

Contents lists available at ScienceDirect

Applied Energy



journal homepage: www.elsevier.com/locate/apenergy

Reinforcement learning for an enhanced energy flexibility controller incorporating predictive safety filter and adaptive policy updates

Siebe Paesschesoone ^{a,b,c,d,*}, Nezmin Kayedpour ^{a,b}, Carlo Manna ^{c,d}, Guillaume Crevecoeur ^{a,b}

^a Dynamics Design Lab D2LAB, Tech Lane Science Park 131, Ghent, 9052, Belgium

^b MIRO Core Lab, Flanders Make, Graaf Karel De Goedelaan 16-18, Kortrijk, 8500, Belgium

^c Flemish Institute for Technological Research VITO, Boeretang 200, Mol, 2400, Belgium

^d EnergyVille, Thor Park, Genk, 3600, Belgium

ARTICLE INFO

Dataset link: https://www.elia.be/en/grid-data

Keywords: Energy flexibility control Safe continual reinforcement learning Predictive safety filter Changepoint detection Policy updating

ABSTRACT

This paper presents a novel data-driven approach that leverages reinforcement learning to enhance the efficiency and safety of existing energy flexibility controllers, addressing challenges posed by the dynamic and uncertain nature of modern energy landscapes. With the increasing integration of renewable energy sources, conventional controllers struggle to maintain both safety and optimality. Our proposed approach introduces two significant contributions to standard RL approaches: a data-driven predictive safety filter and an online changepoint detection and policy updating module. Through continuous constraint satisfaction, the predictive safety filter guarantees absolute safety of the proposed controller. Meanwhile, the changepoint detection and policy updating module. Through controller. Meanwhile, the changepoint detection and policy updating module, inspired by the concept of continual learning, enhances the controller's adaptivity to non-stationary environments. By identifying changes in the environment, it triggers relearning of the agent, making the controller resilient to evolving conditions. Validation of our approach is conducted on a grid-connected PV-battery-load system, demonstrating its effectiveness in simultaneously improving safety and performance over traditional learning methods. More specifically, the proposed solution was able to increase the energy flexibility by reducing energy costs with 9.3%.

1. Introduction

Existing power grid controllers face efficiency challenges in light of a rapidly changing energy landscape. They are often fixed and struggle to adapt to the growing stochastic and unpredictable nature on both the load and supply side. Additionally, the intricate multi-physical nature of the energy system adds complexity to controller deployment and tuning. These challenges render state-of-the-art controllers not just suboptimal but potentially unsafe, presenting substantial technical and economic challenges for power system planning and operation [1,2]. Addressing these challenges requires the implementation of advanced control strategies that are able to increase the grid flexibility by outperforming current controllers in terms of safety and energy efficiency in the face of dynamic and complex environments.

One promising area of focus to increase the grid flexibility is model predictive control (MPC) [3]. A notable advantage of MPC lies in its inherent safety, as it systematically considers and adheres to system constraints throughout the control horizon [4]. However, MPC typically requires a dynamic model to calculate the optimal control actions, making it difficult to deploy swiftly for energy flexibility problems. Furthermore, it is recognized that MPC lacks adaptivity in the face of dynamic environments [5]. To overcome this challenge of adaptivity, recently, adaptive data-driven MPC techniques have been devised that adapt online while performing repetitive tasks [6]. Next to this, recent studies aim to enhance the effectiveness of standard MPC by exploring robust and stochastic MPC techniques for addressing energy flexibility problems. Robust MPC (RMPC) tackles environment uncertainties by incorporating them directly into the optimization problem, envisioning a worst-case scenario. In [7], the authors have designed a novel RMPC algorithm for energy scheduling within a multi-carrier energy system, i.e. an energy system based on gas and electricity. Similarly, in [8], a two-level control scheme based on RMPC to offer frequency reserves for an energy district was proposed. To overcome the challenge of physical modeling, a data-driven model was leveraged. While RMPC has proven to improve the performance of standard MPC approaches with respect to uncertainties and disturbances, this method often compromises performance due to the protection against improbable outliers. Additionally, describing process uncertainties for RMPC is often challenging [9]. Stochastic Model Predictive Control (SMPC) interprets

https://doi.org/10.1016/j.apenergy.2024.123507

Received 31 January 2024; Received in revised form 3 May 2024; Accepted 17 May 2024

0306-2619/© 2024 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

^{*} Corresponding author at: Dynamics Design Lab D2LAB, Tech Lane Science Park 131, Ghent, 9052, Belgium. *E-mail address:* siebe.paesschesoone@ugent.be (S. Paesschesoone).

constraints probabilistically through chance constraints, allowing for a small probability of constraint violation. In [10], a stochastic MPC is built for microgrid optimization in order to address uncertainties in energy forecasts, enabling enhanced flexibility services for system operators. [11] introduces a scenario-based stochastic MPC for residential microgrid control including solar and battery systems, aiming to minimize energy and peak power costs while maximizing flexibility revenue. Unfortunately, solving chance-constrained control problems like SMPC is generally difficult and often requires approximations [12].

As a counterpart to MPC, machine learning techniques like reinforcement learning (RL) are being proposed in the literature to solve energy flexibility problems [13]. In RL-based techniques, agents interact with the environment and gradually learn an optimal control policy based on their experience. This optimal control policy refers to the maximization of the expected cumulative reward [14]. RL can be model-free and function as a data-driven controller, being a main advantage compared to MPC as no prior knowledge of the system is needed. However, RL also faces various challenges, particularly in terms of safety and maintaining optimality in dynamic environments. Safety concerns arise from the explorative actions taken by RL agents [2]. These pose a direct threat to controller safety, a crucial concern given the inherently safety-critical nature of energy systems. Safe RL has emerged as a promising framework to address safety concerns in RL applications. In this framework, the goal is to maximize cumulative rewards while ensuring safety. To this end, the flexibility problem is often modeled as a constrained Markov Decision Process (CMDP), integrating safety constraints as costs in the objective function [15-18]. Although CMDP provides a formalism for safety, it does not guarantee absolute safety [19]. In response, recent studies have aimed at enforcing absolute safety by incorporating an additional layer to constrain the actions of RL agents [20]. For instance, in [18], a rule-based safety filter is introduced for an RL method designed to optimize the charging strategy for electric vehicles in a residential microgrid. Similarly, in [21], authors present a safe RL method for real-time automatic control in a smart energy hub. This method utilizes action clipping, restricting agent actions to maximum allowable values when an unsafe action is proposed. While rule-based methods ensure absolute safety through constraint satisfaction, they often come at the cost of suboptimal performance. Additionally, while rule-based methods can be straightforward and interpretable, defining rules that cover all possible scenarios may become challenging for complex multiphysical problems. In contrast, the authors of [22] propose the use of Control Barrier Functions (CBF) to ensure safety for a controller providing regulation services for district cooling systems. Although CBFs can handle constraint satisfaction in more complex and dynamic systems compared to rule-based systems, the computational complexity of this method leads the authors to approximate the problem using a neural network, impacting the accuracy of their approach. Addressing the computational intricacies associated with barrier functions, [23] suggests implementing a computationally efficient convex safety layer for RL to address real-time optimal power flow problems. However, a significant drawback of their approach lies in its reliance on physicsbased parameters, which are often challenging to measure or estimate. In [24,25], the authors employ supervised learning techniques to train a security layer for the RL controller in a multi-energy microgrid. A drawback of this method is the reliance on substantial training data and learning convergence; otherwise, the safety assurance could again be compromised.

Surprisingly, next to the concerns of safety, RL also faces challenges in maintaining optimality in dynamic environments [26,27]. RL policies often become static post-training, leading to suboptimal performance in non-stationary settings. The concept of Continual Learning (CL) emerges as a promising tool, emphasizing a learning system's ability to adapt and accumulate knowledge over time [28]. In RL, this involves ongoing adaptation and refinement of policies to ensure effectiveness in evolving situations, enabling agents to make optimal decisions over an extended period. Non-stationary changes may not be identifiable from past contexts, requiring online detection. Methods must be developed for flexibility controllers to identify and adapt to changes while online, ensuring optimal performance under continuously changing environments [29].

Changepoint (CP) detection is a statistical technique specifically designed to identify shifts or changes in sequential data [30]. In the context of RL, CP detection algorithms play a crucial role in detecting alterations in the statistical properties of the environment over time. These algorithms analyze sequential data, such as time-series data, to pinpoint points where statistically significant deviations occur compared to a baseline. These deviations could manifest as changes in the mean, variance, distribution, or other statistical properties of the data. For RL agents operating in non-stationary environments, CP provides a valuable mechanism for real-time monitoring and adaptation. By detecting changepoints, RL agents can recognize when the underlying dynamics of the environment have shifted [31]. Upon identifying a changepoint, RL agents can dynamically adapt their policies to accommodate the non-stationary environment. This adaptation may involve updating action-selection strategies, revising reward functions, or retraining the agent using the most recent data. By doing so, RL agents can maintain optimal performance even in the face of evolving environmental conditions.

In the context of energy flexibility problems, CL is essential as energy systems exhibit non-stationary behavior due to factors like consumption patterns, renewable energy output, and energy prices. RL agents employing CL techniques are better equipped to handle these challenges, ensuring adaptivity and optimality in dynamic energy environments. Despite its benefits, the specific application of CL in energy flexibility problems requires further exploration and research.

This paper presents a novel control strategy to address the various challenges of existing controllers when solving flexibility problems. To realize this, a data-driven approach is followed using RL. To alleviate safety issues, an MPC is added as a safety filter. Furthermore, the RL controller is augmented with a changepoint detection and policy updating mechanism to cope with the issue of sub-optimality in the face of non-stationary environments.

The fusion of MPC and RL capitalizes on their unique strengths. While MPC offers inherent safety and optimality, its reliance on precise models poses challenges in dynamic systems like energy systems. In contrast, RL operates model-free, learning optimal policies through interaction with the environment. However, safety concerns persist in RL [32]. By merging MPC and RL, we aim to create a controller that combines safety assurance with the ability to learn optimal policies adaptively. Specifically, MPC provides immediate control benefits by ensuring safety and efficiency from the outset based on existing models and constraints. Concurrently, RL works in the background, learning from the system's performance to enhance future responses and strategies. This dual approach offers both short-term reliability and long-term improvements in control tactics. With respect to this, the main objective of this article is to realize an adaptive and safe control approach that is able to increase energy flexibility within dynamic energy systems. First, the design of the proposed method is established in Section 2. Thereafter, in Section 3, the method's ability to improve the performance and safety of state-of-the-art RL flexibility controllers is validated. To this end, implementation details are presented and results are obtained by applying the proposed solution to a grid-connected PV-battery-load system, being an exemplary use case for energy flexibility. In short, this paper presents the following contributions:

1. Introduction of a novel safe continual reinforcement learning architecture, allowing the use of RL algorithms for safety-critical critical applications and enabling online adaptivity of controllers without prior knowledge of the system dynamics. To realize this, standard RL is augmented with a data-driven MPC safety filter to guarantee absolute safety. Additionally, a changepoint detector and a dynamic policy updating mechanism is built into the design to cope with non-stationary environments.

 Validation of the proposed method on a grid-connected PVbattery-load system, demonstrating its effectiveness in increasing both the safety and performance of standard RL approaches while addressing various real-world flexibility applications.

2. Proposed methodology

2.1. Reinforcement learning

In RL, an agent learns an optimal policy through interactions with the environment [14]. Markov Decision Processes (MDPs) serve as a common mathematical model to describe RL problems. An MDP is defined by the four-tuple (S, A, R, T), where S represents the state space, A the action space, R the reward space, and T the transition function. In the RL framework, an agent observes a state s_t of the environment, takes an action a_t at time t, receives a reward r_t , and updates its policy $\pi(a_t|s_t)$. This is depicted graphically in Fig. 1. The process iterates until policy convergence. The objective is to find an optimal policy π^* maximizing the discounted cumulative reward:

$$L(\theta) = \mathbb{E}_t \left[\sum_{t=0}^{\infty} \gamma^t r_t \right], \tag{1}$$

where θ represents policy parameters, r_t is the reward at time step t, and γ is the discount factor. This study focuses on Proximal Policy Optimization (PPO), a state-of-the-art RL algorithm classified as an actor-critic method, which combines the strengths of both policy gradient and value function estimation approaches. PPO excels in mitigating issues associated with destructive policy updates by introducing a clipping mechanism. Its superior performance in solving multi-dimensional continuous environments makes it particularly relevant for power grid management problems [33]. In the actor-critic paradigm, the actor is responsible for selecting actions based on the current policy, while the critic evaluates the chosen actions by estimating the value or advantage function. PPO employs a policy gradient approach, calculating the gradient of the expected reward with respect to policy parameters θ . Simultaneously, it incorporates a critic component, represented by the advantage function A_t . The gradient of the expected reward with respect to policy parameters θ is mathematically formulated as follows:

$$\nabla L(\theta) = \mathbb{E}_t \left[\sum_{t=0}^{\infty} \nabla_{\infty} \log(\pi_{\theta}(a_t | s_t)) A_t \right],$$
(2)

where $\pi_{\infty}(a_t|s_t)$ is the probability of taking action a_t in state s_t under policy π_{θ} , and A_t is the advantage function. PPO's clipped objective function is expressed as:

$$L_{\text{clipped}}(\theta) = \mathbb{E}_{t} \left[\min\left(\frac{\pi_{\theta}(a_{t}|s_{t})}{\pi_{\theta_{\text{old}}}(a_{t}|s_{t})}A_{t}, \text{clip}(\frac{\pi_{\theta}(a_{t}|s_{t})}{\pi_{\theta_{\text{old}}}(a_{t}|s_{t})}, \frac{1 - \epsilon, 1 + \epsilon)A_{t}}{1 - \epsilon, 1 + \epsilon)A_{t}} \right].$$
(3)

where ϵ is the arbitrarily determined clipping parameter that avoids large policy updates. PPO updates policy parameters θ to maximize this objective. For further details concerning the PPO algorithm, readers are referred to [34].

2.2. Safe continual reinforcement learning (SCRL)

Building on the principles of RL, this study introduces a novel paradigm called Safe Continual Reinforcement Learning (SCRL). It is designed to elevate the efficacy and safety of existing RL methodologies. The proposed SCRL structure, depicted in Fig. 2, encompasses two pivotal enhancements differentiating it from conventional RL approaches. First, by integrating a data-driven MPC safety filter, this method boosts the safety of conventional RL methods leveraging the advantages of RL and MPC [9]. Second, drawing inspiration from the principles of CL, the proposed design incorporates a changepoint detector coupled with a policy updating mechanism to address the



Fig. 1. Illustration of the working principle of reinforcement learning. An agent observes a state s_i of the environment, takes an action a_i at time t, receives a reward r_i , and updates its policy $\pi(a_i|s_i)$ until policy convergence.

intricacies of dynamic environments. The following paragraphs delve into a detailed discussion of each of these contributions.

(1) Data-driven MPC as safety filter for RL: Based on the work in [9], the introduction of a data-driven MPC safety filter for the RL controller is proposed. If a suggested control action of the RL agent is deemed unsafe, the MPC safety filter intervenes, proposing a secure alternative. This is illustrated graphically in Fig. 3. Consequently, absolute safety is guaranteed through continuous constraint satisfaction.

The main state-of-the-art approaches to render RL safe have been discussed in the introduction. While conventional methods such as CBFs or rule-based systems are effective in enforcing safety, the proposed MPC safety filter offers distinct advantages. First, the MPC controller stands out by dynamically selecting the most optimal safe action, driven by a model continuously updated using RL data. This adaptability ensures that our system's response remains finely tuned to real-time conditions, optimizing safety while maintaining efficiency. The MPC safety filter thus not only ensures real-time corrective actions based on predictive outcomes but also optimizes these adjustments to maintain the intent of the RL agent's decisions. This integrated, adaptive approach enhances safety without compromising the learning and performance efficacy of the RL system, making it particularly suitable for environments with rapid or uncertain changes. Second, the MPC filter provides a more general solution compared to constructing rule-based scenarios or functions. Its ability to effortlessly navigate complex multiphysical problems underscores its broader applicability across diverse domains. Specifically, MPC excels in handling complex multi-physical problems because it integrates models of various physical processes and iteratively optimizes control actions over time, allowing it to account for dynamic interactions and adapt to changing conditions effectively. This iterative approach enables MPC to refine predictions and control strategies, addressing the intricacies of interconnected systems with precision and flexibility.

To overcome the challenges of constructing a physical model, a data-driven model is established. To realize this, a multivariate linear regression model is fitted onto the data gathered from the RL process's state and action pairs. This data-driven model therefore captures the system's real-time behavior, remaining up-to-date and offering a more adaptive approach compared to a fixed model. Note that the adoption of a data-driven model, as opposed to a physical model, does not compromise the safety of the controller, given that safety is assured through the constraints alongside the dynamic model. The mathematical formulation of the safety filter closely mirrors a standard MPC formulation. The primary deviation from this formulation lies in the objective function. Specifically, the objective function of the safety filter aims to minimize the absolute difference between the suggested control input and the safe control input, as depicted in Eq. (4). This ensures that the deviation from the suggested RL actions is minimized, aligning the proposed control actions with those deemed safe.

$$\min_{n \in \mathbb{N}} \|u_{\mathrm{RL}} - u_{\mathrm{safe}, 0}\| \tag{4}$$

where u_{RL} and $u_{\text{safe, 0}}$ represent the proposed and safe control actions respectively.



Fig. 2. Overview of the proposed Safe Continual Reinforcement Learning (SCRL) architecture. In the physical domain, the energy system is controlled by an RL controller using PPO. The data-driven MPC reacts to unsafe RL agent actions, adding a safety layer to the controller. In the digital domain, RL agent interactions are stored and used to build a regression model for the MPC. The monitoring and detection unit triggers RL agent retraining and policy updates in response to environmental changes. Simultaneously, the data-driven model of the safety filter is dynamically updated.



Fig. 3. MPC safety filter for RL, based on [9]. If the RL agent proposes an unsafe action, the safety filter intervenes.

(2) Changepoint detection and policy updating: To leverage the concept of CL, a changepoint detector is built into the proposed framework. This detector monitors the environment and can detect changes in the location and scale parameters of the distributed data stream. The algorithm employed in this paper is derived from Ross et al.'s work [35]. In their study, the authors present an innovative technique for detecting changes in the location and scale parameter of continuously distributed univariate data streams. A distinctive aspect of their approach is the introduction of what they term stream discretization. This feature enables rapid computation of test statistics, rendering the technique well-suited for high-frequency data streams. This capability is particularly interesting for energy-like systems, where high-frequency data streams are prevalent. Ross' framework proposes the use of Lepage's statistical test, a non-parametric test designed to detect changes in both location and scale parameters in two samples [36]. It combines the Wilcoxon test statistic for location and the Ansari-Bradley test statistic for scale into a single test statistic L [37]. The Wilcoxon rank-sum statistic is defined by:

$$T_W = \sum_{i}^{N} i \cdot V_i \tag{5}$$

where *i* represents the ranks and V_i is set to 1 if the *i*th smallest observation is from the first sample and 0 otherwise. The total number of observations is denoted as N = n + m, where *n* and *m* represent the sizes of the two samples being compared. The Ansari–Bradley test is a rank test for dispersion, the test statistic is defined by:

$$T_{AB} = \frac{1}{2}n(N+1) - \sum_{i}^{N} \left| i - \frac{1}{2}(N+1) \right| \cdot V_{i}$$
(6)

With $V_i = 1$ if the *i*th smallest observation is from the first sample, and 0 otherwise. Lepage's test statistic combines both statistics as follows:

$$L = \left(\frac{T_W - \mu_W}{\sigma_W}\right)^2 + \left(\frac{T_{AB} - \mu_{AB}}{\sigma_{AB}}\right)^2 \tag{7}$$

where μ and σ are the expected values and variances, respectively, under the null hypothesis. The statistic *L* is asymptotically χ^2 -distributed with two degrees of freedom under H_0 . To decide whether a change

has occurred, one compares the calculated L statistic to a critical value from the distribution of L under the null hypothesis of no change. If L exceeds this critical value, it suggests evidence of change in either location or scale or both. The first main advantage of this method over current state-of-the-art detectors for RL algorithms is that it can detect both changes in scale and location of the distribution, outperforming existing detectors that often only detect for location shifts. The second main advantage of this method is the use of a non-parametric test, meaning that no distribution of the monitored datastream needs to be assumed.

Upon detecting change, the RL policy needs updating. This is crucial to align the agent's actions with the new dynamics. Failure to adapt may lead to suboptimal decisions, rendering the current strategy ineffective or unsafe. Various ways to update the RL policy exist, as detailed in [27]. In this article, the policy is retrained with the updated data. This method eliminates the need to relearn the entire policy from scratch, which not only speeds up the adaptation process but also keeps computational resources to a minimum.

The frequency and duration of retraining the RL model stand as crucial elements within the proposed framework. Firstly, excessive retraining duration may result in the framework persisting with the initial RL policy, yielding suboptimal performance. Secondly, overly frequent retraining may interrupt the training process before the new policy reaches completion, necessitating further retraining cycles. These concerns will be thoroughly examined in Section 3.2.

3. Grid-connected PV-battery-load system

To validate the proposed SCRL algorithm, it is applied to a gridconnected PV-battery-load system, an exemplary use case for energy flexibility. The use case is represented in Fig. 4 and involves the control of a Battery Energy Storage System (BESS) within a grid-connected PVbattery-load configuration. This setup encompasses (a) a PV production profile, (b) an electricity consumption profile or demand, (c) a connection to the grid allowing for energy import or export in case of excess or shortage, and (d) a battery, being the system under control. The key characteristics of the BESS are summarized in Table 1. The response time of the BESS is considered less than 15 min, being a reasonable assumption according to the literature [38]. The objective is to minimize total energy costs while satisfying the demand. The main challenge lies in the unpredictability of solar production, electricity consumption, and fluctuating electricity prices. The dataset used in this study represents actual data for Flanders in 2021-2022 and was sourced from Elia, the Belgian Transmission System Operator (TSO) [39]. Fig. 5 provides a snapshot of the data, illustrating PV generation, electricity demand, and electricity prices, highlighting the multi-source stochastic nature of the problem. The data is scaled with a constant factor. Descriptive statistics for the dataset can be found in Table 2. The granularity of the data points is 15 min.



Fig. 4. Schematic overview of the grid-connected PV-battery-load system use case.



Fig. 5. PV, demand, and electricity price data retrieved from Elia, the Belgian TSO [39]. The data was rescaled with a scaling factor.

Table 2

Descriptive	statistics	of	the	dataset	used	for	simulations.	
-------------	------------	----	-----	---------	------	-----	--------------	--

Statistic	Load [kW]	PV [kW]	Price [€/kWh]
Count	8400	8400	8400
Mean	2277.13	1520.21	32.46
Standard deviation	1681.79	2318.49	15.37
Minimum	300.00	0.00	0.00
25th percentile	1070.00	0.00	21.94
Median	1740.00	47.36	31.61
75th percentile	2970.00	2454.66	41.28
Maximum	16910.00	9134.72	200.04

The optimal solution of this problem can be obtained by formulating the problem as a mixed integer linear programming problem (MILP). However, in reality, achieving the optimal solution is impossible as it presupposes not only a precise understanding of the system dynamics but also a perfect knowledge of future data, such as PV generation, demand, and solar generation.

min.
$$\sum_{t} p_{t}^{\text{import}} \cdot P_{t}^{\text{import}} - p_{t}^{\text{export}} \cdot P_{t}^{\text{export}}$$
(8)

st.
$$E_{t+1} = E_t + P_t^{\text{charge}} \cdot \Delta t$$
 (9)

$$P_t^{\text{import}} - P_t^{\text{export}} = P_t^{\text{demand}} - P_t^{\text{solar}} + P_t^{\text{charge}}$$
(10)

$$M \cdot S_t^{\text{import}} \ge P_t^{\text{import}} \tag{11}$$

$$M \cdot S_t^{\text{export}} \ge P_t^{\text{export}}$$
 (12)

$$S_t^{\text{import}} + S_t^{\text{export}} \ge 0 \tag{13}$$

$$S_t^{\text{import}} + S_t^{\text{export}} \le 1 \tag{14}$$

$$E^{\text{lower limit}} \le E_t \le E^{\text{upper limit}} \tag{15}$$

$$-P^{\text{charging capacity}} \le P_t^{\text{charge}} \le P^{\text{charging capacity}}$$
(16)

$$E_0 = 0 \tag{17}$$

$$P_t^{\text{import}} \ge 0, \quad P_t^{\text{export}} \ge 0$$
 (18)

$$S_t^{\text{export}}, S_t^{\text{import}} \in \{0, 1\}$$
(19)

where $p_t^{\text{import}}, p_t^{\text{export}}$ are the import and export price at time t, P_t^{import} and $P_{t}^{P_{t}^{T}}$ the import and export power at time t, E_{t} and E_{t+1} the battery state of charge at time t and t + 1, Δt the time step, P_t^{charge} the charging power, P_t^{demand} the demand, P_t^{solar} the generated solar power at time t. M is a large constant value, facilitating big M constraints (11) and (12). S_t^{export} and S_t^{import} are binary variables, that are equal to one if P_t^{export} and P_t^{import} are bigger than 0 respectively. $E^{\text{lower limit}}$ and $E^{upper limit}$ the upper and lower limits of the battery state of charge. P^{charging capacity} represents the maximum (dis)charging power of the battery. Constraint (8) represents the objective function, i.e. the minimization of total energy cost. The variable to be controlled is the (dis)charging power of the battery P_t^{charge} . Constraint (9) represents the battery dynamics. Constraint (10) reflects the energy balance of the system. Constraints (11)-(14) ensure that at most one variable of P_t^{import} or P_t^{export} is equal to 1, as importing and exporting electricity cannot occur simultaneously. Constraint (15) confines the state of charge of the battery within the limits of $E^{\text{lower limit}}$ and $E^{\text{upper limit}}$. The charging power is further restricted by constraint (16). Constraint (17) is the initial energy constraint of the battery, set to zero. Finally, P_t^{import} and P_t^{export} must be greater than or equal to zero. S_t^{export} and S_{t}^{import} are binary variables.

3.1. SCRL implementation details

In this section, implementation details related to the proposed framework are expounded.

(1) *MDP formulation:* The BESS is taken as an agent. Its optimal decision-taking in the uncertainty of PV generation, electricity demand and price can be formulated as an MDP. Consequently, the system state at a certain point in time *t* includes the following variables:

$$s_{t} = \left(t, E_{t}^{\text{batt}}, P_{t}^{\text{demand}}, P_{t}^{\text{solar}}, p_{t}^{\text{import}}, p_{t}^{\text{export}}, P_{t+1}^{\text{demand}}, P_{t+1}^{\text{solar}}, p_{t+1}^{\text{export}}, p_{t+1}^{\text{export}}\right)$$
(20)

where E_t^{batt} is the battery state of charge at time *t*, P_t^{demand} the demand at time step *t*, P_t^{solar} the generated solar power at time step *t*. p_t^{import} and p_t^{export} represent the import and export price at time step *t*. Analogously, P_{t+1}^{demand} , P_{t+1}^{solar} , p_{t+1}^{import} , p_{t+1}^{export} are the predicted demand, PV generation and prices at time step *t* + 1.

The action an agent can take corresponds to the (dis)charging power.

$$a_t = \left(P_t^{\text{charge}}\right) \tag{21}$$



Fig. 6. Performance comparison of PPO, QR-DQN, TRPO, and DDPG algorithms. Ultimately, PPO was chosen based on its rapid and stable convergence, together with its computational efficiency. A 95% confidence interval is added around the curve (100 runs).

Table 3 PPO hyperparameter tuning.	
Hyperparameter	Value
Learning rate	0.0005
Discount factor (γ)	0.95
Entropy coefficient (β)	0.02
PPO clip parameter (ϵ)	0.2
Mini-batch size	64
Episode length	24
Training episodes	20 000

where P_t^{charge} is bounded by the maximum (dis)charging power (cf. Eq. (16)). The reward $r_t \in \mathcal{R}$ consists of two parts: the energy cost and a penalty for taking an unsafe action. In this case, an unsafe action is an action that drives the battery state of charge out of its predefined limits (cf. Eq. (15)).

$$r_{t} = \underbrace{-c_{1} \cdot p_{t}^{\text{import}} \cdot P_{t}^{\text{import}} + c_{2} \cdot p_{t}^{\text{export}} \cdot P_{t}^{\text{export}}}_{\text{Energy cost}} - \underbrace{-c_{3} \cdot unsafe_action}_{\text{Safety penalty}}$$
(22)

where c_1 , c_2 , c_3 are weighting parameters, and unsafe_action is a boolean variable equal to 1 if the proposed action by the RL controller is unsafe, and 0 otherwise. The tuning of parameters c_1 , c_2 , and c_3 requires a nuanced balance between the safety and performance considerations of the controller. Achieving this balance, normalization was implemented across the reward components. As detailed in Section 2, this paper focuses on Proximal Policy Optimization (PPO) [34]. However, for comprehensive analysis, a comparison was drawn with three other state-of-the-art RL algorithms-Quantile Regression Deep Q-Network (DQN) [40], Trust Region Policy Optimization (TRPO) [41], and Deep Deterministic Policy Gradient (DDPG) [42]. The learning curves presented in Fig. 6 provide a comparison of their performance. TRPO and PPO outperformed QR-DQN and DDPG, demonstrating quicker and more stable convergence. Considering computational efficiency, PPO was selected over TRPO for this study [34]. For this particular use case, the PPO parameters are tuned as indicated in Table 3.

(2) Safety filter: The data-driven model used for the MPC is devised as a multivariate regression model, which predicts the battery state of charge in the next time step, based on the current system state s_t and the action a_t that is taken. The data for fitting this regression model is gathered online, by keeping track of the agent's interactions with the environment. Let S_t represent the current system state, P_t^{charge} denote the (disc)charging action, and E_{t+1} represent the predicted battery state of charge in the next time step. The multivariate linear regression model is then formulated as:

$$E_{t+1} = S_t \cdot W + P_t^{\text{charge}} \cdot V + b \tag{23}$$

where W and V denote the fitted weight matrices associated with the state and action variables S_t and P_t^{charge} , respectively. The term b represents the fitted intercept. The coefficients of the weight matrices and intercept term are initially fitted in the training phase. Subsequently, during online operation, the model undergoes updates triggered by the online changepoint detection mechanism, ensuring its continual adaptation to changes in the environment. This updating mechanism enables the model to stay up-to-date and enhances the overall adaptivity of the proposed control strategy. For instance, consider the scenario where the battery experiences degradation over time, leading to an impact on the efficiency of both charging and discharging processes. This would have an impact on the battery dynamics model. The proposed strategy allows to accounts for this degradation, by considering the actual updated and measured dynamics of the controlled system.

(3) *CP* detection and policy updating: The online changepoint detection mechanism is implemented by utilizing the *ocpdet* Python library, which incorporates the two-sample test proposed by Lepage [36,43]. The analysis focuses on the live datastream represented by $[(P_t^{demand} - P_t^{solar}) * p_t]$. The detector is illustrated onto 350 days of data in Fig. 8. The black dotted lines show points in time where the detector identified a change in either scale or location of the data stream. When a change is detected, the RL agent is retrained on the changed data stream, encompassing ten days, for a total of 7500 training episodes. Once the training process is completed, the updated policy is pushed online. The confidence level α for Lepage's test is set at 0.05.

3.2. Results and discussion

(1) Performance aspects: The first achievement is the RL agent's ability to efficiently control the battery in order to minimize the total energy costs. This means the battery is being charged during low-price periods and discharged during high-price periods. The learned policy closely aligns with the optimal policy, obtained by solving the MILP problem. This is illustrated in Fig. 7. This is a major achievement given that the learning agent bases its decisions solely on the current system state and predictions about the next state. In contrast, the optimal solution assumes perfect knowledge of system dynamics and future data.

The numerical results of the PPO, TRPO, QR-DQN, and DDPG algorithms are presented in Table 4. The results are represented relatively to the optimal solution, which is obtained by solving the MILP. These results confirm indeed that PPO is the preferred RL algorithm for the proposed controller, as was already clear from the analysis in Fig. 6. Comparing these numerical results, also noted as the optimality gap, to literature findings proves challenging and often unfair as outcomes are heavily influenced by the specific system characteristics, such as input features like solar generation and the battery system itself. However, a fairer comparison can be made by evaluating the algorithm against a standard RL approach. In Fig. 9, the performance of the final proposed controller, which integrates PPO, is juxtaposed with that of a standard RL approach. The comparison reveals that the proposed algorithm surpasses the standard RL method, highlighting the efficacy of incorporating dynamic policy updating mechanisms alongside an MPC safety filter. More specifically, the employed policy and updating mechanism contribute to performance enhancement, as reflected by the reduced rewards, indicating lower overall energy costs. The adaptive approach demonstrates an average reduction of 9.3% in energy costs. Analogous to Fig. 8, the dotted lines in the graph indicate instances where the changepoint detector identified changes in the environment, leading to subsequent agent retraining and a corresponding policy update. This outcome underscores the performance benefits of integrating a



(c) Zoomed in comparison between the optimal and learned battery control schedule.

Fig. 7. Comparisons between the optimal and learned policy in testing phase.



Fig. 8. Illustration of the changepoint detection method. P_d represents the demand, P_s the solar output and p_t the electricity price. The dotted lines represent points in time where the detection method identified a change in the data.

changepoint detector within energy flexibility controllers in the face of dynamic environments.

In this simulation, retraining occurs 7 times over a 200-day long test dataset, as can be seen from the vertical dotted lines in both Figs. 9 and 11. Retraining intervals are determined by the variance of the input data, $(P_{demand} - P_{solar}) * p_t$, the statistical test performed, i.e. the Lepage's test, and the used significance level 0.05. The retraining process is executed in a brief timeframe, typically less than 5 min,

Table 4						
Average	energy	cost	comparison	of	the	different
algorithm	is compa	ared to	o the optimal	sol	ution	

Algorithm	Relative energy cost
Optimal solution	1
PPO [34]	1.79
TRPO [41]	1.82
QR-DQN [40]	1.91
DDPG [42]	2.11
Random agent	10.71

Table 5

0

Retraining time (minutes:seconds)							
4:15	3:47	4:01	4:34	2:27	4:36	3:08	

as illustrated in Table 5. This swift retraining mitigates any potential impact on control results and ensures the model remains responsive to evolving dynamics.

(2) Safety aspects: During training, the agent learns to avoid unsafe actions over time, as illustrated in Fig. 10. In the early training stages, the agent explores and frequently suggests unsafe actions, prompting the MPC safety filter to intervene, thereby enforcing absolute safety. As a result of penalization in the reward function, the agent progressively learns to refrain from proposing these unsafe actions.



Fig. 9. Performance comparison between the proposed adaptive RL strategy and a standard non-adaptive RL strategy. Rewards represent negative energy costs, indicating superior performance of the adaptive strategy. Black dotted lines show instances where the detector identified an environment change and the algorithm was retrained. A 95% confidence interval is shown around the curve (100 runs).



Fig. 10. Illustration of MPC safety filter interventions during the training process. The agent effectively learns to avoid unsafe actions due to penalization in the reward function, leading to less interruptions by the safety filter over time. A 95% confidence interval is added around the curve (100 runs).

Next, the frequency of proposed unsafe actions during operation is compared between the adaptive framework and a conventional, nonadaptive RL framework, as depicted in Fig. 11. The figure suggests an improvement in online safety within the proposed strategy, indicated by a reduced occurrence of suggested unsafe actions during operation. This enhancement stems from the dynamic nature of the RL policy, which is updated following changes, providing the RL agent with a better understanding of actions that may lead to unsafe situations in specific environments. Enhancing online safety results in reduced interventions by the MPC filter during its operation. This is advantageous because it minimizes the real-time usage of the safety filter, thereby reducing the computational effort required.

4. Conclusions and future work

In this paper, SCRL was proposed, a method to improve the performance and safety of current flexibility controllers. Current controllers are often fixed and lack the capability to deal with dynamics involved in modern power systems caused by, for instance, the increasing introduction of RES into the grid. These non-stationarities render existing controllers suboptimal and potentially unsafe. The SCRL algorithm introduces a changepoint detection method that allows the controller to adapt to changes in the environment. Moreover, a dynamic datadriven MPC safety filter is introduced in the proposed framework to



Fig. 11. Safety comparison between the proposed adaptive RL strategy and a standard non-adaptive RL strategy in testing phase. The adaptive method clearly shows less unsafe actions compared to the conventional RL strategy. Black dotted lines show instances where the detector identified an environment change and the algorithm was retrained. A 95% confidence interval is added around the curve (100 runs).

guarantee absolute safety of the controller. The model employed in the model predictive controller is data-driven as it is based on data from RL interactions with the environment. The effectiveness of the framework was validated on a grid-connected PV-battery-load system, showcasing improved flexibility and safety compared to a conventional, non-adaptive framework. By increasing the energy flexibility of the system, the energy costs were reduced by 9.3%. Future work will focus on exploring more complex use cases to further validate the framework's applicability and performance.

CRediT authorship contribution statement

Siebe Paesschesoone: Writing – review & editing, Writing – original draft, Visualization, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. Nezmin Kayedpour: Writing – review & editing, Validation, Supervision, Conceptualization. Carlo Manna: Writing – review & editing, Validation, Supervision, Project administration, Conceptualization. Guillaume Crevecoeur: Writing – review & editing, Validation, Supervision, Project administration, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The photovoltaic (PV), demand and price data utilized in this study are sourced from the official website of Elia, the Belgian electricity transmission system operator. The data is publicly available on the Elia Grid Data platform at https://www.elia.be/en/grid-data [39].

Acknowledgments

The authors gratefully acknowledge the support and funding provided by Flemish Institute for Technological Research (VITO), Belgium for this research under grant number VITO_UG_PhD_2302.

References

- [1] Gong X, Wang X, Cao B. On data-driven modeling and control in modern power grids stability: Survey and perspective. Appl Energy 2023;350:121740. http: //dx.doi.org/10.1016/j.apenergy.2023.121740, URL https://www.sciencedirect. com/science/article/pii/S0306261923011042.
- [2] Chen X, Qu G, Tang Y, Low S, Li N. Reinforcement learning for selective key applications in power systems: Recent advances and future challenges. IEEE Trans Smart Grid 2022;13(4):2935–58. http://dx.doi.org/10.1109/tsg.2022.3154718.
- [3] Babayomi O, Zhang Z, Dragicevic T, Hu J, Rodriguez J. Smart grid evolution: Predictive control of distributed energy resources—A review. Int J Electr Power Energy Syst 2023;147:108812. http://dx.doi.org/10.1016/j.ijepes.2022.108812, URL https://www.sciencedirect.com/science/article/pii/S0142061522008080.
- [4] Aswani A, Gonzalez H, Sastry SS, Tomlin C. Provably safe and robust learningbased model predictive control. Automatica 2013;49(5):1216–26. http://dx.doi. org/10.1016/j.automatica.2013.02.003.
- [5] Arroyo J, Helsen L, Spiessens F. Synergy between control theory and machine learning for building energy management. 2022.
- [6] Rosolia U, Zhang X, Borrelli F. Data-driven predictive control for autonomous systems. Annual Rev Control Robot Auton Syst 2018;1(1):259–86. http://dx.doi. org/10.1146/annurev-control-060117-105215.
- [7] Carli R, Cavone G, Pippia T, De Schutter B, Dotoli M. Robust optimal control for demand side management of multi-carrier microgrids. IEEE Trans Autom Sci Eng 2022;19(3):1338–51. http://dx.doi.org/10.1109/TASE.2022.3148856.
- [8] Bünning F, Heer P, Smith RS, Lygeros J. Increasing electrical reserve provision in districts by exploiting energy flexibility of buildings with robust model predictive control. Adv Appl Energy 2023;10:100130. http://dx.doi.org/10.1016/j.adapen. 2023.100130.
- [9] Zeilinger MN, Jones CN, Morari M. Real-time suboptimal model predictive control using a combination of explicit MPC and online optimization. IEEE Trans Autom Control 2011;56(7):1524–34. http://dx.doi.org/10.1109/TAC.2011. 2108450.
- [10] Garcia-Torres F, Bordons C, Tobajas J, Real-Calvo R, Santiago I, Grieu S. Stochastic optimization of microgrids with hybrid energy storage systems for grid flexibility services considering energy forecast uncertainties. IEEE Trans Power Syst 2021;36(6):5537–47. http://dx.doi.org/10.1109/TPWRS.2021.3071867.
- [11] Antoniadou-Plytaria K, Steen D, Tuan LA, Carlson O, Mohandes B, Ghazvini MAF. Scenario-based stochastic optimization for energy and flexibility dispatch of a microgrid. IEEE Trans Smart Grid 2022;13(5):3328–41. http://dx.doi.org/10. 1109/TSG.2022.3175418.
- [12] Hewing L, Zeilinger MN. Stochastic model predictive control for linear systems using probabilistic reachable sets. In: 2018 IEEE conference on decision and control. 2018, p. 5182–8. http://dx.doi.org/10.1109/CDC.2018.8619554.
- [13] Li Y, Yu C, Shahidehpour M, Yang T, Zeng Z, Chai T. Deep reinforcement learning for smart grid operations: Algorithms, applications, and prospects. Proc IEEE 2023;111(9):1055–96. http://dx.doi.org/10.1109/JPROC.2023.3303358.
- [14] Sutton RS, Barto AG. Reinforcement Learning: An Introduction. 2nd ed.. The MIT Press; 2018, URL http://incompleteideas.net/book/the-book-2nd.html.
- [15] Sayed AR, Zhang X, Wang Y, Wang G, Qiu J, Wang C. Online operational decision-making for integrated electric-gas systems with safe reinforcement learning. IEEE Trans Power Syst 2023;1–14. http://dx.doi.org/10.1109/TPWRS. 2023.3320172.
- [16] Shi X, Xu Y, Chen G, Guo Y. An augmented Lagrangian-based safe reinforcement learning algorithm for carbon-oriented optimal scheduling of EV aggregators. IEEE Trans Smart Grid 2024;15(1):795–809. http://dx.doi.org/10.1109/TSG. 2023.3289211.
- [17] Ding H, Xu Y, Chew Si Hao B, Li Q, Lentzakis A. A safe reinforcement learning approach for multi-energy management of smart home. Electr Power Syst Res 2022;210:108120. http://dx.doi.org/10.1016/j.epsr.2022.108120, URL https://www.sciencedirect.com/science/article/pii/S0378779622003443.
- [18] Zhang S, Jia R, Pan H, Cao Y. A safe reinforcement learning-based charging strategy for electric vehicles in residential microgrid. Appl Energy 2023;348:121490. http://dx.doi.org/10.1016/j.apenergy.2023.121490, URL https://www.sciencedirect.com/science/article/pii/S0306261923008541.
- [19] Daneshvar Garmroodi A, Nasiri F, Haghighat F. Optimal dispatch of an energy hub with compressed air energy storage: A safe reinforcement learning approach. J Energy Storage 2023;57:106147. http://dx.doi.org/10.1016/j.est.2022.106147, URL https://www.sciencedirect.com/science/article/pii/\$2352152X22021363.
- [20] 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. http://dx.doi.org/10.1016/j.egyai.2022.100227, URL https://www.sciencedirect.com/science/article/pii/S2666546822000738.

- [21] Qiu D, Dong Z, Zhang X, Wang Y, Strbac G. Safe reinforcement learning for real-time automatic control in a smart energy-hub. Appl Energy 2022;309:118403. http://dx.doi.org/10.1016/j.apenergy.2021.118403, URL https://www.sciencedirect.com/science/article/pii/S030626192101638X.
- [22] Yu P, Zhang H, Song Y. District cooling system control for providing regulation services based on safe reinforcement learning with barrier functions. Appl Energy 2023;347:121396. http://dx.doi.org/10.1016/j.apenergy.2023.121396, URL https://www.sciencedirect.com/science/article/pii/S0306261923007602.
- [23] Sayed AR, Wang C, Anis HI, Bi T. Feasibility constrained online calculation for real-time optimal power flow: A convex constrained deep reinforcement learning approach. IEEE Trans Power Syst 2023;38(6):5215–27. http://dx.doi. org/10.1109/TPWRS.2022.3220799.
- [24] Wang Y, Qiu D, Sun M, Strbac G, Gao Z. Secure energy management of multi-energy microgrid: A physical-informed safe reinforcement learning approach. Appl Energy 2023;335:120759. http://dx.doi.org/10.1016/j. apenergy.2023.120759, URL https://www.sciencedirect.com/science/article/pii/ S030626192300123X.
- [25] 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. http://dx.doi. org/10.1016/j.segan.2023.101202, URL https://www.sciencedirect.com/science/ article/pii/S2352467723002102.
- [26] Staessens T, Lefebvre T, Crevecoeur G. Adaptive control of a mechatronic system using constrained residual reinforcement learning. IEEE Trans Ind Electron 2022;69(10):10447–56. http://dx.doi.org/10.1109/TIE.2022.3144565.
- [27] Vantilborgh V, Staessens T, De Groote W, Crevecoeur G. Dual regularized policy updating and shiftpoint detection for automated deployment of reinforcement learning controllers on industrial mechatronic systems. Control Eng Pract 2024;142:105783. http://dx.doi.org/10.1016/j.conengprac.2023.105783.
- [28] Khetarpal K, Riemer M, Rish I, Precup D. Towards continual reinforcement learning: A review and perspectives. J Artificial Intelligence Res 2022;75:1401–76.
- [29] Naug A, Quinones-Grueiro M, Biswas G. Deep reinforcement learning control for non-stationary building energy management. Energy Build 2022;277:112584. http://dx.doi.org/10.1016/j.enbuild.2022.112584, URL https: //www.sciencedirect.com/science/article/pii/S0378778822007551.
- [30] Basseville M, Nikiforov I. Detection of Abrupt Change Theory and Application, vol. 15, 1993.
- [31] Padakandla S, Prabuchandran KJ, Bhatnagar S. Reinforcement learning algorithm for non-stationary environments. Appl Intell 2020;50(11):3590–606. http://dx. doi.org/10.1007/s10489-020-01758-5.
- [32] García J, Fernández F. A comprehensive survey on safe reinforcement learning. J Mach Learn Res 2015;16(1):1437–80.
- [33] Arwa EO, Folly KA. Reinforcement learning techniques for optimal power control in grid-connected microgrids: A comprehensive review. IEEE Access 2020;8:208992–9007. http://dx.doi.org/10.1109/ACCESS.2020.3038735.
- [34] Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal policy optimization algorithms. 2017, arXiv:1707.06347.
- [35] Ross GJ, Tasoulis DK, Adams NM. Nonparametric monitoring of data streams for changes in location and scale. Technometrics 2011;53(4):379–89.
- [36] Lepage Y. A combination of Wilcoxon's and Ansari-Bradley's statistics. Biometrika 1971:58:213–7.
- [37] Neuhäuser M, Leuchs A-K, Ball D. A new location-scale test based on a combination of the ideas of levene and lepage. Biom J 2011;53(3):525–34. http:// dx.doi.org/10.1002/bimj.201000162, arXiv:https://onlinelibrary.wiley.com/doi/ pdf/10.1002/bimj.201000162.
- [38] Muriithi G, Chowdhury S. Optimal energy management of a grid-tied solar PV-battery microgrid: A reinforcement learning approach. Energies 2021;14(9). http://dx.doi.org/10.3390/en14092700, URL https://www.mdpi. com/1996-1073/14/9/2700.
- [39] Elia. Open data elia.be. 2023, https://www.elia.be/en/grid-data/open-data. [Accessed 21 December 2023].
- [40] Dabney W, Rowland M, Bellemare MG, Munos R. Distributional reinforcement learning with quantile regression. 2017, arXiv:1710.10044.
- [41] Schulman J, Levine S, Moritz P, Jordan MI, Abbeel P. Trust region policy optimization. 2017, arXiv:1502.05477.
- [42] Lillicrap TP, Hunt JJ, Pritzel A, Heess N, Erez T, Tassa Y, et al. Continuous control with deep reinforcement learning. 2019, arXiv:1509.02971.
- [43] Khamesi V. Ocpdet: A python package for online changepoint detection in univariate and multivariate data. 2022, http://dx.doi.org/10.5281/zenodo. 7632721.