characteristics of the output indicators and the analytical needs of the decision context.

Keywords Discrete event simulation · Electric vehicle batteries · Remanufacturing · Metamodeling · Emulation · Time-dependency

✉ Sandra Huster
sandra.huster@kit.edu

1 Institute for Industrial Production, Karlsruhe Institute of Technology, Hertzstr. 16, 76187 Karlsruhe, Germany

2 Warwick Manufacturing Group, University of Warwick, Coventry CV4 7AL, UK

3 The Faraday Institution, Harwell Science and Innovation Campus, Quad One, Didcot, UK

Published online: 12 July 2026

Abbreviations

aao	All-at-once estimation
ABM	Agent-based modeling
ALAMO	Automated Learning of Algebraic Models
ANN	Artificial Neural Network
ARIMA	Autoregressive Integrated Moving Average
DT	Decision Tree
EoL	End-of-Life
EV	Electric vehicle
EVB	Electric vehicle battery
GBT	Gradient-Boosted Trees
GPR	Gaussian Process Regression
gpr_stand	Gaussian Process Regression with a Constant Kernel combined with a Radial Basis Function Kernel
gpr_matern	Gaussian Process Regression with a Matern Kernel
gpr_ratQuad	Gaussian Process Regression with a Rational Quadratic Kernel
gpr_comb	Gaussian Process Regression with a composite Kernel integrating Constant, RBF, and Matern kernels
KPI	Key performance indicator
LHS	Latin Hypercube Sampling
LSTM	Long Short-Term Memory
MARS	Multivariate Adaptive Regression Splines
ML	Machine Learning; M&S: Modeling and Simulation
MSE	Mean Square Error
NCO	Nickel–Cobalt-Oxide
NCx	NCO or NMC battery
NMC	Nickel-Manganese-Cobalt
OEM	Original Equipment Manufacturer
RBF	Radial Basis Function
rec	Recursive estimation
RF	Random Forest
RMSE	Root Mean Square Error
SVM	Support Vector Machine
XGB	XGBoost

1 Introduction

The number of end-of-life (EoL) electric vehicle batteries (EVBs) that will become available for recycling, repurposing, or reconditioning in the coming years depends on numerous factors, such as the development of electric vehicle (EV) sales, technological advances in EoL pathways, and decisions by stakeholders along the value chain (Kadner et al., 2021). Recycling refers to recovering material; repurposing denotes reusing the battery in an application other than the original one, i.e., an application outside the vehicle; and reconditioning means that the battery is restored to a state suitable for usage in an EV as a spare battery (DeRousseau et al., 2017; European Commission, 2008; Potting et al., 2017). While recycling and repurposing are already practiced to some extent (Faessler, 2021; Slattery et al., 2024),

reconditioning is currently only carried out on a pilot scale (e.g., Stellantis, 2023). To predict EoL quantities, material flow analyses are commonly used. These estimate EVB returns based on EV sales data and assume that a certain share of batteries is suitable for repurposing or reconditioning (Sanclemente Crespo et al., 2022; L. Yang et al., 2024). More advanced simulation models include effects of legislative measures or market dynamics such as supply-driven price effects (Pratap et al., 2024; Seika & Kubli, 2024). These models allow for more realistic and detailed representations of the EVB value chain but come with longer runtimes. One such model is a discrete event and agent-based simulation model by Huster et al. (2025), which uses 49 input parameters, including expected EV and EVB lifetimes, future EV sales, and assumptions on the profitability of EoL pathways. Based on these inputs, the model estimates quantities of returned, recycled, repurposed, reconditioned, and demanded EVBs for the period from 2025 to 2050 in five-year intervals, resulting in 30 output variables. With the large number of input factors and a runtime of about ten minutes per simulation run, extensive sensitivity studies are hardly feasible. For instance, testing only three levels per input factor would require around 817 days of computation time, not including replications to account for stochasticity.

To address such limitations, it has been proposed to approximate the input–output behavior of simulation models using simplified models, often referred to as surrogate models, response surfaces, metamodels, or emulators (McBride & Sundmacher, 2019; Pietzsch et al., 2020). These terms originate from different research communities, such as response surface methodology in statistics (Box & Wilson, 1951, as cited by Khuri & Cornell, 1987) and simulation metamodeling in operations research (Barton, 1992; Kleijnen, 1982), but they generally describe models that approximate the behavior of experiments or simulations. In this study, the term surrogate model is used as a generic term for such approximations. The basic idea of those models is that, first, sample points are generated with the original simulation model as training data, then a surrogate model is fitted, and finally, the surrogate model is evaluated (Thombre et al., 2015). Those surrogates can be created with various techniques, for example, artificial neural networks (ANNs), Gaussian process regression (GPR) or random forests (RFs) (Angione et al., 2022). The model selection depends on many factors, including the sampling size for training and the number of input factors (Williams & Cremaschi, 2021). Furthermore, the type of input and output data is crucial. While some methods are primarily used for static outputs, others, like long short-term memory (LSTM), are more suitable for time series output (Zhao et al., 2023). Surrogate modeling has been applied to different domains, from chemical process engineering to social studies (Angione et al., 2022; McBride & Sundmacher, 2019), for agent-based simulations, finite element simulations and techno-economic analyses (Hürkamp et al., 2021; ten Broeke et al., 2021; Wiesberg et al., 2021).

Typically, it is evident whether data is of a time series nature or not. Cerqueira et al. (2020) define time series as "a temporal sequence of values $Y = \{y_1, y_2, \ldots, y_t\}$, where y_i is the value of Y at a time i and t is the length of Y", and time series forecasting as "the task of predicting the next value of the time series, y_{t+1}, given the previous observations of Y". Consequently, time series prediction always relies on previous time series values, potentially augmented by additional data. In emerging markets, however, predicting product sales or product return quantities must be conducted without historical precedents. Predictions must rely on static input data, such as product lifetimes and estimated consumer preferences, while the predictions are of a time series nature in that they are sequential and time-dependent. Therefore, approximating a simulation that predicts EVB recycling, repurposing, and reconditioning quantities comes with particularities:

- The input data, such as EVB lifetimes and consumer preferences, are static, but the outputs are time-related.
- The time-related outputs have no historical predecessors due to the novelty of the EV market.
- The data originate from a simulation model, allowing for repeated sampling.

It remains a gap in research which type of surrogate model and modeling technique is the most appropriate for this type of problem. While the characteristics of EoL EVB forecasting seem quite particular, the general issue of time-dependent outputs without predecessors is likely to appear in other circular economy domains. For instance, currently only 1% of used textiles in the EU are recycled, but the share is expected to rise with mandatory separate collection starting in January 2025 (EEA, 2022). Therefore, methods for approximating simulation models with time-dependent outputs but no historical time series may gain relevance beyond the EV sector.

More generally, this type of approach can be understood as part of a broader research trend that combines modeling and simulation (M&S) with data-driven techniques. In recent years, growing attention has been given to hybrid modeling approaches that integrate simulation models with machine learning methods in order to improve computational efficiency, enable faster scenario exploration, or support decision-making (Schweidtmann et al., 2024; Von Rueden et al., 2020). Such approaches combine the strengths of both paradigms: simulation models represent structural relationships and system behavior (Law, 2015), while data-driven models can approximate complex input–output relationships based on generated or observed data (Goodfellow et al., 2016).

The aim of this study is to reduce order of the EoL EVB simulation model by Huster et al. (2025), making it more agile and practically applicable. As the methodology, different machine learning techniques are evaluated for their suitability to predict the time-dependent EoL EVB output. One strategy treats the output as static and predicts all values at once. Presumably, that might lead to inaccurate results because it might omit trends in the output time series. Alternatively, a recursive approach is applied that accounts for the temporal structure by training one model per time step, using previously predicted values as additional inputs. The performance of the approaches is compared by common metrics like R^2 and (root) mean square error ((R)MSE), and the training time.

By combining a discrete event and agent-based simulation model of EV battery end-of-life flows with machine-learning-based surrogate models, the study contributes to research at the intersection of modeling and simulation and data science. In particular, it provides insights into how surrogate modeling strategies can be used to approximate simulation models that produce time-dependent outputs but lack historical time series observations.

The remainder of the study is structured as follows: First, a literature overview of methods for approximating simulation models is given in Sect. 2. Subsequently, in Sect. 3, the EoL EVB simulation model that shall be approximated is introduced, and the selection of the surrogate models is outlined. In Sect. 4, the metrics of the selected surrogate models are presented and compared, followed by a discussion of the results in Sect. 5. Finally, conclusions are drawn in Sect. 6.

2 Literature review

Computer simulation has long been used to represent simplified abstractions of real-world or conceptual systems in order to study their behavior under different conditions (Banks,

1998; Shannon, 1975). As systems become more complex and involve stochastic behavior, nonlinear interactions, or large numbers of components, simulation provides a means of exploring system dynamics and interactions (Banks, 1998). With the development of digital computing in the second half of the twentieth century, simulation methods became increasingly powerful and widely applied (Goldsman et al., 2009). Today, simulations are used to represent phenomena ranging from natural and physical processes, e.g., climate systems, to socio-technical systems such as manufacturing, logistics, and energy networks (Law, 2015; Obaidat & Papadimitriou, 2003).

Over time, a variety of modeling paradigms have emerged, including discrete-event simulation, system dynamics, agent-based modeling, Monte Carlo simulation, and physics-based simulation approaches such as finite element models (Cheong & Butscher, 2019; Law, 2015). The increasing availability of computational resources has enabled simulation models to represent systems in greater detail and across larger spatial and temporal scales (Pan et al., 2025). However, this increased realism has also led to growing computational demands, particularly when large parameter spaces must be explored for sensitivity analysis, scenario evaluation, or optimization (X.-S. Yang et al., 2014).

Within the simulation literature, considerable research has focused on techniques for analyzing and approximating simulation models. In particular, the concept of simulation metamodels has been developed to describe statistical or machine learning models that approximate the input–output behavior of computer experiments (Sacks et al., 1989). Such surrogates have been applied to various simulation types, including agent-based models (ABMs), discrete event simulations, finite element models, and process chains (Hürkamp et al., 2021; ten Broeke et al., 2021). For instance, Angione et al. (2022) compared eight machine learning (ML) methods, including ANN, GPR, support vector machine (SVM), and gradient-boosted trees (GBTs), to approximate the outputs of an ABM simulating social care provision. They found that ANNs performed best in terms of predictive accuracy, especially for larger datasets, while GBTs were more efficient to train. Hürkamp et al. (2021) developed surrogates for a high-dimensional finite element method and a discrete process chain simulation using RFs and decision trees (DTs). In both cases, RFs performed slightly better in terms of prediction accuracy, while maintaining relatively low training times. In groundwater management, Christelis et al. (2019) compared single and multiple surrogate models for optimization of pumping strategies. They found that while ensembles slightly outperformed single surrogates, simple models like cubic radial basis functions (RBFs) and Gaussian Kriging also delivered good solutions with minimal computational overhead. Stevanović et al. (2024) contrasted XGBoost (XGB) and AANs in the context of building energy simulations. While AANs had slightly better predictive accuracy, XGB was more efficient and required less tuning. Mukhtar et al. (2023) explored seven surrogate methods for aerospace design and found that k-nearest neighbors and RF offered high accuracy with low training costs, while GPR provided good uncertainty estimates but longer runtimes.

In the context of simulation-based optimization, Williams and Cremaschi (2021) evaluated eight surrogate modeling techniques on benchmark functions with different shapes and input dimensions. Automated learning of algebraic models (ALAMO), multivariate adaptive regression splines (MARS), and GPRs were most accurate for surface approximation, while RFs and SVMs were more reliable for identifying optimal values. The results depended on the shape of the functions, the sample size and the input dimensionality. Wang et al. (2014) similarly demonstrated that model performance depended on both sample size and problem structure, with GPR and SVMs generally outperforming ANNs on their benchmark problems.

Also, Bashiri et al.'s (2022) study on surrogate model and sampling combinations on a multi-objective car front crash scenario supports that no single method consistently outperforms others.

These examples illustrate that there is no universally best surrogate modeling technique. Instead, performance depends on factors such as input and output dimensionality, the size and structure of the training data, the nature of the simulated system, and the modeling goal (e.g., prediction or optimization) (McBride & Sundmacher, 2019; Pietzsch et al., 2020; Williams & Cremaschi, 2021). This observation is reinforced by several overview studies. McBride and Sundmacher (2019) summarized surrogate modeling in chemical engineering and emphasized the importance of selecting models based on the application context. Pietzsch et al. (2020) reviewed surrogate modeling for ABMs and identified model selection, uncertainty handling, and implementation effort as key trade-offs. They concluded that GPR models perform well for prediction tasks but may require more effort than simpler ML models like RFs. Trinchero and Canavero (2021) further found that model performance varied with noise levels and output structure, with SVM and GPR showing strong performance across applications. The nature of the data, whether static or time-related, also influences model choice. While most surrogate models are trained on static input and output values, recent work has explored time series-based modeling (e.g., Dearing, 2024; Zhao et al., 2023), particularly using deep learning methods like LSTM and convolutional neural network-LSTM architectures.

The present study is related to several strands of the surrogate modeling literature. Like Hürkamp et al. (2021), it approximates an agent-based and discrete event simulation model using static inputs, and like Angione et al. (2022), it focuses on predictive performance. The present study builds on the insights from those studies but addresses two underexplored challenges in combination. First, the output of the EoL EVB simulation model consists of sequences of future quantities predicted for multiple time points. However, unlike classical time series forecasting, there are no historical values of the output variables to use as inputs. As a result, no time series-specific methods such as recurrent neural networks or autoregressive integrated moving average (ARIMA) models can sensibly be applied (Hyndman & Athanasopoulos, 2021). Instead, standard machine learning techniques are used to either predict all time steps at once or recursively: first predicting the outputs for the initial period, then feeding those predictions into the model to predict subsequent periods. To the authors' knowledge, it remains unclear how well such recursive strategies perform when no actual predecessors exist, and all temporal structure must be inferred from static input features. Second, this study addresses a multi-output prediction problem. Rather than focusing on a single key performance indicator (KPI) or using aggregated outcomes over time, multiple KPIs are predicted simultaneously, each disaggregated over five future time points. This increases the modeling complexity substantially. Angione et al. (2022) have identified multi-output prediction as a research gap in the context of surrogate modeling for ABMs. By exploring both an all-at-once and a recursive approach in this setting, this study contributes to closing this gap and offers guidance for applications where simulation outputs are time-related but not based on sequential observations.

3 Material and methods

This section begins with a description of the simulation model that serves as the basis for surrogate modeling. The overall methodological procedure is then outlined (cf. Figure 1). 1First, the experimental design is defined, and simulation runs are carried out accordingly.

Fig. 1 Outline of the procedure

The resulting data is then preprocessed to prepare it for machine learning modeling. The final step, which is analyzing and comparing the performance of the surrogate models, is addressed in the subsequent results section.

3.1 Simulation model for approximation

The simulation model to be approximated is introduced in Huster et al. (2025) and briefly summarized in the following. It is a mixed-methods model consisting of discrete event and agent-based elements (cf. Figure 2). The model takes a vector of 149 parameters as an input. This vector contains assumptions on EV sales numbers from 2015 to 2050, on the lifespan of EVs and EVBs, on recycling, repurposing and reconditioning costs, recycled content requirements and other aspects. As EV adoption is beyond the scope of the simulation study, external projections of EV sales are used as input parameters. Starting the simulation in the past (2015) acts as a warm-up period and ensures that the model is populated with a realistic stock of EVs at the beginning of the analyzed period. The time horizon until 2050 was chosen because EVs and their batteries have long lifetimes and circular activities such as repurposing and recycling occur with considerable delays. In addition, many policy targets and regulatory frameworks in the energy and mobility sector, including EU climate targets, are defined with a horizon up to 2050 (European Commission, n.d.). A complete table of the inputs is included in the Supplementary Material. The discrete event logic depicts the material flow of EVs and EVBs through the lifecycle stages of production, usage, rebuilding, repurposing and finally recycling. The agent-based elements account for the behavior of consumers, business entities and legislative bodies. Consumer behavior is

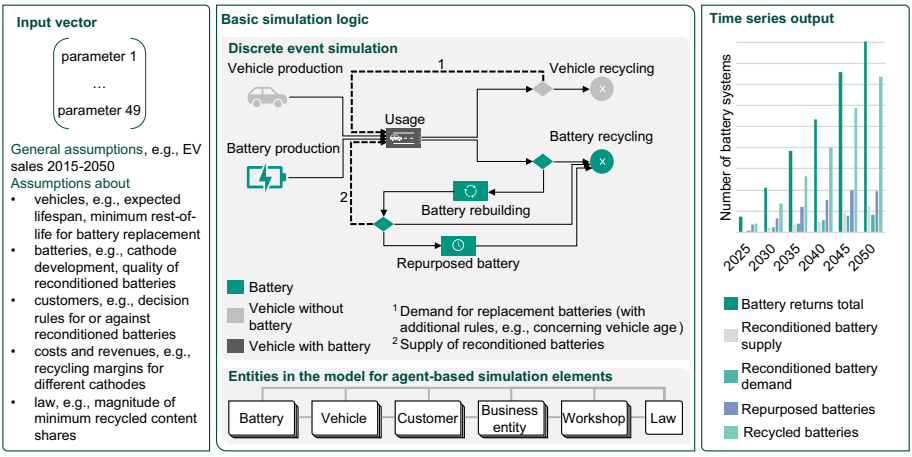

Fig. 2 General simulation logic (reprinted from Huster et al. (2025) with minor adjustments to the arrow labels, License: CC BY 4.0)

represented through decisions regarding battery replacement in case of battery failure. Based on a survey with 837 complete responses (Huster et al., 2024b), utility functions were derived for the options of replacing the battery with a new battery, replacing it with a reconditioned battery, or replacing the entire vehicle. These utility functions depend on the vehicle and battery price, as well as the expected lifespan and environmental impact of the respective battery option. Business entities, mainly OEMs and car workshops, are modeled as decision-makers primarily driven by economic considerations, which were derived from 19 interviews and a survey with 58 participants. These considerations include, for instance, reconditioning costs and warranty periods. The legislative actor considered in the model is the EU through the EU Battery Regulation, thereby setting the regional focus to the EU. The regulation requires minimum recycled content quotas in new batteries from 2031 onwards (European Parliament and European Council, 2023), and the simulation model allows these percentages to be varied. Hence, the stakeholder behavior and key model parameters are grounded in empirical data collected through surveys and interviews, as well as in European legislation, thereby supporting the face validity and realism of the model. More detailed information about the stakeholders and their modeling can be found in Huster et al. (2025). The outputs of the model are the numbers of returning batteries, reconditioned battery supply, reconditioned battery demand, repurposed and recycled batteries, each of the KPIs for the years 2025, 2030, 2050. The intermediate years can be interpolated. To further validate the model, the assumptions and simulation results were presented and discussed in an expert workshop involving 12 participants from industry, industry associations, and academia. The experts confirmed the overall plausibility of the modeled mechanisms and identified no major conceptual gaps. In addition, key trends observed in the simulation results are consistent with findings reported in related studies (e.g., Kastanaki and Giannis (2023) on the supply of reconditioned batteries; Engel et al. (2019) on the supply of repurposed batteries).

The model is implemented using the simulation software AnyLogic 8.9.3 University. The duration of each simulation run depends on the input parameters, particularly concerning the projected EV sales from 2015 to 2050. On average, a simulation run is expected to take approximately ten minutes on a workstation equipped with an Intel Core i9-14900 K processor (24 cores, 32 threads, 3.2 GHz) and 192 GB of RAM. Due to the stochastic nature of the model, multiple replications are required to reduce stochastic variation in the output KPIs. In this study, three replications were conducted for each input vector, resulting in an evaluation time average of 30 min per vector. This choice represents a compromise between reducing stochastic variability and maintaining computational feasibility. Furthermore, each simulation run represents the lifecycle of thousands of EVs and EV batteries, meaning that stochastic elements such as battery lifetimes are sampled many times within a single run, which contributes to stabilizing the aggregated KPIs.

3.2 Design-of-experiments

To generate a dataset for training and testing surrogate models, various experimental design approaches can be considered, such as (fractional) factorial designs or space-filling designs. The choice of design should take into account the specific problem, the number of input factors, and the expected structure of the data (Montgomery, 2012). For example, the objective may be to screen influential factors or to optimize inputs toward a predefined target, and factor effects may be assumed to be linear or nonlinear depending on the modeling context (Montgomery, 2012). In the case of generating training data for surrogate models based on the discrete event and agent-based simulation described in Sect. 3.1, the relationships

between the 49 input factors and the multiple output indicators are not known in advance, and linearity, as assumed in 2^k-factorial designs, cannot be assumed a priori. Following the recommendation by Santner et al. (2018) that *"[b]ecause one does not know the true relationship between the response and inputs, designs should allow one to fit a variety of models and should provide information about all portions of the experimental region*", a space-filling design is chosen. This aligns with findings by Choi et al. (2021), who identify Latin Hypercube Sampling (LHS), a space-filling design, as more effective than factorial designs for metamodel construction. Among space-filling strategies such as LHS, Sobol, and Halton sampling, Williams and Cremaschi (2021) report minimal approximation performance differences across designs. Therefore, LHS was selected for this study. LHS allows comprehensive coverage of the input space and is well-suited for high-dimensional problems (Huntington & Lyrintzis, 1998; Santner et al., 2018). Specifically, a maximin LHS design was selected to maximize the minimum distance between points (Sheikholeslami & Razavi, 2017). This design was created using Matlab's "lhsdesign" function (The MathWorks, Inc., n.d.-a). For each simulation run, one value is randomly sampled from a predefined range of each input parameter. The resulting set of sampled values forms a parameter vector that is used for one simulation run. Thus, the parameter ranges are not varied within a simulation run but define the domain from which the input values are sampled across runs. The ranges of the values for the 49 parameters are detailed in the Supplementary Information. Following the guideline by Loeppky et al. (2009), which recommends using at least ten times as many experiment runs n as input dimensions d ($n = 10d$), 490 design points were generated for 49 input parameters. To account for stochasticity in the simulation model, each configuration was repeated three times, resulting in a total of 1,470 simulation runs. The computing time was approximately eleven days. KPIs of the replicated simulation runs were averaged, resulting in a dataset with 490 observations for training and testing.

3.3 ML modeling and algorithm

As delineated in Sect. 1, various methodologies for approximating the input–output relationship of the EVB EoL simulation model are evaluated. These methodologies include:

1. Treating both input and output features as static, whereby all output features are predicted simultaneously using machine learning techniques (case name: all-at-once (aao)).
2. Conducting multiple one-step-ahead forecasts, in which several static models are trained, each predicting one timestep ahead, thereby incorporating the temporal component (case name: recursive (rec)).

In all methodologies, the data is preprocessed by splitting it into training and testing sets, and by scaling the input features. Splitting is done using an 80:20 ratio, with 80% of the 490 instances allocated for training and 20% for testing. This approach is widely adopted (e.g., Hürkamp et al., 2021; Nakano & Liu, 2025; Pietukhov et al., 2023). To enhance the reliability of the metrics derived from the model testing, the procedure of data splitting, model training on the 80% training subset, testing on the 20% testing subset, and metric acquisition is conducted thrice. Subsequently, the metrics from each independent train-test iteration are averaged. This approach is similar to Wang et al. (2019), who trained three models on the same train-test split to evaluate the reliability of predictions made with the approximated models. For one of the three 80:20 splits, the histograms of the training and testing data for each output parameter are depicted in the Supplementary Information. The metrics monitored include the R^2 value, MSE, RMSE, and computational time. These metrics are then averaged across the three model replications. For each split, the input and output features are scaled based on the

training data to facilitate accurate model estimation. Given that no single scaling method is universally optimal (De Amorim et al., 2023; Mueller et al., 2023), various methods including the StandardScaler, the QuantileTransformer and a log transformation were evaluated. The StandardScaler produced the most favorable results concerning R^2.

3.3.1 All-at-once estimations (aao)

When considering the output features as static and estimating them simultaneously (cf. Figure 3), a variety of machine learning methods can be employed, such as ANNs, GPR or RFs (Angione et al., 2022). To obtain a preliminary assessment of the suitability of different models for the given dataset, a pre-screening was conducted using MATLAB's built-in Regression Learner (The MathWorks, Inc., n.d.-b). 28 models were estimated for each of the KPIs for a single timestep (e.g., "Battery returns total" for 2025, "Recycled batteries" for 2030, "Reconditioned batteries" for 2035, etc.). In terms of both R^2 and RMSE, various types of ANNs consistently demonstrated the best fit, followed by GPRs with different kernels and tree-based models. SVMs were found to be the least suitable models.

Following the preliminary screenings, ANNs, GPR, RFs and XGB were selected as the models under investigation. For comparative purposes with a simpler model, a DT model was also evaluated. Except for XGB, all estimations utilized the scikit-learn libraries (Pedregosa et al., 2011). Specifically, the MLPRegressor was employed for ANNs, the GaussianProcessRegressor for GPRs, the RandomForestRegressor for RFs, and the DecisionTreeRegressor for DTs. The XGBRegressor was used for XGBs (xgboost developers, 2022). Regarding other model parameters, the following configurations were examined:

ANN: The size of the dataset is comparably small with 490 input–output combinations. On small datasets, overly complex models can degrade model performance (Goodfellow et al., 2016). Consequently, this study evaluates simple architectures featuring two, three, and four hidden layers, initially configured with 128, 64, and 32 neurons, respectively, with the number of neurons per layer being halved in subsequent configurations. For example, in the case of a two-layer ANN, the architectures (128, 64), (64, 32) and (32, 16) are examined. The activation functions tested include all those available in the MLPRegressor library, specifically *identity*, *logistic*, *tanh* and *relu*. Similarly, all implemented solvers (*lbfgs*, *sgd* and *adam*) are used.

GPR: The primary parameter of GPR is the selection of the kernel, which defines the covariance function of the Gaussian Process (scikit-learn, n.d.). The kernels evaluated include: (1) a constant kernel combined with an RBF kernel, serving as a straightforward default configuration (abbreviation: gpr_stand); (2) a Matern kernel, suitable for high-dimensional input spaces (abbreviation: gpr_matern); (3) a rational quadratic kernel, offering flexibility in length scales (abbreviation: gpr_ratQuad); and (4) a composite kernel integrating constant, RBF, and Matern kernels to capture more complex data patterns (abbreviation: gpr_comb) (Eric P. Natsume, 2022; Xing et al., 2015).

RF and DT: For both RF and DT estimation, the squared error is employed as the criterion for assessing quality. In the case of RF, it is also necessary to specify the number of trees within the forest. Scikit-learn's default value 100 is tested as well as the doubled quantity, i.e., 200.

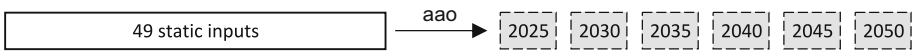

Fig. 3 Schematic representation of the aao logic

XGB: Given that multi-output regression remains in an experimental phase in XGBoost 3.0 (xgboost developers, n.d.), a separate model is constructed for each KPI and period using Scikit-learn's MultiOutputRegressor. This approach remains consistent with the other models, as only the 49 input features are utilized for fitting each model.

3.3.2 Recursive estimations (rec)

The recursive estimations are designed to identify patterns in the output data by training six models (rec 1 to rec 6), each tasked with predicting the KPIs for specific timesteps: 2025, 2030, 2035, 2040, 2045 and 2050. Each model for a given timestep t utilizes the 49 static input features and the *simulated* time-dependent features up to $t-1$ for training. For testing, the 49 static input features are used and the *predicted* time-dependent features up to $t-1$. The architecture is schematically illustrated in Fig. 4. Note that the same 80:20 split is used for fitting all models rec 1 to rec 6, ensuring no data leakage occurs between models for each timestep. The same ML methodologies as those used in the aao approach are evaluated.

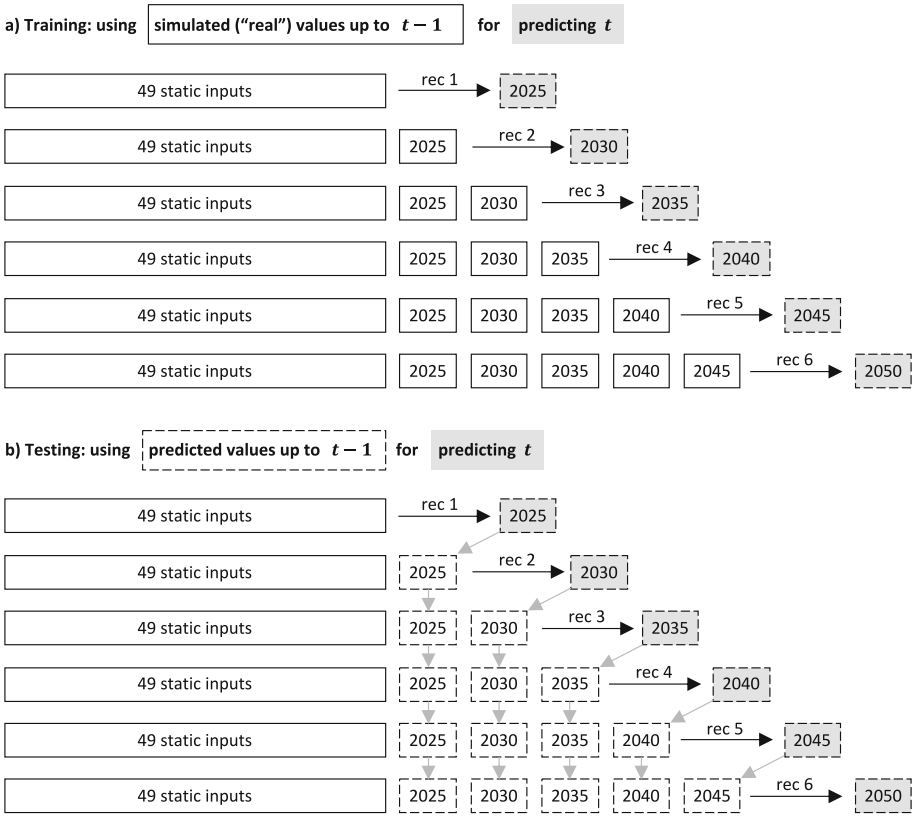

Fig. 4 Schematic representation of the rec logic

4 Results

This section presents a comparative analysis of the results obtained from the aao and rec approaches, evaluated at an aggregated level across all KPIs and time periods, as well as at the level of individual KPIs or periods. Subsequently, the feature importance for each KPI within each period is assessed using an XGB model. These results are then employed to estimate new surrogate models, utilizing only the ten most significant features for each KPI.

4.1 Comparisons and analysis of the aao and rec approaches

A total of 232 models were fitted, with 116 utilizing the aao approach and 116 employing the rec approach. In 50% of the instances, the aao models demonstrated higher R^2 values, while in the remaining 50%, the rec model exhibited better outcomes. On average, the aao models required 11.6 s to fit and achieved an R^2 of 0.34 (minimum R^2: -0.18 (DT); maximum R^2: 0.69 (XGB)). The rec models took an average of 54.9 s to develop, and showed an average R^2 of 0.41 (minimum R^2: -0.30 (DT); maximum R^2: 0.68 (GPR with composite kernel)). The low R^2 values indicate that among the 232 models, some produced exceedingly poor results. This is exemplified by the DT models, which even resulted in negative R^2 values, indicating outcomes inferior to using the mean as a predictor (Wall, 2022). Additionally, there were notable discrepancies in the accuracy of the ANN models. For further examination, only the four aao models with the highest average R^2 values and their rec counterparts, as well as the four rec models with the highest average R^2 values and their aao counterparts, are considered. For the aao approach, this includes the XGB model (average R^2: 0.69), an ANN with two hidden layers (128 and 64 neurons), a logistic activation function and the adam solver (average R^2: 0.68), and two GPR models, one with a Matern kernel and one with a combination of kernels, both with an average R^2 of 0.67 For the rec approach, the four most accurate models, each achieving an average R^2 of 0.68, are GPR models with the four examined kernel types (cf. Sections 3.3.1 and 3.3.2). Furthermore, those ANN models that deliver the best results concerning at least one KPI in a given year are included for both the aao and rec approaches (cf. Figure 5 and Table 1). RF models are not considered further, as their average R^2 values remain below 0.5 for both the aao and rec approaches.

Figure 5 presents an overview of the remaining models for further analysis. It is clear that models excelling in the aao approach do not necessarily perform equally well in aao estimations. For instance, nn29 and nn32 are neural networks utilizing logistic (nn29) and relu (nn32) activation functions, both employing lbfgs as a solver and comprising two hidden layers with 128 and 64 neurons. While nn29 and nn32 rank among the top-performing neural networks for the rec approach, they exhibit relatively weak performance for the aao approach. Conversely, the aao XGB model emerges as the highest-performing model when averaged across all KPIs and periods, yet it is significantly less effective for the rec approach. The performance of the GPR models remains nearly constant across all examined kernels for both the aao and rec approaches.

The observed differences between model types can partly be explained by their modeling characteristics. In the rec approach, separate models are fitted for each period, allowing the models to adapt more easily to period-specific relationships between inputs and outputs. In contrast, the aao approach estimates all periods simultaneously with a single model and therefore learns a joint relationship across time. Tree-based ensemble methods such as XGBoost are well suited for modeling complex relationships and interactions between variables (Chen & Guestrin, 2016), which may explain their stronger performance when

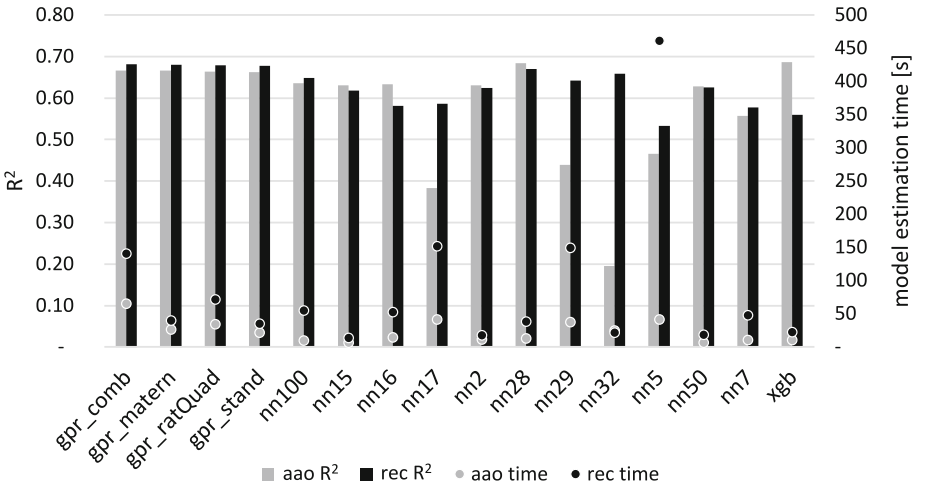

Fig. 5 R^2 of selected surrogate models, averaged over all KPIs and all periods, and fitting time of each surrogate model. gpr_comb: GPR with a combination of constant, RBF and matern kernels; gpr_matern: GPR with matern kernel; gpr_ratQuad: GPR with rational quadratic kernel; gpr_stand: GPR with constant kernel with a radial basis function kernel (RBF); nn100: neural network with two layers (64, 32 neurons), activation: logistic, solver: adam; nn15: neural network with two layers (64, 32 neurons), activation: identity, solver: sgd; nn16: neural network with three layers (128, 64, 32 neurons), activation: logistic, solver: adam; nn17: neural network with three layers (128, 64, 32 neurons), activation: logistic, solver: lbfgs; nn2: neural network with four layers (128, 64, 32, 16 neurons), activation: identity, solver: lbfgs; nn28: neural network with two layers (128, 64 neurons), activation: logistic, solver: adam; nn29: neural network with two layers (128, 64 neurons), activation: logistic, solver: lbfgs; nn32: neural network with two layers (128, 64 neurons), activation: relu, solver: adam; nn5: neural network with four layers (128, 64, 32, 16 neurons), activation: logistic, solver: lbfgs; nn50: neural network with three layers (32, 16, 8 neurons), activation: identity, solver: lbfgs; nn7: neural network with four layers (128, 64, 32, 16 neurons), activation: relu, solver: adam; xgb: XGBoost

all periods are estimated simultaneously. GPR, in contrast, often assumes relatively smooth functional relationships between inputs and outputs (Rasmussen & Williams, 2008), which may make it well suited for modeling individual periods separately in the rec approach. Neural network models provide a flexible framework for approximating complex nonlinear functions, but their performance depends on the chosen network architecture and training procedure (Goodfellow et al., 2016), which may explain the heterogeneous results observed across models and approaches in this study.

The estimation times vary considerably between the models. While estimations using a matern and a standard kernel configuration (cf. Section 3.3.1) require approximately 27 and 22 s in the aao case and 40 and 35 s in the rec case, the rec estimation with a combination of kernels takes approximately 140 s, representing an increase by a factor of 3.5–4 compared to the aforementioned kernels. Overall, it is noteworthy that rec model estimations are more time-intensive than aao model estimations. On average, among the models depicted in Fig. 5, it took 21 s to construct the aao models, and 51 s for the rec models. This 140% increase in model building time resulted in an average R^2 increase of 5% and an RMSE reduction of 3%. In the following, results are reported using R^2. corresponding figures and tables for MSE and RMSE are provided in the Supplementary Information.

Even the best-performing models exhibit relatively low R^2 values when averaged across all KPIs and years. Figure 6 a-c provides an overview of the R^2 values for three of the KPIs across the six examined periods, averaged over all models. On average, the models predict

Table 1 Best models per KPI, averaged over all periods

KPI	aao/rec	Model name	R^2	Estimation time
all	aao	xgb	0.6870	10.31
	rec	gpr_comb	0.6820	140.71
Battery returns total	aao	nn28	0.8070	12.83
	rec	gpr_comb	0.8105	140.71
Recycled batteries	aao	nn28	0.7266	12.83
	rec	gpr_comb	0.7026	140.71
Repurposed batteries	aao	nn28	0.6891	12.83
	rec	nn29	0.6861	37.80
Demand for reconditioned batteries	aao	xgb	0.6157	10.31
	rec	nn32	0.5021	21.53
Supply of reconditioned batteries	aao	nn28	0.4820	12.83
	rec	nn100	0.4857	54.99

aao, all-at-once approach; gpr_comb, Gaussian process regression with a combination of constant, radial basis function and matern kernels; nn28, neural network with two layers (128, 64 neurons); activation, logistic; solver, adam; nn29, neural network with two layers (128, 64 neurons); activation, logistic; solver, lbfgs; nn32, neural network with two layers (128, 64 neurons); activation, relu; solver, adam; nn100, neural network with two layers (64, 32 neurons); activation, logistic; solver, adam; rec, recursive approach; xgb, XGBoost

the total number of battery returns (Fig. 6 a, d) with significantly greater accuracy than the demand for reconditioned batteries (Fig. 6 c, f). Specifically, the total number of battery returns is predicted with an R^2 of approximately 0.8 for the years 2025 and 2030 for rec models, and an R^2 of 0.7 for aao models (c.f. Figure 6 a). The average R^2 of the KPI "Total number of returning batteries" decreases over the years to $R^2 \approx 0.67$ for rec models and 0.62 for aao models. Figure 6 d illustrates the corresponding simulated-versus-predicted graph for the 2035 period and the gpr_stand model, showing that the ML predicted values fall within an acceptable range compared to the simulated values. The decline in prediction accuracy over time can be attributed to the structure of the simulation model being approximated. This model takes 49 parameters as input, which remain constant throughout the simulation runtime (cf. Section 3.1). However, certain parameters only become relevant at later points in time. For instance, recycling quotas are introduced in 2031 and 2036, in accordance with the EU Battery regulation. Consequently, the surrogate models may disregard some information in earlier periods but must account for their influence and the noise they introduce in later periods. Also, as the simulated EV fleet grows and more lifecycle pathways become possible, the output variability increases, which makes the approximation task more challenging for the surrogate models.

Figure 6 c illustrates that the ML models possess limited predictive power regarding the KPI "Demand for reconditioned batteries." The average R^2 values exhibit a modest increase from approximately 0.3 in 2025 to 0.45 by 2050, yet they remain insufficient. Figure 6 f complements this by displaying the simulated-versus-predicted plot for 2035 using the gpr_stand model. Interestingly, the predictive capability improves over time rather than diminishes. This trend is likely due to the fact that reconditioning is considered viable only from 2030 onwards in some training and testing datasets, leading to a weaker non-zero data foundation in earlier years. Various combinations of input features can result in minimal or no reconditioning. For example, OEMs may choose not to use reconditioned

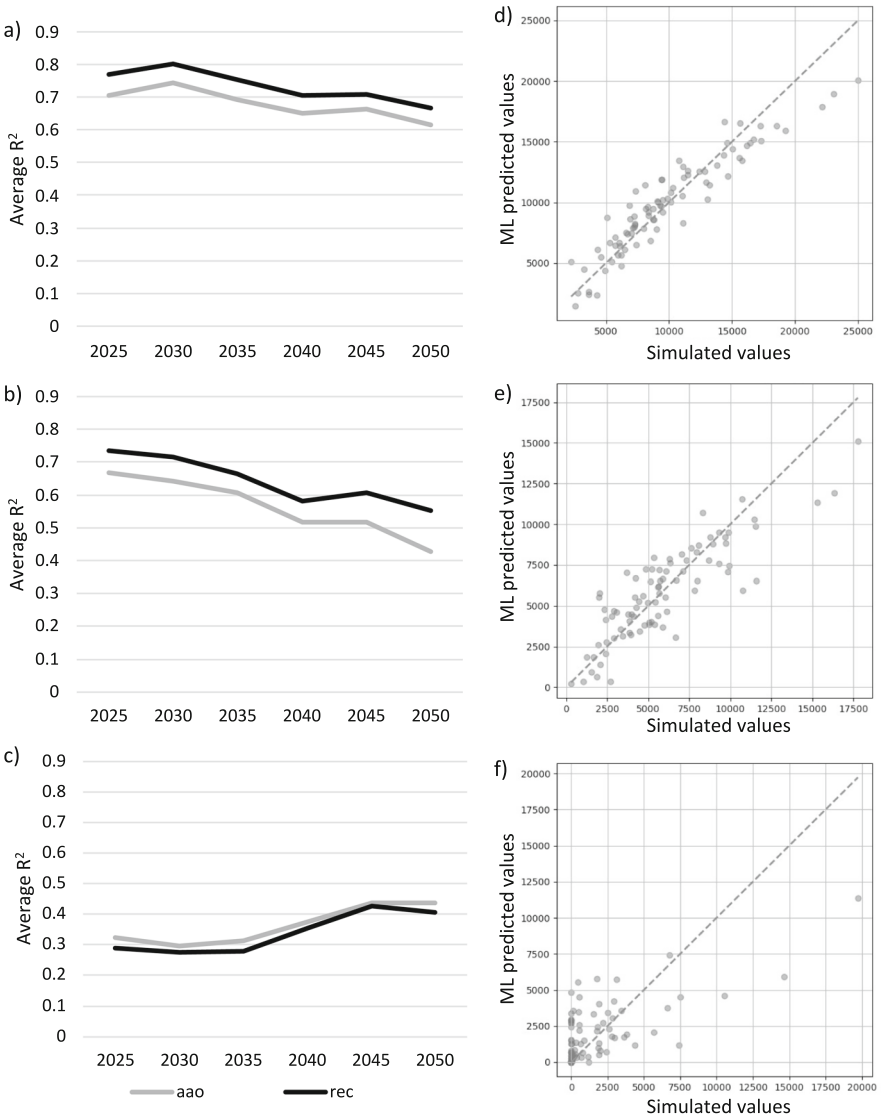

Fig. 6 a-c: Average R^2 over time of all models for **a** total number of returning batteries, **b** number of recycled batteries, **c** demand for reconditioned batteries. d-e: simulated-vs-predicted plots for period 2035, determined with the rec gpr_stand model, for d) total number of returning batteries, e) number of recycled batteries, f) demand for reconditioned batteries. aao: all-at-once approach; rec: recursive approach

batteries for warranty purposes, repurposing or recycling might be more lucrative, or the average battery lifespan could surpass that of vehicles. These scenarios are realistic and should be included in the training and testing datasets. However, this inclusion results in numerous zero or near-zero values (cf. Supplementary Information for the histograms of the KPI values), and given the high number of input features, it is challenging for surrogate models to extract meaningful insights from the remaining data during training. The authors

opted not to enhance the dataset with additional non-zero values for reconditioning KPIs through adaptive sampling, as the data scarcity also highlights the limitations of sampling for multiple KPIs (cf. Section 3.2), which should not be concealed. The prediction accuracy of the number of recycled batteries (Fig. 6 b and e) is intermediate between that of the total battery return numbers and the demand for reconditioned batteries. The moderate decline in R^2 over time can similarly be explained by increasing variability in recycling flows with a growing EV fleet and with additional lifecycle pathways becoming available. As more combinations of input parameters influence recycling outcomes in later years, the surrogate models face a more complex mapping between inputs and outputs. Thus, the underlying simulation dynamics help explain the patterns observed in Fig. 6. Not shown are the figures for the KPIs "Repurposed batteries", which have accuracy measures similar to those of recycled batteries, and "Supply of reconditioned batteries", which perform similarly to "Demand for reconditioned batteries".

Although the rec and aao approaches are expected to yield similar results for the first period, small but systematic differences can still occur. One reason is that the training processes differ slightly between the two approaches: in the rec approach, models are fitted separately for each period, while in the all-at-once approach, a single model is trained to predict all periods simultaneously. Moreover, even slight variations in the treatment of training data, such as the independent standardization of input features and target variables for each period compared to a collective standardization across all periods, can result in moderate differences in model performance.

Table 1 presents the models with the highest performance, as determined by the R^2 values averaged across all periods, for each of the evaluated KPIs. The data indicate that the models with the best overall performance (XGB for the aao approach and gpr_comb for the rec approach) also demonstrate superior accuracy for certain, though not all, KPIs. Notably, neural networks in various configurations are among those achieving the highest R^2 values. These networks consistently feature only two dense layers, either (128, 64) or (64, 32), supporting the literature's findings that excessively complex models may compromise accuracy due to overfitting (Goodfellow et al., 2016). Of the AANs, three utilize a logistic activation function, while one employs a relu activation. Additionally, two networks use the adam solver, and two utilize lbfgs.

Evaluating KPIs individually across periods shows that certain models consistently perform better than others. Figure 7 presents the maximum R^2 values achievable for each KPI and year. The xgb model, fitted through the aao approach, emerges as the most effective model when averaged across all KPIs and is also the leading model for nine out of the 30 targets. Equally frequently, the aao nn28 model, a neural network with two layers (128 and 64 neurons), activated by a logistic function and utilizing the adam solver, is identified as the top performer. The rec GPR model, which employs a combination of kernels (cf. Section 3.3.1), ranks as the next most successful model. Additional neural networks using a rec approach are recognized as the most effective for the remaining targets.

4.2 Feature engineering

While conducting a comprehensive examination of the significance of features for the original simulation model (cf. Section 3.1) through extensive sensitivity analysis would be exceedingly time-consuming, certain surrogate models offer inherent feature importance rankings. For instance, XGB models assess importance as the "fractional contribution of each feature to the model based on the total gain of this feature's splits" (Chen et al., n.d.). To extract the

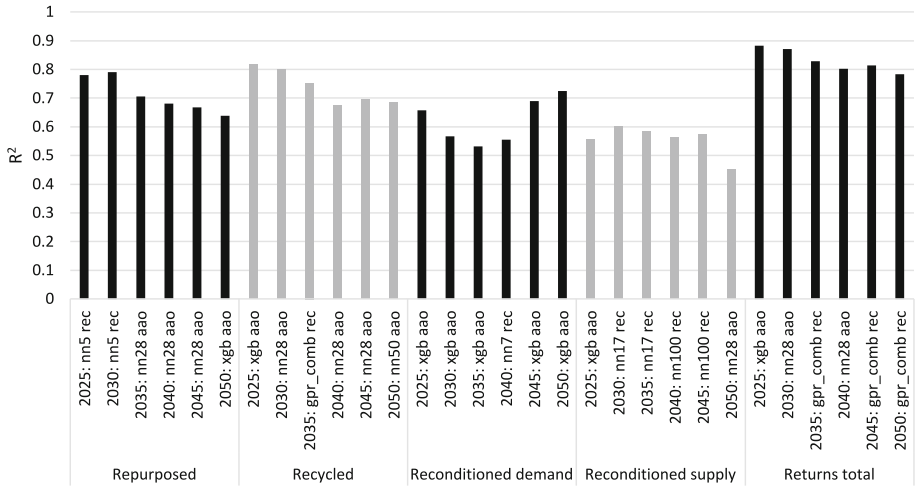

Fig. 7 R^2 of the best performing surrogate model per KPI and year. aao: all-at-once approach; gpr_comb: Gaussian process regression with a combination of constant, radial basis function and matern kernels; nn5: neural network with four layers (128, 64, 32, 16 neurons), activation: logistic, solver: lbfgs; nn7: neural network with four layers (128, 64, 32, 16 neurons), activation: relu, solver: adam; nn17: neural network with three layers (128, 64, 32 neurons), activation: logistic, solver: lbfgs; nn28: neural network with two layers (128, 64 neurons), activation: logistic, solver: adam; nn50: neural network with three layers (32, 16, 8 neurons), activation: identity, solver: lbfgs; nn100: neural network with two layers (64, 32 neurons), activation: logistic, solver: adam; rec: recursive approach; xgb: XGBoost

feature importance for each of the five KPIs across six periods, an XGB model was estimated for each KPI and period using the aao approach, resulting in 30 models. Subsequently, the feature importance for each model was determined. This analysis revealed that among the 49 features, 16 never appear among the top ten most important, whereas the remaining 33 do at least once. The feature most frequently appearing in the top ten list (29 out of 30 instances) is the life expectancy of NCx batteries, referring to batteries with Nickel-Manganese-Cobalt (NMC) or Nickel–Cobalt-Oxide (NCO) cell chemistry. Additionally, the decisions by OEMs regarding the use of reconditioned batteries for warranty cases and the proportion of batteries damaged beyond repair are also significant, appearing 23 and 20 times, respectively, in the top ten for various KPIs and periods. Furthermore, EV sales data from 2020 to 2035 and the average vehicle lifetime are mentioned in the top ten at least ten times. The relevance of these KPIs appears justified. The simulation encompasses a timeframe extending to the year 2050. In the initial years, NCx batteries are among the most frequently sold EVBs, making their average lifespan critical for the period when these batteries are returned and potentially repurposed for a second-life application. The average battery lifespan varies across the experiments, ranging from 6 to 20 years. Consequently, EV sales data from years immediately preceding the end of the examined timeframe are less important than those from earlier years.

To mitigate noise potentially introduced by an excessive number of features, each period of each KPI is approximated using an XGB surrogate model that incorporates only the ten most significant features for that specific KPI and period. The outcomes are depicted in Fig. 8 and are compared with the results derived from estimations utilizing all 49 features. On average, the models employing ten features exhibited a 6.6% decrease in performance compared to the 49-feature models. However, this overall average does not accurately represent performance at

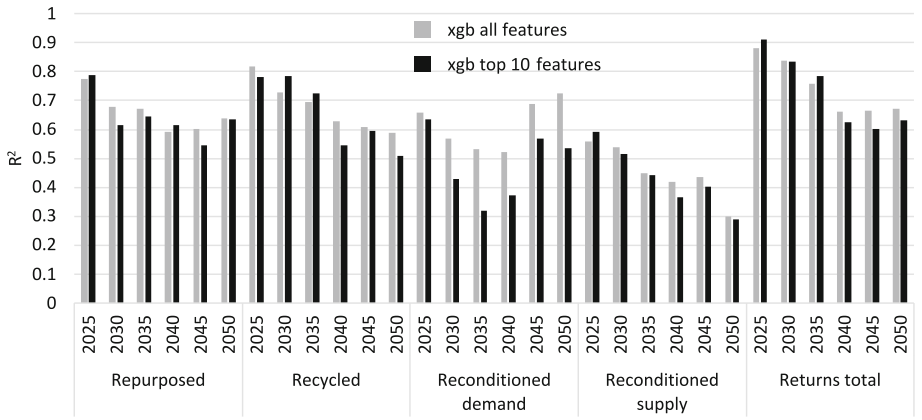

Fig. 8 R^2 of the xgb surrogate model estimated with all 49 features and with the 10 most important features of each KPI and year. xgb: XGBoost

the KPI level. The differences are relatively minor for repurposed batteries (−2.9%), recycled batteries (−3.0%), the supply of reconditioned batteries (−3.6%), and the total number of returning batteries (−1.8%). In contrast, the performance for the demand for reconditioned batteries declines significantly (−22.6%). This indicates that for most KPIs, a reduced number of features suffices without substantial information loss, whereas for reconditioned battery demand, a greater number of features is necessary. The reduced-feature approach necessitates the determination of feature importance, which involves fitting models with all features. Therefore, it would only be considered beneficial if the additional estimation step enhances accuracy rather than diminishes it.

5 Discussion

The results of this study highlight both the opportunities and limitations of using ML techniques to approximate a simulation model for predicting EoL EVB quantities. While the overall performance of the surrogate models varied considerably across KPIs and periods, several consistent insights emerged and suggest avenues for further research.

First, the generally low R^2 values across models and KPIs point to the inherent difficulty of the problem at hand. The input data is static, while the outputs are sequential but lack temporal predecessors. This creates a setting that cannot be fully addressed by conventional static regression or standard time series forecasting techniques. Moreover, the dataset size (490 samples) represents a limitation of this study and may have constrained model performance. Preliminary experiments with additional preprocessing steps, such as logarithmic transformations, did not lead to consistent improvements in predictive performance. This suggests that the main limitation lies less in the preprocessing of the dataset and more in the limited number of simulation samples and the complexity of the underlying input–output relationships. Although the number of samples is in line with surrogate modeling literature (e.g., Loeppky et al., 2009), this size may be insufficient for capturing complex, nonlinear relationships, especially when outputs interact strongly or are dominated by zero values, as seen in the reconditioning-related KPIs, which represents an additional challenge and limitation for surrogate modeling in this setting. Several contributions in the literature emphasize

the value of adaptive sampling, where training points are iteratively added in regions of high model uncertainty or near optima. For example, Wang et al. (2014) showed that adaptive sampling significantly improved model performance for both surface approximation and optimization tasks. Similarly, ten Broeke et al. (2021) employed adaptive sampling in ABM analysis to efficiently explore tipping points. Future work could explore whether larger simulation datasets, adaptive sampling strategies, more extensive data preprocessing, or alternative machine learning approaches enhance surrogate performance in settings like ours. The presented approach can serve as a basis for further developments in data-driven circular economy modeling.

Second, although recursive models achieved slightly higher average performance than the aao approach, they required longer estimation times. However, all models in this study were trained within a few seconds to two and a half minutes, meaning time savings are not practically relevant in this setting. The benefits of faster training times become more important with larger datasets or extended feature sets.

Third, model performance varied across KPIs, and the results also reveal differences in the suitability of individual surrogate modeling techniques. Predictions for total battery returns were relatively accurate across most models, likely due to more stable relationships between inputs and outputs and the comparatively large number of non-zero observations. In contrast, demand for reconditioned batteries proved harder to model, likely because it depends on multiple interacting features, including policy assumptions, OEM decisions, and competing end-of-life pathways. These assumptions are inherited from the underlying simulation model described in Huster et al. (2025). Looking at the individual surrogate models, GPR models achieved comparatively stable results across KPIs and years and were frequently among the better-performing models, particularly within the recursive approach. For the aao approach, the XGBoost model reached the highest average accuracy across KPIs. Neural network models with relatively simple architectures also performed well for several outputs and showed particularly good performance for KPIs related to repurposed and reconditioned batteries. These outputs depend on a larger number of interacting input factors and contain fewer informative observations than more aggregate outputs such as total battery returns or recycled batteries, which may explain why more flexible model structures achieved better results for these targets. Overall, these findings suggest that GPR models represent a robust general-purpose choice for approximating the outputs of the simulation model, while XGBoost provides a strong alternative for the aao approach. For outputs characterized by sparse observations and stronger nonlinear dependencies, such as repurposing and reconditioning demand, neural network models may offer advantages. At the same time, the unequal accuracy among the KPIs suggests that surrogate modeling is not equally effective across all types of outputs, and more specialized modeling strategies may be required for sparse or volatile targets. This also highlights that the suitability of a given surrogate model ultimately depends on the characteristics of the underlying data and the specific prediction task. The present study therefore provides a structured approach for comparing and selecting surrogate models in similar settings, rather than identifying a universally optimal method.

Fourth, feature selection results indicate that dimensionality reduction can be useful, though its benefits are KPI-specific. Using only the ten most important features preserved model accuracy for several outputs, such as recycled and repurposed batteries. However, this strategy significantly reduced performance for reconditioning demand, pointing to a need for richer input representations when modeling outputs are shaped by multiple interdependent factors.

Overall, the results highlight both the potential and the limitations of using surrogate models to approximate complex simulation outputs in this setting. The relatively small training dataset, the high dimensionality of the input space, and the sparsity of several KPIs, particularly those related to reconditioning, represent important limitations that likely constrained predictive performance. In addition, the surrogate models approximate the behavior of a single simulation model and therefore inherit the assumptions embedded in that model, including policy assumptions and behavioral rules of stakeholders. Despite these limitations, the surrogate models developed in this study can substantially accelerate simulation-based evaluations. Given that each simulation run takes approximately ten minutes, surrogate models, especially those with acceptable accuracy, can drastically reduce computation time in tasks such as optimization, sensitivity analysis, or policy testing to the fraction of a second. This opens the door for more agile planning in contexts where historical data is sparse, but future-oriented decisions are necessary, such as in early-stage circular economy strategies.

Future work should explore enhancements to model architecture and training. This includes experimenting with hybrid approaches that combine ML with system-specific knowledge, testing architectures more suited to sparse or structured outputs, and evaluating performance on larger datasets. Furthermore, applying the surrogate models within optimization frameworks or interactive scenario tools would demonstrate their practical value more fully. Expanding this research to other domains with similar data challenges, such as resource planning, critical material recovery, or other product reuse systems, would help test the robustness and generalizability of the approach.

6 Conclusion

This study addressed the challenge of accelerating scenario analysis in a discrete event and agent-based simulation model used to forecast end-of-life quantities of EV batteries. The simulation model estimates the availability of batteries for recycling, repurposing, and reconditioning from 2025 to 2050 based on 49 static input parameters, including EV sales projections, battery lifetimes, and economic assumptions. While this approach offers a detailed representation of second-life battery pathways, its computational demands limit the feasibility of extensive sensitivity analyses and parameter optimization. To reduce runtime and support broader scenario exploration, machine learning-based surrogate models were tested as an approximation of the simulation's input–output relationships. Two modeling strategies were compared: predicting all time steps simultaneously, and a recursive approach that forecasts time steps sequentially using prior predictions.

Results showed that neither approach was universally superior. Recursive models achieved slightly higher *average* accuracy (R^2 of 0.41 vs. 0.34) but required longer training times (on average 140% longer). Still, with training durations under three minutes for all models, this trade-off was practically irrelevant for the given dataset. Among the algorithms, XGBoost performed best in the all-at-once setup with regard to accuracy (R^2: 0.69), while Gaussian process regression was most accurate in the recursive approach (R^2: 0.68). Neural networks also performed well in both setups when kept relatively simple, whereas decision trees consistently underperformed. Prediction accuracy varied strongly across output indicators. Battery return volumes were estimated with high reliability, while reconditioning demand and supply proved more difficult due to sparse and volatile simulation outputs. Feature reduction using XGBoost importance rankings worked well for most KPIs but significantly hurt performance for reconditioning demand.

The results suggest that surrogate models are a viable tool for reducing simulation complexity in second-life battery studies. They may support early-stage planning and scenario analysis where empirical data is scarce and a wide range of possible developments must be considered. Future research should investigate adaptive sampling strategies, hybrid architectures, and optimization performance to extend these insights to more complex and dynamic circular economy applications.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10479-026-07342-3.

Acknowledgements Parts of this work were carried out during a research stay of the first author at the University of Warwick, Warwick Manufacturing Group (WMG), which was supported by funding from the Karlsruhe House of Young Scientists (KHYS). The support for the research stay is gratefully acknowledged.

Author contributions SH: Methodology, Conceptualization, Writing—Original Draft; MFN: Conceptualization, Supervision, Writing—Review & Editing; AR: Writing—Review & Editing; FS: Writing—Review & Editing, Supervision.

Funding Open Access funding enabled and organized by Projekt DEAL. Karlsruhe House of Young Scientists

Data availability Additional data supporting the findings of this study, beyond those provided in the Supplementary Information, are available from the corresponding author (SH) upon reasonable request.

Declarations

Conflict interest The authors have no relevant financial or non-financial interests to disclose.

Use of generative artificial intelligence During the preparation of this work the authors used the tool "Paperpal" in order to improve readability and language. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

Angione, C., Silverman, E., & Yaneske, E. (2022). Using machine learning as a surrogate model for agent-based simulations. *PLoS ONE, 17*(2), Article 0263150. https://doi.org/10.1371/journal.pone.0263150

Banks, J. (Ed.). (1998). *Handbook of simulation: Principles, methodology, advances, applications, and practice* (1st ed.). Wiley. https://doi.org/10.1002/9780470172445

Barton, R. R. (1992). Metamodels for simulation input-output relations. Proceedings of the 24th Conference on Winter Simulation - WSC '92, 289–299. https://doi.org/10.1145/167293.167352

Bashiri, S., Yasari, E., & Tayyebi, S. (2022). Comparison of different sampling and surrogate modelling approaches for a multi-objective optimization problem of direct dimethyl ether synthesis in the fixed-bed reactor. *Chemometrics and Intelligent Laboratory Systems, 230*, Article 104683. https://doi.org/10.1016/j.chemolab.2022.104683

Box, G. E. P., & Wilson, K. B. (1951). On the experimental attainment of optimum conditions. Journal of the Royal Statistical Society: Series B (*Methodological*), 13(1), 1–45.

Cerqueira, V., Torgo, L., & Mozetič, I. (2020). Evaluating time series forecasting models: An empirical study on performance estimation methods. *Machine Learning, 109*(11), 1997–2028. https://doi.org/10.1007/s10994-020-05910-7

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., Li, Y., Yuan, J., & Cortes, D. (n.d.). Feature importance. Retrieved 25 March 2025, from https://xgboost.readthedocs.io/en/stable/r_docs/R-package/docs/reference/xgb.importance.html

Cheong, H., & Butscher, A. (2019). Physics-based simulation ontology: An ontology to support modelling and reuse of data for physics-based simulation. *Journal of Engineering Design, 30*(10–12), 655–687. https://doi.org/10.1080/09544828.2019.1644301

Choi, Y., Song, D., Yoon, S., & Koo, J. (2021). Comparison of factorial and Latin hypercube sampling designs for meta-models of building heating and cooling loads. *Energies, 14*(2), 512. https://doi.org/10.3390/en14020512

Christelis, V., Kopsiaftis, G., & Mantoglou, A. (2019). Performance comparison of multiple and single surrogate models for pumping optimization of coastal aquifers. *Hydrological Sciences Journal, 64*(3), 336–349. https://doi.org/10.1080/02626667.2019.1584400

De Amorim, L. B. V., Cavalcanti, G. D. C., & Cruz, R. M. O. (2023). The choice of scaling technique matters for classification performance. *Applied Soft Computing, 133*, Article 109924. https://doi.org/10.1016/j.asoc.2022.109924

Dearing, M. T. (2024). Deep Learning Surrogate Models for Network Simulation. Proceedings of the 38th ACM SIGSIM Conference on Principles of Advanced Discrete Simulation, 65–66. https://doi.org/10.1145/3615979.3662151

DeRousseau, M., Gully, B., Taylor, C., Apelian, D., & Wang, Y. (2017). Repurposing used electric car batteries: A review of options. *JOM Journal of the Minerals Metals and Materials Society, 69*(9), 1575–1582. https://doi.org/10.1007/s11837-017-2368-9

EEA. (2022). Textiles and the environment: The role of design in Europe's circular economy [Briefing]. European Environment Agency. https://www.eea.europa.eu/publications/textiles-and-the-environment-the/textiles-and-the-environment-the

Engel, H., Hertzke, P., Siccardo, G., & McKinsey & Company. (2019). Second-life EV batteries: The newest value pool in energy storage. https://www.mckinsey.com/industries/automotive-and-assembly/our-insights/second-life-ev-batteries-the-newest-value-pool-in-energy-storage

European Commission. (n.d.). 2050 long-term strategy—Climate Action—European Commission. Retrieved 22 March 2026, from https://climate.ec.europa.eu/eu-action/climate-strategies-targets/2050-long-term-strategy_en

European Commission. (2008). Directive 2008/98/EC of the European Parliament and of the Council of 19 November 2008 on waste and repealing certain Directives: 2008/98/EC. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32008L0098

European Parliament and European Council. (2023). Regulation (EU) 2023/1542 of the European Parliament and of the Council of 12 July 2023 concerning batteries and waste batteries, amending Directive 2008/98/EC and Regulation (EU) 2019/1020 and repealing Directive 2006/66/EC. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv%3AOJ.L_.2023.191.01.0001.01.ENG&toc=OJ%3AL%3A2023%3A191%3ATOC

Faessler, B. (2021). Stationary, second use battery energy storage systems and their applications: A research review. *Energies, 14*(8), 2335. https://doi.org/10.3390/en14082335

Goldsman, D., Nance, R. E., & Wilson, J. R. (2009). A brief history of simulation. *Proceedings of the 2009 Winter Simulation Conference (WSC)*, 310–313. https://doi.org/10.1109/WSC.2009.5429341

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Huntington, D. E., & Lyrintzis, C. S. (1998). Improvements to and limitations of Latin hypercube sampling. *Probabilistic Engineering Mechanics, 13*(4), 245–253. https://doi.org/10.1016/S0266-8920(97)00013-1

Hürkamp, A., Gellrich, S., Dér, A., Herrmann, C., Dröder, K., & Thiede, S. (2021). Machine learning and simulation-based surrogate modeling for improved process chain operation. *International Journal of Advanced Manufacturing Technology, 117*(7–8), 2297–2307. https://doi.org/10.1007/s00170-021-07084-5

Huster, S., Heck, R., Rudi, A., Rosenberg, S., & Schultmann, F. (2025). Will electric vehicle battery reconditioning succeed? – A flexible simulation approach considering stakeholders' interests. *Journal of Environmental Management, 389*, Article 125783. https://doi.org/10.1016/j.jenvman.2025.125783

Huster, S., Rosenberg, S., Hufnagel, S., Rudi, A., & Schultmann, F. (2024). Do consumers want reconditioned electric vehicle batteries? – A discrete choice experiment. *Sustainable Production and Consumption*. https://doi.org/10.1016/j.spc.2024.05.027

Hyndman, R. J., & Athanasopoulos, G. (2021). Forecasting: Principles and practice (Third print edition). Otexts, Online Open-Access Textbooks.

Kadner, S., Kobus, J., Hansen, E. G., Akinci, S., Elsner, P., Hagelüken, C., Jaeger-Erben, M., Kick, M., Kwade, A., Müller-Kirschbaum, T., Kühl, C., Obeth, D., Schweitzer, K., Stuchtey, M., Vahle, T., Weber, T., Wiedemann, P., Wilts, H., Wittken, R., Circular Economy Initiative Deutschland. (2021). *Circular Economy Roadmap für Deutschland*. acatech - Deutsche Akademie der Technikwissenschaften. https://doi.org/10.48669/ceid_2021-3

Kastanaki, E., & Giannis, A. (2023). Dynamic estimation of end-of-life electric vehicle batteries in the EU-27 considering reuse, remanufacturing and recycling options. *Journal of Cleaner Production, 393*, Article 136349. https://doi.org/10.1016/j.jclepro.2023.136349

Khuri, A. I., & Cornell, J. A. (1987). *Response surfaces: Designs and analysis* (2nd ed.). Dekker.

Kleijnen, J. P. C. (1982). *Regression metamodel summarization of model behaviour: FEW 120*. https://repository.tilburguniversity.edu/bitstreams/43d6e3ea-86f7-4450-84af-3110337f4b32/download

Law, A. M. (2015). *Simulation modeling and analysis* (Fifth edition). McGraw Hill Education.

Loeppky, J. L., Sacks, J., & Welch, W. J. (2009). Choosing the sample size of a computer experiment: A practical guide. *Technometrics, 51*(4), 366–376. https://doi.org/10.1198/TECH.2009.08040

McBride, K., & Sundmacher, K. (2019). Overview of surrogate modeling in chemical process engineering. *Chemie Ingenieur Technik, 91*(3), 228–239. https://doi.org/10.1002/cite.201800091

Montgomery, D. C. (2012). *Design and analysis of experiments* (8th ed.). John Wiley & Sons.

Mueller, J., Sprave, J., & Frey, G. (2023). The Effect of Target Variable Rescaling on Energy Consumption Prediction in a Vehicle Powernet using Multi-Target Regression Trees: A study from the automotive industry. Proceedings of the 2023 8th International Conference on Machine Learning Technologies, 144–150. https://doi.org/10.1145/3589883.3589905

Mukhtar, A., Yasir, A. S. H. M., & Nasir, M. F. M. (2023). A machine learning-based comparative analysis of surrogate models for design optimisation in computational fluid dynamics. *Heliyon, 9*(8), Article e18674. https://doi.org/10.1016/j.heliyon.2023.e18674

Nakano, S., & Liu, Y. (2025). Interpreting temporal shifts in global annual data using local surrogate models. *Mathematics, 13*(4), Article 626. https://doi.org/10.3390/math13040626

Natsume, Y. (2022). *Gaussian Process Kernels*. Towards Data Science. https://towardsdatascience.com/gaussian-process-kernels-96bafb4dd63e/

Obaidat, M. S., & Papadimitriou, G. I. (Eds.). (2003). *Applied System Simulation: Methodologies and Applications*. Springer US. https://doi.org/10.1007/978-1-4419-9218-5

Pan, X., Chen, D., Pan, B., Huang, X., Yang, K., Piao, S., Zhou, T., Dai, Y., Chen, F., & Li, X. (2025). Evolution and prospects of Earth system models: Challenges and opportunities. *Earth-Science Reviews, 260*, Article 104986. https://doi.org/10.1016/j.earscirev.2024.104986

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Pietukhov, R., Ahtamad, M., Faraji-Niri, M., & El-Said, T. (2023). A hybrid forecasting model with logistic regression and neural networks for improving key performance indicators in supply chains. *Supply Chain Analytics, 4*, Article 100041. https://doi.org/10.1016/j.sca.2023.100041

Pietzsch, B., Fiedler, S., Mertens, K. G., Richter, M., Scherer, C., Widyastuti, K., Wimmler, M.-C., Zakharova, L., & Berger, U. (2020). Metamodels for evaluating, calibrating and applying agent-based models: A review. *Journal of Artificial Societies and Social Simulation*. https://doi.org/10.18564/jasss.4274

Potting, J., Hekkert, M., Worrell, E., Hanemaaijer, A., & PBL Netherlands Environmental Assessment Agency, The (with PBL Netherlands Environmental Assessment Agency, The). (2017). Circular economy: Measuring innovation in the product chain.

Pratap, B., Mohan, T. K., Amit, R. K., & Venugopal, S. (2024). Evaluating circular economy strategies for raw material recovery from end-of-life lithium-ion batteries: A system dynamics model. *Sustainable Production and Consumption, 50*, 191–204. https://doi.org/10.1016/j.spc.2024.07.027

Rasmussen, C. E., & Williams, C. K. I. (2008). *Gaussian processes for machine learning* (3. print). MIT Press.

Sacks, J., Welch, W. J., Mitchell, T. J., & Wynn, H. P. (1989). Design and Analysis of Computer Experiments. *Statistical Science*, *4*(4). https://doi.org/10.1214/ss/1177012413

Sanclemente Crespo, M., Van Ginkel González, M., & Talens Peiró, L. (2022). Prospects on end of life electric vehicle batteries through 2050 in Catalonia. *Resources, Conservation and Recycling, 180*, Article 106133. https://doi.org/10.1016/j.resconrec.2021.106133

Santner, T. J., Williams, B. J., & Notz, W. I. (2018). *The design and analysis of computer experiments*. Springer New York. https://doi.org/10.1007/978-1-4939-8847-1

Schweidtmann, A. M., Zhang, D., & Von Stosch, M. (2024). A review and perspective on hybrid modeling methodologies. *Digital Chemical Engineering, 10*, Article 100136. https://doi.org/10.1016/j.dche.2023.100136

Scikit-learn. (n.d.). GaussianProcessRegressor. Scikit-Learn. Retrieved 7 March 2025, from https://scikit-learn.org/stable/modules/generated/sklearn.gaussian_process.GaussianProcessRegressor.html

Seika, J., & Kubli, M. (2024). Repurpose or recycle? Simulating end-of-life scenarios for electric-vehicle batteries under the EU battery regulation. *Sustainable Production and Consumption*. https://doi.org/10.1016/j.spc.2024.09.023

Shannon, R. E. (1975). *Systems simulation: The art and science*. Prentice-Hall.

Sheikholeslami, R., & Razavi, S. (2017). Progressive Latin hypercube sampling: An efficient approach for robust sampling-based analysis of environmental models. *Environmental Modelling & Software, 93*, 109–126. https://doi.org/10.1016/j.envsoft.2017.03.010

Slattery, M., Dunn, J., & Kendall, A. (2024). Charting the electric vehicle battery reuse and recycling network in North America. *Waste Management, 174*, 76–87. https://doi.org/10.1016/j.wasman.2023.11.018

Stellantis. (2023II, January 1). Stellantis Inaugurates its First Circular Economy Hub in Turin, Italy. https://www.stellantis.com/en/news/press-releases/2023/november/stellantis-inaugurates-its-first-circular-economy-hub-in-turin-italy

Stevanović, S., Dashti, H., Milošević, M., Al-Yakoob, S., & Stevanović, D. (2024). Comparison of ANN and XGBoost surrogate models trained on small numbers of building energy simulations. *PLoS ONE, 19*(10), Article e0312573. https://doi.org/10.1371/journal.pone.0312573

ten Broeke, G., van Voorn, G., Ligtenberg, A., & Molenaar, J. (2021). The use of surrogate models to analyse agent-based models. *Journal of Artificial Societies and Social Simulation*. https://doi.org/10.18564/jasss.4530

The MathWorks, Inc. (n.d.-b). *Regression Learner*. Retrieved 7 March 2025, from https://de.mathworks.com/help/stats/regressionlearner-app.html

The MathWorks, Inc. (n.d.-a). Lhsdesign. Retrieved 6 March 2025, from https://de.mathworks.com/help/stats/lhsdesign.html

Thombre, M. N., Preisig, H. A., & Addis, M. B. (2015). Developing surrogate models via computer based experiments. *Computer aided chemical engineering* (Vol. Vol. 37, pp. 641–646). Elsevier. https://doi.org/10.1016/B978-0-444-63578-5.50102-X

Trinchero, R., & Canavero, F. (2021). Machine learning regression techniques for the modeling of complex systems: An overview. *IEEE Electromagnetic Compatibility Magazine, 10*(4), 71–79. https://doi.org/10.1109/MEMC.2021.9705310

Von Rueden, L., Mayer, S., Sifa, R., Bauckhage, C., & Garcke, J. (2020). Combining machine learning and simulation to a hybrid modelling approach: Current and future directions. In M. R. Berthold, A. Feelders, & G. Krempl (Eds.), *Advances in intelligent data analysis XVIII* (Vol. Vol. 12080, pp. 548–560). Springer International Publishing. https://doi.org/10.1007/978-3-030-44584-3_43

Wall, K. (2022). Negative R2: Where Did You Go Wrong? *TDS Archive*. https://medium.com/data-science/negative-r2-where-did-you-go-wrong-9d4f2aa84cfb

Wang, C., Duan, Q., Gong, W., Ye, A., Di, Z., & Miao, C. (2014). An evaluation of adaptive surrogate modeling based optimization with two benchmark problems. *Environmental Modelling & Software, 60*, 167–179. https://doi.org/10.1016/j.envsoft.2014.05.026

Wang, S., Fan, K., Luo, N., Cao, Y., Wu, F., Zhang, C., Heller, K. A., & You, L. (2019). Massive computational acceleration by using neural networks to emulate mechanism-based biological models. *Nature Communications, 10*(1), Article 4354. https://doi.org/10.1038/s41467-019-12342-y

Wiesberg, I. L., De Medeiros, J. L., Paes De Mello, R. V., Santos Maia, J. G. S., Bastos, J. B. V., & Araújo, O. D. Q. F. (2021). Bioenergy production from sugarcane bagasse with carbon capture and storage: Surrogate models for techno-economic decisions. *Renewable and Sustainable Energy Reviews, 150*, Article 111486. https://doi.org/10.1016/j.rser.2021.111486

Williams, B., & Cremaschi, S. (2021). Selection of surrogate modeling techniques for surface approximation and surrogate-based optimization. *Chemical Engineering Research and Design, 170*, 76–89. https://doi.org/10.1016/j.cherd.2021.03.028

xgboost developers. (2022). *XGBoost Python Package—Xgboost 3.0.0 documentation*. https://xgboost.readthedocs.io/en/stable/python/index.html

xgboost developers. (n.d.). *Multiple Outputs—Xgboost 3.0.0 documentation*. Retrieved 25 March 2025, from https://xgboost.readthedocs.io/en/stable/tutorials/multioutput.html

Eric P. Xing, Konstantin Genin, & Yutong Zheng. (2015). Probabilistic Graphical Models—21: Advanced Gaussian Processes. https://www.cs.cmu.edu/~epxing/Class/10708-15/notes/10708_scribe_lecture21.pdf

Yang, L., Yanhui, L., Yiyi, J., & Kiyoshi, F. (2024). Forecasting the echelon utilization potential of end-of-life electric vehicle batteries: A province-level investigation in China. *Clean Technologies and Environmental Policy*. https://doi.org/10.1007/s10098-024-03072-x

Yang, X.-S., Koziel, S., & Leifsson, L. (2014). Computational optimization, modelling and simulation: Past, present and future. *Procedia Computer Science, 29*, 754–758. https://doi.org/10.1016/j.procs.2014.05.067

Zhao, Y., Jiang, C., Vega, M. A., Todd, M. D., & Hu, Z. (2023). Surrogate modeling of nonlinear dynamic systems: A comparative study. *Journal of Computing and Information Science in Engineering, 23*(1), Article 011001. https://doi.org/10.1115/1.4054039