Alves *et al. J Artif Intell Smart Energy Syst* *(2023) 2:29*
https://doi.org/10.68159/r802231902

ORIGINAL RESEARCH **Open access**

Experimental Evaluation of Safe Reinforcement Learning and Advanced Control Strategies for Home Energy Management Systems

Ricardo Alves[1*], Bruno Costa[2], Helena Martins[1]

Abstract

The field of machine learning for energy management has lately seen encouraging developments, particularly with reinforcement learning augmented by a safety filter (OptLayerPolicy) and a metaheuristic procedure that yields a decision-tree-based control law (TreeC). So far, however, their capabilities have only been confirmed within simulated environments. The present work offers real-world verification of these two techniques, benchmarking them alongside model predictive control and basic rule-driven controllers. The experimental setup consisted of the electrical infrastructure of four replica residential units, each incorporating a dedicated battery, a photovoltaic array, and a dynamic load emulator replicating both an uncontrollable household demand and a steerable electric vehicle charging station. The data indicates that the rule-based logic, TreeC, and model predictive control attained near-identical operational costs, separated by a mere 0.6%. The reinforcement-learning controller, which remained in a learning phase, produced costs 25.5% above those of the competing approaches. Follow-up simulation studies suggest that further cost reductions are attainable by supplying TreeC with a more representative training corpus and by correcting inaccuracies in the MPC setup that stem from its dependence on reliable data streams from disparate sources. The OptLayerPolicy module makes it feasible to train a reinforcement learning agent safely and online in a physical setting, provided the constraint specification is accurate. Although the safety-layer concept remains susceptible to faults, it nevertheless delivered measurable benefits for every technique examined. TreeC, which depends on a faithful simulation environment for its training, posted the most secure real-world behavior, surpassing the grid limit by a mere 27.1 Wh, compared with 593.9 Wh recorded for reinforcement learning.

Keywords Reinforcement learning, Machine learning, Model predictive control, Energy management system, Decision tree, Hardware-in-the-loop

*Correspondence:
Ricardo Alves
ricardo.alves@ua.pt
[1] Department of Artificial Intelligence for Energy Engineering, School of Engineering, University of Aveiro, Aveiro, Portugal
[2] Department of Smart Grid Analytics, Faculty of Engineering, University of Porto, Porto, Portugal

Introduction

Home energy management systems (EMS) are acquiring ever greater importance as societies shift away from fossil-based energy toward widespread electrification. A key driver of this transition is the uptake of affordable photovoltaic (PV) technology [1], which promotes the generation of renewable electricity right at the dwelling level. In parallel, steerable residential assets—batteries, heat pumps, and electric vehicles (EVs)—create opportunities to optimize the local consumption of that self-generated power. Meanwhile, electricity pricing models are transforming [2]; numerous jurisdictions are rolling out dynamic household tariffs meant to nudge consumers toward periods of plentiful, inexpensive

This is an open access article distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0), which permits non-commercial use, distribution, reproduction, and adaptation in any medium or format, provided that appropriate credit is given to the original author(s) and source, a link to the license is provided, and any changes made are indicated. Adapted material must be distributed under the same license.

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

energy. Such tariffs may shift within a single day on spot markets, as well as between days on day-ahead markets. Expecting occupants to constantly and manually tune their consumption to these fluctuating price signals over extended timeframes is unrealistic [3], especially when faced with unpredictable factors such as the highly variable solar output on partly overcast days. Thus, as controllable electrical devices and dynamic pricing become widespread, automated residential EMS become indispensable for pursuing either economic or environmental aims [4] without sacrificing occupant comfort.

Driving any steerable system to and keeping it at near-optimal operation constitutes a formidable challenge. Respecting every real-time system constraint invariably takes precedence (e.g., preventing electrical overloads or thermal discomfort in climate-control contexts). Only then does the pursuit of a specific objective become relevant (e.g., reducing monetary cost or CO2-equivalent footprint). The mathematical representation can also call for various temporal dependencies—energy storage and deferrable loads, for instance, theoretically require an unbounded planning horizon [5]. All of this must be accomplished under multiple sources of uncertainty—variable demand, weather, and prices—making the task one of optimizing the expected outcome in what are typically continuous-state systems.

Model predictive control (MPC) has risen to broad acceptance as an optimal-control technology across a range of industrial domains [6], building management among them [7]. The MPC framework is mature and extensively studied with respect to feasibility, stability margins, robustness guarantees, and constraint enforcement [8]. That said, it demands a detailed upfront model (encompassing plant dynamics, input/output disturbance characterizations, and measurement noise descriptions), a requirement that is prone to inaccuracies and cannot generally self-adjust over time. For smaller-scale installations especially, the effort and cost of constructing, calibrating, and maintaining such models can be prohibitive [9]. Recent breakthroughs in machine learning [10] have begun to show considerable promise on the fronts of modeling burden, model convexity, adaptivity, constraint satisfaction, and integration demands. Two particular directions are of interest here: (1) coupling reinforcement learning (RL) with a safety filter capable of accommodating arbitrary constraint types and of falling back on a known safe default policy where one exists (the OptLayerPolicy framework [11]), and (2) using metaheuristic search to derive an intelligible control policy in the form of a decision tree (the TreeC method [12]). Up until now, the viability of both approaches had been examined exclusively in simulation.

Contribution and outline

To the authors' best knowledge, the original contributions put forward in this paper are without precedent and can be condensed into the following points:

- First real-world experimental confirmation of RL integrated with a safety layer that can enforce any class of constraints and exploit an a priori safe fallback policy (OptLayerPolicy [11]).

- First real-world experimental confirmation of a metaheuristic-driven technique that yields an interpretable decision-tree-encoded control policy (TreeC [12]).

- A side-by-side assessment of rule-based control (RBC), MPC, safety-augmented RL (via OptLayerPolicy), and TreeC—on any case study, simulated or physical—conducted using an evaluation protocol engineered to be as equitable as achievable.

Related Work

The literature on energy management systems (EMS) has been comprehensively surveyed from multiple angles—spanning mathematical modeling paradigms [13] as well as objectives, system architecture, stakeholder roles, and implementation hurdles [14]—with dedicated reviews addressing the residential EMS domain [15]. A consistent observation, pointedly emphasized in the review articles by Rathor and Saxena [14] and Allwyn *et al*. [16], is that investigations predominantly rest on simulation-based case studies, whereas contributions reporting practical, real-world

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

deployment outcomes remain sparse. That said, several prominent exemplars deserve mention. Zhao *et al*. [17] deployed a nonlinear MPC scheme at the Hong Kong Zero Carbon Building that orchestrated an electric chiller and a combined cooling and power unit, leveraging a stratified chilled-water thermal storage tank as a flexibility buffer alongside a PV array, all operating under day-ahead electricity tariffs. Viot *et al*. [18, 19] commissioned an MPC system within the SYNERGI building of the University of Bordeaux, France, to regulate thermally activated building systems; the outcome was a pronounced 40% drop in energy use accompanied by thermal comfort levels that were either improved or unchanged. Later, Arroyo *et al*. [20] juxtaposed white-box, grey-box, and black-box modeling strategies in an MPC-centric home EMS applied to the thermally activated building systems of KU Leuven's Vliet building in Heverlee, Belgium. Elkazaz *et al*. [21] put together a two-layer EMS architecture on a lab-scale, grid-tied microgrid comprising an uncontrollable load, solar PV generation, a wind turbine, and battery storage (BESS). The upper-layer optimization determined optimal reference signals for the lower-layer MPC, yielding daily cost reductions of up to 30%. More recently, Zhang *et al.* [22] shared findings from a twelve-month trial of an MPC-driven EMS at a small commercial site in California, United States, furnished with solar PV, steerable space conditioning, commercial refrigeration, and a BESS. Two categories of demand flexibility were explored: (1) minimizing electricity expenditure under time-of-use rate structures and (2) reacting to grid flexibility requests. Relative to a supervised-learning-trained baseline, the results indicated a 12% annual reduction in electricity bills and a 34% curtailment of peak demand. At the same time, thermal comfort and food safety requirements remained satisfied. Yang *et al*. [23] likewise turned to supervised learning, albeit as a stand-in for an MPC, and ran experiments across two actual testbeds—an office level and a lecture theatre—in Singapore. The EMS handled the air-conditioning plant and still captured 52% of the cooling energy savings achieved by the original MPC (which reached 58%). Yet, it did so with a hundredfold decrease in computational demand.

When one seeks real-world, experimentally validated deployments of machine-learning-driven EMS, the pool of available studies shrinks considerably, a point underscored by Fu *et al*. [24] in their survey. A host of practical challenges still await empirical verification and thus persist as open issues, as articulated by Nweye *et al*. [25]. Combined with recent progress in imitation learning, this makes imitation-based approaches a more pragmatically viable route [26]. Nonetheless, a handful of concrete real-world deployments can be highlighted. Costanzo *et al.* [27] exhibited a model-assisted batch RL agent tasked with operating heating, ventilation, and air conditioning units subject to time-varying electricity prices for room conditioning in a living lab setting; within a span of 10 to 20 days, sensible control behaviors—further sculpted by domain-specific insights—could be attained. Kazmi *et a*l. [28] afterwards showcased a model-based RL agent targeting the optimization of domestic hot water provision, modulating the reheat cycles of a storage tank linked to an air-source heat pump across 19 occupied homes in the Netherlands. An efficiency uplift of 20% was documented, with no perceived erosion of occupant comfort (as observed on a subset of 5 residences). During roughly the same timeframe, Zhang and Lam [29] installed and assessed a deep RL agent for the fine-tuned regulation of a radiant heating installation within an occupied office building. The agent underwent offline pre-training on a physics-grounded model calibrated against actual operational measurements, after which it was placed online on the real heating apparatus without undertaking any further exploration. When benchmarked against an RBC approach across a three-month evaluation span, heating demand reductions ranging from 16.6% to 18.2% were observed. Turning to distributed energy assets, Touzani *et al*. [30] put forward a deep RL agent that optimized the combined operation of heating, ventilation, air conditioning, and a BESS in the presence of solar PV at the FLEXLAB facility of Lawrence Berkeley National Laboratory in Berkeley, California, United States. Offline training on a calibrated physics-driven simulation model preceded online rollout onto the physical installation. The results showed cost reductions as high as 39.6% while thermal comfort remained broadly equivalent over a three-week test window. Naug *et al.* [31] subsequently introduced a deep RL agent evaluated both on a simulated five-zone testbed and on an actual three-story edifice at Vanderbilt University, United States, where it continuously dictated heating setpoints in air-handling units for one year. Consistent with prior cases, offline training was followed by online activation; whenever a drop in performance was identified, data-driven models capturing the building's thermal dynamics were refreshed, and an offline relearning cycle subsequently updated the RL agent based on those revised models. As a closing instance, Gokhale *et al*. [32] executed a four-week experimental investigation into demand response for a cluster of eight residential units, coupling an RL-based energy coordination layer with a PI-type power dispatch mechanism. The dispatch layer held the authority to countermand any commands that threatened to breach thermal comfort boundaries, exploiting the structural thermal mass of the dwellings as a no-cost flexibility asset.

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

Although we have surveyed a broadly similar volume of real-world contributions, experimental case studies centered on machine-learning-based energy management remain far less numerous. Across almost the entire body of RL-oriented work, agents are trained in a risk-free offline fashion using meticulously constructed a priori simulators. Such simulators may neither be readily obtainable nor economically justifiable to develop on a per-project basis, and they can embed substantial modeling biases that translate into weak online performance. Even when exploratory actions are withheld during online execution, there is no guarantee that the deployed policy will behave safely over the full expanse of the state–action space. The limited number of studies that do undertake direct online RL training incorporate simplistic override logic that is both highly intrusive—particularly when set against approaches that nudge actions toward the nearest feasible solutions—and poorly generalizable. Moreover, while EMS formulations based on learnable decision trees have been described in the literature [33, 34], we are unaware of any prior demonstration of such methods on physical hardware. In response, we sought to close this gap by (1) fielding an online RL-based EMS augmented with a safety layer that accommodates arbitrary constraint categories and can fall back on an a priori safe default policy when one is available (i.e., OptLayerPolicy [11]), and (2) fielding a learned, interpretable decision-tree-based EMS (i.e., TreeC [12]) in an actual real-world environment.

Materials and Methods

Hardware setup

Four physical replicas of a residential electrical circuit form the experimental platform. Each unit incorporates a BESS, a PV system, a simulated household electrical demand, and simulated EV charging periods. The dwellings are linked to the grid and outfitted with official digital meters supplied by the local Belgian distribution system operator, which log both imported and exported energy flows. The BESS and PV configurations are unique to each unit, whereas the emulated consumption and EV charging schedules are uniform across all four. This configuration reproduces the electrical environment of a present-day home served by a non-electric heating solution.

The electrical demand representing consumption and EV charging is delivered by a dynamic load bank built from seven Perel® PHP2000 units, each rated at 2000 W and driven by a Showtec Single DP-1 dimmer module. A calibration procedure matches the commanded setpoints to the actual measured active power as closely as achievable; any remaining offset is presumed to stem from measurement error and production variability in the load elements.

Sensing equipment captures power, voltage, and current at the PV arrays, the BESS units, and the grid tie points. The power drawn by the dynamic loads and any parasitic losses within the rig are inferred from the difference among the PV, BESS, and grid readings.

A block diagram and a photographic record of the platform appear in **Figures 1 and 2**. Extended hardware characteristics —encompassing the BESS units, inverter models, PV panels, and the electrical topology of each house.

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

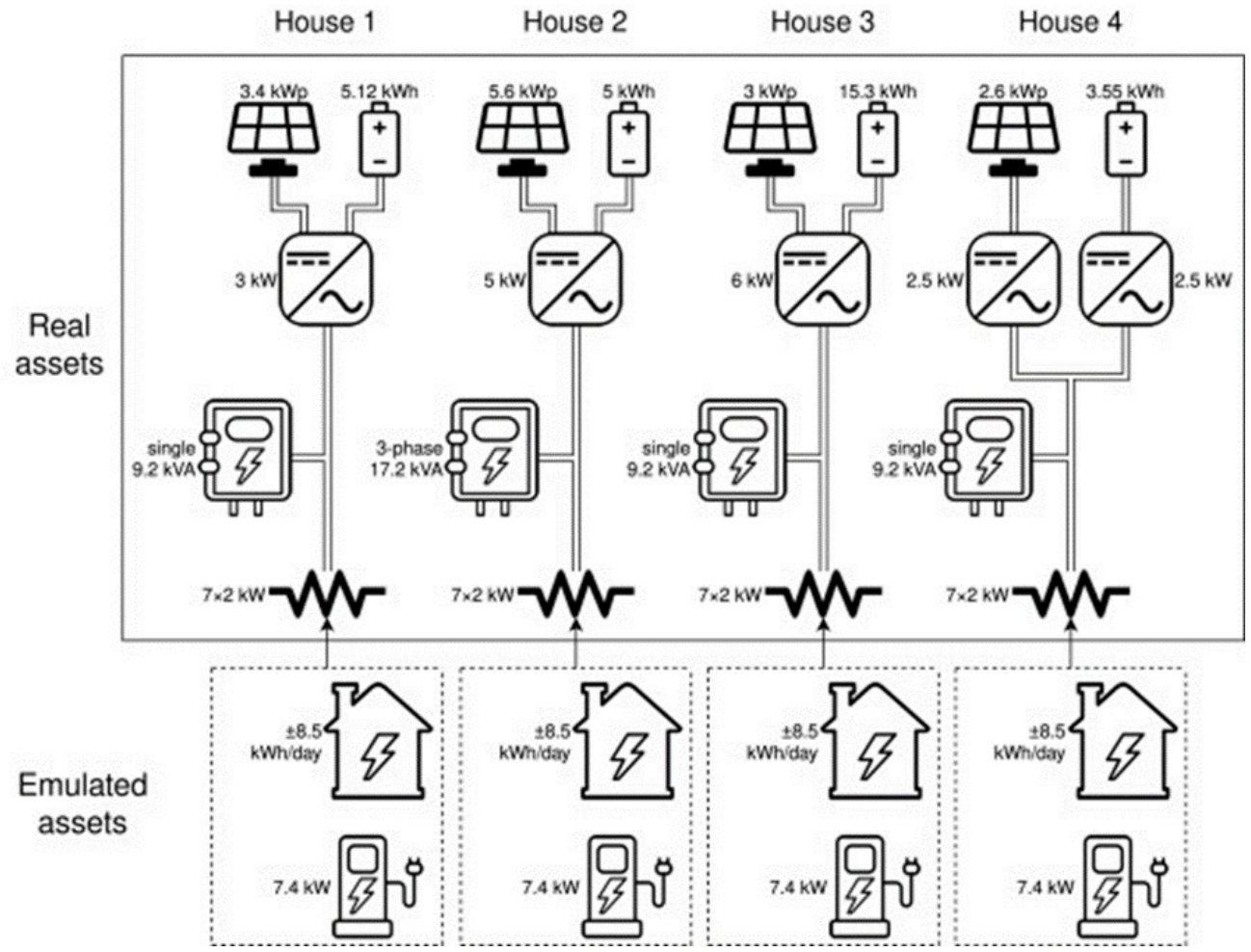

Figure 1. Schematic of the experimental setup per house with from top to bottom: the PV installation with its peak power, the BESS with its capacity, the inverter with its maximum power, the grid connection with its phase specification and maximum power, the dynamic loads with their maximum power, the emulated house consumption with its average daily consumption and the EV charger with its maximum charge power.



© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

Figure 2. Photograph of the experimental setup: with G being the grid connection equipped with the official digital meter of the local distribution system operator, B the BESS, S the solar PV installation inverters (solar panels equipped on the roof) and L the connectors to the dynamic loads (dimmers in a separate cabinet and electrical loads stationed outdoors). The numbers represent the house numbers.

Experimental setup

Consumption profile

The real-time (10 s) residential electricity consumption measurements from Schlemminger *et al*. [35] serve as the foundation, with the analysis confined to the 2019 subset that precedes the COVID-19 outbreak. The experiments adopt the records from single-family house (SFH) 19, a residence lacking solar PV, BESS, EV charging capability, and heat pump demand. This particular household was singled out for its superior data completeness and integrity, together with broadly representative consumption characteristics—an annual draw of 3425 kWh **(Figure 3).**

The raw timestamps were shifted to align weekdays correctly (so that a Monday during live operation corresponds to a Monday in the source records) and then deployed identically on every hardware-in-the-loop dynamic load bank, yielding the same real-time uncontrollable demand profile across all four houses.

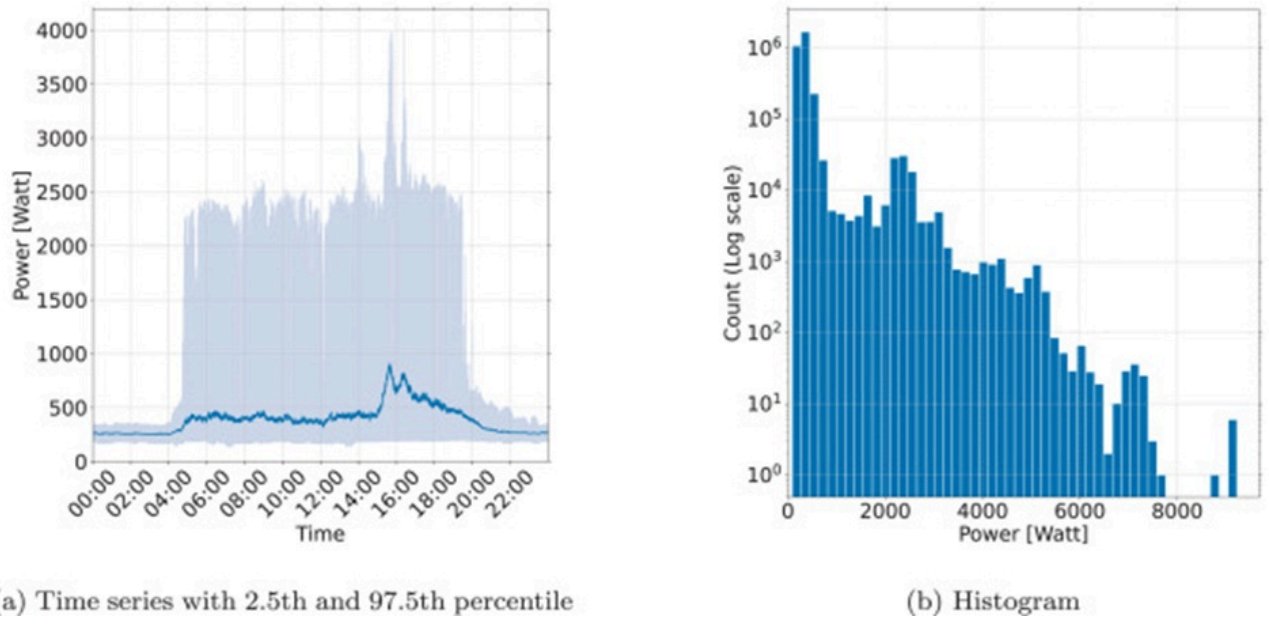

Figure 3. Household consumption profile: 10 s SFH 19 data [35] without PV, BESS, EV charging or heat pump having 3425.75 kWh/year, Std: 445.85 W, max: 9229 W.

Electric vehicle and driver charging behaviour

The residential e-mobility demand consists of two separate layers: (a) a behavioural layer capturing when and how the driver plugs in—arrival moment, departure moment, and required energy—and (b) a technical layer encoding the power delivery curve during charging.

For the behavioural layer, an existing open-source repository made available in Sørensen *et al*. [36] is used to drive the simulation directly. This collection documents 97 EV users and 6878 residential charging events recorded between

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

December 2018 and January 2020. From this pool, user "Bl2-2" was chosen because its arrival distribution, dwell time, and per-session energy need closely match the population average, yet it exhibits an unusually dense charging frequency. That density proves valuable for experimentation, as it multiplies the occasions for the EMS to intervene. In concrete numbers, this user accumulated 315 charging events and an aggregated energy transfer of 3274.74 kWh across 386 days.

For the technical layer, the charging power profile is assembled from a defined set of variables and operational bounds. The power ceiling is expressed by (1).

$$0 \leq P_t^{ev} \leq P_{max}^{ev} \quad (1)$$

Where P_{max}^{ev} denotes the maximum permissible charging power in [kW], fixed at 7.4 kW in line with the IEC 61851 charging framework [37]. The power trajectory follows the Constant-Current/Constant-Voltage (CC-CV) uncontrolled charging methodology laid out in Vagropoulos and Bakirtzis [38], under which power becomes a piecewise linear mapping of the state of charge (SOC), as summarized in **Eq. 2.**

$$P_t^{ev} = \begin{cases} P_{max}^{ev} & \text{if } SOC_t^{ev} \leq SOC_{cc,cv}^{ev} \\ P_{max}^{ev} - (P_{max}^{ev} - P_{min}^{ev}) * \dfrac{SOC_t^{ev} - SOC_{cc,cv}^{ev}}{1 - SOC_{cc,cv}^{ev}} & \text{if } SOC_t^{ev} > SOC_{cc,cv}^{ev} \end{cases} \quad (2)$$

Where SOC_t^{ev} is the SOC at instant t in [%], $SOC_{cc,cv}^{ev}$ is the transition SOC at which the Constant-Current regime hands over to the Constant-Voltage regime in [%], and P_{min}^{ev} is the charging power floor at 100% SOC, specified as 1 kW. The SOC and charging power variables are coupled through the balancing **Eq. 3.**

$$SOC_{t+1}^{ev} = SOC_t^{ev} + \eta^{charge} \times \frac{P_t^{ev} \times \Delta t}{E^{ev}} \quad (3)$$

Where SOC_{t+1}^{ev} is the projected SOC at the subsequent interval t + 1, Δt is the interval duration in [h], E^{ev} is the EV's energy capacity in [kWh], taken as 60 kWh, and η^{charge} is the charging efficiency in [%], assumed to be 95%, conforming to typical values for lithium-ion chemistries [39].

Tariff

The real-time pricing structure employed in this work is taken from a one-year dynamic supply contract of a major Belgian energy retailer, priced as though activated in April 2024.2 The overall bill is assembled from four cost components, each governed by a distinct pricing mechanism: the day-ahead energy cost, the offtake surcharges, the monthly peak cost, and the annual fixed cost.

The day-ahead component derives from the price definitions in **Eqs. 4 and 5.**

$$Vo_t = \begin{cases} (Vd_t + 0.011 \frac{€}{\text{kWh}}) * 1.06 & \text{if } (Vd_t + 0.011 \frac{€}{\text{kWh}}) \geq 0 \\ (Vd_t + 0.011 \frac{€}{\text{kWh}}), & \text{otherwise} \end{cases} \quad (4)$$

$$Vi_t = Vd_t - 0.009 \frac{€}{\text{kWh}} \quad (5)$$

Where Vot and Vit represent the per-unit offtake and injection prices at step t, respectively. Vdt is the Belgian day-ahead hourly market clearing price. A value-added tax of 6% is added to the offtake price whenever the latter exceeds zero.

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

Additional levies tied to energy withdrawn from the grid—encompassing transport tariffs, distribution fees, and government surcharges—are grouped and designated as the offtake extras cost. The aggregated per-unit price for these supplementary offtake charges, Vf, stands at 0.114 €/kWh.

The monthly peak offtake charge Cpm is computed via **Eq. 6.**

$$Cp_m = \frac{|B_m|}{N_m} Vp * \max_{t \in B_m}\left(\frac{Eo_t}{0.25h}, 2.5\ \text{kW}\right) \quad (6)$$

Where Vp is the peak tariff of 3.5 €/kW; this term penalizes the maximum offtake power, evaluated as a 15-minute rolling average, observed during each calendar month. For this study, the monthly peak charge is prorated according to the fraction of the month that falls within the billing window, so that partial months can also be accounted for. Bm denotes the set of all billed time steps belonging to month m, and Nm is the total step count within that month. Eot signifies the grid energy extraction at step t. The extracted energy is divided by the step duration to convert it to an average power in kW. A floor of 2.5 kW is imposed on the monthly peak power for billing purposes, even when the actual measured peak lies below that value.

The annual fixed charge derives from the yearly tariff Vy of €115.84 and is likewise prorated to the billing interval. The total electricity expense is computed by **Eq. 7.**

$$C_{tot} = \underbrace{\sum_{t \in T} (Eo_t Vo_t - Ei_t Vi_t)}_{\text{day-ahead cost}} + \underbrace{\sum_{t \in T} Eo_t Vf}_{\text{offtake extras cost}} + \underbrace{\sum_{m \in M} Cp_m}_{\text{peak cost}} + \underbrace{\frac{|T|}{N_y} Vy}_{\text{yearly cost}} \quad (7)$$

Where T collects all billed time steps, Eit is the energy fed back into the grid at step t, M is the collection of billed months, and Ny is the total step count spanning a full year.

House switching

The four houses possess non-identical BESS and PV characteristics. Assigning a different EMS to each house and judging their relative merit by operating them concurrently would confound the comparison; one house might register superior figures purely because it benefits from a more capable BESS or a larger PV array. To eliminate this source of systematic error, the EMS assignments are rotated among houses on a daily cycle.

With 4 houses and 4 EMSs under evaluation, the total number of distinct house–EMS pairings amounts to 24. Every pairing is executed twice, bringing the total trial length to 48 days. The 48-day sequence is constructed by randomly shuffling the 24 pairings until two independent 24-day schedules are obtained. This randomisation enforces the constraint that the same reinforcement learning or MPC EMS never occupies the same house on two successive days. The rule has a practical motivation: the RL and MPC EMS behaviours depend on their prior operation within a house. When test-day failures occur—which they occasionally do—preventing an MPC or RL EMS from staying on the same house for consecutive days creates a larger detection window. It stops a failure from contaminating the subsequent day's run.

The rotation is performed daily at 15:00. By this hour, each house's BESS has been replenished to a full 100% state of charge, ensuring uniform post-switch initial conditions. The EV schedule is modified to avoid any charging session straddling the 15:00 changeover moment. Overlapping sessions are truncated—either by advancing the end time or

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

delaying the start time to 15:00—with the option that retains the longest possible charging window being selected. The start-of-session and end-of-session SOC values prescribed by the original schedule are preserved.

This particular switching hour was adopted because 15:00 sees fewer charging events relative to the morning and evening peaks. By that time, the BESS has also had sufficient opportunity to absorb surplus PV generation from the preceding sunlit hours. Moreover, the day-ahead electricity prices for the coming day are already published by 15:00, sparing the MPC EMS from having to forecast them.

Enforced charging behaviour

A mandatory charging protocol is put in place to drive the BESS units to exactly 100% SOC when house switching occurs, and likewise to bring EV batteries to their prescribed terminal SOC at charging session completion. The protocol operates by commanding full charging power so that the SOC objective is met at the latest possible moment, diminished by a headroom period designated Btime. This headroom serves as a safeguard against shortfalls traceable to battery model inaccuracies, safety layer interventions (capable of downwardly adjusting battery commands), or any other unanticipated disruptions.

A larger discrepancy between the SOC objective and the actual SOC calls for a more generous headroom, since a wider deficit multiplies the potential for intervening events to derail timely completion. **Eq. 8** captures this dependency.

$$B_{time} = (SOC_g - SOC) * (B_{max} - B_{min}) + B_{min} \quad (8)$$

Where the symbols denote: B_{time}, the headroom duration; SOC the battery's SOC objective; SOC, the battery's present SOC; B_{max}, the headroom imposed when the battery is fully depleted and requires a complete recharge; and B_{min}, the minimum headroom applied irrespective of conditions. At a 5-second cadence, the battery controller evaluates—by means of **Eqs. 3 and 11**—whether the battery remains on track to satisfy its SOC objective by a checkpoint instant, computed as the target instant less the headroom. Should the evaluation return a negative answer, a 1-minute period of maximum-rate charging is triggered. Both the stationary storage and the EV batteries share a B_{min} of 3 minutes. For B_{max}, however, the values diverge: 1 hour for the stationary units versus 4 hours for the EV packs. These parameter choices were settled by simulating the full 3-month training corpus and selecting values that yielded consistent, punctual SOC target attainment.

Safety layer

In light of our intention to demonstrate energy management strategies that lack inherent safety guarantees, the OptLayerPolicy safety filter from Ceusters *et al*. [11] is incorporated—regarded as representing the current state of the art. The filter works by minimally perturbing proposed control actions to project them onto the nearest safe, feasible set, without compromising optimality considerations. A further capability is quantifying how far a proposed action is from feasibility, enabling a switchover to a pre-established safe fallback policy should that distance grow too large. The logic underpinning this is that actions remote from the feasible region are probably remote from optimality too, whereas actions near feasibility may well be near-optimal. The distance value at which fallback occurs is a tunable parameter of OptLayerPolicy, and its utility was demonstrated on a simulated benchmark by Ceusters *et al*. [11].

Although the safety layer's principal beneficiary is the RL-based EMS, the constraint satisfaction assurances it delivers are generic in nature, extending protection even to the elementary RBC reference controller—which, on its own, is not guaranteed to remain safe under all circumstances. The sampling period chosen for the safety layer matches the 5-second hardware-in-the-loop cycle. By adopting this rate, the safety layer is effectively enlisted as a real-time optimisation stage inside a cascaded (equivalently, hierarchical) control structure [40]—a role it fulfills for the MPC-based EMS as well.

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

A practical consequence is that the EMS modules are free to operate at a substantially coarser 15-minute rhythm; their computed setpoints then act as reference signals for the safety layer, which departs from them solely when constraint adherence leaves no alternative, factoring in the most recent measurement updates—such as when solar PV injection suffers a sudden collapse.

In the context of the experiments reported here, the OptLayerPolicy formulation is cast as:

$$a_t^{safe,bess}, a_t^{safe,ev} = \arg\min_{a_t^{bess}, a_t^{ev}} \frac{1}{2}\|(a_t^{bess} - \tilde{a}_t^{bess}) + (a_t^{ev} - \tilde{a}_t^{ev})\|^2 \quad (9a)$$

$$\text{s.t.} \quad -P_{max}^{grid} \leq P_t^{grid} \leq P_{max}^{grid} \qquad \forall t \quad (9b)$$

$$0 \leq P_t^{ev} \leq P_{max}^{ev} \qquad \forall t \quad (9c)$$

$$P_{min}^{bess} \leq P_t^{bess} \leq P_{max}^{bess} \qquad \forall t \quad (9d)$$

A quadratic program whose internal decision variables a_t^{bess} and a_t^{ev}, associated with the BESS and EV respectively, yield the nearest feasible, safe actions $a_t^{safe,bess}$ and $a_t^{safe,ev}$ —the actual directives transmitted to the physical assets—originating from the proposed actions and delivered by the EMS (the intended setpoints). The program enforces the grid exchange (power balance) bound and the individual power envelopes of the BESS and EV charger, expressible in expanded form as:

$$P_t^{grid} = -P_t^{load} - P_t^{ev} + P_t^{pv} + P_t^{bess} \qquad \forall t \quad (10a)$$

$$P_t^{ev} = f(a_t^{ev}, P_{max}^{ev}, SOC_t^{ev}) \qquad \forall t \quad (10b)$$

$$P_t^{bess} = \gamma_t^{bess} \cdot P_t^{ch} + (1 - \gamma_t^{bess}) \cdot P_t^{di} \qquad \forall t \;,\; \gamma_t^{bess} \in \{0, 1\} \quad (10c)$$

$$P_t^{ch} = \begin{cases} a_t^{bess} \cdot P_{min}^{bess} & ,\text{if } SOC_t^{bess} < 1 \\ 0 & ,\text{if } SOC_t^{bess} = 1 \end{cases} \leq 0 \quad \forall t \quad (10d)$$

$$P_t^{di} = \begin{cases} a_t^{bess} \cdot P_{max}^{bess} & ,\text{if } SOC_t^{bess} > 0 \\ 0 & ,\text{if } SOC_t^{bess} = 0 \end{cases} > 0 \quad \forall t \quad (10e)$$

Where the quantities Pt designate the electrical powers at step t attributed, sequentially, to grid interchange, uncontrollable demand, EV charging, PV generation, and BESS operation. The functional relationship f (.) appearing in **Eq. 10b** is supplied by **Eq. 2.** A binary decision term, γ_t^{bess} , is appended to the BESS power expression to forbid concurrent charge and discharge. The scope of this case study is confined to upholding the electrical energy balance and the associated grid ceiling; further constraint categories—ramping limitations, minimum uptime and downtime—are outside its remit.

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

The safe fallback rule π_{safe}, triggered in the present study whenever the feasibility gap d_{safe} surpasses 1e6, reverts the BESS to its self-consumption operating mode and re-invokes the safety layer to solve for the maximum allowable EV charging power. The latter re-invocation is needed because an unqualified maximum-power instruction could still violate the grid interchange bound and cause the main breaker to open.

Simulation

A simulation replica of the residential setup was developed to provide a training environment for the TreeC EMS and a benchmark against which the physical experimental findings could be juxtaposed. The simulation clock advances in 15-minute jumps. It emulates the behaviour of the BESS, EV charging dynamics, PV output, electrical demand, and grid coupling, as well as the way these components react to the distinct control policies enacted by the EMS.

The BESS representation within the simulation follows **Eq. 11.**

$$SOC_{t+1}^{bess} = SOC_{t}^{bess} + \frac{P_t^{ch} * \eta - P_t^{di} / \eta^{bess}}{E^{bess}} \Delta t \quad (11)$$

Where SOC_t is the state of charge at step P_t^{ch} and P_t^{di} are the charge and discharge powers at step t, E^{bess} is the nominal energy storage capacity, η^{bess} is the round-trip charge/discharge efficiency, and Δt is the temporal granularity of a step. The efficiency η^{bess} is taken to be 95% [39] for BESSs 1, 2, and 4, and 96% for BESS 3, as attested by the unit's technical documentation.

The EV is modelled via **Eqs. 1-3**. Measured PV output traces and measured electrical load traces are injected straight into the simulation. Grid power is determined from **Eq. 10a.** For modelling simplicity, the restriction that the hybrid inverter imposes on the combined instantaneous power from the PV array and the BESS is left unenforced. Neglecting this detail is not expected to distort the simulation outcomes in any meaningful way, because the hybrid inverter ceiling is practically never encountered under the configurations tested here.

The mandatory charging behaviour is replicated in the simulation with reasonable fidelity. At each step, **Eq. 8** recomputes the headroom. If following the EMS-specified charge or discharge power would leave the SOC objective unreachable by the checkpoint moment, the battery is instead forced to charge at the lowest power that restores feasibility.

Energy management systems

The present section introduces the four EMS approaches evaluated and the dataset underpinning the training of the MPC, RL, and TreeC variants. Every 15 minutes, each EMS dispatches a setpoint pair—one for the BESS, one for the EV charger. How these setpoints are derived is the subject of the following exposition.

Training data

The MPC, RL, and TreeC EMS formulations all call for training data; a three-month window stretching from 2024/01/01 00:00 to 2024/04/01 00:00 was employed for this purpose. The dataset encompasses real PV generation measurements, the household demand signal presented and the EV charging episodes. The MPC exploited these records to calibrate its forecasting modules. The RL agents were pre-trained over the same temporal span, supplemented with artificial BESS and EV charger power trajectories generated as though an RBC-type controller had been in command. TreeC used the same data within the simulation framework to synthesise its decision tree policy.

Rule-based control

The RBC EMS fielded in this study implements the default logic typically pre-programmed into off-the-shelf devices. The BESS rule prioritises behind-the-meter self-consumption: it stores energy when PV output outstrips household demand

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

and releases that stored energy when demand overtakes generation. The EV charger rule is to start charging instantly, at the maximum power the vehicle permits. This RBC controller constitutes the reference baseline against which the three other EMS candidates are assessed.

Model predictive control

Within this experimental framework, MPC serves as an EMS technique reliant upon an explicit plant representation together with projected future exogenous signals to drive optimal decision-making. The particular instantiation formulated here takes the shape of a Mixed-Integer Linear Program, recomputed afresh at each 15-minute interval. Rather than employing a fixed-length horizon, the MPC looks ahead only as far as the next scheduled house-switching event. The optimizer is invoked every quarter hour to determine the power trajectories for the two steerable elements—the BESS and the EV charging station.

Eq. 12 formalises the cost functional that the MPC seeks to minimise.

$$O = \underbrace{\sum_{t\in H}(Eo_t Vo_t - Ei_t Vi_t)}_{\text{day-ahead cost}} + \underbrace{\sum_{t\in H} Eo_t Vf}_{\text{offtake extras cost}} + \underbrace{Vp * \max_{t\in H}\left(\frac{Eo_t}{0.25h}, Pp\right)}_{\text{peak cost}} \quad (12)$$

In this expression, H denotes the collection of time indices falling within the prediction window. At the same time, Pp represents the historical maximum grid draw recorded for the particular house while governed by the MPC policy. At the outset of experimentation, Pp is seeded with a value of 2.5 kW. The cost structure is designed to heavily penalise any power draw that surpasses this previously established peak Pp.

The information stream entering the MPC models comprises: the hour-of-day timestamp, the state-of-charge levels of both the stationary storage unit and the EV battery, the stream of Belgian day-ahead hourly clearing prices Vdt covering the span until the next switching occurrence, the forecast trajectories (also extending to the next switching event) for PV generation and household demand, plus the projected end-of-session SOC and the anticipated unplug time for whichever EV is currently attached.

Regarding the constraint set imposed on the MPC, it coincides with those introduced in **Eqs. 1, 3, 11, 9b, 9d, and 10a**. One deviation, however, arises in the treatment of the EV charging characteristic originally specified by **Eq. 2**; the adapted formulation appears in **Eq. 13.**

$$P_t^{ev} \le \begin{cases} P_{max}^{ev} & \text{if} \quad SOC_{t+1}^{ev} \le SOC_{cc,cv}^{ev} \\ P_{max}^{ev} - (P_{max}^{ev} - P_{min}^{ev}) * \dfrac{SOC_{t+1}^{ev} - SOC_{cc,cv}^{ev}}{1 - SOC_{cc,cv}^{ev}} & \text{if} \quad SOC_{t+1}^{ev} > SOC_{cc,cv}^{ev} \end{cases} \quad (13)$$

The rationale behind this adaptation is that computing the mean maximum permissible EV charging rate over the upcoming interval once the threshold $SOC_{cc,cv}^{ev}$ has been crossed would inject an exponential dependency into the MPC, compromising its linear formulation. Linearity is preserved by instead capping the EV's allowable power based on the SOC value anticipated one step ahead, at t + 1.

A mandatory condition forces the modelled BESS SOC to attain 100% precisely at the upcoming switching instant. An analogous stipulation guarantees that the EV battery reaches its forecast final SOC by the forecast departure moment.

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

Because these forecasted quantities—final SOC and departure time—can be underestimated, the enforced charging protocol may subsequently be activated. To account for this, the approximated enforced EV charging power anticipated for the next interval is imposed as a lower bound on the EV power setpoint within the MPC formulation. Lastly, a ceiling constraint ensures the modelled grid power never breaches the 9.2 kW hard limit.

A Light Gradient Boosting Machine (LightGBM) model [41] is entrusted with generating forecasts of both electrical load and PV output. From the immediately preceding 96 quarter-hourly observations, the model predicts the next 96 such values (equivalent to a full 24-hour window), supplemented with cyclical temporal features—time of day and day of the week. Hyperparameter tuning is conducted via the Tree-structured Parzen Estimator [42], with the search boundaries enumerated in **Table 1**. A deliberate choice was made to omit externally sourced meteorological data as an input feature for the PV forecasting component, thereby insulating the system from potential failure points such as remote server outages during maintenance. For load prediction, a single model is trained on the original consumption series and applied uniformly across the four houses; PV generation forecasting, by contrast, is performed on a per-house basis, with distinct models trained for each installation. The forecasting of EV charging duration and energy requirements employs k-Nearest Neighbors (k-NN) [43] regression. These k-NN models operate on two features—the initial SOC and the hour of the day—and identify the most analogous historical occurrences through cosine similarity. Prior work has attested to the superiority of this approach in data-sparse regimes, including the estimation of EV charging session attributes [44].

Table 1. Hyperparameter ranges for LightGBM model.

Hyperparameter	Evaluated values/Range
Learning rate	0.001, 0.01, 0.1, 0.2, 0.3
Number of leaves	32, 64, 128, 256
Maximum tree depth	2, 4, 6, 8, 10
Column sampling per tree (ColSampleByTree)	0.6, 0.8, 1.0
Minimum child samples	1, 2, 3, 4, 5

Reinforcement learning

The RL agent tasked with energy management is formalised as a fully observable discrete-time Markov Decision Process, encapsulated by the tuple $\langle S, A, Pa, Ra \rangle$, structured as follows:

$$S_t = (P_t^{load},\ P_t^{pv},\ SOC_t^{bess},\ SOC_t^{ev},\ Vd_t,\ H_t,\ D_t) \qquad S_t \in S \quad (14a)$$

$$a_t^{bess} = (a_{min}^{bess} \rightarrow a_{max}^{bess}) \qquad a_t^{bess} \in A \quad (14b)$$

$$a_t^{ev} = (0 \rightarrow a_{max}^{ev}) \qquad a_t^{ev} \in A \quad (14c)$$

$$P_a(s, s') = Pr(s_{t+1} = s' \mid s_t = s, a_t = a) \quad (14d)$$

$$R_a(s, s') = -(\underbrace{Eo_t Vo_t - Ei_t Vi_t}_{\text{day-ahead cost}} + \underbrace{Eo_t Vf}_{\text{offtake extras cost}}) - z \quad (14e)$$

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

Where P_t^{load} denotes the uncontrollable household demand, P_t^{pv} the solar generation infeed, followed by the state of charge of the stationary battery and of the EV, Vdt the pricing signal (the Belgian day-ahead market price), Ht the hour within the day, and Dt the day within the week—all referenced at time index t and collectively defining the state space S. The continuous action space A encompasses the control directives a_t^{bess} destined for the stationary storage and a_t^{ev} for the EV charger. The symbol Pa encodes a state-transition kernel that depends exclusively on the current state and is conditionally independent of all earlier states (thereby satisfying the Markov Property): should the system reside in state s ∈ S at step t and action a ∈ A be selected, the successor state at t + 1 will be some s′ ∈ S.

The learning objective, embodied in the reward function, is architected such that maximising this function corresponds to minimising its positive counterpart. In operational terms, the agent seeks to drive down the sum of the day-ahead and offtake surcharge components (extracted from **Eq. 7** but evaluated over the immediately preceding 15-minute step), denominated in EUR, plus an auxiliary penalty term z = 1 intended to sculpt the reward signal further when the raw, uncorrected proposed action a˜ was predicted to fall outside the feasible constraint set. It should be borne in mind that the RL agent's goal is maximisation of the expected sum of discounted future rewards (the return).

The algorithmic machinery selected for this task is a twin delayed deep deterministic policy gradient agent (TD3 [45], an extension of DDPG incorporating twin critics and delayed policy updates), adopted from the stable baselines [46] library, configured according to the hyperparameter specification collected in **Table 2.**

A distinct RL agent is instantiated per house. All agents undergo a pre-training phase based on Behavioral Cloning (BC) [47], using the 3-month archive of historical data. With an elementary random train–test partition reserving 20% of the data for evaluation, executed across 10 distinct random seeds, the performance indicators summarised in **Table 3** are obtained.

Table 2. TD3 hyperparameters. The values stem from an optimisation investigation documented in Ceusters *et al*. [48], albeit with train_freq increased and learning_starts decreased relative to the original recommendations.

Parameters	TD3
gamma	0.7
learning_rate	0.000583
batch_size	16
buffer_size	1e6
train_freq	(4, “step”)
noise_type	normal
noise_std	0.183
learning_starts	96

Table 3. Pre-trained TD3 BC metrics over 10 random seeds and using 20% randomly split test data and with: epochs = 3, gamma = 0.7, learning_rate = 1.0, batch_size = 64 and test_batch_size = 100.

Metric	House 4	House 3	House 2	House 1
R2-score [–]	0.4617	0.5837	0.4997	0.4746

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

MAE [W]	299	481	344	372
RMSE [W]	721	855	734	770

TreeC

TreeC designates an EMS paradigm in which the control logic is encoded as a decision tree, itself the product of a metaheuristic synthesis procedure. Candidate trees are scored by simulating the governed system across a training dataset. Over successive iterations, the metaheuristic uncovers progressively improved trees; the tree achieving the highest score is subsequently deployed as the operational EMS. Reference [12] provides an exhaustive account of the TreeC methodology.

Within each house, one tree is evolved to regulate the BESS and a second, separate tree to govern the EV charger. The trees intended for experimental deployment are built from the training data specified in **Eq. 14a**, the simulation layer and the cost functional stated in **Eq. 7.**

Relative to the baseline TreeC recipe originally set out in Ruddick *et al*. [12], two adjustments have been introduced:

- The Covariance Matrix Adaptation Evolution Strategy exhibited a persistent tendency to converge to local minima across multiple independent training runs. It was therefore supplanted by particle swarm optimisation [49], which yielded reliably better training outcomes. The variant adopted here is that provided by the Python library pygmo [50] under its default parameterisation, which corresponds to the canonical particle swarm formulation of Poli *et al*. [51].

- Upon termination of the training phase, the best tree undergoes a pruning stage: leaves are excised so long as the resulting deterioration in the objective score remains bounded by 1% of the score achieved by the intact tree. Leaves that were traversed least frequently during evaluation are prioritised for removal.

The pool of candidate split features available to the decision trees corresponds to the state space of the RL agent, enlarged by one additional signal: the day-ahead price, shifted such that the minimum price observed over the course of the current experiment day is anchored at zero. This supplementary feature is constructed by taking the instantaneous day-ahead price and subtracting from it the lowest price registered within the experiment-day window delimited by the preceding switching instant and the forthcoming switching instant.

The action catalogue of the TreeC EMS spans charging the BESS at assorted power levels, discharging it at assorted power levels, and entering a self-consumption operating mode. The action variable is normalised to the interval [0, 1]. Whenever the normalised action value falls below 0.1, self-consumption is engaged; for values at or above this threshold, the BESS is either charged or discharged at a power rate obtained by mapping the normalised value back to the original scale.

During the training phase, TreeC invokes particle swarm optimisation over 1000 generations with a swarm comprising 1000 particles. This amounts to an evaluation of 1,000,000 distinct decision-tree–based EMS candidates in simulation over the three-month training window, with the single highest-scoring candidate retained. The choice of a 1,000-particle swarm was motivated by its superior performance on real-world and non-unimodal benchmarks drawn from the IEEE Competitions in Evolutionary Computation, relative to the more conventional swarm sizes of 20–50 individuals [52]. In keeping with standard TreeC practice, five separate EMS candidates are generated per house via repeated training and pruning cycles. The candidate attaining the best performance among the five is then selected for the live experiment.

The TreeC EMS obtained after training for house 1 is depicted in **Figure 4.**

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

The resulting control logic admits straightforward human interpretation. The BESS opts to charge only when the prevailing electricity price lies near the daily minimum for the experiment day; at all other times, it defaults to self-consumption. The EV charger adopts a high charging rate when the BESS is nearly replenished; otherwise, it selects a low charging rate after 15:30 and a moderate charging rate before 15:30.

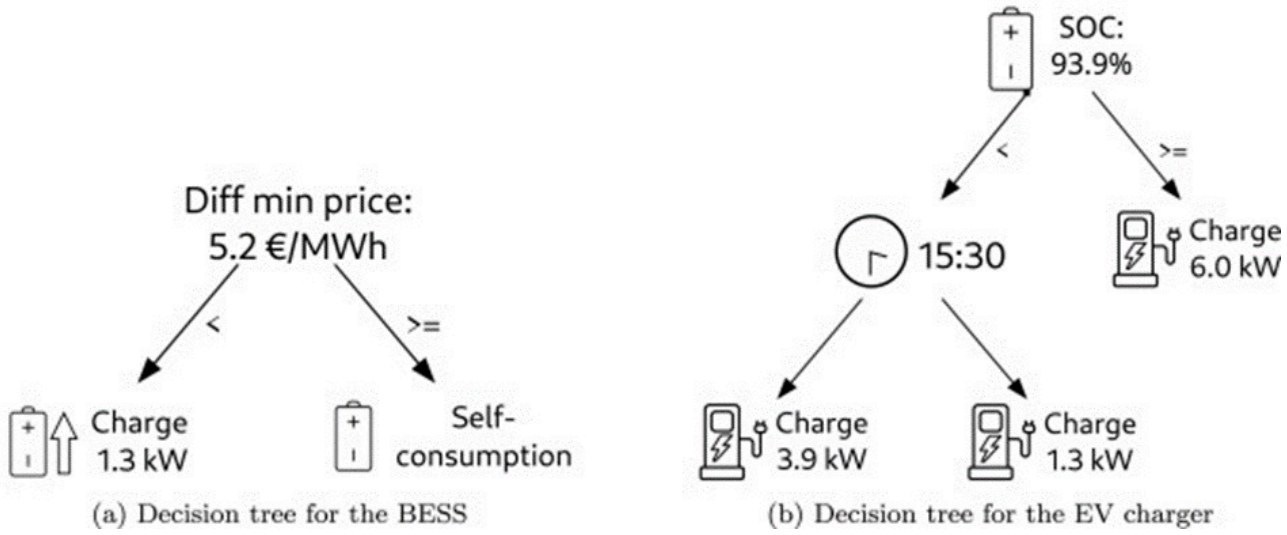

Figure 4. The TreeC EMS obtained after training and pruning for house 1.

Experiment execution

Data collection commenced on 2024/04/11 at 15:00 and concluded on 2024/06/17 at 15:00, targeting a yield of 48 experiment days suitable for comparison. Across this time span, 19 days proved unusable and were excluded because BESS charge targets were missed at the changeover hour, measurement gaps occurred, or protection device trips occurred. The specific failure modes break down as follows:

- A server reboot disrupted the execution routine (2 days)

- The BESS at house 1 entered an unresponsive state, persisting until a manual reset was performed (2 days)

- Upon attaining a full charge, the BESS at house 4 became unresponsive; a patch was introduced to cap the state of charge at 95% (1 day)

- The initial patch for house 4 did not resolve the underlying issue. A replacement fix was rolled out that forced self-consumption operation whenever SOC rose above 95% (2 days)

- A branch circuit breaker on the dynamic load circuit of house 2 tripped, its interrupting rating proving inadequate for sustained high loading; a unit with a higher breaking capacity was substituted (6 days)

- A previously overlooked software exception halted the control program (2 days)

- The main breaker of house 3 opened after apparent power draw remained excessive for an extended duration (2 days)

- Grid metering data from house 1 became unavailable due to a fault in its digital meter (2 days)

The experiment underwent several mid-run modifications. As noted in the list above, the hardware platform **(Figure 1)** exhibited an unexpectedly large reactive power contribution from the dynamic load banks, sufficient to cause a breaker

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

trip. The underlying issue was that the safety layer had been constructed using active power quantities alone, in effect assuming unity power factor. To address this, at the midway point of the experiment, reactive power was folded into the constraint definitions (**Eqs. 9b, 10a**), so that limits were placed on apparent power at the grid interface rather than on active power exclusively. This revision turned **Eq. 10a** into a quadratic constraint, elevating the mathematical program to a Mixed-Integer Quadratically Constrained Quadratic Program. The benefit, however, was that the main breaker current was now correctly bounded. Additional readings taken at the main grid breaker established that the 9.2 kVA apparent power threshold equates to approximately 8.7 kW of active power. Accordingly, the maximum grid power enforced inside both the simulation environment and the MPC constraint set was adjusted downward from 9.2 kW to 8.7 kW. This rectification was applied on 2024/5/21 at 15:00.

Two additional minor corrections were introduced during the trial. The SOC value supplied to the TreeC EMS for the EV battery was found to be underscaled by a factor of 100. Fixing this affected only one charging session and left overall performance largely unchanged. The scaling error was also backported into the simulation (corrected 2024/4/15 15:00:00). Separately, the maximum allowable SOC for the BESS of house 4 was clamped to 95%, as already described in the bullet points above (first correction 2024/04/22 15:00:00; second correction 2024/04/30 15:00:00).

Figure 5 provides a visual record of one experimental day during which the RBC EMS governed house 3. Displayed are the power traces of the individual assets, the state-of-charge evolution of the EV and BESS, the enforced charging sequence applied to the BESS in the day's final hours, and two windows during which the safety layer overrode the EV and BESS power commands.

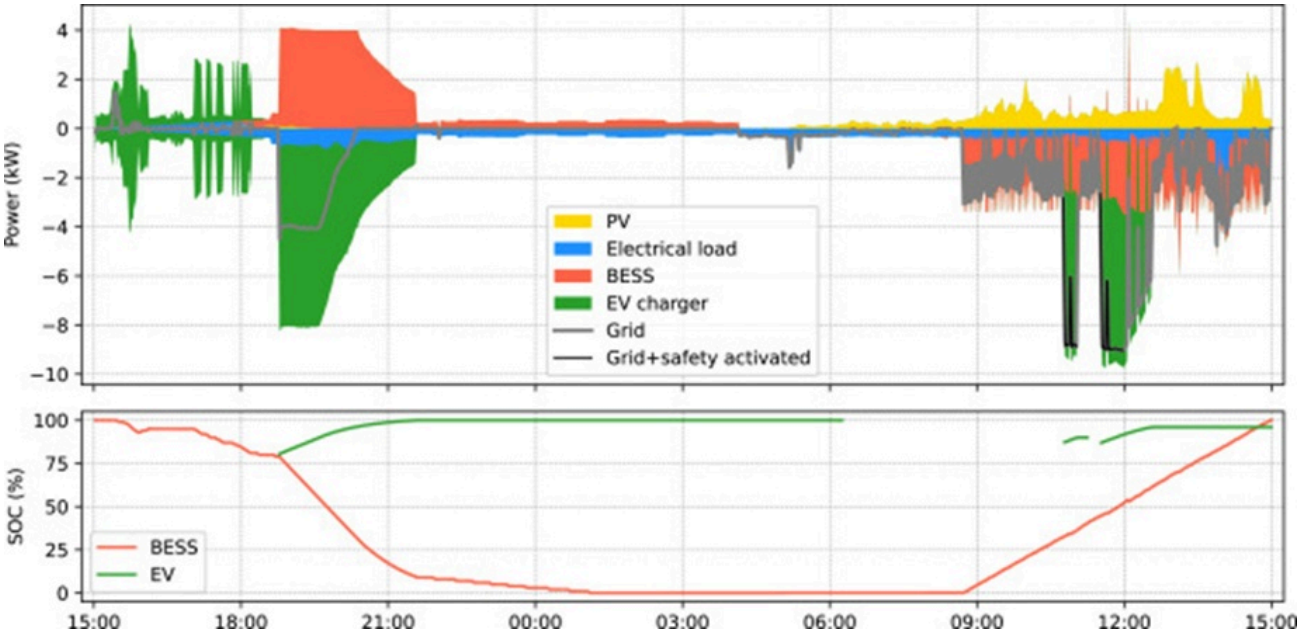

Figure 5. Representation of the experiment day starting on the 30th of May at 15:00. The upper plot shows the power profiles of the different assets of house 3 controlled by the RBC EMS. The grid power is represented in grey when the safety layer is not activated and in black when it is. The bottom plot shows the SOC of the BESS and EV. The enforced charging behaviour of the BESS is executed at the end of the experiment day to reach the 100% SOC goal at 15:00.

Results and Discussion

EMS performance comparison

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

Experimental outcomes across the RBC, TreeC, and MPC controllers proved broadly comparable, as depicted in **Figure 6**. The RL-based EMS delivered inferior results, a predictable consequence of its having undertaken a substantially larger number of exploratory moves during the trial. The tight clustering of the other three approaches was not anticipated; the explanations are developed in the subsections that follow.

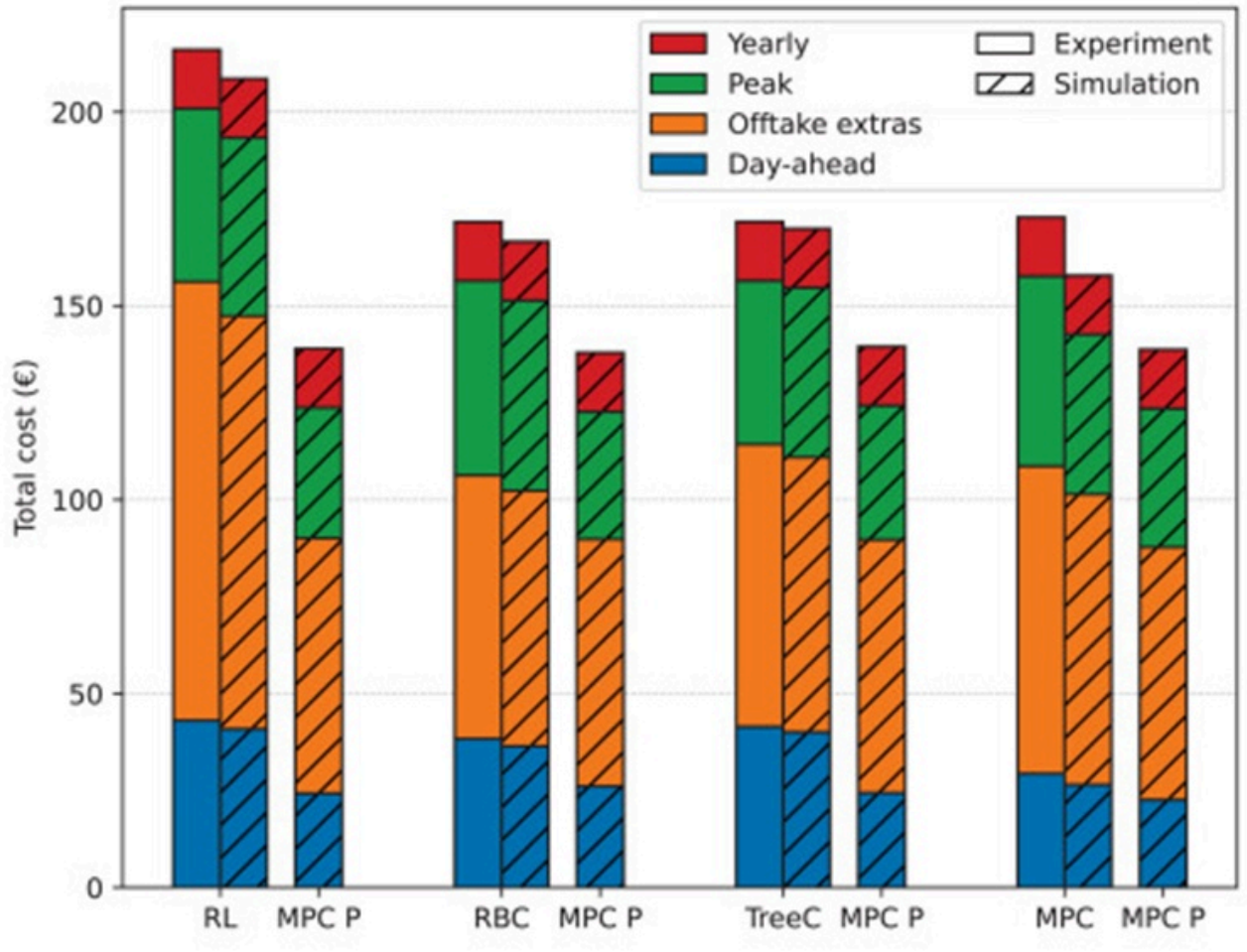

Figure 6. Costs of electricity for each EMS over the 48-day experiment from the real measurements and simulation. The total cost of electricity is divided into the four different costs from Eq. (7). In addition to EMS costs, the simulated costs of an MPC EMS with a perfect forecast are also shown with the label MPC P. The MPC P costs are simulated using the same set of 48 days and house combinations as the EMS on its left and serve as benchmarks.

None of the fielded EMSs approached the performance level of the MPC P reference, a clear indication that improved implementations or altogether different techniques remain discoverable. The MPC P controllers yielded total costs clustered between €137.9 and €139.5, notwithstanding having been assessed across distinct day–house assignments (matching those of the EMS for which each functioned as benchmark). The modest spread among these reference values supports the fairness of the comparison methodology.

The figures compiled in **Table 4** indicate that the RL EMS imported more energy from the grid than the competing controllers and registered a higher net consumption figure. Greater net consumption is most naturally attributed to heavier battery cycling, entailing additional round-trip efficiency losses.

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

Table 4. Electrical energy metrics for each EMS over the 48-day experiment from the real measurements. Net consumption is defined as the difference between grid-imported and grid-exported energy.

Result	MPC	TreeC	RBC	RL
Grid imported [kWh]	694.9	638.6	595.3	991.7
Grid exported [kWh]	190.5	142.2	91.3	458.3
Net consumption [kWh]	504.4	496.3	504.0	533.4

Rule-based control

The RBC approach centres on maximising self-consumption—put differently, on drawing as little energy from the grid as feasible, a tendency clearly reflected in **Table 4**. This operational stance translates into a comparatively low offtake extras bill, which, it should be noted, constitutes the dominant cost category for every EMS examined.

TreeC

Among the tested controllers, TreeC turned in the lowest peak cost, yet its offtake extras and day-ahead costs did not descend to the levels attained by the MPC and RBC alternatives. A look at the decision trees produced during training **(Figure 4)** suggests that the favourable peak cost figure is attributable to the assignment of reduced EV charging setpoints.

On an intuitive level, TreeC's inability to better the elementary RBC seems puzzling. **Figure 7** reveals that, for houses 1, 2, and 4, the MPC P benchmark costs sit 10%–15% below those recorded by TreeC. For house 3, however, the gap balloons: the MPC P benchmark undercuts TreeC by 33%, a far starker differential than observed elsewhere. Comparing the decision trees obtained from training **(Figure 4)**, the feature that most clearly separates house 3 from the others is the BESS charging logic. Houses 1, 2, and 4 charge their stationary storage only during episodes of very low electricity price, defaulting to self-consumption the remainder of the time. House 3, in contrast, remains in self-consumption mode until midnight, after which it charges the BESS without interruption until the 15:00 switching hour. While such a policy might have been entirely rational during the January–March training period, when solar generation is sparse, during the April–June experimental window it would be expected to drive higher grid offtake—and therefore elevated offtake extras charges—relative to the self-consumption-favouring RBC strategy. When the BESS policy of house 3 is replaced, in simulation, with the RBC self-consumption rule, the MPC P benchmark cost for that house drops to only 15% below TreeC, as illustrated in **Figure 7**. Adopting this TreeC + RBC hybrid for house 3 would have lowered the aggregate TreeC cost by €10.5, enough to place it ahead of the RBC EMS in total cost summed over the four houses **(Figure 6)**.

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

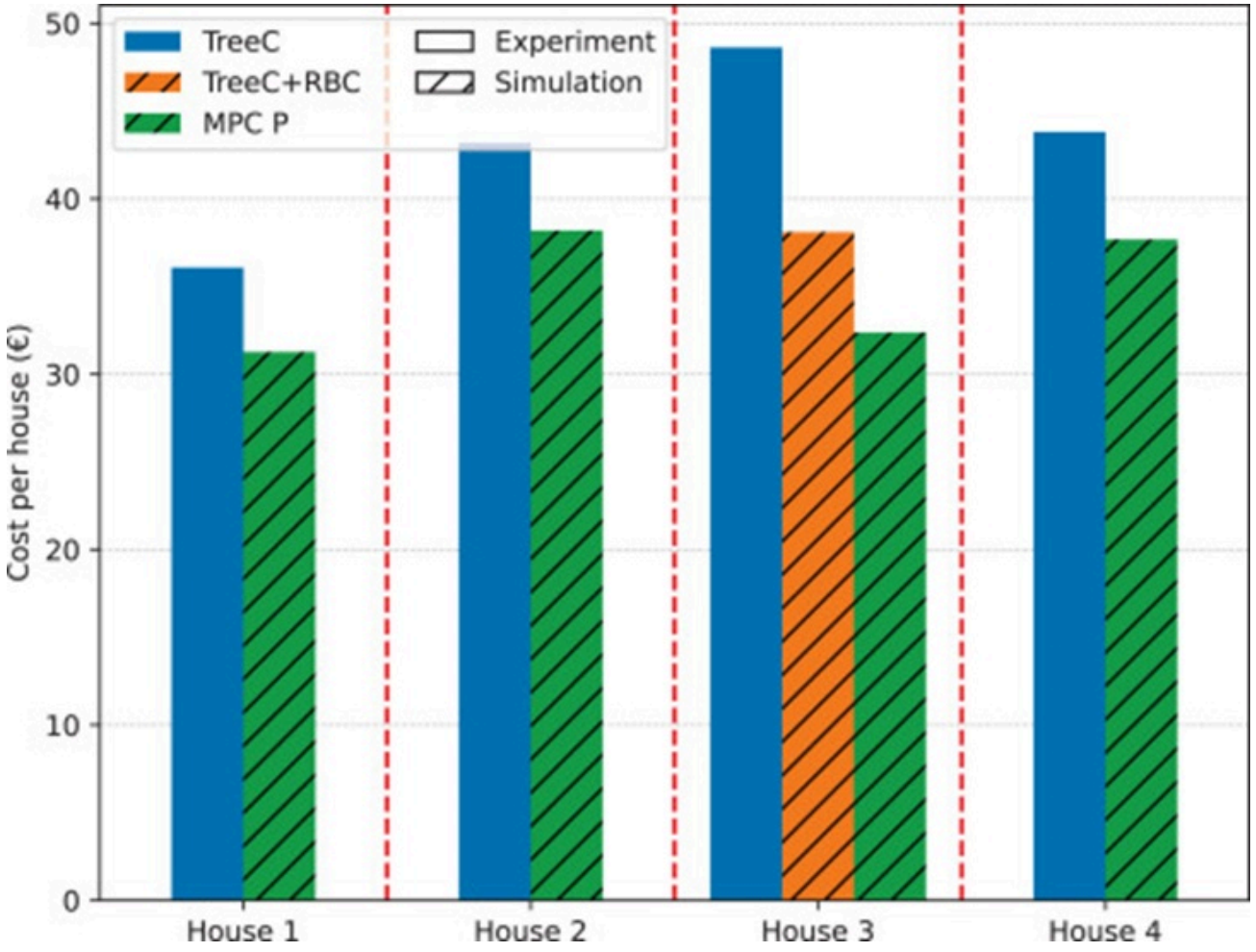

Figure 7. Comparison for each house of the total cost obtained with the TreeC EMS and the benchmark MPC P EMS. For house 3, the TreeC + RBC EMS is added to the comparison and represents the TreeC EMS using the self-consumption control strategy of the RBC EMS to control the BESS.

This analysis demonstrates how vital it is that the training data span be representative of the deployment period. Here, the training window was not sufficiently aligned with the experimental conditions, and as a result, TreeC converged on an unsuitable BESS strategy for house 3.

Model predictive control

The MPC EMS recorded the lowest day-ahead costs among the evaluated controllers, though its peak and offtake extras costs did not match the levels achieved by the RBC and TreeC approaches. A low day-ahead cost is unsurprising: the MPC possesses full knowledge of the future stream of Belgian day-ahead prices and can plan accordingly. That it failed to manage the remaining cost categories more adroitly is, at first glance, unexpected.

Figure 6 highlights a wider cost disparity between simulation and experiment for the MPC than for any of the other EMSs, with the simulated MPC also registering lower peak, offtake extras, and day-ahead costs. On select days, the divergence surpassed €1, a pattern visible in **Figure 8**. A more detailed inquiry was launched into three such days, covering houses 1 and 3, to uncover the drivers of these discrepancies. House 4 also had two days on which the simulation performed

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

worse than the experiment by over €1; however, a third day saw the simulation beat the experiment by roughly €2, partially balancing the two underperformance instances.

In each of the three deep-dive cases, peak cost was the primary source of the gap.

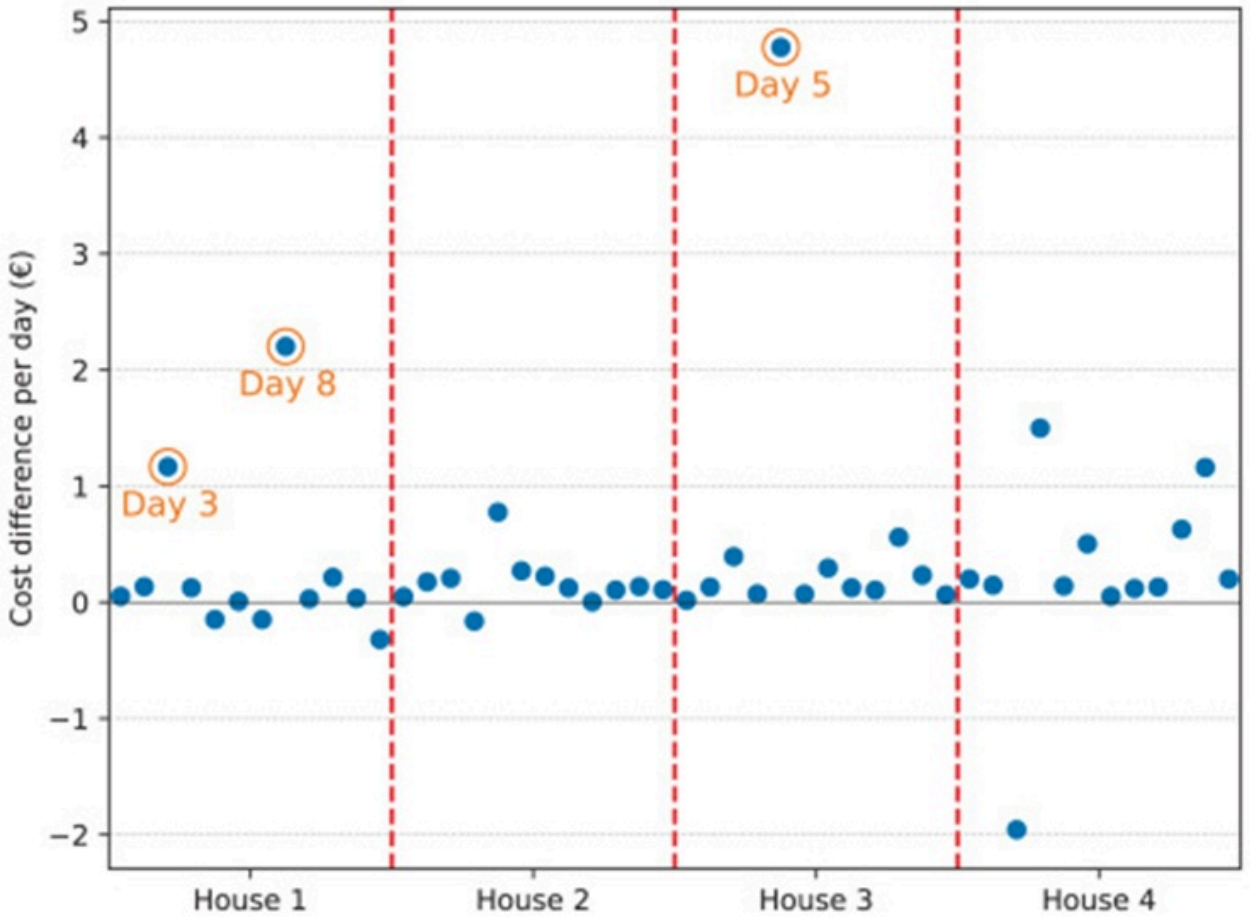

Figure 8. Difference in costs between the experiment and the simulation for the MPC EMS for each day.

- Day 5, house 3, a gap of €4.8: A brief charging session that began right at the day's first time step compelled the EV charger to draw maximum power. In the simulated environment, this peak was neutralised by concurrently discharging the BESS; the physical experiment did not mirror this behaviour. The experimental MPC set its BESS and EV charger commands without having received notification that an EV session would start at that time step, and hence selected an unsuitable BESS setpoint. The fault is a coding error triggered whenever an EV session commences at the very first step of a day.

- Day 8, house 1, a gap of €2.2: The daily peak rose higher in the physical run than in simulation because the MPC specified a larger EV charging power at midnight than its simulated equivalent. The MPC's forecasting module had incorrectly projected some PV generation at that hour, having ingested the previous day's PV data in a scrambled temporal order. The misordering originated from a device failure that interrupted grid measurement logging. The device stopped collecting data from the 29th of May at 10:45 until the 30th of May at 06:45; the day under investigation started on the 30th of May at 15:00. When the MPC forecaster queried the preceding 24 hours of data, the missing stretch, coupled with formatting functions that did not sort chronologically by default, caused the data stream to be delivered beginning from 06:45 up to the query moment, then continuing from 24 hours before the query up to 06:45. This

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

scrambled sequence led the forecaster to project PV output between 300 W and 400 W during nighttime—coinciding with the MPC's decision to charge the EV at a power level near the existing peak—thereby producing a higher peak in the experiment than in the simulation for the day in question.

- Day 3, house 1, a gap of €1.7: Toward the end of the investigated day, approaching the switching boundary, the simulated MPC managed to replenish the BESS to 100% SOC and complete an EV session without elevating the peak cost. The experimental counterpart, however, saw the peak rise due to poor coordination between the EV session and the BESS switching-time constraint. The root of the mismatch was a divergence between the experimental and simulated MPC code. In simulation, whenever the forecaster returned a departure time later than 15:00, the MPC would override it, resetting the departure time to 15:00. This safeguard was inadvertently absent from the experimental implementation.

All three faults were reproduced inside the simulation framework, causing the total simulated cost to shift from €157.8 to €165. This reduces the distance to the experimental total of €172.8 by half the original margin. A more carefully engineered MPC deployment in the experimental setting would therefore enhance its performance. On the flip side, the episode illustrates how heavily the peak cost structure can penalise an MPC whose implementation harbours defects. EMS formulations that depend on less intricate models and fewer external inputs—the RBC and TreeC controllers being cases in point—were immune to errors of this nature. It is also worth observing that pinpointing these faults would have been far more arduous absent a simulation capable of reproducing the MPC's behaviour.

Reinforcement learning

Looking at **Figure 6**, the RL-driven EMS posted the poorest result of the set, with the gap traceable above all to an oversized offtake extras burden (distribution plus taxes). That the RL agent found itself in an early learning regime despite having undergone pre-training was evident from its interaction with genuinely novel (real) experience tuples, so the outcome fits expectation. On top of that, each agent–house pair was allotted a mere 1.7 weeks (12 days) of experimental clock time, notwithstanding the three-month historical dataset used for offline pre-training. Parallel simulated case studies reinforce how much additional training time is typically needed: Ceusters *et al*. [5] documented that their RL EMS reached 95% of MPC-level performance only after roughly 18.6 weeks and did not fully equal the MPC until about 2 years had passed—also working at a 15-minute granularity.

Clues about the learning trajectory and convergence behaviour can be extracted from the actor and critic loss profiles together with the per-episode returns, displayed in **Figure 9**, respectively. The actor loss makes clear that behaviour cloning pre-training conferred a genuine head start, given the low loss values right at the outset. This indicates the agents were initialised with a policy already resembling the one embedded in the pre-training corpus (the RBC policy). The loss subsequently climbs, a signature of ongoing policy adjustment and exploratory deviation once online learning commences. Turning to the critic**,** one sees a very high initial loss that plummets as the critic network swiftly learns to appraise the pre-trained policies with fidelity. This steep drop once again confirms the utility of the behaviour cloning stage—the critic is refining its value estimates rather than constructing them from nothing. Over successive updates, the critic loss settles into a lower, flatter band, consistent with convergence and gradual sharpening of its value function approximations.

To dig further into agent performance, the rewards collected in the physical experiment were measured against those of a simulated oracle—the perfect-foresight MPC, labelled MPC P—run over the identical day/house pairings encountered by the real RL agents. The motivation for doing so is that raw episodic rewards, examined on their own, exhibit large swings driven simply by environmental variability, i.e., the inherent differences from one day to the next, above all whether an EV was connected. The differential plotted in **Figure 9** therefore supplies a cleaner signal for judging whether, and how quickly, the agents narrow the gap to the optimal benchmark.

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

Rates of progress differ noticeably from house to house. House 3 registers the strongest upward movement, with a trend slope of 4.34, though it also began from the least favourable policy, indicating genuine learning gains. Houses 1 and 4 likewise show positive slopes—3.03 and 2.48, respectively—pointing to effective policy refinement. House 2, in contrast, exhibits a shallow negative slope of −0.66, suggesting its performance has plateaued or mildly regressed. A particularly poor day 11 for House 2 exerts a downward pull on the fitted line. Across the board, the benefit of pre-training is underscored: the agents started from a markedly better position on day 2 than on day 1, and they managed to close a portion of the gap to the optimal policy relatively briskly—despite having taken only a very modest 1152 real environment steps.

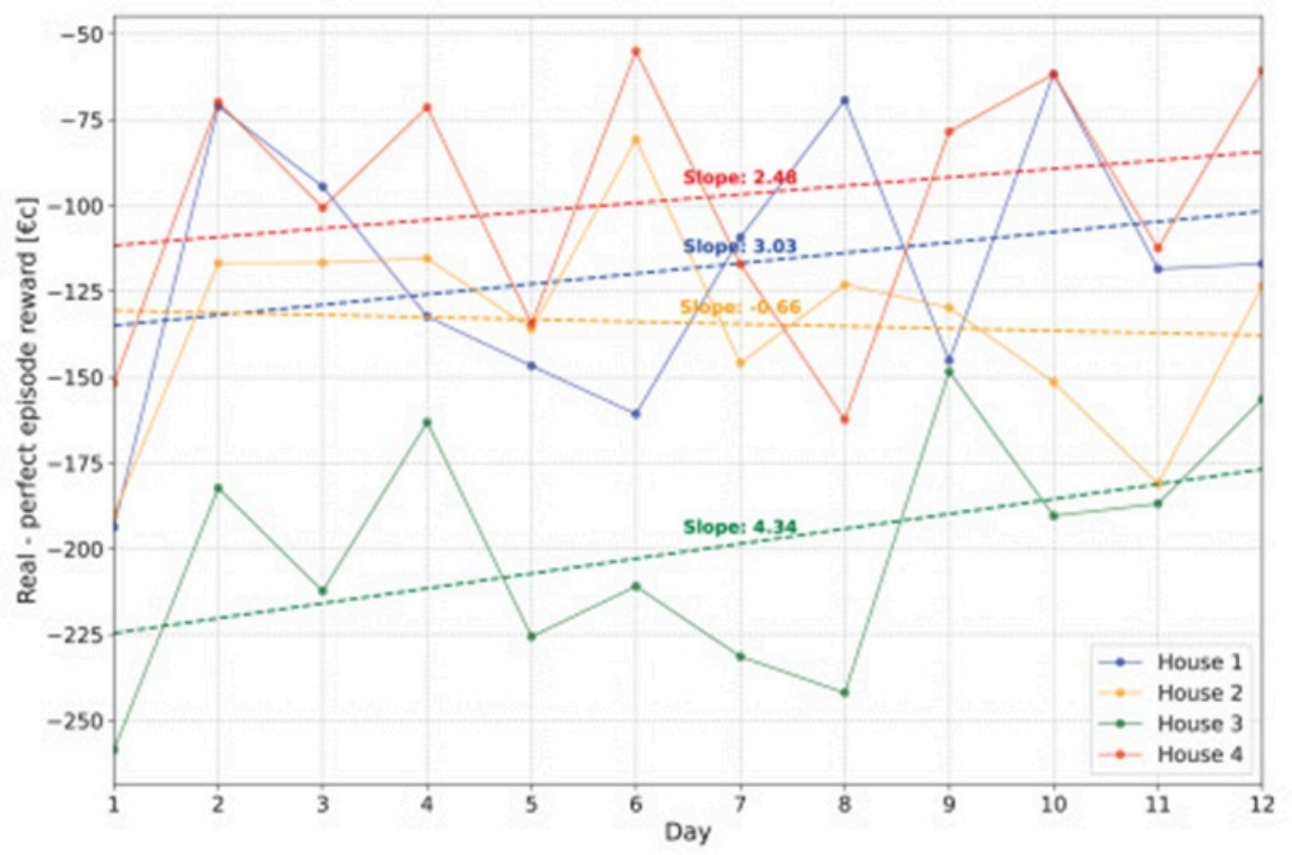

Figure 9. Difference in episodic rewards (in EUR cents) between real RL agents and the simulated MPC with perfect foresight (MPC P) that shows the learning rate and policy improvement of the RL agents. Note that these episodic rewards are without the peak and yearly costs and that even though only 12 episodes are observed per agent, the policy networks are trained every 4 steps **(Table 2).**

Extrapolating the observed trends forward, the RL agents would nonetheless demand a lengthy total online training period. Whether this transient underperformance—and its accompanying monetary training expense—is justified depends on what incremental costs, if any, are attached to the alternative methods (for instance, the effort of model construction). In the residential context studied here, the absolute money at stake is fairly small: across 12 days, the RL EMS incurred an average of 10.8 EUR more per house than the other EMSs, all of which delivered similar results. If one postulates linear improvement from this point and assumes, following [5], that the RL agent catches up to MPC after 2 years, the projected training cost per house works out to 333.9 EUR. Once the additional engineering effort demanded by MPC (and likewise the effort of arranging offline training for the tree-based EMS) is factored in, it does not seem unreasonable to suppose that, under these conditions, the total cost of ownership of an RL-based EMS could be lower—all the more so if the RL EMS eventually outperforms the competing methods post-training, as [5] reports (a benefit the simple arithmetic here omits). For sites with a much higher absolute energy bill—industrial or commercial installations—this training outlay

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

would scale up, eroding the RL EMS's attractiveness. The present paper does not attempt to furnish conclusive proof regarding the relative economic case for the different EMS approaches. One straightforward lever for compressing training time and thus cost would be to artificially raise the RL agents' sampling rate, i.e., to pack more experience tuples into the same wall-clock interval—a direction we put forward for future study.

Safety layer

A breakdown of the safety layer's behaviour **(Figure 10)** across the four EMS strategies brings out distinct signatures, both in how often the safety mechanism intervened and in what drove those interventions for each controller.

The RBC approach was responsible for the greatest number of safety layer activations. The root cause appears to be a vulnerability in its simplistic rule-based design. Because the BESS stayed locked in self-consumption mode, any moment of grid offtake would cause it to discharge aggressively—effectively attempting to nullify all grid exchange—especially during stretches of negligible solar output coinciding with high household demand. Once the BESS had been emptied and a sizeable residual load persisted, the EV charging setpoint had to be corrected. Approaching the day's end, the compulsory recharging of the BESS to 100% would kick in; with an EV load still active, the simultaneous charging of both the BESS and the EV could all too easily violate the grid ceiling in the absence of sufficient solar feed-in, prompting the safety layer to adjust the setpoints preemptively.

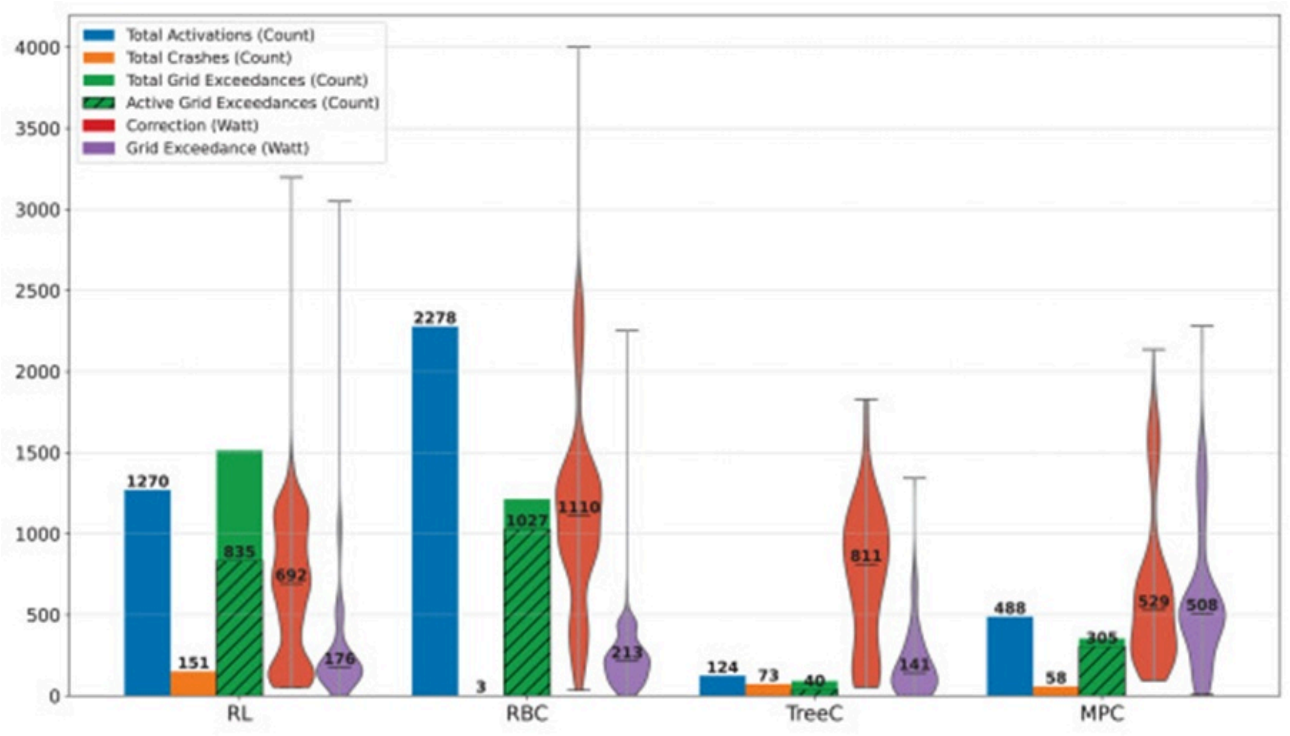

Figure 10. Safety layer results per EMS. Comparing both the number of activations and crashes, as well as how many times the grid limit still was exceeded, together with how much (in Watt) got corrected by the safety layer and how much (in Watt) the grid limit still was exceeded.

The RL strategy likewise provoked a large tally of safety layer interventions—an expected outcome, considering its exploratory and constraint-agnostic design. This occurred with particular frequency when the agent pursued aggressive simultaneous charging of the BESS and EV; that said, the agent's exploration was not confined to high-power regimes, so safety layer correction was not uniformly required. As online learning progressed, the RL policies also developed an early-day pattern of rapidly draining the BESS, which by day's end produced the same predicament described above, necessitating enforced concurrent charging of both storage and vehicle.

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

The MPC controller triggered safety layer activations at a rate that was not anticipated. Sifting through the raw data logs uncovered mis-specifications in the constraint models used by both the MPC and the safety layer, concentrated in the low-SOC operating region of the BESS. In actuality, the BESS power output tapered off more sharply than the models encoded, while both the MPC and safety layer continued to assume that further discharging was possible. For BESS units 1 and 2, a hard floor was encountered at 5% SOC, and BESS 1 additionally suffered a discrete 500 W power drop at the 10% SOC mark. Meanwhile, BESS 4 permitted uninterrupted discharging only down to 10%; if discharging stopped at, say, 8% SOC, the unit could not resume discharging until it had first been recharged back above the 10% threshold. None of these nuanced rules and behaviours appeared in any manufacturer documentation. In fact, a large share of the grid limit exceedances—including those for which the safety layer failed to intervene entirely (the unpatterned portion of the total exceedance count)—can be laid at the door of this very mismatch in low-SOC BESS dynamics. The active exceedance count (the patterned portion) designates time steps where the grid limit was breached (active power above 9.2 kW). At the same time, the safety layer was active, meaning the correction applied was insufficient rather than absent. The violin plot of grid exceedances, for its part, captures the magnitude of the overshoot (in Watts) beyond the 9.2 kW mark, irrespective of safety layer engagement.

What remains of the grid exceedance total is due, in descending order of contribution, to load calibration errors, a time synchronisation problem, and residual modelling inaccuracies in the safety layer. Calibration of the dynamic loads rested on a polynomial mapping known to be non-stationary over time (cold loads behaved differently from warmed-up ones, and the load banks are sited outdoors, exposed to the elements) and to vary between houses. Yet, a common function was applied universally, yielding an error band of roughly 150–200 W. Measurement streams flowed from heterogeneous hardware over different communication buses and at mismatched internal sampling cadences, causing temporal offsets that occasionally produced errors large in magnitude but fleeting in duration—for example, when cloud passage caused a sharp drop in PV infeed, or during ramping of the BESS, EV, and household loads. Lastly, the constraint equations themselves are imperfect, even outside the low- and high-SOC extremes of the BESS and EV; within normal operating bands, however, the error is far smaller (on the order of 10–20 W), except for BESS 3, whose pronounced 35–40 s settling time transiently yields larger observable deviations. The aggregate grid exceedance energy sums to 593.9 Wh for the RL EMS, 408.4 Wh for the RBC, 27.1 Wh for TreeC, and 295.5 Wh for the MPC.

Although TreeC stands apart with by far the fewest activations and the smallest grid exceedance tally, it was not beyond reproach: a rare, brief interval (roughly 30 minutes) of unusually intense uncontrollable household load (5 kW) still obliged the safety layer to step in.

As a final note, no specific trigger could be pinpointed for any of the safety layer solver crashes—measurement gaps were absent, and no clear infeasibility conditions were identified—leading us to attribute them largely to solver instability. It is worth keeping in mind that the safety layer was invoked slightly fewer than 829,440 times per EMS (every 5 seconds, 12 days, 4 houses, minus a handful of initialisation steps) over the course of the experiment. The crash incidence is thus extremely low (99.97% uptime). We underscore that these occasional failures pertain to the safety layer solver alone, not to the EMS solvers.

What these results bring into sharp relief is the double-edged nature of the safety layer: it is simultaneously susceptible to human modelling errors and absolutely dependent on accurate constraint functions. Still, because only a limited subset of equations populates the safety layer (exclusively those touching the constraints), its exposure to this weakness is narrower than that of, say, MPC—which demands a larger set of equations specified a priori, thereby multiplying the opportunities for error and imprecision. Ceusters *et al*. [11], for their part, addressed this very limitation by endowing the safety layer's constraint functions with adaptivity.

Difference experiment vs simulations

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

This section delves into divergences between the simulation and the physical experiment that go beyond the MPC-specific issues already dissected.

Where the experiment progresses in continuous real time, the simulation condenses system behaviour into 15-minute discrete jumps. A further discrepancy lies in how grid energy is registered: the physical meter logs both imported and exported energy within a given time step, whereas the simulation collapses this to a single net consumption figure per step—the simple arithmetic difference between offtake and injection. Computing the cost for a time step from separate import and export values invariably yields a higher number than using the net figure, because offtake extras charges are partially cancelled out when the energy difference is taken. **Table 5** lays out, for each EMS, the cost calculated from the experimental import/export values side by side with the cost obtained using net consumption, as a means of judging whether the simulation's simplifying convention materially distorts the comparison.

The wedge between import/export costing and net-consumption costing is wider for the RL and MPC EMSs than for the TreeC and RBC variants. A fundamental behavioural split separates the two groups: MPC and RL issue an explicit, fixed setpoint to the BESS, whereas TreeC and RBC lean heavily—or entirely—on self-consumption mode. Self-consumption gives the BESS the latitude to track real-time fluctuations, resulting in negligible simultaneous import and export so long as the battery is neither full nor entirely drained. Imposing a constant setpoint over a full 15-minute interval, by contrast, denies the BESS the ability to respond to intra-step variations and can readily produce both import and export within the same step.

Table 5. Costs of electricity obtained for each EMS in the experiment when taking the export and import energy values and when calculating net consumption.

EMS	Simulation	Experiment	Experiment
Empty cell	(Net consumption)	(Net consumption)	(Import/export)
RL	€208.6	€207.2	€216.1
RBC	€166.6	€171.2	€171.7
TreeC	€169.9	€170.7	€171.7
MPC	€157.8	€170.1	€172.8

A possible fix would be to have the BESS track a target grid power level instead of holding a static BESS power command across the whole interval. Additional rounds of simulation and physical testing would be required to verify whether this approach yields the hoped-for benefit.

When the MPC-specific divergences analysed are combined with the grid energy accounting mismatch covered here, a portion of the gap between simulated and experimental outcomes becomes explicable. Yet further mismatches were noted that resist straightforward translation into monetary terms. These include:

- Actual BESS storage capacities and round-trip efficiencies that stray from the nominal values printed in the datasheets.

- A hard discharge floor at 5% SOC exhibited by BESS units 1 and 2.

- A drifting minimum SOC threshold seen in BESS 4, an artificial 95% SOC ceiling imposed to keep the unit responsive, and a periodic abrupt jump from 85% to 95% SOC attributable to internal recalibration logic.

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

- The continuous-time safety layer operating in the physical experiment being crudely approximated at 15-minute granularity inside the simulation.

- The absence, within the simulation, of the hybrid inverter's power transfer cap.

Once the MPC faults detailed in before were recreated inside the simulation, the resulting total-cost mismatch between the two environments settled between 1% and 4.5% for every EMS. Recalculating the experiment's total from net consumption shrinks the spread to between 0.5% and 3%. What remains of the discrepancy can be laid at the door of the items listed above, supplemented by other, unnoticed divergences; taken as a whole, however, their collective influence on the bottom line is modest. No independent metering separates the electrical loads from whatever parasitic losses are present within the rig, obliging the simulation to work with a single fused data feed. The magnitude of those losses may well scale with BESS and EV utilisation intensity—a coupling the simulation does not reflect and that warrants dedicated attention in later studies. Even so, the parsimonious BESS model **(Eq. 11)** paired with the manufacturer-supplied capacity and efficiency parameters proved accurate enough to underpin simulations of genuine analytical value.

Conclusion

What this paper has delivered is the first practical, real-world corroboration of two machine-learning-driven EMS concepts that had previously been tested only in simulation: (1) an RL agent paired with a safety filter able to enforce arbitrary constraints and to activate a pre-computed safe fallback rule when available (the OptLayerPolicy framework [11]), and (2) a metaheuristic procedure that distils control logic into a transparent decision tree (the TreeC methodology [12]). The work also contributes what is, to our knowledge, the first fair-protocol comparison of MPC, safety-filtered RL, RBC, and an interpretable tree-based EMS applied to any case study—simulated or experimental. The evidence supports the following conclusions:

- Across the RBC, TreeC, and MPC controllers, economic performance was nearly indistinguishable, with the total cost separating the three by no more than 0.6%. This result strongly emphasises that representative training data for TreeC and a carefully engineered MPC implementation are prerequisites for unlocking superior real-world behaviour.

- The OptLayerPolicy safety filter renders online RL agent training physically safe, subject to the caveat that the constraint equations are faithful to reality—a condition that, like MPC, remains susceptible to specification errors.

- Even with pre-training in place, a model-free RL agent continues to need a protracted period of online learning (this holds even under hyperparameters tuned for faster adaptation) and therefore accumulates a non-negligible monetary training bill, which would have to be weighed against any incremental cost—or absence thereof—attached to alternative methods (modelling time, for instance).

- The TreeC method posted the most secure operational record in the experiments reported here (total grid overshoot of 27.1 Wh versus 593.9 Wh for RL). It does, however, demand an a priori simulation environment and a heavy offline training investment. While less error-prone, it still calls for domain insight to predetermine its practical value and its safety margins (since, absent a safety layer, it executes its policy without any runtime constraint checks).

- A real-time safety layer can deliver value across a spectrum of (optimal) control strategies and may simplify the broader (hierarchical) control architecture, a possibility underscored by the complete absence of PID controllers from the experimental configuration.

- With the MPC's experimental faults reproduced inside the simulation, the cost mismatch between experiment and simulation ranged from 1% to 4.5% across all EMSs. Part of this stemmed from the simulation's representation of grid energy as a single net-consumption figure per time step. In contrast, the experimental meter registers both import and

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

export within each step. Overall, the simulations tracked reality with sufficient accuracy in cost terms and proved vital in diagnosing the MPC implementation errors and the inadequacy of the TreeC training dataset.

We conclude by putting forward the following directions for subsequent research:

- Test whether deliberately increasing the RL agents' sampling frequency would accelerate the overall pace of learning by compressing a larger count of experience tuples into an unchanged wall-clock span (i.e., generating more time steps per unit of real time).

- Perform a robustness audit under corrupted or noisy sensor feeds and observations, and establish whether—and what kind of—state estimation machinery would be necessary.

- Compare alternative hierarchical control topologies along the dimensions of optimality, computational demand, modelling prerequisites, adaptivity, and constraint enforcement (including Lyapunov-based guarantees).

- Explore whether EMS formulations reliant on more elaborate models and richer input data are systematically more fragile when crossing from simulation to real deployment, following the pattern observed here with the MPC EMS.

- Re-run the EMS comparison incorporating the improvements suggested by the findings: supplying TreeC with more seasonally representative training data, eliminating the MPC bugs identified, and switching the battery control objective from a fixed power setpoint to active regulation of the grid power.

Nomenclature

BC = Behavioral Cloning

BESS = Battery energy storage system

CC = Constant Current

CV = Constant Voltage

EMS = Energy management system

EV = Electric vehicle

MPC = Model predictive control

PID = Proportional Integral Derivative

PV = Photovoltaic

RBC = Rule-based control

RL = Reinforcement learning

SOC = State of charge

Acknowledgements

None

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

Conflict of interest

None

Financial support

This work has been supported by the ECOFLEX project funded by FOD Economie, K.M.O., Middenstand en Energie, by the ICON project OPTIMESH (FLUX50 ICON Project Collaboration Agreement - HBC.2021.0395) funded by VLAIO and by the Baekeland project SLIMness (HBC.2019.2613) funded by ABB n.v. and VLAIO in equal parts.

Ethics statement

None

Received: 02 Sep 2022 Revised: 08 Nov 2022 Accepted: 25 Nov 2022
Published online: 15 November 2023

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at https://creativecommons.org/licenses/by-nc-sa/4.0/.

References

1. Luderer G, Madeddu S, Merfort L, Ueckerdt F, Pehl M, Pietzcker R, et al. Impact of declining renewable energy costs on electrification in low-emission scenarios. Nat Energy. 2022;7(1):32-42. https://doi.org/10.1038/s41560-021-00937-z.

2. Hampton H, Foley A, Del Rio DF, Smyth B, Laverty D, Caulfield B. Customer engagement strategies in retail electricity markets: A comprehensive and comparative review. Energy Res Soc Sci. 2022;90:102611. https://doi.org/10.1016/j.erss.2022.102611.

3. Matisoff DC, Beppler R, Chan G, Carley S. A review of barriers in implementing dynamic electricity pricing to achieve cost-causality. Environ Res Lett. 2020;15(9):093006. https://doi.org/10.1088/1748-9326/ab9a69.

4. Fabrizio E, Filippi M, Virgone J. Trade-off between environmental and economic objectives in the optimization of multi-energy systems. Build Simul. 2009;2(1):29-40. https://doi.org/10.1007/S12273-009-9202-4.

5. Ceusters G, Rodríguez RC, García AB, Franke R, Deconinck G, Helsen L, et al. Model-predictive control and reinforcement learning in multi-energy system case studies. Appl Energy. 2021;303:117634. https://doi.org/10.1016/j.apenergy.2021.117634.

6. Forbes MG, Patwardhan RS, Hamadah H, Gopaluni RB. Model predictive control in industry: Challenges and opportunities. IFAC-PapersOnLine. 2015;48(8):531-8.

https://doi.org/10.1016/j.ifacol.2015.09.022.

7. Drgoňa J, Arroyo J, Cupeiro Figueroa I, Blum D, Arendt K, Kim D, et al. All you need to know about model predictive control for buildings. Annu Rev Control. 2020;50:190-232. https://doi.org/10.1016/j.arcontrol.2020.09.001.

8. Görges D. Relations between model predictive control and reinforcement learning. IFAC-PapersOnLine. 2017;50(1):4920-8. https://doi.org/10.1016/j.ifacol.2017.08.747.

9. Sturzenegger D, Gyalistras D, Morari M, Smith RS. Model predictive climate control of a swiss office building: Implementation, results, and cost–benefit analysis. IEEE Trans Control Syst Technol. 2016;24(1):1-12. https://doi.org/10.1109/TCST.2015.2415411.

10. Schwenzer M, Ay M, Bergs T, Abel D. Review on model predictive control: an engineering perspective. Int J Adv Manuf Technol. 2021;117(5):1327-49. https://doi.org/10.1007/s00170-021-07682-3.

11. Ceusters G, Putratama MA, Franke R, Nowé A, Messagie M. An adaptive safety layer with hard constraints for safe reinforcement learning in multi-energy management systems. Sustain Energy Grids Netw. 2023;36:101202. https://doi.org/10.1016/j.segan.2023.101202.

12. Ruddick J, Ramirez Camargo L, Putratama MA, Messagie M, Coosemans T. TreeC: A method to generate interpretable energy management systems using a metaheuristic algorithm. Knowl-Based Syst. 2024;112756. https://doi.org/10.1016/j.knosys.2024.112756.

13. Beaudin M, Zareipour H. Home energy management systems: A review of modelling and complexity. Renew Sustain Energy Rev. 2015;45:318-35. https://doi.org/10.1016/j.rser.2015.01.046.

14. Rathor SK, Saxena D. Energy management system for smart grid: An overview and key issues. Int J Energy Res. 2020;44(6):4067-109. https://doi.org/10.1002/er.4883.

15. Zhou B, Li W, Chan KW, Cao Y, Kuang Y, Liu X, et al. Smart home energy management systems: Concept, configurations, and scheduling strategies. Renew Sustain Energy Rev. 2016;61:30-40. https://doi.org/10.1016/j.rser.2016.03.047.

16. Allwyn RG, Al-Hinai A, Margaret V. A comprehensive review on energy management strategy of microgrids. Energy Rep. 2023;9:5565-91. https://doi.org/10.1016/j.egyr.2023.04.360.

17. Zhao Y, Lu Y, Yan C, Wang S. MPC-based optimal scheduling of grid-connected low energy buildings with thermal energy storages. Energy Build. 2015;86:415-26. https://doi.org/10.1016/j.enbuild.2014.10.019.

18. Viot H, Sempey A, Mora L, Batsale JC, Malvestio J. Model predictive control of a thermally activated building system to improve energy management of an experimental building: Part I—Modeling and measurements. Energy Build. 2018;172:94-103. https://doi.org/10.1016/j.enbuild.2018.04.055.

19. Viot H, Sempey A, Mora L, Batsale JC, Malvestio J. Model predictive control of a thermally activated building system to improve energy management of an experimental building: Part II - Potential of predictive strategy. Energy Build. 2018;172:385-96. https://doi.org/10.1016/j.enbuild.2018.04.062.

20. Arroyo J, Spiessens F, Helsen L. Comparison of model complexities in optimal control tested in a real thermally activated building system. Buildings. 2022;12(5):539.

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

https://doi.org/10.3390/buildings12050539.

21. Elkazaz M, Sumner M, Thomas D. Energy management system for hybrid PV-wind-battery microgrid using convex programming, model predictive and rolling horizon predictive control with experimental validation. Int J Electr Power Energy Syst. 2020;115:105483.
https://doi.org/10.1016/j.ijepes.2019.105483.

22. Zhang K, Prakash A, Paul L, Blum D, Alstone P, Zoellick J, et al. Model predictive control for demand flexibility: Real-world operation of a commercial building with photovoltaic and battery systems. Adv Appl Energy. 2022;7:100099.
https://doi.org/10.1016/j.adapen.2022.100099.

23. Yang S, Wan MP, Chen W, Ng BF, Dubey S. Experiment study of machine-learning-based approximate model predictive control for energy-efficient building control. Appl Energy. 2021;288:116648.
https://doi.org/10.1016/j.apenergy.2021.116648.

24. Fu Q, Han Z, Chen J, Lu Y, Wu H, Wang Y. Applications of reinforcement learning for building energy efficiency control: A review. J Build Eng. 2022;50:104165.
https://doi.org/10.1016/j.jobe.2022.104165.

25. Nweye K, Liu B, Stone P, Nagy Z. Real-world challenges for multi-agent reinforcement learning in grid-interactive buildings. Energy AI. 2022;10:100202.
https://doi.org/10.1016/j.egyai.2022.100202.

26. Dey S, Marzullo T, Zhang X, Henze G. Reinforcement learning building control approach harnessing imitation learning. Energy AI. 2023;14:100255.
https://doi.org/10.1016/j.egyai.2023.100255.

27. Costanzo GT, Iacovella S, Ruelens F, Leurs T, Claessens BJ. Experimental analysis of data-driven control for a building heating system. Sustain Energy Grids Netw. 2016;6:81-90.
https://doi.org/10.1016/j.segan.2016.02.002.

28. Kazmi H, Mehmood F, Lodeweyckx S, Driesen J. Gigawatt-hour scale savings on a budget of zero: Deep reinforcement learning based optimal control of hot water systems. Energy. 2018;144:159-68.
https://doi.org/10.1016/j.energy.2017.12.019.

29. Zhang Z, Lam KP. Practical implementation and evaluation of deep reinforcement learning control for a radiant heating system. In: Proceedings of the 5th Conference on Systems for Built Environments. BuildSys '18. New York (NY): Association for Computing Machinery; 2018. p. 148-57.
https://doi.org/10.1145/3276774.3276775.

30. Touzani S, Prakash AK, Wang Z, Agarwal S, Pritoni M, Kiran M, et al. Controlling distributed energy resources via deep reinforcement learning for load flexibility and energy efficiency. Appl Energy. 2021;304:117733.
https://doi.org/10.1016/j.apenergy.2021.117733.

31. Naug A, Quinones-Grueiro M, Biswas G. Deep reinforcement learning control for non-stationary building energy management. Energy Build. 2022;277:112584.
https://doi.org/10.1016/j.enbuild.2022.112584.

32. Gokhale G, Tiben N, Verwee MS, Lahariya M, Claessens B, Develder C. Real-world implementation of reinforcement learning based energy coordination for a cluster of households. In: Proceedings of the 10th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation. BuildSys '23. New York (NY): Association for Computing Machinery; 2023. p. 347-351.
https://doi.org/10.1145/3600100.3625681.

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

33. Huo Y, Bouffard F, Joós G. Decision tree-based optimization for flexibility management for sustainable energy microgrids. Appl Energy. 2021;290:116772. https://doi.org/10.1016/j.apenergy.2021.116772.

34. Dai X, Cheng S, Chong A. Deciphering optimal mixed-mode ventilation in the tropics using reinforcement learning with explainable artificial intelligence. Energy Build. 2023;278:112629. https://doi.org/10.1016/j.enbuild.2022.112629.

35. Schlemminger M, Ohrdes T, Schneider E, Knoop M. Dataset on electrical single-family house and heat pump load profiles in Germany. Sci Data. 2022;9(1):56. https://doi.org/10.1038/s41597-022-01156-1.

36. Sørensen AL, Lindberg KB, Sartori I, Andresen I. Residential electric vehicle charging datasets from apartment buildings. Data Brief. 2021;36:107105. https://doi.org/10.1016/j.dib.2021.107105.

37. IEC 61851-1:2017. Electric vehicle conductive charging system - part 1: General requirements. IEC; 2017.

38. Vagropoulos SI, Bakirtzis AG. Optimal bidding strategy for electric vehicle aggregators in electricity markets. IEEE Trans Power Syst. 2013;28(4):4031-41. https://doi.org/10.1109/TPWRS.2013.2274673.

39. Farhad S, Nazari A. Introducing the energy efficiency map of lithium-ion batteries. Int J Energy Res. 2019;43(2):931-44. https://doi.org/10.1002/er.4332.

40. Scattolini R. Architectures for distributed and hierarchical Model Predictive Control – a review. J Process Control. 2009;19(5):723-31. https://doi.org/10.1016/j.jprocont.2009.02.003.

41. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. Lightgbm: A highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst. 2017;30.

42. Bergstra J, Bardenet R, Bengio Y, Kégl B. Algorithms for hyper-parameter optimization. Adv Neural Inf Process Syst. 2011;24.

43. Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans Inf Theory. 1967;13(1):21-7.

44. Genov E, De Cauwer C, Van Kriekinge G, Coosemans T, Messagie M. Forecasting flexibility of charging of electric vehicles: Tree and cluster-based methods. Appl Energy. 2024;353:121969. https://doi.org/10.1016/j.apenergy.2023.121969.

45. Fujimoto S, van Hoof H, Meger D. Addressing function approximation error in actor-critic methods. arXiv [Preprint]. 2018. arXiv:1802.09477. Available from: http://arxiv.org/abs/1802.09477.

46. Raffin A, Hill A, Gleave A, Kanervisto A, Ernestus M, Dormann N. Stable-Baselines3: Reliable reinforcement learning implementations. J Mach Learn Res. 2021;22(268):1-8.

47. Bain M, Sammut C. A framework for behavioural cloning. In: Machine Intelligence 15, Intelligent Agents. Oxford (UK): Oxford University; 1999. p. 103-29.

48. Ceusters G, Camargo LR, Franke R, Nowé A, Messagie M. Safe reinforcement learning for multi-energy management systems with known constraint functions. Energy AI. 2023;12:100227. https://doi.org/10.1016/j.egyai.2022.100227.

49. Kennedy J, Eberhart R. Particle swarm optimization. In: Proceedings of ICNN'95 - International Conference on Neural Networks. Vol. 4. 1995. p. 1942-8.

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

https://doi.org/10.1109/ICNN.1995.488968.

50. Biscani F, Izzo D. A parallel global multiobjective framework for optimization: pagmo. J Open Source Softw. 2020;5(53):2338. https://doi.org/10.21105/joss.02338.

51. Poli R, Kennedy J, Blackwell T. Particle swarm optimization. Swarm Intell. 2007;1(1):33-57. https://doi.org/10.1007/s11721-007-0002-0.

52. Piotrowski AP, Napiorkowski JJ, Piotrowska AE. Population size in Particle Swarm Optimization. Swarm Evol Comput. 2020;58:100718. https://doi.org/10.1016/j.swevo.2020.100718.

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press