

---

# OPTIMAL CONTROL OF RENEWABLE ENERGY COMMUNITIES SUBJECT TO NETWORK PEAK FEES WITH MODEL PREDICTIVE CONTROL AND REINFORCEMENT LEARNING ALGORITHMS

---

**Samy Aittahar**  
Department of Computer Science  
University of Liège  
Liège, Belgium  
saittahar@uliege.be

**Adrien Bolland**  
Department of Computer Science  
University of Liège  
Liège, Belgium

**Guillaume Derval**  
Department of Computer Science  
University of Liège  
Liège, Belgium

**Damien Ernst**  
Department of Computer Science  
University of Liège  
Liège, Belgium

## ABSTRACT

We propose in this paper an optimal control framework for renewable energy communities (RECs) equipped with controllable assets. Such RECs allow its members to exchange production surplus through an internal market. The objective is to control their assets in order to minimise the sum of individual electricity bills. These bills account for the electricity exchanged through the REC and with the retailers. Typically, for large companies, another important part of the bills are the costs related to the power peaks; in our framework, they are determined from the energy exchanges with the retailers. We compare rule-based control strategies with the two following control algorithms. The first one is derived from model predictive control techniques, and the second one is built with reinforcement learning techniques. We also compare variants of these algorithms that neglect the peak power costs. Results confirm that using policies accounting for the power peaks lead to a significantly lower sum of electricity bills and thus better control strategies at the cost of higher computation time. Furthermore, policies trained with reinforcement learning approaches appear promising for real-time control of the communities, where model predictive control policies may be computationally expensive in practice. These findings encourage pursuing the efforts toward development of scalable control algorithms, operating from a centralised standpoint, for renewable energy communities equipped with controllable assets.

## 1 Introduction

Decentralisation of renewable electricity production, close to its consumption place, is gaining increasing attention as a promising approach to decarbonise the electricity demand [1; 2; 3]. In this context, the European Union (EU) has provided a legal context introducing *renewable energy communities* (REC) [4]. A REC is a local electricity market where the members, equipped with renewable energy generation and storage assets, can share local and decarbonised electricity production with the other members; we refer to this aggregated electricity production as the *REC production*. In practice, the members have joined the REC in order to make cost savings in their electricity bills, where they are charged by their retailers proportionally to their electricity consumption. Moreover, in many parts of the EU, for large companies, the electricity bills also include another (but prominent) grid fees related to their consumption and production peaks in the main electrical grid, which we refer to as *offtake peaks* and *injection peaks*, respectively. These large companies thus expect to significantly maximise their cost savings by joining a REC. This can be achieved by efficiently managing the controllable assets (e.g., batteries) within the REC; we explore this decision-making issue in this paper.

According to the European directives, the local market clearing process of RECs is performed at fixed periods, to which we refer as *market periods*, through the reallocation of the REC production by a central entity, namely the Energy Community Manager (ECM). Among his responsibilities, the ECM must communicate for longer periods (e.g., every month) to the Distribution System Operator (DSO) the results of the clearings performed during this period. We refer to the latter as the *billing period*. The ECM must ensure the compliance of the reallocation of the REC production with the regional (or national) regulations, which notably requires that the REC production is fully allocated to the members in the limits of the amount of energy they have consumed, and that no energy is bought from the retailers to be sold to a member through the REC (and vice-versa). Once the DSO has accepted the clearing results (in the shed of light of the regulation), they transfer them to the retailers to compute the electricity bills of each REC member, accounting for the reallocation of the REC production. We refer to these electricity bills as *ex-post* electricity bills. The ECM is often remunerated by a fraction of the cost savings realised in the *ex-post* electricity bills of the members. We thus assume, in this paper, that the ECM aims to minimise the sum of the *ex-post* electricity bills; note that we neglect the remuneration of the ECM in these bills. We refer to this minimised sum of *ex-post* electricity bills as the *global REC bill*.

*Ex-post* electricity bills include costs related to electricity exchanges with the retailers and through the REC (as determined by the ECM for each market period). On one hand, the tariffs of buying electricity from the retailers mainly include the price of the energy consumed, the distribution and transmission fees (the tariffs related to these fees are fixed by the DSO for all members), and other taxes; for prosumers, the electricity sold to the retailers, at a price fixed by contract, is deducted from the electricity bills. On the other hand, the electricity production shared through the REC is only subject to distribution fees; we note that they differ from the fees fixed by the DSO. As mentioned above, grid fees related to offtake and injection peaks are also part of these *ex-post* electricity bills. However, the European regulation does not specify whether these grid fees should be computed as if no REC has been implemented. We argue that the injection and offtake peaks should be only computed from the energy exchanged with the retailers in order to accelerate the development of RECs; this is also an assumption that we will make throughout this paper. Indeed, the overall REC electricity consumption and production are physically seen by the main electrical grid as already aggregated. Furthermore, these costs often constitute a significant fraction of the electricity bills. Beyond the decarbonation of the demand and the cost savings in the electricity bills, investing into assets that foster consumption from local renewable sources reduces the risks of outages of the main electrical grid, particularly in winter periods [5; 6]. We note however that the optimal control framework proposed in this paper can be easily adapted to the case where power peaks costs would be computed without accounting for the REC.

The scientific literature, discussed in Section 2, does not provide, to date and at the best of our knowledge, any existing modelling framework that explicitly optimise the controllable assets usage in renewable energy communities towards the minimisation of global REC bills over time. To address this gap, we propose in this paper a framework for modelling the optimal usage of controllable assets in RECs. In Section 3, we describe the problem of acting on the controllable assets to minimise the sum of global REC bills over time as a Partially Observable Markov Decision Process (POMDP). The dynamics of this POMDP represent the electricity power flows of the members, which are influenced by external events, and the reward function reflects the (negated) global REC bill as computed by the ECM. We describe, in Section 4, two practical control algorithms that aim to minimise the sum of the global REC bills over time. The first one follows a *model predictive control* scheme [7] by solving an internal optimisation problem built from the POMDP specifications and predictions of incoming external events. The second one is built through a *reinforcement learning* scheme, resulting from a simulation-based policy-search procedure that seeks to approximate the optimal policy [8; 9]. We also introduce variants of these policies that somehow neglect the costs related to the peaks. In Section 5, we compare these policies with rule-based strategies against RECs with 2 members built from synthetic data and with 7 members derived from historical data of an existing REC located in Belgium. We conclude this paper in Section 6 with a detailed discussion on recommendations of research directions to pursue towards the development of scalable control algorithms dedicated to RECs.

## 2 Related work

The work carried out in this paper is related to decision-making issues in REC. We partition this related work into three parts. The first part is related to the main generic decision-making problems associated to REC. The second part focuses on model predictive control techniques (MPC) [7] and reinforcement learning techniques (RL) [8] in microgrids, which can be seen as single-entities RECs. The third part focuses on the scarce literature of the applications of these techniques in renewable energy communities, highlighting their limitations with respect to the decision-making problem we address in this paper.

**Decision-making issues for RECs.** In a broader perspective, many decision-making problems related to RECs are challenging. There is a particular focus on investment strategies. These strategies are influenced by energy costs, which must be carefully set to foster investments into renewable energy generation as well as devices helping to increase the self-consumption rate (e.g., storage devices)[10; 11; 12; 13]. In [14], the authors show that maximising the benefits of a REC operation does not necessarily lead to peak reductions in the main electrical grid. However, they show that fostering investments towards storage devices helps to sensitively decrease the load consumed from the main electrical grid. In [15], the authors introduce the concept of *repartition keys* to reallocate the REC production to the members. Altogether, the results of all these works advocate for pursuing the efforts towards solving these decision-making challenges to support the deployment of the RECs around the world.

**MPC and RL for microgrids.** Control algorithms derived from MPC and RL classes of techniques have been widely used in the past to operate microgrids (see review in [16]). These control algorithms are still popular due to their promising performances as well as their simplicity to implement and deploy in practice [17; 18; 19]. In [20], two algorithms derived from MPC and RL were tested on a single building with PV panels and a battery, with similar performances reported in the results, even though the authors noted that MPC algorithms are more suitable since they can be adapted more quickly for a new building configuration. In [21], the authors propose a hierarchical model predictive control scheme to simultaneously minimise the electricity bills and maximise the lifetime within a microgrid equipped with wind turbines, PV panels and batteries. In [22] and [23], the authors compare several reinforcement learning algorithms against instances of microgrids and show they are able to extract near-optimal policies. Many of these works share similarities with the optimal REC control framework proposed in this paper.

**MPC and RL for REC.** At the best of our knowledge, the literature about the design of control algorithms specifically dedicated to renewable energy communities is rather scarce. In [24], various MPC strategies have been tested on RECs. However, unlike the framework proposed in this paper, the peak powers are not taken into account. In [25], a multi-agent reinforcement learning (MARL) algorithm is designed to optimise the controllable assets of a REC, and is compared to an algorithm combining MPC and supervised learning. Their approach allowed to decrease energy bills as well as mitigate offtake and injection peaks. However, their results are restricted by the specific energy community structure (the members cannot possess their own individual storage devices). In [26], the authors propose another MARL approach to optimise the balance between consumption and production, thus neglecting the overall electricity bill. In [27], the authors propose a framework, based on bi-level optimisation, where peak demand and energy bills are jointly optimised with controllable assets for a single billing period. In this framework, the peaks are computed *before* the reallocation of the REC production.

### 3 Optimal REC Control Problem

This section details the optimal control framework of RECs with some members owning controllable assets. In these RECs, the electricity bill of each member is the sum of the fees related to the exchanges through the REC, the energy bought and sold with prices fixed by the retailers, and grid costs related to their power peaks. The sum of these electricity bills, minimised by the ECM through the determination of the energy exchanged through the REC, forms the *global REC bill*. In this section, we describe how a REC is structured and managed through time by the ECM. Notably, the latter periodically clears the local market through the reallocation of the REC production by following a repartition scheme that is explained in Section 3.2. In practice, the ECM is remunerated by charging members either through fixed fees (e.g., annual subscription plans) or proportionally to the amount of energy exchanged through the REC. We leave the business model of the ECM out of the scope of this paper. However, the repartition scheme that we describe further in this section can easily be adapted to account for the ECM remuneration. We conclude that section by providing a full formulation of the dynamical system as a (particular case of) Partially Observable Markov Decision Process (POMDP), notably providing a time discretisation that accounts for the billing or market periods, as well as smaller time periods for a more granular usage of the controllable assets.

#### 3.1 REC Structure and Management

A member of the REC is characterized by (i) its tariffs for energy bought (sold) from (to) the retailers and (ii) its means of consumption and electricity production (some of them being controllable). The tariffs related to energy exchanges with the retailers are composed of the price of (bought or sold) energy, distribution and transmission fees, and taxes; recall that we do not consider the fixed terms in these tariffs (e.g., subscription plans). According to the contract fixed by retailers with the members (and more generally with end users), the price of energy itself is either fixed or variable through time (e.g., peak and off-peak hours or real-time markets such as BELPEX for Belgium). To maintain a reasonable complexity in this paper, we only consider contracts with fixed price plans. However, the framework

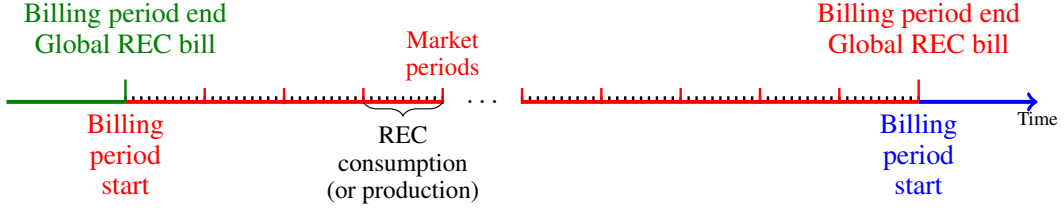


Figure 1: Renewable energy community timeline. During a billing period, each member consumes and produces in real time electricity on their own and by the usage of their controllable assets; in the timeline, black ticks refer to time steps during which control actions are taken. At the end of a billing period, the ECM computes the (optimal) reallocation of the REC production for each market period. This reallocation creates new meter readings and are emitted by the ECM to the DSO and the retailers to compute the global REC bill.

that we propose in this paper may easily be adapted for variable prices of energy through time. Energy exchanges through the REC are subject to distribution fees that are proportional to the amount of energy exchanged. Unlike the tariffs of energy exchanges with the retailers, the tariffs related to distribution fees within the REC are fixed for all members. Note that these distribution fees, to which we concisely refer in the remainder of this paper as *REC fees*, are different than those that are charged by the DSO (to the users of its electrical network) for electricity exchanges with the retailers. Additional (but often prominent) costs are grid fees related to offtake and injection peaks. Similarly to the REC fees, their respective tariffs are fixed for all the members by the DSO. In this paper, we assume that these peaks correspond, for each billing period and for each member of the REC, to maximum values of energy exchanged with the retailers across the market periods elapsed in that billing period. We also assume that peak tariffs are the highest in the electricity bills of the members of the REC; however, this is not a strict hypothesis, but rather a design choice that follows the typical structure of electricity bills.

The net consumption and production of the members (i.e., after accounting for self-consumption at electrical level) are independently metered in real time. At the end of fixed time periods, that we denote as *billing periods* (e.g., every month), we assume that the ECM clears the local market by following a known reallocation scheme, which is described in Section 3.2, to share the REC production to the other members for each market period. This reallocation scheme aims to minimise the sum of the *ex-post* electricity bills while ensuring its compliance with the European Union directives. The result of this reallocation scheme, which may be concisely represented as *repartition keys* [15], is emitted by the ECM to the retailers and the DSO (along with the meter readings). Once the reallocation scheme is accepted by the DSO, the retailers emit the *ex-post* electricity bills to their respective members. These bills account for the energy bought (sold) from (to) the retailers, the distribution fees related to the energy exchanges through the REC, the costs related to the offtake and injection peaks (accounting for the reallocation scheme). Figure 1 illustrates the timeline of the REC.

### 3.2 Optimal reallocation scheme

For a member, sharing a fraction of its energy production or receiving a fraction of the REC production impacts its *ex-post* electricity bill. More precisely, if a member has a surplus of electricity production while another has a positive net consumption, selling the former to the retailer and buying the latter from another retailer will cost more than sharing the production via the REC, notably because of the peaks measured after sharing the REC production (retailer's/main grid point of view). We assume that the ECM applies an *optimal reallocation* scheme. This scheme, resulting in the global REC bill, is formalised hereafter.

Let us consider a REC with  $M$  members for which the electricity bills are computed for  $R$  market periods within a billing period. Let  $C_{m,r}^-$  and  $C_{m,r}^+$  be the values of the meter readings, both non-negative and expressed in kWh, of a member  $m \in \{1, \dots, M\}$  that respectively correspond to its net consumption and its net production for the market period  $r \in \{1, \dots, R\}$ . Traditionally, this member is billed by its retailer, for the entire billing period, at a cost of  $B_m^-$  for its net consumption and remunerated at a price of  $B_m^+$  for its net production. Recall that these cost coefficients are sums of proportional terms (to energy amounts, we neglect fixed terms) related to prices of buying and selling energy, to distribution/transmission fees, and to other taxes. Electricity flows create offtake and injection peaks, which respectively correspond to the greatest values of the net consumption and the net production meter readings measured during the billing period. These values are billed to the members through the unique tariffs  $P^-$  and  $P^+$  that are applied to offtake and injection peaks, respectively. In this paper, we do not convert the peak values into power units (kW) to avoid overburdening the reallocation scheme modelling. If the member was not participating in the REC, its electricity bill, to which we refer as  $EB_m$ , would be computed as follows:

$$EB_m = \left( \sum_{r=1}^R B_m^- C_{m,r}^- - B_m^+ C_{m,r}^+ \right) + P^- \max_{r \in \{1, \dots, R\}} C_{m,r}^- + P^+ \max_{r \in \{1, \dots, R\}} C_{m,r}^+. \quad (1)$$

In others words, energy consumption and production are proportionally billed and remunerated at the prices fixed by the retailers for each market period and peaks are proportionally billed at unique prices for all members, We denote as  $EB = \sum_1^M EB_m$  the total combined value of the electricity bills of the members. Inside a REC, the members can exchange their electricity production (to the other members) at each market period; these exchanges, characterised by either sharing a fraction of the net production of a member to the REC or sharing a fraction of the REC production to a member, are subject to REC fees that are defined by  $\Lambda^+$  and  $\Lambda^-$ , respectively. Note that these distribution fees are not necessarily equal to the hidden distribution fees that fall within prices fixed by the retailers. To model the energy exchanges through the REC, we introduce the decision variable  $e_{m,r}^-$ , which is the quantity of electricity production allocated to the member  $m$  for the market period  $r$ . Similarly, we introduce the decision variable  $e_{m,r}^+$  which is the quantity of electricity production shared by the member  $m$  for the market period  $r$ . Recall that the ECM cannot shape its repartition scheme in order to buy energy from the retailer of a member to share it to another member through the REC. To that end, the value of the former ( $e_{m,r}^-$ ) is bounded by the value of the consumption meter reading after accounting of the production meter reading, and vice versa for the latter ( $e_{m,r}^+$ ):

$$0 \leq e_{m,r}^- \leq \max(C_{m,r}^- - C_{m,r}^+, 0), \quad \forall m \in \{1, \dots, M\}, \quad \forall r \in \{1, \dots, R\}, \quad (2)$$

$$0 \leq e_{m,r}^+ \leq \max(C_{m,r}^+ - C_{m,r}^-, 0), \quad \forall m \in \{1, \dots, M\}, \quad \forall r \in \{1, \dots, R\}. \quad (3)$$

Any quantity of net electricity production exchanged through the REC is shared with a member:

$$\sum_{m=1}^M e_{m,r}^- = \sum_{m=1}^M e_{m,r}^+, \quad \forall r \in \{1, \dots, R\}. \quad (4)$$

Within the REC, the offtake and injection peaks are measured after the reallocation of the REC production, and the energy exchanges in the REC are still subject to the network fees. Once the REC production has been reallocated (by the ECM), the ex-post electricity bill of the current billing period is computed (at the end of this billing period) for a given member as follows:

$$EB'_m(e_m^-, e_m^+) = \left( \sum_{r=1}^R B_m^- (C_{m,r}^- - e_{m,r}^-) - B_m^+ (C_{m,r}^+ - e_{m,r}^+) + \Lambda^- e_{m,r}^- + \Lambda^+ e_{m,r}^+ \right) + \quad (5)$$

$$P^- \left( \max_{r \in \{1, \dots, R\}} C_{m,r}^- - e_{m,r}^- \right) + P^+ \left( \max_{r \in \{1, \dots, R\}} C_{m,r}^+ - e_{m,r}^+ \right),$$

where  $e_m^- = (e_{m,1}^-, \dots, e_{m,R}^-)$ ,  $e_m^+ = (e_{m,1}^+, \dots, e_{m,R}^+)$ . Note that intermediate ex-post electricity bills can be computed at any time during the billing period (accounting for the elapsed time within that billing period). More formally, let  $\tau \in ]0; 1]$  be a ratio value of the time elapsed in the billing period;  $\tau = 1$  corresponds to a fully elapsed billing period (in which case ex-post electricity bills are computed by Equation (5)). If  $\tau < 1$ , we compute, for each member, its intermediate ex-post electricity bill as follows:

$$EB''_m(e_m^-, e_m^+, \tau) = \left( \sum_{r=1}^{\lceil \tau R \rceil} B_m^- (C_{m,r}^- - e_{m,r}^-) - B_m^+ (C_{m,r}^+ - e_{m,r}^+) + \Lambda^- e_{m,r}^- + \Lambda^+ e_{m,r}^+ \right) + \quad (6)$$

$$\tau \left[ P^- \left( \max_{r \in \{1, \dots, \lceil \tau R \rceil\}} C_{m,r}^- - e_{m,r}^- \right) + P^+ \left( \max_{r \in \{1, \dots, \lceil \tau R \rceil\}} C_{m,r}^+ - e_{m,r}^+ \right) \right].$$

Note that, if  $\tau < 1$ , values of meter reading inputs for all members ( $C^+$  and  $C^-$ ) that comes after the market period indexed by  $\lceil \tau R \rceil$  are undefined, and so are corresponding variables ( $e^+$  and  $e^-$ ).

The goal of the ECM is to minimise the sum of ex-post electricity bills by identifying the optimal energy exchanges through the REC. The (possibly intermediate) global REC bill, accounting for the elapsed time during the billing period and referred to as  $GRB_\tau$ , is the result of the following minimisation problem:

$$GRB_\tau = \min_{(e^{*,-}, e^{*,+})} \sum_{m=1}^M EB'_m(e_m^{*,-}, e_m^{*,+}, \tau) \quad (7)$$

where  $e^{*,-} = (e_1^{*,-}, \dots, e_M^{*,-})$  and  $e^{*,+} = (e_1^{*,+}, \dots, e_M^{*,+})$ . We provide, in Appendix B, illustrative examples of this optimal reallocation scheme in order to provide some intuition about this formulation and its complexities and to show the importance of accounting for the peak costs during the computation of the optimal reallocation scheme.

### 3.3 Decision Process associated to RECs

We formalise the above-mentioned REC dynamical system as an infinite-time decision process. The latter is a particular instance of Partially Observable Markov Decision Processes (POMDP) in that the agent interacting with this decision process fully observes the electrical state of the system (in this case, controllable assets state and meter readings) but has a partial view on external events that are not influenced by its decision. However, these external events have an impact, altogether with the decisions of the agent, on the state transitions (meter readings increments) and reward function (the global REC bill). We recall that a billing period covers a fixed number of market periods. Typically, market periods have a fixed duration (e.g., fifteen minutes). During a market period, there is a need to take actions on the controllable assets. Thus, we introduce a time discretisation inside a market period such that, at any time step, an action can be applied to controllable assets for a fixed duration. More formally, this POMDP provides a time discretisation from a time step ( $t \in \mathcal{T} = \mathbb{N}$ ) to the next, which also accounts for the duration of market periods and billing periods, respectively expressed as a number of time steps and a number of market periods. We provide below the components of this decision process that we describe from a high-level perspective (see Appendix A.1 for more mathematical details):

**State space.** A state of the system contains, for each member, the current state of their controllable assets, the elapsed time of the current billing period and their meter readings, collected at the end of each elapsed market period (including the current one) for the current billing period. We denote as  $\mathcal{S}$  the set of all such states, and the unique initial state as  $S_0 \in \mathcal{S}$ .

**Exogenous space.** An exogenous variable contains, for each member, their consumption and production powers generated from their non-controllable assets, averaged during the current time interval. We denote as  $\mathcal{E}$  the set of such exogenous variables. We assume that the initial exogenous variable is a random variable following a probability distribution  $P_0^\mathcal{E}(\cdot)$  and that exogenous variables transitions from any time step  $t$  to the next are steered by an unknown non-Markovian conditional distribution that we denote as  $P^\mathcal{E}(e_{t+1}|e_{0:t})$  for all  $t \in \mathcal{T}$ , where  $e_{0:t} = (e_0, \dots, e_t)$ .

**Action space.** An action contains the (control) decisions to be applied to controllable assets. We denote the set of such actions as  $\mathcal{U}$ . The set of actions that can be applied in a given state  $s \in \mathcal{S}$  may be restricted, and are referred to as *admissible actions*. These admissible actions are given by the function  $U(s)$ , such that  $U(s) \subseteq \mathcal{U}$ .

**Transition dynamics.** The transition dynamics of the state space are defined as follows. Controllable assets states are updated for each discrete time step  $t \in \mathcal{T}$  according to a model depending on the asset specifications, their current state  $s_t \in \mathcal{S}$  and the action  $u_t \in U(s_t)$ . After the electricity flows have been computed by accounting for the controllable assets usage, the meter readings are updated for the market period to which the next time step is associated. The state transition is concisely expressed by the function  $f$  as follows:

$$s_{t+1} = f(s_t, e_t, u_t). \quad (8)$$

**Cost function.** The cost function describes the global REC bill computation as described in Section 3.2. More formally, this function provides a zero cost at all time steps except at the end of each billing period where the solution  $GRB$  of the optimal reallocation scheme is returned. In short, the cost signal is defined as follows:

$$\rho_t = \begin{cases} GRB_1 & \text{if } t \text{ is the last timestep of a billing period,} \\ 0 & \text{otherwise.} \end{cases} \quad (9)$$

Note that this cost signal can be easily modified to output intermediate global REC bills at other time steps:

$$\rho_t = \begin{cases} GRB_{\tau(t)} & \text{if an intermediate global REC bill is needed at time step } t, \\ 0 & \text{otherwise,} \end{cases} \quad (10)$$

where  $\tau(t)$  is the value of  $\tau$  in Equation (6) that is computed from (the information contained in) state  $s_t$ .

To maintain a reasonable complexity, we neglect the operational costs related to the controllable assets (e.g., discharging fees for storage devices). In practice, the cost function can be extended to incorporate these operational costs at each time step depending on the actions performed.

## 4 Optimal control of RECs

A process that chooses the next action to execute based on the current information of the system is called a *policy*. This information is composed of the current state and the history of past exogenous variables values. In this section, we describe how to select a policy that optimally reduces the sum of the global REC bills of the REC through time, as follows. Since these optimal policies cannot be computed directly as  $P_0^\mathcal{E}$  and  $P^\mathcal{E}$  are not known, we propose to approximate them through practical policies derived from model predictive control and reinforcement learning schemes [7; 8], for which we provide a high-level description in this section. To assess the efficiency of policies accounting for the peak power costs to minimise the sum of global REC bills over time, we introduce simpler variants which neglect the costs associated to the peaks when choosing the action. All the mathematical details about these policies are available in Appendix A.2.

### 4.1 Optimal policies

In this paper, we call admissible policies for a REC policies that provide actions that are feasible in the REC dynamical system. Formally, let  $\Pi$  be the set of such policies:

$$\Pi = \left\{ \pi : S \times \mathcal{H}_\mathcal{E} \rightarrow \mathcal{U} \mid \pi(s, e_{0:t}) \in U(s, e_t), \forall (s, e_{0:t}) \in S \times \mathcal{H}_\mathcal{E} \right\}.$$

where  $\mathcal{H}_\mathcal{E}$  is the set of all possible ordered sequences of exogenous variables defined as

$$\mathcal{H}_\mathcal{E} = \bigcup_{n \in \mathbb{N}^+} \mathcal{E}^n. \quad (11)$$

We measure the performance of a policy when executed starting from a given state and a given sequence of past exogenous variable as the opposite of the expected discounted sum of the future costs arising from the actions chosen by the policy through time [9]:

$$V_\pi(s_k, e_{0:k}) = \lim_{T \rightarrow \infty} \mathbb{E}_{e_{t+1} \sim P^\mathcal{E}(\cdot | e_{0:t})} \sum_{t=k}^{T-1} -\gamma^t \rho(s_t, e_t, u_t, s_{t+1}), \quad (12)$$

where  $\gamma \in ]0, 1[$  is a discount factor that indicates the relative importance of future costs compared to present costs. For any  $t \geq k$ , the next states are computed as  $s_{t+1} = f(s_t, e_t, u_t)$  and the related action as  $u_t = \pi(s_t, e_{0:t})$ . We define the expected return of a policy  $\pi \in \Pi$ , starting from the initial state  $S_0$ , as follows:

$$J(\pi) = \mathbb{E}_{e_0 \sim P_0^\mathcal{E}(\cdot)} V_\pi(S_0, e_{0:0}). \quad (13)$$

The *optimal REC control problem* is to identify a policy  $\pi^* \in \Pi$  that maximises Equation (13):

$$\pi^* \in \arg \max_{\pi \in \Pi} J(\pi). \quad (14)$$

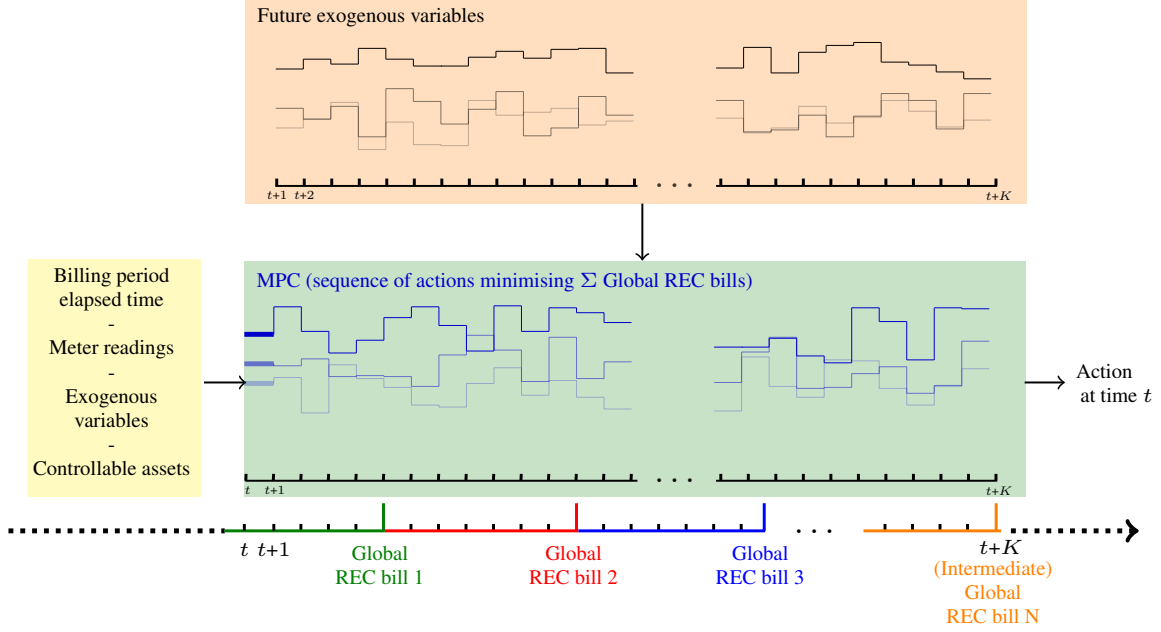


Figure 2: Illustration of how MPC policies compute the next action to apply in the REC dynamical system from a given state and a sequence of exogenous variables. The latter is composed of the current exogenous variable and values for the future ones (up to the policy horizon  $t + K$ ). Colors are used to differentiate the billing periods that the MPC policies consider during its optimisation process.

## 4.2 Model Predictive Control (MPC) policies

The *MPC policy* solves an internal optimisation problem minimising a discounted sum of the future costs (possibly over several billing periods) starting from a state and an exogenous variable sequence, and arising from a sequence of admissible actions up to a fixed time horizon  $K$  that we denote as *policy foresight*. This optimisation problem expects, as inputs, a sequence of future exogenous variables; they are usually computed with external methods, which are left out of the scope of this paper. If the last time step of this optimisation problem does not coincide with the end of a billing period, no cost signal is associated to the actions applied during this period. In that case, the objective function does not quantify the impact of these actions in the associated global REC bill. To avoid this pitfall, an intermediate global REC bill accounting for the elapsed time of that billing period, as defined in Section 3.2, is added to the objective function at the last time step. To that end, we introduce a cost signal, namely  $\widehat{GRB}$ , which outputs this intermediate bill for time step  $t + K$  if it is not located at the end of a billing period, and otherwise 0. In short, the MPC policy computes the next (admissible) action  $u_t^*$  at the time step  $t$  such that

$$\widehat{GRB}_t = \arg \max_{u_t} \left[ \max_{(u_{t'+1}, \dots, u_{t'+K})} - \left( \widehat{GRB} + \sum_{t'=t}^{t+K} \gamma^{t'} \rho(s_{t'}, \hat{e}_{t'}, u_{t'}, s_{t'+1}) \right) \right], \quad (15)$$

where  $\hat{e}_t = e_t$  is the current exogenous variable and  $\hat{e}_{t+1:t+K}$  are future values of exogenous variables up to the policy foresight. We model this optimisation problem as a mixed-integer linear program (MILP). We also consider a variant of this policy, namely the *MPC retail* policy, which only differs from the MPC policy in that the objective function of the MILPs does not include the costs related to the peaks. We have used CPLEX to solve the MILPs built by the MPC policies. Figure 2 provides an illustration of the computation steps of MPC policies.

## 4.3 Reinforcement Learning (RL) policy

In Reinforcement Learning, one typically considers differentiable policies parameterised by a vector of parameters  $\theta \in \Theta$ . The goal is thus to identify the optimal parameters to maximize the expected return of the resulting policy:

$$\theta^* \in \arg \max_{\theta \in \Theta} J(\pi_\theta). \quad (16)$$



However, this policy cannot be computed directly as we do not possess the distributions  $P_0^\mathcal{E}$  and  $P^\mathcal{E}$ . Instead, we focus on reinforcement learning algorithms performing stochastic gradient ascent on the objective function  $J$ , where the gradient is estimated through simulations of the policies in the POMDP. In particular, we use the Proximal Policy Optimization [28] (PPO) algorithm. This algorithm learns, simultaneously with the parameterised policy, another parametric function, called the *critic*. The latter is used, in combination with heuristic rules defined in the PPO algorithm, to mitigate the high variance issues associated to simulation-based gradient estimates. Neural networks are used as parametric functions, along with recurrent layers to consume sequences of exogenous variables values [29; 30].

Policy gradient algorithms are often more efficient when costs are defined at every time steps. However, in the cost function defined in Section 3, sparsity increases with the duration of a billing period (the longer the billing period gets, the more we have zeros in the costs). To attempt to mitigate this issue, we consider another version of this policy, denoted *RL dense*, in which zero costs are replaced by intermediate global REC bills during the policy search.

Similarly to the policy *MPC retail*, we consider a variant of this policy that we denote as *RL retail*. In this variant, cost signals are modified, before updating the parameters of the policy, as if there were no power peak costs (i.e., assuming that  $P^- = P^+ = 0$ ). We also define a policy which combines the cost modifications applied in the policy search procedure for the RL dense and RL retail policies. We refer to that policy as *RL dense retail*. To implement the RL policies with the above-mentioned procedure, we have used the PPO implementation of the python programming library *RLlib* [31] to implement this policy search procedure, along with the deep learning python library *PyTorch* [32] to implement the neural network models.

## 5 Experiments

In this section, we simulate the policies described in Section 4. Note that, in practice, we do not simulate algorithms for infinite time; in our experiments, we always choose a time horizon that is sufficiently long to cover several billing (and consequently, market) periods. Due to the scarcity of historical data associated with existing RECs, we sample exogenous variables (up to the fixed time horizon of the simulations) by applying a time-correlated white noise to (scarcely) existing data; see Appendix C.1 for more details about this sampling procedure. We first describe how future exogenous values are computed for the MPC policies in the context of our simulations. We then introduce baseline policies in Section 5.2 that we compare with MPC and RL policies against two REC instances, referred to as *REC-2* and *REC-7*; they respectively count 2 and 7 members, see Appendices C.3 and C.4 for more details about the experimental protocol settings. The first one is an illustrative example built from synthetic data, and the second one from historical data from an existing REC located in Wallonia, Belgium. We compare the results of all the simulated policies, respectively for each REC instance, in Sections 5.3 and 5.4. The settings related to the procedure to build the RL policies follow the usual recommendations related to policy gradient algorithms [33; 34], as detailed in Appendix C.

### 5.1 Computing future exogenous values for MPC policies

MPC policies expect future values of exogenous variables to compute its next action. In a realistic settings, these future values, usually computed with forecasting methods [35; 36], differ from the actual future values; this difference often increases with the length of these values, and has an (often negative) impact in the performance of the MPC policies. As mentioned earlier in this section, sequences of exogenous variables are sampled, from existing data, at once before simulating the policies. We thus propose the following procedure to compute the future values of the exogenous variables (during the simulation of the MPC policies). At each time step  $t$ , the future exogenous value at the next time step  $t + 1$ , that we provide to the MPC policies, corresponds to the actual one. From  $t + 2$  (if  $K > 1$ ), the time series of future exogenous values progressively converges (from the true future values) to the existing data (again, from  $t + 2$ ). We characterise the speed of this convergence by a parameter  $\alpha \in ]0, 1]$ , such that if  $\alpha = 0$ , then the time series provided to the MPC policies is equals to the existing data from  $t + 2$ ; and that if  $\alpha = 1$ , then all the next exogenous variables are equals to the true future values (again, from  $t + 2$ ). In this procedure, another characteristic is that, as long as  $\alpha$  gets closer to 0, the convergence speed of the predicted time series (to the existing data) dramatically increases; see Appendix C.2 for a more detailed description of this sampling procedure. We refer to this parameter ( $\alpha$ ) as the *foresight efficiency*.

### 5.2 Baseline policies

We introduce the following baseline policies to provide boundaries in which the expected returns of the MPC and RL policies are located:

**Self-consumption.** The two following rule-based policies output the next actions to be applied to the controllable assets by maximising some self-consumption criterion. The first one, denoted as *REC policy*, compares the total net consumption and the total net production of the non-controllable assets of the REC. It then uses the controllable assets, to either absorb the excess of electricity consumption or the surplus of electricity production. The second one, denoted as *SELF policy*, uses the same approach but individually for each member equipped with controllable assets.

**Perfect-foresight.** We consider the particular case of MPC policies where  $\alpha = 1$  and  $K = T$  with  $T$  a finite time horizon that is fixed for the simulations. In that case, as defined in Section 4.2, the MPC policies perfectly "predict" the future exogenous values at each time step until the time horizon  $T$ , thus leading to the minimal discounted sum of the global REC bills for that horizon. Under these conditions, the MPC retail policy leads to the minimal discounted sum of the global REC bill with  $P^+ = P^- = 0$ . We refer to these respective policies as *OPT policy* and *OPT retail*.

### 5.3 REC-2

REC-2 is composed of 2 members. One of them is only equipped with non-controllable assets that consume electricity; the other one is equipped with PV panels, that always produce more electricity than consumed by the other non-controllable assets, and is also equipped with a battery. Their corresponding consumption and production profiles (from non-controllable assets) are derived from synthetic historical data; see Appendix C.3 for more details about these profiles. We set the time horizon of the simulations at  $T = 101$ . Expected returns of baseline policies and MPC policies are estimated over 1024 simulations. In order to evaluate the RL policies, we first train 16 policy functions using different random seeds. This training procedure is carried as follows. Each policy function is trained for 600 iterations with the PPO algorithm, each iteration being comprised of 64 simulations of the policy functions. At the end of these simulations, we have 16 instances of trained RL policies. We estimate their expected returns by running 64 simulations for each of these RL policies, and we average them.

Figure 3 shows the expected returns of all the simulated policies in REC-2; see Appendix C.3 for more details about the simulation settings of REC-2. We first notice that the expected returns of MPC policies quickly converge as  $K$  grows. Furthermore, MPC policies with perfect foresight efficiency ( $\alpha = 1$ ) yield optimal expected returns (with respect to the sampled exogenous variables time series). We observe that the SELF policy has the worst expected return. Since the owner of the battery only produces electricity from its non-controllable assets, this policy will only charge the battery until full capacity is reached. The profiles show that no electricity production is available for the first time steps. By consequence, the producer turned to consume electricity which could not be absorbed by the REC, thus increasing the energy consumption bills as well as the offtake peaks, at least for the first billing period. The REC policy mitigates this issue by accounting for the electricity consumption of all the members. However, the other policies (eventually) yield better expected returns than these rule-based policies. The performance of RL policies neglecting peak costs are close to REC policy; they might have learnt to operate the battery similarly to the latter. Although the RL policies have a worse expected return than the OPT retail policy, it manages to get slightly better than some of the MPC policies with noisy prediction. This illustrates the advantage of having access to predictions of future exogenous variables values (given that the foresight efficiency is reasonably high). Note that the expected return of RL dense is slightly worse than the RL policy.

During the simulations of the PPO algorithm (see Appendix C.3 for detailed results), we also noticed that the RL dense policy got a slightly better expected return, for the first iterations, than the RL policy (we even have observed a higher gap for RL retail and RL retail dense). Both observations are due to the modified cost signal (by adding intermediate global REC bills, with or without peak costs). Indeed, having cost signals at every time steps allowed to speed up the policy updates (in terms of expected return) for the first few iterations. However, the discounted sum of the original cost signals is not equal to the discounted sum of the modified cost signals. By consequence, the RL dense policy might have converged too quickly to a local optimum due to the difference between the two formulations.

### 5.4 REC-7

REC-7 is composed of 7 members. Some of them are equipped with PV panels. Similarly to REC-2, one of them is equipped with a (controllable) battery. Consumption and production profiles are derived from available historical data from an existing REC located in Wallonia, Belgium; see Appendix C.4 for more details. We set the time horizon of the simulations at  $T = 721$ . Expected returns of baseline policies and MPC policies (with their respective policy horizons and foresight efficiencies) are estimated over 4096 simulations. Expected returns of baseline policies and MPC policies are estimated over 1024 simulations. We evaluate the RL policies exactly like for REC-2, excepted that we run the PPO algorithm on 32 policies for 1000 iterations and that we perform, for each policy, 128 simulations at each iteration to train them, and another 128 simulations to estimate the expected return of the RL policies.

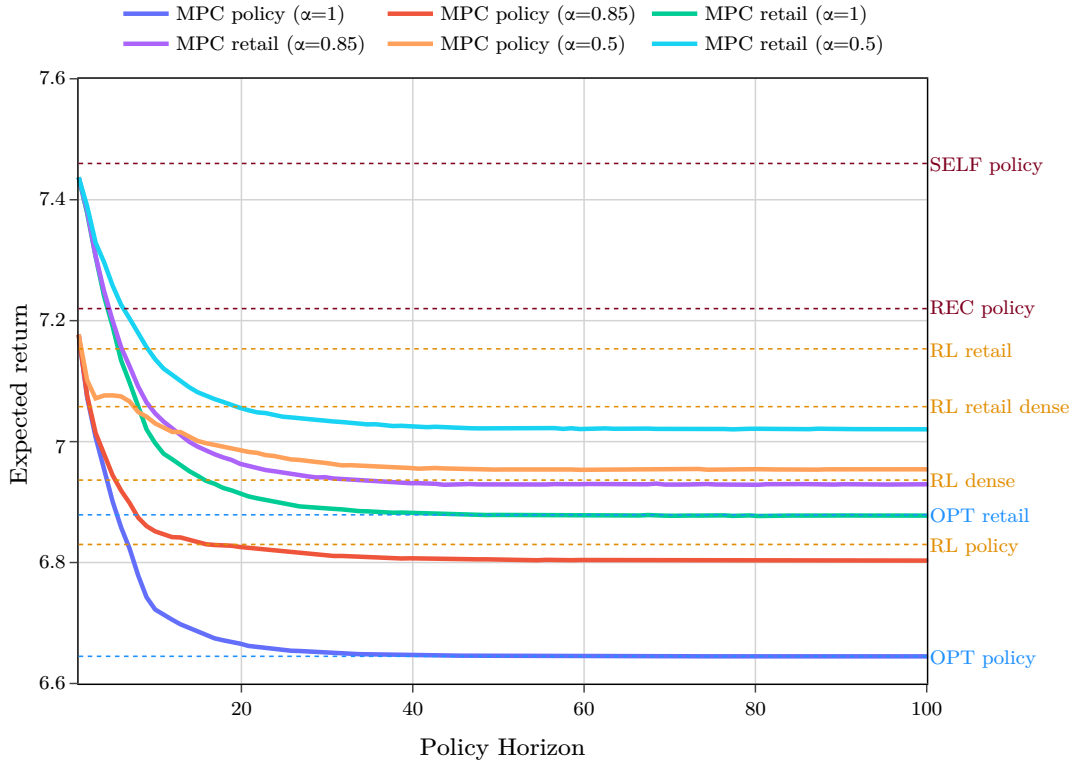


Figure 3: Expected returns of MPC policies (averaged over 16 runs) in REC-2, given policy horizon and foresight efficiency, along with expected returns of RL and baseline policies. Recall that  $\alpha$  is the foresight efficiency defined in Section 4.2, with  $\alpha = 1$  corresponding to perfect foresight.

Similarly to Section 5.3, we report the results of simulating the policies in REC-7 in Figure 4. We first notice that REC policy has a worse expected return than the SELF policy. This is surprising, and might be due to the composition of this REC. More surprisingly, RL policies neglecting peak costs also have a worse expected return than the SELF policy. The RL policy barely managed to get better than the latter. They might have suffered from either the high sparsity of the cost function, or the dense formulation (with intermediate global REC bills) of the discounted sum of costs (for the RL retail dense). Only RL dense managed to get slightly better than the OPT retail policy. While MPC policies are still the best policies in REC-7 (beating RL dense with a rather low policy foresight), their variants neglecting the peak costs are still better than most of the policies.

## 5.5 General comments

Overall, MPC policies and RL policies are better than self-consumption rule-based policies, excepted for REC-7 where the SELF policy is better than some RL policies, but notably not the RL dense one. Note that, at the exception of the REC policy in REC-7, the expected return of the policies were not very far from each other. This might be due to the limited impact of the controllable assets in the cost savings [37]. As expected, MPC policies accounting for the peak costs got the best expected returns with a reasonably low policy foresight. However, as shown by the results, they are dependent on both the policy horizon and the foresight efficiency. In our results, the expected return of RL policies were, at best, close to the OPT retail policy.

Both densifying the reward signals and accounting for the peak costs prove to be useful. Indeed, in the case of REC-7, RL dense obtains better results than the other RL policies. Indeed, densifying the reward signals without peak costs has a limited effect (RL retail dense), as well as taking into account the peak costs without densifying the reward signals (RL policy). This is not true for REC-2 where the non-densified RL policy (accounting for the peak costs) beats RL dense. This is probably mainly due to the limited time horizon, providing an already quite dense reward signal.

The required computation time for computing the next action with MPC policies dramatically increases with the number of members and the policy horizon, as shown in Table 1. Conversely, and similarly to the baseline policies,

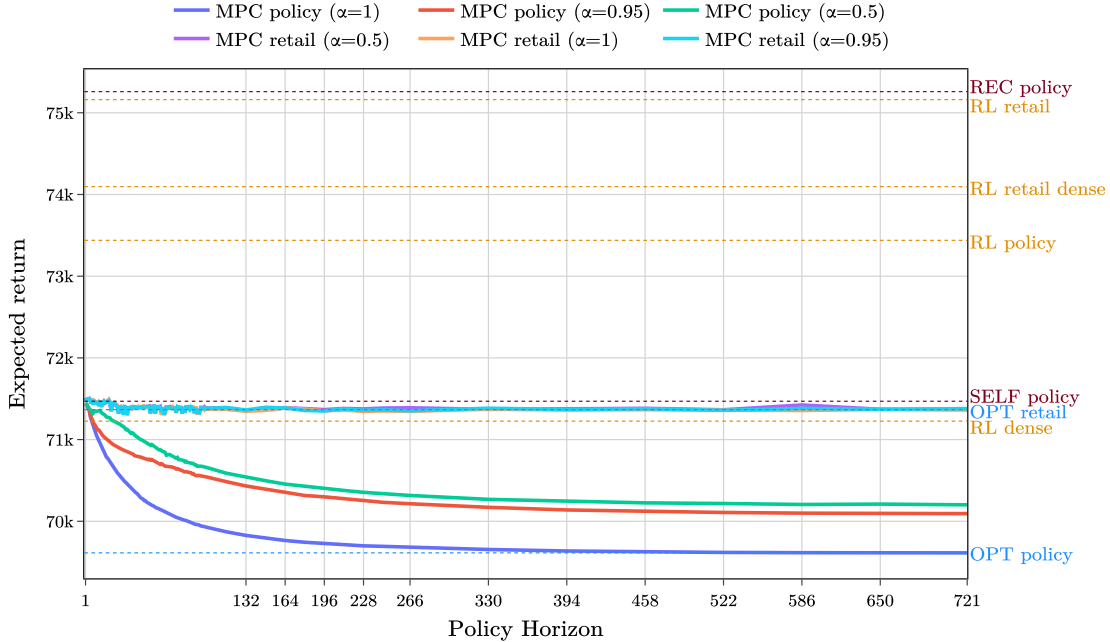


Figure 4: Expected returns of MPC policies (averaged over 16 runs) in REC-7, given policy horizon and foresight efficiency, along with expected returns of RL and baselines policies.

REC instance	MPC policies				RL policies		Baselines	
	$\frac{T}{8}$	$\frac{T}{4}$	$\frac{T}{2}$	T	Update	Action	REC	SELF
REC-2	0.01	0.02	0.04	0.82	43.20	6e-4	1e-3	
REC-7	0.05	0.11	0.20	0.40	321.27	1e-4	1e-3	

Table 1: Sample computation times required to compute the next action for MPC policies (given policy horizons), RL policies (also including update iteration of the PPO algorithm) and baselines (SELF and REC) policies. Times are expressed in seconds, and are averaged over 32 independent runs. These computation times have been carried on a laptop equipped with 32 GB of RAM and a (Intel Gen Core i7) CPU with 12 cores and 16 threads.

computing the next action with RL policies (once trained) is much faster and scales with large sizes of RECs. However, the overall runtime of training RL policies may dramatically increase (with the size of RECs), especially with the frequency of the reward function computation (e.g., at every time steps).

There is thus a trade-off between MPC policies (very good results but slow) and RL policies (less good results but very fast at runtime while very slow to train) that must be made depending on the context of the REC specific needs.

## 6 Conclusion and Future work

In this paper, we have proposed a generic modelling framework to optimise RECs with controllable assets towards the minimisation of the sum of the *ex-post* electricity bills (accounting for the energy exchanges through the REC), which we call the global REC bill. One of the prominent costs of these electricity bills are the offtake and injection peak fees, which are computed from the energy exchanges with the retailers. In this framework, we have formalised how the Energy Community Manager optimally reallocates (to minimise the sum of the *ex-post* electricity bills), at regular time intervals, the surplus of electricity production to the members. This formalisation is integrated in the cost function of a Partially Observable Markov Decision Process, which also encapsulates the dynamics related to the electricity power flows (accounting for the controllable assets). We have tested practical policies that approximate the optimal ones, from the model predictive control and reinforcement learning classes of techniques. These policies have been compared to baseline algorithms, including variants of these policies that neglect the peak costs, or rule-based policies that operate

the controllable assets to maximise self-consumption criteria. The tests were carried against a synthetic REC instance, and against a REC instance derived from historical data of an existing REC located in Wallonia, Belgium. The results obtained from these tests show how policies, accounting for the peak costs, may sensitively decrease the global REC bills. They also strongly encourage pursuing the efforts towards the development of REC frameworks to optimise the controllable assets from a centralised standpoint, possibly by addressing the following limitations of the approach proposed in this paper.

The optimal reallocation scheme, described in Section 3.2, is computed by solving an optimisation problem, which is done in practice through third-party solvers. As long as the size of the RECs stays relatively low, solving this reallocation scheme is quite fast. But as RECs grow, with dozens, even hundreds of members, commercial third-party solvers coupled with heavy-duty hardware begin to be required to scale the computation time, especially when simulations are needed for e.g., running business cases for REC investments, or training algorithms (like RL policies) to optimally operate RECs with controllable assets. However, these solvers usually exploit interior point methods to solve convex optimisation problems. Running these algorithms with GPUs may be much faster compared to non-commercial third-party solvers [38][39][40]. Another research direction would be to pursue the efforts to identify closed-form approaches to compute approximations for optimal reallocation schemes (accounting for the peak costs) which still lead to efficient policies that rely on this approximation.

Similar issues arise for MPC policies. Indeed, they require to solve mixed-integer linear programs (see Appendix A.3 for more details). As the size of the REC and the policy horizon both grow, their computation time dramatically increases. These policies also require predictions of the future electricity flows of the non-controllable assets of the REC members. Scarcity of historical data, controllable assets with complex dynamics, and the difficulty to build algorithms dedicated to predictions of these flows make it challenging to scale MPC techniques with big sized RECs. Like for the optimal reallocation scheme, adapting these MPC policies to run on GPUs might be an alternative to solve MILPs [ref needed]. As for the electricity flows predictions, to face the scarcity of historical data for a given REC, transfer learning techniques might help, provided the existence of supervised learning models predicting similar electricity flows [41].

The RL policies have been shown to be a promising alternative to MPC policies, especially when the reward signals are densified, as they keep a rather low complexity to be used to operate RECs with controllable assets once they have been trained. For the training procedure, since the REC control problem is centralised (by the ECM), we have used a single-agent policy gradient algorithm called PPO [28]. While building RL policies through this single-agent configuration should be the proper approach, they are in practice difficult to train due to both the presence of recurrent layers [34] and the amount of information needed to feed to the parametric functions used by the RL policies, which quadratically grows with the number of members in the REC and the length of a billing period. The complexity of the cost signal, which is mainly due to the peak costs and requires solving an optimisation problem, further plague the scaling of the training procedure for RECs with large sizes. Yet, according to our results, the presence of these peak costs sensitively impacted the expected return of these RL policies, particularly in REC-7, our largest REC in our simulations having 7 members. To mitigate these scaling issues, a surrogate cooperative multi-agent POMDP, where each member could be an agent, might be designed to build the RL policies, while keeping them compatible with the original (centralised) POMDP [42; 26; 25].

## References

- [1] Susana Soeiro and Marta Ferreira Dias. Renewable energy community and the european energy market: main motivations. *Heliyon*, 6(7):e04511, July 2020.
- [2] Maria Luisa Di Silvestre, Mariano Giuseppe Ippolito, Eleonora Riva Sanseverino, Giuseppe Sciumè, and Antony Vasile. Energy self-consumers and renewable energy communities in italy: New actors of the electric power systems. *Renewable and Sustainable Energy Reviews*, 151(111565):111565, November 2021.
- [3] Robin Sudhoff, Sebastian Schreck, Sebastian Thiem, and Stefan Niessen. Operating renewable energy communities to reduce power peaks in the distribution grid: An analysis on grid-friendliness, different shares of participants, and economic benefits. *Energies*, 15(15):5468, July 2022.
- [4] European Union. Directive 2018/2001 of the European Parliament and of the Council of 11 december 2018 on the promotion of the use of energy from renewable sources. *Official Journal of the European Union*, 2018.
- [5] Ruben Laleman and Johan Albrecht. Belgian blackout? estimations of the reserve margin during the nuclear phase-out. *International Journal of Electrical Power and Energy Systems*, 81:416–426, October 2016.
- [6] Hassan Haes Alhelou, Mohamad Hamedani-Golshan, Takawira Njenda, and Pierluigi Siano. A survey on power system blackout and cascading events: Research motivations and challenges. *Energies*, 12(4):682, February 2019.

- [7] James B Rawlings, David Q Mayne, and Moritz Diehl. *Model Predictive Control: Theory, Computation, and Design*. Nob Hill Publishing, 2017.
- [8] Richard S Sutton and Andrew G Barto. *Reinforcement Learning*. Adaptive Computation and Machine Learning series. Bradford Books, Cambridge, MA, 2 edition, November 2018.
- [9] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.
- [10] Na Li and Özge Okur. Economic analysis of energy communities: Investment options and cost allocation. *Applied Energy*, 336(120706):120706, April 2023.
- [11] Longxi Li, Xilin Cao, and Sen Zhang. Shared energy storage system for prosumers in a community: Investment decision, economic operation, and benefits allocation under a cost-effective way. *Journal of Energy Storage*, 50(104710):104710, June 2022.
- [12] Rolf Wüstenhagen and Emanuela Menichetti. Strategic choices for renewable energy investment: Conceptual framework and opportunities for further research. *Energy Policy*, 40:1–10, January 2012.
- [13] Francisco Belmar, Patrícia Baptista, and Diana Neves. Modelling renewable energy communities: assessing the impact of different configurations, technologies and types of participants. *Energy, Sustainability and Society*, 13(1), June 2023.
- [14] Robin Sudhoff, Sebastian Schreck, Sebastian Thiem, and Stefan Niessen. Operating renewable energy communities to reduce power peaks in the distribution grid: An analysis on grid-friendliness, different shares of participants, and economic benefits. *Energies*, 15(15):5468, July 2022.
- [15] Miguel Manuel De Villena, Samy Aittahar, Sebastien Mathieu, Ioannis Boukas, Eric Vermeulen, and Damien Ernst. Financial optimization of renewable energy communities through optimal allocation of locally generated electricity. *IEEE Access*, 10:77571–77586, 2022.
- [16] Damien Ernst, Mevludin Glavic, Florin Capitanescu, and Louis Wehenkel. Reinforcement learning versus model predictive control: a comparison on a power system problem. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2):517–529, April 2009.
- [17] Karan Singh Joshal and Neeraj Gupta. Microgrids with model predictive control: A critical review. *Energies*, 16(13):4851, June 2023.
- [18] Ting Yang, Liyuan Zhao, Wei Li, and Albert Y Zomaya. Reinforcement learning in sustainable energy and electric systems: a survey. *Annual Reviews in Control*, 49:145–163, 2020.
- [19] Erick O Arwa and Komla A Folly. Reinforcement learning techniques for optimal power control in grid-connected microgrids: A comprehensive review. *IEEE Access*, 8:208992–209007, 2020.
- [20] Sicheng Zhan, Yue Lei, and Adrian Chong. Comparing model predictive control and reinforcement learning for the optimal operation of building-PV-battery systems. *E3S Web Conference*, 396:04018, 2023.
- [21] Xiaobing Kong, Xiangjie Liu, Lele Ma, and Kwang Y Lee. Hierarchical distributed model predictive control of standalone wind/solar/battery power system. *IEEE Transactions on Systems, Man, and Cybernetics: Systems.*, 49(8):1570–1581, August 2019.
- [22] Taha A Nakabi and Pekka Toivanen. Deep reinforcement learning for energy management in a microgrid with flexible demand. *Sustainable Energy, Grids and Networks*, 25(100413):100413, March 2021.
- [23] Ding Liu, Chuanzhi Zang, Peng Zeng, Wanting Li, Xin Wang, Yuqi Liu, and Shuqing Xu. Deep reinforcement learning for real-time economic energy management of microgrid system considering uncertainties. *Frontiers in Energy Research*, 11, March 2023.
- [24] Samy Aittahar, Miguel Manuel de Villena, Guillaume Derval, Michael Castronovo, Ioannis Boukas, Quentin Gemine, and Damien Ernst. Optimal control of renewable energy communities with controllable assets. *Frontiers in Energy Research*, 11, February 2023.
- [25] Bo-Chen Lai, Wei-Yu Chiu, and Yuan-Po Tsai. Multiagent reinforcement learning for community energy management to mitigate peak rebounds under renewable energy uncertainty. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(3):568–579, June 2022.
- [26] Amit Prasad and Ivana Dusparic. Multi-agent deep reinforcement learning for zero energy communities, September 2019.
- [27] Nikita Tomin, Vladislav Shakirov, Aleksander Kozlov, Denis Sidorov, Victor Kurbatsky, Christian Rehtanz, and Electo E S Lora. Design and optimal energy management of community microgrids with flexible renewable energy sources. *Renewable Energy*, 183:903–921, January 2022.

- [28] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv*, 2017.
- [29] Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. A review of recurrent neural networks: LSTM cells and network architectures. *Neural computation*, 31(7):1235–1270, July 2019.
- [30] Ronald J Williams and Jing Peng. An efficient gradient-based algorithm for on-line training of recurrent network trajectories. *Neural computation*, 2(4):490–501, December 1990.
- [31] Eric Liang, Richard Liaw, Robert Nishihara, Philipp Moritz, Roy Fox, Ken Goldberg, Joseph Gonzalez, Michael Jordan, and Ion Stoica. Rllib: Abstractions for distributed reinforcement learning. In *International conference on machine learning*, pages 3053–3062. PMLR, 2018.
- [32] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.
- [33] Marcin Andrychowicz, Anton Raichuk, Piotr Stańczyk, Manu Orsini, Sertan Girgin, Raphael Marinier, Léonard Hussenot, Matthieu Geist, Olivier Pietquin, Marcin Michalski, Sylvain Gelly, and Olivier Bachem. What matters for on-policy deep actor-critic methods? A large-scale study. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. International Conference on Learning Representations, 2021.
- [34] Marco Pleines, Matthias Pallasch, Frank Zimmer, and Mike Preuss. Generalization, mayhems and limits in recurrent proximal policy optimization. *CoRR*, abs/2205.11104, 2022.
- [35] Kangji Li, Wenping Xue, Gang Tan, and Anthony S Denzer. A state of the art review on the prediction of building energy consumption using data-driven technique and evolutionary algorithms. *Building Services Engineering Research and Technology*, 41(1):108–127, January 2020.
- [36] Wen-Chang Tsai, Chia-Sheng Tu, Chih-Ming Hong, and Whei-Min Lin. A review of state-of-the-art and short-term forecasting models for solar PV power generation. *Energies*, 16(14):5436, July 2023.
- [37] Alex Felice, Lucija Rakocevic, Leen Peeters, Maarten Messagie, Thierry Coosemans, and Luis Ramirez Camargo. Renewable energy communities: Do they have a business case in flanders? *Applied Energy*, 322(119419):119419, September 2022.
- [38] Akshay Agrawal, Brandon Amos, Shane Barratt, Stephen Boyd, Steven Diamond, and Zico Kolter. Differentiable convex optimization layers. *Advances in neural information processing systems*, 32, 2019.
- [39] Tu-Hoa Pham, Giovanni De Magistris, and Ryuki Tachibana. OptLayer - practical constrained optimization for deep reinforcement learning in the real world. *2018 IEEE International Conference on Robotics and Automation (ICRA)*, May 2018.
- [40] Ga Wu, Buser Say, and Scott Sanner. Scalable planning with tensorflow for hybrid nonlinear domains. *Advances in Neural Information Processing Systems*, 30, 2017.
- [41] Elissaios Sarmas, Nikos Dimitropoulos, Vangelis Marinakis, Zoi Mylona, and Haris Doukas. Transfer learning strategies for solar power forecasting under data scarcity. *Scientific Reports*, 12(1):14643, August 2022.
- [42] Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in Neural Information Processing Systems*, 35:24611–24624, 2022.
- [43] E M L Beale and J J H Forrest. Global optimization using special ordered sets. *Operational Research*, 10(1):52–69, December 1976.
- [44] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016.
- [45] Vincent François-Lavet, Raphael Fonteneau, and Damien Ernst. How to discount deep reinforcement learning: Towards new dynamic strategies. *arXiv*, 2015.
- [46] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv*, 2014.
- [47] George B Dantzig. Discrete-variable extremum problems. *Operational Research*, 5(2):266–288, April 1957.

## A Mathematical details

### A.1 Decision Process

In this section, we provide a detailed description of the decision process introduced in Section 3 as a particular instance of Partially Observable Markov Decision Process (POMDPs). This POMDP provides the following time discretisation. The duration between two consecutive time steps (for controllable assets and thus meter reading transitions) is defined by the value  $\delta \in \mathbb{R}^+$ . The duration of a market period, expressed in number of time steps, is defined by the value  $\Delta_M \in \mathbb{N}^+$ . The duration of a billing period, expressed in number of market periods, is defined by the value  $\Delta_B \in \mathbb{N}^+$ . We describe below, in detail, the components of the POMDP. Let  $M$  be the number of members of the REC. A state  $s \in \mathcal{S}$  contains the following information:

- the number of time steps elapsed in the current market period  $s^{\Delta_M} \in \mathbb{N}$ ,
- the number of market periods elapsed in the current billing period  $s^{\Delta_B} \in \mathbb{N}$ ,
- the state of the controllable assets  $s_m^c$  of each member  $m \in \{1, \dots, M\}$ ,
- the production meter reading  $s_{(m,n)}^{r+} \geq 0$  of each member  $m \in \{1, \dots, M\}$  at each market period  $n \in \{1, \dots, \Delta_B\}$ ,
- the consumption meter reading  $s_{(m,n)}^{r-} \geq 0$  of each member  $m \in \{1, \dots, M\}$  at each market period  $n \in \{1, \dots, \Delta_B\}$ .

An exogenous variable  $e \in \mathcal{E}$  contains the following information:

- the electricity net production of each member  $e_m^p \geq 0$ ,
- the electricity net consumption of each member  $e_m^c \geq 0$ .

Since these net electricity flows account for self-consumption during the time interval between  $t$  and  $t + 1$ , we have that  $e_m^p e_m^c = 0$ . The dynamics of the controllable assets are derived from their specifications. The dynamics related to the states memorising the elapsed time in the current billing period are defined as follows:

$$s_0^{\Delta_M} = s_0^{\Delta_B} = 0, \quad (17)$$

$$s_{t+1}^{\Delta_M} = \begin{cases} 1, & \text{if } s_t^{\Delta_M} = \Delta_M, \\ s_t^{\Delta_M} + 1, & \text{otherwise,} \end{cases} \quad (18)$$

$$s_{t+1}^{\Delta_B} = \begin{cases} 0, & \text{if } s_t^{\Delta_B} = \Delta_B, \\ s_t^{\Delta_B}, & \text{if } s_t^{\Delta_M} < \Delta_M, \\ s_t^{\Delta_B} + 1, & \text{otherwise.} \end{cases} \quad (19)$$

Let  $q_m^c$  be a function giving the amount of electricity consumption (positive) or production (negative) generated by applying the action  $u_m^c$  in the controllable asset state for each member  $s_m^c$ . The dynamics related to the production and consumption meter readings are defined for all members  $m \in \{1, \dots, M\}$  as follows:

$$s_{(m,n),0}^{r+} = s_{(m,n),0}^{r-} = 0, \quad \forall n \in \{1, \dots, \Delta_B\} \quad (20)$$

$$s_{(m,n),t+1}^{r+} = \begin{cases} 0 & \text{if } s_t^{\Delta_B} = \Delta_B, \\ s_{(m,n),t+1}^{r+} & \text{if } s_t^{\Delta_B} > n - 1, \\ s_{(m,n),t+1}^{r+} + l_{m,t}^+ & \text{otherwise,} \end{cases} \quad (21)$$

$$s_{(m,n),t+1}^{r-} = \begin{cases} 0 & \text{if } s_t^{\Delta_B} = \Delta_B, \\ s_{(m,n),t+1}^{r-} & \text{if } s_t^{\Delta_B} > n - 1, \\ s_{(m,n),t+1}^{r-} + l_{m,t}^- & \text{otherwise,} \end{cases} \quad (22)$$

where

$$l_{m,t}^+ = \max(q_m^c(s_{m,t}^c, u_{m,t}^c) - e_{m,t}^c + e_{m,t}^p, 0.0), \quad (23)$$

$$l_{m,t}^- = \max(q_m^c(s_{m,t}^c, u_{m,t}^c) + e_{m,t}^c - e_{m,t}^p, 0.0). \quad (24)$$



are respectively the net electricity production (i.e., after accounting of self-consumption) and the net electricity consumption (i.e., after accounting of self-consumption) of the member  $m$  at time step  $t$ .

As defined in Section 3.3, the cost signal of the POMDP is computed by solving the minimisation problem defined in Equation (7), resulting in what we call the global REC bill, from input values (related to meter readings and elapsed time in the billing period) provided by states at last time steps of billing periods; otherwise the cost signal is zero. However, intermediate global REC bills can be computed during the simulation of the POMDP by adapting the value of  $\tau$ , introduced in Equation (6), to correspond to the time elapsed during the billing period. More formally, given decision variables  $e_m^+$  and  $e_m^-$  for each member  $m \in \{1, \dots, M\}$ , related to the energy exchanged by the members through the REC as defined in Section 3.2, we rewrite Equation (6) to account for the input values provided by a state  $s \in \mathcal{S}$  as follows:

$$EB_m''(e_m^+, e_m^-, s) = \left( \sum_{n=1}^{\delta_B} B_m^-(s_{(m,n)}^{r^-} - e_{m,n}^-) - B_m^+(s_{(m,n)}^{r^+} - e_{m,n}^+) + \Lambda^- e_{m,n}^- + \Lambda^+ e_{m,n}^+ \right) + \quad (25)$$

$$\frac{s^{\Delta_{BM}}}{\Delta_B \Delta_M} \left[ P^- \left( \max_{n \in \{1, \dots, \delta_B\}} s_{(m,n)}^{r^-} - e_{m,n}^- \right) + P^+ \left( \max_{n \in \{1, \dots, \delta_B\}} s_{(m,n)}^{r^+} - e_{m,n}^+ \right) \right],$$

where  $\delta_B = \min(s^{\Delta_B} + 1, \Delta_B)$ , and  $s^{\Delta_{BM}}$  is the number of time steps elapsed during the billing period corresponding to the state  $s$ :

$$s^{\Delta_{BM}} = \begin{cases} \Delta_M \Delta_B & \text{if } s^{\Delta_B} = \Delta_B, \\ s^{\Delta_B} \Delta_M & \text{if } s^{\Delta_M} = \Delta_M, \\ s^{\Delta_B} \Delta_M + s^{\Delta_M} & \text{otherwise.} \end{cases} \quad (26)$$

## A.2 Policies

In this section, we provide a detailed description of the policies introduced in Sections 4 and 5.2 in the context of the POMDP described above. Namely, (i) the MPC policies which decide, given an approximate prediction of the future exogenous values, the next optimal action to apply (with respect to the time horizon and the values of the prediction); (ii) the RL policies which approximate the optimal policies through parameterised functions that output the next action by exploiting only the current state and history of exogenous values; (iii) the SELF policy and the REC policies that outputs the next action which maximises some self-consumption criteria given the current state and exogenous variables values.

## A.3 MPC Policies

The MPC policies compute the next action by solving an optimisation problem accounting for the structure of the POMDP introduced in Section 3.3 and future exogenous variables values; we refer to this sequence of values as  $\hat{e}$ . More formally, these policies compute the optimal sequence of actions that minimises, with respect to the current state and the predicted exogenous values, the discounted sum of the upcoming global REC bills. To that end, it solves a mixed-integer linear (MILP) program that is modelled as follows. Let  $\mathcal{T}_{t:t+K}^{\Delta_M}$  be the last time steps of market periods between  $t$  and  $t + K$ :

$$\mathcal{T}_{t:t+K}^{\Delta_M} = \left\{ t' \in \left\{ t - s_{t'}^{\Delta_{BM}}, \dots, t + K \right\} \mid s_{t'}^{\Delta_M} = \Delta_M \text{ or } t' = t + K \right\}. \quad (27)$$

Let  $(g_{m,t\Delta_M}^-, g_{m,t\Delta_M}^+, r_{m,t\Delta_M}^-, r_{m,t\Delta_M}^+)$  be the non negative auxiliary variables that respectively correspond, for each member  $m \in \{1, \dots, M\}$  and for each last time step market period  $t^{\Delta_M} \in \mathcal{T}_{t:t+K}^{\Delta_M}$ , to the amount of energy bought(sold) from(to) the retailer, as well as energy bought(sold) from(to) the REC (local market). Let  $\mathcal{T}_{t+1:t+K}^{\Delta_B}$  be the time steps that correspond to the last time step of billing periods between  $t + 1$  and  $t + K$ :

$$\mathcal{T}_{t+1:t+K}^{\Delta_B} = \left\{ t' \in \{t + 1, \dots, t + K\} \mid s_{t'}^{\Delta_B} = \Delta_B \text{ or } t' = t + K \right\}. \quad (28)$$

Let  $p_{m,t\Delta_B}^-$  and  $p_{m,t\Delta_B}^+$  be the non negative auxiliary variables that respectively correspond to the offtake and injection peaks for each member  $m \in \{1, \dots, M\}$  and for each end of billing period  $t^{\Delta_B} \in \mathcal{T}_{t:t+K}^{\Delta_B}$ . We concisely denote the

energy exchanges summed over an entire billing period, which ends at  $t^{\Delta_B} \in \mathcal{T}_{t+1:t+K}^{\Delta_B}$ , as follows:

$$\begin{aligned}
g_{m,t^{\Delta_B}}^- &= \sum_{t^{\Delta_M} \in \mathcal{T}_{t^{\Delta_B}}^{\Delta_B \leftarrow M}} g_{m,t^{\Delta_M}}^-, \\
g_{m,t^{\Delta_B}}^+ &= \sum_{t^{\Delta_M} \in \mathcal{T}_{t^{\Delta_B}}^{\Delta_B \leftarrow M}} g_{m,t^{\Delta_M}}^+, \\
r_{m,t^{\Delta_B}}^- &= \sum_{t^{\Delta_M} \in \mathcal{T}_{t^{\Delta_B}}^{\Delta_B \leftarrow M}} r_{m,t^{\Delta_M}}^-, \\
r_{m,t^{\Delta_B}}^+ &= \sum_{t^{\Delta_M} \in \mathcal{T}_{t^{\Delta_B}}^{\Delta_B \leftarrow M}} r_{m,t^{\Delta_M}}^+,
\end{aligned} \tag{29}$$

where  $\mathcal{T}_{t^{\Delta_B}}^{\Delta_B \leftarrow M} = \mathcal{T}_{t^{\Delta_B} - s_{t'}^{\Delta_{BM}} : t^{\Delta_B}}^{\Delta_M}$  is the set of time steps corresponding to end of market periods that belong to the billing period ending at time step  $t^{\Delta_B}$ . To compute the next action, the MPC policies select an optimal sequence (of decisions) minimising the following objective function (discounted sum of the upcoming global REC bills):

$$\begin{aligned}
\widehat{GRB}(s_t, \hat{e}_{t:t+K}, \hat{u}_{t:t+K}) &= \\
&\min \left( g_{m,t^{\Delta_M}}^-, g_{m,t^{\Delta_M}}^+, r_{m,t^{\Delta_M}}^-, r_{m,t^{\Delta_M}}^+ \right), \forall t^{\Delta_M} \in \mathcal{T}_{t:t+K}^{\Delta_B}, t^{\Delta_B} \in \mathcal{T}_{t:t+K}^{\Delta_B} \\
&\left[ B_m^- g_{m,t^{\Delta_B}}^- - B_m^+ g_{m,t^{\Delta_B}}^+ + \Lambda^- r_{m,t^{\Delta_B}}^- + \Lambda^+ r_{m,t^{\Delta_B}}^+ + \frac{P^+ p_{m,t^{\Delta_B}}^+ + P^- p_{m,t^{\Delta_B}}^-}{s_{t^{\Delta_B}}^{\Delta_{BM}}} \right]
\end{aligned} \tag{30}$$

where  $\gamma$  is the discount factor associated to the POMDP. Note that MPC retail only differs from MPC policy in that it sets  $P^+ = P^- = 0$  in its internal optimisation problem. The decision variables of the MILP, which correspond to a sequence of actions in the POMDP, is constrained accordingly to the function  $U$  defined in Section 3. Since these constraints depend on the current state (of the controllable assets), we introduce auxiliary variables for the sequence of states for each member  $m$  equipped with controllable assets. In these sequences, the first state corresponds to  $s_t$ , while the subsequent states are defined through the transition function  $f_m^c$ . For each member, the variables related to the offtake and injection peaks correspond, for a given billing period ending at time step  $t^{\Delta_B} \in \mathcal{T}_{t^{\Delta_B}}^{\Delta_B}$ , to the maximum values of the energy exchanged with the retailer across the market periods within this billing period:

$$p_{m,t^{\Delta_B}}^- \geq g_{m,t^{\Delta_M}}^-, \tag{31}$$

$$p_{m,t^{\Delta_B}}^+ \geq g_{m,t^{\Delta_M}}^+, \tag{32}$$

for all  $m \in \{1, \dots, M\}$  and  $t^{\Delta_M} \in \mathcal{T}_{t^{\Delta_B}}^{\Delta_B \leftarrow M}$ . We introduce, in the optimisation problem modelled by the MPC policies, the constraints related to the energy exchanges described in Section 3.2 as follows:

$$\begin{aligned}
g_{m,t^{\Delta_M}}^- + r_{m,t^{\Delta_M}}^- &= \begin{cases} s_t^{r,-} + l_{m,t:t^{\Delta_M}}^- & \text{if } t^{\Delta_M} > t \text{ and } t^{\Delta_M} - \Delta_M < t, \\ l_{m,(t-t^{\Delta_M}):t^{\Delta_M}}^- & \text{if } t^{\Delta_M} > t, \\ s_{t^{\Delta_M}}^{r,-} & \text{otherwise,} \end{cases} \\
g_{m,t^{\Delta_M}}^+ + r_{m,t^{\Delta_M}}^+ &= \begin{cases} s_t^{r,+} + l_{m,t:t^{\Delta_M}}^+ & \text{if } t^{\Delta_M} > t \text{ and } t^{\Delta_M} - \Delta_M < t, \\ l_{m,(t-t^{\Delta_M}):t^{\Delta_M}}^+ & \text{if } t^{\Delta_M} > t, \\ s_{t^{\Delta_M}}^{r,+} & \text{otherwise,} \end{cases} \\
\sum_{m=1}^M r_{m,t^{\Delta_M}}^- &= \sum_{m=1}^M r_{m,t^{\Delta_M}}^+,
\end{aligned} \tag{33}$$

for all  $t^{\Delta M} \in \mathcal{T}_{t:t+K}^{\Delta M}$ , where

$$l_{m,t_{min}:t_{max}}^- = \max \left( \sum_{t'=t_{min}}^{t_{max}} l_{m,t'}^- - \sum_{t'=t_{min}}^{t_{max}} l_{m,t'}^+, 0 \right) \quad (34)$$

and

$$l_{m,t_{min}:t_{max}}^+ = \max \left( \sum_{t'=t_{min}}^{t_{max}} l_{m,t'}^+ - \sum_{t'=t_{min}}^{t_{max}} l_{m,t'}^-, 0 \right) \quad (35)$$

are respectively the net electricity consumption and production from time step  $t$  to  $t' > t$ . We introduce auxiliary variables to compute the net electricity flows (for members that are equipped with controllable assets). To account for self-consumption, we define the value of these variables as follows. If  $l^- > l^+$ , then the variable corresponding to the net consumption is equals to  $l^+ - l^-$ . Similarly, if  $l^+ > l^-$ , then the variable corresponding to the net production is equals to  $l^- - l^+$ . These two variables are mutually exclusive, only one of them can be non-zero at any time ( $l^- l^+ = 0$ ). For net electricity flows involving controllable assets, we have implemented this mutual exclusion constraint by using *special ordered sets* [43] (of type 1, which themselves require to introduce binary variables). We have followed the same approach to implement Equations (34) and (35).

#### A.4 RL Policies

Unlike the MPC policies, the RL policies compute the next action without an explicit approximation of the future exogenous values by using parameterised closed-form functions which are differentiable (with respect to their parameters). In this section, we describe how this policy computes the next action through parameterised functions. Thereafter, we describe how these parameters are determined through reinforcement learning to approximate the optimal policies.

We consider stochastic and parameterised functions which are differentiable with respect to their parameters (usually, deep neural networks). These functions take as input the history of exogenous variables and the states, which are usually transformed (e.g., to decrease dimensionality), and output the parameters of a Gaussian distribution that are used to sample the next action to be applied in the POMDP. We refer to these transformed inputs as *observations*, and we introduce the observation space  $\mathcal{O}$ . We optimise the parameter of one of these functions to maximise its expected return with PPO [28]. Furthermore, we rely on an additional differentiable parameterised function, called the *critic*, which we use in combination with PPO. We respectively denote the two functions as  $\pi_\theta$  and  $v_\phi$ , where  $\theta$  is the parameter of the function  $\pi$  acting as the policy and  $\phi$  is the parameter of the critic function  $v$ . The policy obtained after the last iteration of PPO is what we call a RL policy. After initialising the parameters  $\theta$  and  $\phi$ , this algorithm iterates on the following steps (over a fixed number of iterations). It runs the policy over several independent simulations of the POMDP over a fixed time horizon  $T$ . The result of these simulations are sequences of transitions that are associated with reward signals. We refer to these sequences as *episodes*. From these episodes, stored in a training set that we denote as  $\mathcal{TS}$ , it performs  $N_{\text{upd}} \in \mathbb{N}_+$  updates of the policy  $\pi_\theta$  (with respect to its parameters). For each update step  $1 \leq i \leq N_{\text{upd}}$ , a (small) subset of the transitions of size  $BS \in \mathbb{N}_+$  is sampled, which is denoted as  $TS_i \subseteq \mathcal{TS}$ . Then, for each of these subsets, the update of the policy is computed as follows. Let

$$\hat{A}_\phi(o_t) = \sum_{t'=t}^T (\gamma' \lambda_{\text{GAE}})^{t'-t} (r_{t'} + \gamma v_{\phi'}(o_{t'}) - v_\phi(o_t)) \quad (36)$$

be the so-called *generalised advantage estimation* of the policy  $\pi_\theta$  [44] where  $v_{\phi'}$  and  $v_\phi$  are respectively the prior and the current critic functions,  $\gamma' \leq \gamma$  is the discount factor used in the PPO algorithm (that can be set lower than the one fixed for the POMDP [45]) and  $\lambda_{\text{GAE}} \in ]0; 1]$  is a hyperparameter that further increases the importance of the first reward signals. The descending gradient used to update the policy is computed from the following loss function:

$$\mathcal{L}^\pi(\theta) = -\frac{1}{BS} \sum_{(o_t, u_t, r_t) \in \mathcal{TS}} \left[ \min(r_\theta(o_t, u_t), \hat{A}_\phi(o_t)), \max(1 - \epsilon, \min(r_\theta(o_t, u_t), 1 + \epsilon)) \hat{A}_\phi(o_t) \right], \quad (37)$$

where  $\theta' = \theta$  before the first update,  $r_\theta(o_t, u_t) = \frac{\pi_\theta(u_t|o_t)}{\pi_{\theta'}(u_t|o_t)}$  is the probability ratio between the initial policy and the updated one and  $\epsilon$  is a hyperparameter that somehow limits the value of the loss function. The PPO algorithm updates the critic through another loss function which penalises the mean squared error of the critic function:

$$\mathcal{L}^v(\phi) = \frac{\lambda_{v_\phi}}{BS} \sum_{(o_t, u_t, r_t) \in \mathcal{TS}} (v_{\phi'}(o_t) - v_\phi(o_t))^2, \quad (38)$$

where  $\lambda_{v_\phi}$  is a hyperparameter that scales the importance of this loss function for updating the critic function. The gradient ascent updates are computed with ADAM [46], which dynamically modifies the learning rate, starting from an initial one to which we refer as  $\eta$ . These updates are usually clipped by a global norm defined by a hyperparameter that we denote as  $\beta > 0$ .

We implement parameterised policy and critic functions with deep neural networks, containing recurrent layers [34; 29] to memorise the past exogenous variables in the form of hidden states. These hidden states are incorporated to the observation of the policies. We use the TBPTT algorithm [30] to backpropagate the gradients of the recurrent layers up to a limited time horizon in the past transitions that we denote as  $T_{\text{rnn}}$ .

### A.5 Baseline policies

The *REC policy* and the *SELF policy* computes the next action by respectively maximising the global self-consumption rate of the REC (as if the whole REC is composed of one member only) and the individual self-consumption rate for each member equipped with controllable assets. To that end, they respectively solve the optimisation problem defined below with  $c_1 = 1$  and  $c_2 = 1$ , and setting the other coefficient to 0:

$$\min_{\substack{u_{m,t}^{c,*} \in U(s_{m,t}^c), \\ \forall m \in \{1, \dots, M\}}} c_1 l_t^{REC} + c_2 \sum_{m \in \mathcal{M}_c} l_{m,t} \quad (39a)$$

$$\text{s.t.} \quad l_{m,t} \geq l_{m,t}^- - l_{m,t}^+ \quad \forall m \in \{1, \dots, M\}, \quad (39b)$$

$$l_{m,t} \geq l_{m,t}^+ - l_{m,t}^- \quad \forall m \in \{1, \dots, M\}, \quad (39c)$$

$$l_t^{REC} \geq \sum_{m=1}^M l_{m,t}^- - l_{m,t}^+, \quad (39d)$$

$$l_t^{REC} \geq - \sum_{m=1}^M l_{m,t}^- - l_{m,t}^+. \quad (39e)$$

The terms enabled by the coefficient value  $c_1$  refer to the global self-consumption rate of the REC and the terms enabled by the coefficient value  $c_2$  refer to the individual self-consumption rates.

## B Illustrative examples of the optimal reallocation scheme

In this section, we provide three illustrative examples of the optimal reallocation scheme described in Section 3.2. The first one is a REC composed of 2 members. The second one is a REC composed of 3 members, with peak costs coefficients set to 0. The third one is another REC composed of 3 members with large peak costs (compared than the other cost coefficients). We consider, for all these RECs, a single billing period with 2 market periods. Note that, expected for the second REC, peak cost coefficients are always (sensitively) larger than the other ones (prices of energy and distributions fees).

### B.1 REC composed of 2 members

When there are only 2 members in the REC, that we denote  $m_1$  and  $m_2$ , the optimal solution can be computed directly as follows for each meter reading  $r$ . When one of the members has a positive net production and the other one has a positive net consumption (i.e., that  $C_{m_1,r}^+ > 0$  or  $C_{m_2,r}^- > 0$  or vice versa), that member allocates as much as possible of its net production to the REC, and the REC reallocates the whole net electricity production to the other member. In other situations, no REC production is reallocated. In this REC composition, either buying energy from the retailers instead of buying it from the REC if possible or selling energy instead of sharing it with the REC if possible is always suboptimal in both cases due to the assumptions on the inputs of this optimal reallocation scheme problem. Table 2 shows an example of an optimal reallocation scheme in this REC.

### B.2 REC without peaks costs

When the offtake and injection peaks costs coefficients are neutralised (i.e.,  $P^+ = P^- = 0$ ), the optimal reallocation scheme can be solved with greedy algorithms for each market period [47]. More precisely, this greedy algorithm works as follows. Firstly, this algorithm identifies whether the net consumption (of the REC) is greater than the net production. If that is the case, it allocates the whole net production to the members in decreasing order of their energy buying prices. Otherwise, in a similar manner, it shares the net production of the members to the REC in the increasing

Buying €/kWh		Selling €/kWh		Network €/kWh			
$M_1$	$M_2$	$M_1$	$M_2$	$P^-$	$P^+$	$\Lambda^-$	$\Lambda^+$
<b>0.20</b>	<b>0.22</b>	<b>0.04</b>	<b>0.05</b>	<b>1.00</b>	<b>1.00</b>	<b>0.02</b>	<b>0.03</b>

(a)

Market period	Net consumption kWh		Retail kWh		REC kWh	
	$M_1$	$M_2$	$M_1$	$M_2$	$M_1$	$M_2$
1	<b>252.59</b>	<b>-596.18</b>	<b>0.00</b>	<b>-343.59</b>	<b>252.59</b>	<b>-252.59</b>
2	<b>811.43</b>	<b>-244.02</b>	<b>567.41</b>	<b>0.00</b>	<b>244.02</b>	<b>-244.02</b>
Billing period	Offtake peak kWh		Injection peak kWh		Global REC Bill €	
	$M_1$	$M_2$	$M_1$	$M_2$	NO-REC	REC
1	<b>567.41</b>	<b>0.00</b>	<b>0.00</b>	<b>343.59</b>	<b>1578.40</b>	<b>1032.13</b>

(b)

Table 2: Optimal reallocation scheme problem over 2 market periods for a REC composed of 2 members. Retail prices and network costs are specified in (a). Energy exchanges, peaks and electricity bills are reported in (b). The sum of the individual electricity bills (before accounting for energy exchanges through the REC) is reported in red. The global REC bill (after accounting for energy exchanges through the REC) is reported in blue.

order of their energy selling prices. Table 3 shows the optimal reallocation scheme following that REC composition (i.e., a REC with 3 members where offtake and injection peak costs coefficients are set to 0). Note that repartition schemes that are computed by the above-mentioned greedy algorithm are optimal only if all buying and selling prices are strictly positive.

### B.3 REC with peaks costs

Despite our efforts, we did not identify any closed form solution of the optimal reallocation scheme problem. However, as this optimisation problem is repeatedly solved with different inputs, we have observed during our tests that reusing the previous solutions sensitively speeds-up the solving process. Table 4 shows results of optimal reallocation schemes that either neglect or account for the peak costs.

## C Experimental protocol details

In this section, we provide the details of the experimental protocol (including numerical values) under which the policies, described in detail in Appendix A.2, have been simulated in REC-2 and REC-7. More precisely, we outline the configuration on the exogenous variables sampling for both REC instances, as well as the structure of each REC instance, and the configuration of the MPC and RL policies with respect to these REC instances.

### C.1 Sampling exogenous variables

Weather and daytime related time series, notably solar-based energy production and the companies' energy consumption, are typically time correlated. As an attempt to replicate these time series dynamics, we propose the following approach to simulate  $P_0^e$  and  $P^e$  from historical data. Let  $e$  be an exogenous time series. Let  $\omega$  be a white noise time series (centred on zero) sampled with a standard deviation that we denote  $\sigma$ . To take into account time correlation, we

Buying €/kWh			Selling €/kWh			Network €/kWh			
$M_1$	$M_2$	$M_3$	$M_1$	$M_2$	$M_3$	$P^-$	$P^+$	$\Lambda^-$	$\Lambda^+$
<b>0.20</b>	<b>0.22</b>	<b>0.24</b>	<b>0.04</b>	<b>0.05</b>	<b>0.06</b>	<b>0.00</b>	<b>0.00</b>	<b>0.02</b>	<b>0.03</b>

(a)

Market period	Net consumption kWh			Retail kWh			REC kWh		
	$M_1$	$M_2$	$M_3$	$M_1$	$M_2$	$M_3$	$M_1$	$M_2$	$M_3$
1	<b>368.10</b>	<b>-608.36</b>	<b>-564.67</b>	<b>0.00</b>	<b>-240.26</b>	<b>0.00</b>	<b>368.10</b>	<b>-368.10</b>	<b>0.00</b>
2	<b>486.34</b>	<b>186.40</b>	<b>-162.35</b>	<b>486.34</b>	<b>24.05</b>	<b>-162.35</b>	<b>0.00</b>	<b>162.35</b>	<b>-162.35</b>
Billing period	$M_1$	$M_2$	$M_3$	$M_1$	$M_2$	$M_3$	NO-REC		REC
1	/	/	/	/	/	/	<b>137.86</b>		<b>83.19</b>

(b)

Table 3: Optimal reallocation scheme over 2 market periods for a REC composed of 3 members which are not subject to peak costs. Retail prices and network costs are specified in (a). Energy exchanges, peaks and electricity bills are reported in (b). The sum of the individual electricity bills (before accounting for energy exchanges through the REC) is reported in red. The global REC bill (after accounting for energy exchanges through the REC) is reported in blue.

derive a new noise  $x$ , to which we refer as *red noise*<sup>1</sup>, as follows:

$$x_0 = \omega_0, \quad (40)$$

$$x_{t+1} = rx_t + \sqrt{(1-r^2)}w_{t+1}, \quad \forall t \geq 1, \quad (41)$$

where  $r \in ]0, 1]$  controls the time correlation of the red noise. We then define the two distributions  $P_0^\mathcal{E}$  and  $P^\mathcal{E}$  such that the value of a sample  $\tilde{e}_t$  of the two distributions is given at time step  $t$  by

$$\tilde{e}_t = x_t + e_t. \quad (42)$$

## C.2 Future exogenous variables values for MPC policies

Let  $\tilde{e}$  be the historical sequence of exogenous variables from which new sequences are sampled through the procedure described in Section C.1. Let  $e$  be the exogenous time series containing the future values from the time step  $t$  (as sampled through the procedure described in Section C.1). We thus define the time series  $\hat{e}_{t:t+K}$  to be the prediction of the exogenous variables from  $t$  to  $t+K$ , computed as follows:

$$\hat{e}'_t = \alpha^{t'-t}e_{t'} + (1 - \alpha^{t'-t})\tilde{e}_{t'}, \quad \forall t' \in \{t, \dots, t+K\}, \quad (43)$$

where  $\alpha$  controls the convergence speed of the exogenous time series to the mode of the two distributions  $P_0^\mathcal{E}$  and  $P^\mathcal{E}$ . Note that, in practice, these predictions are computed through algorithms built with supervised learning techniques [35; 36], provided that a reasonably large training dataset is available. This is not the case for the context of our research work, where the scarcity of historical data for RECs challenges the building process of such forecasting algorithms.

## C.3 REC-2

REC-2 is composed of a consumer and a producer, which we denote as  $M1$  and  $M2$ , respectively; the first one is equipped with a load only, and the second one is equipped with loads and PV panels, with the latter always producing more than what the load consumes. The duration between two time steps  $\delta$  is set to 1 hour. A billing period occurs every 5 market periods. A market period lasts 4 time steps. In other words, optimal reallocation schemes are computed for the last 20 hours at each end of billing period. Figure 5 shows the consumption and production profiles of the

<sup>1</sup><https://atmos.washington.edu/~breth/classes/AM582/lect/lect8-notes.pdf>

Buying €/kWh			Selling €/kWh			Network €/kWh			
$M_1$	$M_2$	$M_3$	$M_1$	$M_2$	$M_3$	$P^-$	$P^+$	$\Lambda^-$	$\Lambda^+$
<b>0.20</b>	<b>0.22</b>	<b>0.24</b>	<b>0.04</b>	<b>0.05</b>	<b>0.06</b>	<b>1.00</b>	<b>1.00</b>	<b>0.02</b>	<b>0.03</b>

(a)

Market period	Net consumption kWh			Retail kWh			REC kWh		
	$M_1$	$M_2$	$M_3$	$M_1$	$M_2$	$M_3$	$M_1$	$M_2$	$M_3$
1	-642.66	644.85	748.11	0.00	644.85	105.44	-642.66	0.00	642.66
2	-666.00	142.05	-150.40	-523.94	0.00	-150.40	-142.05	142.05	0.00
3	232.98	-111.48	813.45	232.98	0.00	701.97	0.00	-111.48	111.48
4	-538.31	542.80	-579.49	0.00	0.00	-575.00	-538.31	542.80	-4.50

Billing period	Offtake peak kWh			Injection peak kWh			Global REC Bill €	
	$M_1$	$M_2$	$M_3$	$M_1$	$M_2$	$M_3$	NO-REC	REC
1	232.98	644.85	701.97	523.94	0.00	575.00	<b>3638.90</b>	<b>3068.45</b>

(b)

Market period	Net consumption kWh			Retail kWh			REC kWh		
	$M_1$	$M_2$	$M_3$	$M_1$	$M_2$	$M_3$	$M_1$	$M_2$	$M_3$
1	-642.66	644.85	748.11	0.00	2.18	748.11	-642.66	642.66	0.00
2	-666.00	142.005	-150.4	-523.94	0.00	-150.4	-142.05	142.05	0.00
3	232.98	-111.48	813.45	186.85	0.00	748.11	46.13	-111.48	65.34
4	-538.31	542.8	-579.49	-424.59	0.00	-150.4	-113.71	542.8	-429.09

Billing period	Offtake peak kWh			Injection peak kWh			Global REC Bill €	
	$M_1$	$M_2$	$M_3$	$M_1$	$M_2$	$M_3$	NO-REC	REC
1	186.85	2.18	748.11	523.94	0.00	150.4	<b>3638.9</b>	<b>2024.38</b>

(c)

Table 4: Optimal reallocation scheme over 4 market periods for a REC composed of 4 members. Retail prices and network costs are specified in (a). Energy exchanges, peaks and electricity bills (that results from neglecting the peak costs) are reported in (b). The sum of the individual electricity bills (before accounting for energy exchanges through the REC) is reported in red. The global REC bills (after accounting for energy exchanges through the REC) computed by neglecting and accounting for the peak costs are reported in purple and blue, respectively.

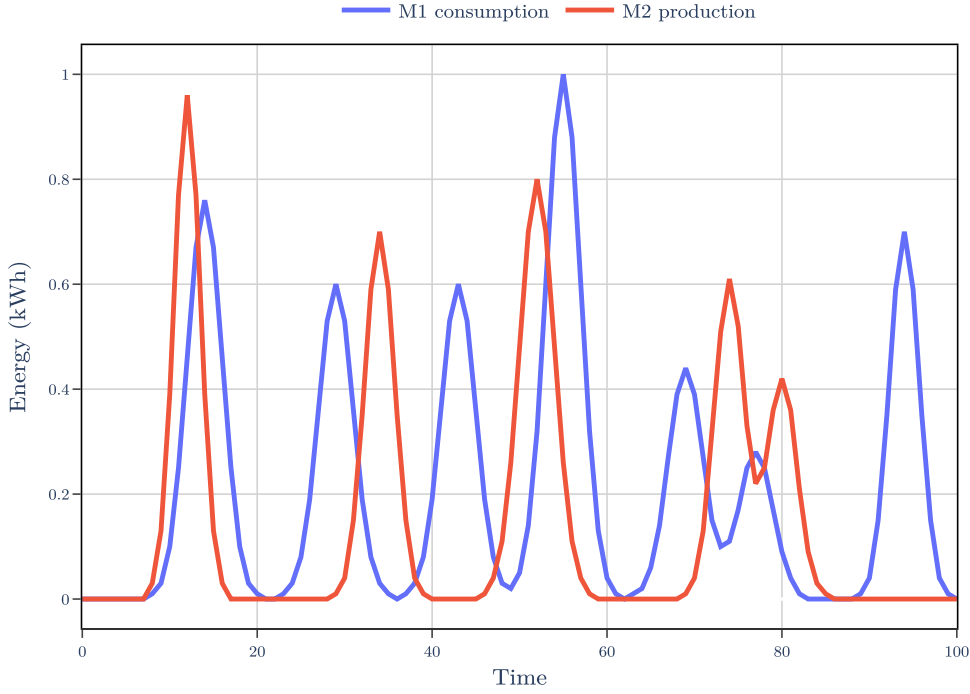


Figure 5: Consumption/production profiles of members in REC-2. Member M1 does not produce electricity and member M2 does not consume electricity (through their respective non-controllable assets).

REC Member	Buying €/kWh	Selling €/kWh
M1	0.10	0.01
M2	0.12	0.01

Table 5: Buying and selling prices of members in REC-2.

members. Table 5 shows the buying and selling prices as defined by retailer contracts in REC-2. Offtake and injection peak cost coefficients are both set to 1 €/kWh. Network fees are set to  $\Lambda^- = 0.03$  €/kWh and  $\Lambda^+ = 0.01$  €/kWh. The member M2 is equipped with a battery following linear charging power dynamics. More formally, let  $s_{M2}^c$  be the state of charge of the battery,  $u_{M2}$  be the charging power applied to the battery (negative is discharging),  $U^-$  and  $U^+$  be the respective discharging and charging powers bounds, and  $S_{M2}^c$  be its maximum capacity. The dynamics of the state of charge is defined as follows:

$$s_{M2,0}^c = \frac{S_{M2}^c}{2}, \quad (44)$$

$$s_{M2,t+1}^c = s_{M2,t}^c + \delta \left[ \nu^+ u_{M2,t}^+ - \frac{u_{M2,t}^-}{\nu^-} \right],$$

where  $u_{M2,t}^+$  and  $u_{M2,t}^-$  are respectively charging and discharging power of the battery, and  $\nu^+$  and  $\nu^-$  are respectively charging and discharging efficiencies. Table 6 shows the producer's battery specifications. The time horizon of the simulations is fixed to 101 time steps, with a discount factor of 0.9995.



Specification	Value	
Maximum capacity ( $S_{M2}^c$ )	1.00	kWh
Maximum charging power ( $U^+$ )	0.05	kW
Maximum discharging power ( $U^-$ )	0.10	kW
Charging efficiency ( $\nu^+$ )	1.00	
Discharging efficiency ( $\nu^-$ )	1.00	

Table 6: Specifications of the battery owned by member M2 in REC-2.

### C.3.1 Exogenous variables sampling

Exogenous variables are sampled from the consumption and production profiles of the REC members, as shown in Figure 5. To generate the red noise, we set the correlation parameter to  $r = 0.5$  and the standard deviation of the white noise to  $\sigma = 0.3$ .

### C.3.2 MPC and baseline policies

To evaluate the MPC and the baseline policies, we sample 64 exogenous time series through the procedure described in Appendix C.1, and we run the simulations with each of these time series independently. More precisely, they are evaluated by averaging the sample expected returns resulting from these simulations through 16 distinct random seeds. We run the MPC policies with the values of  $\alpha$  in  $\{1.0, 0.85, 0.5\}$  and the values of  $K \in \{1, 2, \dots, 101\}$ .

### C.3.3 RL policies

The RL policies transform an input data to its corresponding observation (as specified in Appendix A.4) as follows. The states of the controllable assets and the counters (as continuous values) are kept. For each member, the values of the consumption(production) meter readings and the net consumption(production) – before the controllable assets usage – are summed up for the current time step; the other values (of exogenous variables and meter readings) of the previous time steps are discarded. Afterwards, the consumption and production meter reading are replaced, for each member, by a net consumption meter reading (negative value is net production meter reading). Since standardisation of the observation (and the rewards) is recommended for training policies in a reinforcement learning setting [33], we proceed to that end as follows. Let  $O$  be an observation vector of size  $N > 1$ . Each value  $o_i$  of this observation vector, with  $i$  between 1 and  $N$ , is shifted by a fixed mean value and divided by a fixed standard deviation value. These statistics are computed by a simulation of the OPT policy (described in Section 5.2) through the historical exogenous variables. Note that, for the RL dense policies, we add in the observation the last reward computed during the current billing period. These rewards are standardised in the same fashion as the observations.

Thereafter, the observation is fed to the underlying deep neural network, for which the architecture is shown in Figure 6. The input hidden state is either the last output of the recurrent layers, or an initial vector of zero values (before the first step of an episode). The result of this forward computation is a pair of values corresponding to the parameters of a Gaussian distribution, namely the mean and the log standard deviation. Before sampling a value from this distribution, the standard deviation is transformed by the exponential function and shifted with a value  $\epsilon = 1e^{-6}$  (to avoid exploding gradients). The sampled value is then clipped between  $-1$  and  $1$  and projected into the bounds of the action space (in this case, between the maximum discharging and the maximum charging powers of the battery). If this action is not admissible (e.g., discharging power greater than the content of the battery), it is projected to the closest feasible value (e.g., discharging power value corresponding to the content of the battery).

The weights of the underlying neural network of the prior policy (before the first update of the PPO algorithm) are initialised accordingly to the programming library RLLib [31]. Additionally, as recommended by [33], we divide the weights of the last layer outputting the action distribution by a hyperparameter that we denote as  $W$ . Through the simulation process of policy functions, the training set records the transitions along with the respective parameters of the (Gaussian) distribution that sampled the actions and the critic values. It also records the standardised values of the rewards, which are used instead of the original reward signals to compute the loss function of the policy. The policy is then updated as explained in Appendix A.2. Table 7 shows the hyperparameters (including the grid search space) used to train the RL policies.

To evaluate the RL policies, we have run the PPO algorithm for 600 iterations through 16 distinct random seeds. For each iteration, the current policy is run through 64 simulations in the environment (by sampling 64 exogenous time series). At the end of this procedure, 64 additional simulations are sampled, and the expected return estimate is

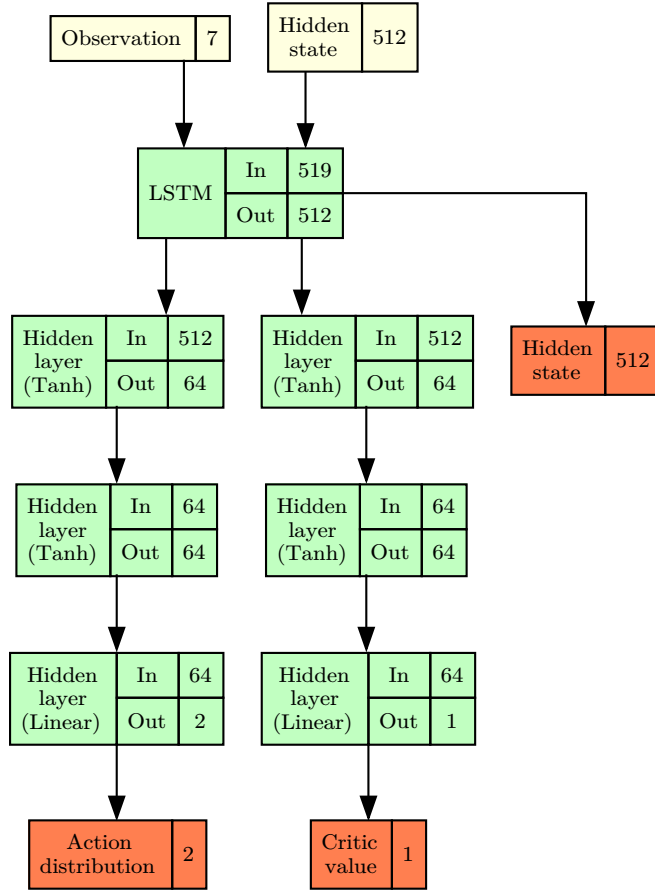


Figure 6: Forward view of the underlying deep recurrent learning model of RL policies for REC-2 for a single observation. Yellow nodes are the inputs (including the RNN hidden states). Green nodes are hidden layers with activation functions. Orange nodes are output (including the next RNN hidden states).

computed by averaging the expected return samples across the random seeds. Figure 7 shows the evaluation of the RL policies at each iteration of the PPO algorithm.

#### C.4 REC-7

From now on, as the experimental settings share many similarities with those in REC-2, we only describe the parts that differ from the latter. REC-7 is composed of 7 members, which are denoted from  $M1$  to  $M7$ . The duration between two time steps  $\delta$  is set to 3 minutes. A billing period occurs every 45 market periods. Figure 8 shows the consumption and production profiles of the members, derived from historical data of an existing REC in Wallonia, Belgium. Table 8 shows the buying and selling prices of members as defined by their retailer contracts. Offtake and injection peak cost coefficients are both set to 1.210 €/kWh. Network fees are set to  $\Lambda^- = 0.143$  €/kWh and  $\Lambda^+ = 0.126$  €/kWh. Similarly to REC-2, only the member  $M1$  is equipped with a battery following the same charging dynamics as described by Equation (44). Table 9 shows the producer’s battery specifications (by reusing the notation introduced in Section C.3). The time horizon of the simulations is fixed to 720 time steps, with a discount factor of 0.99993.

##### C.4.1 MPC and baseline policies

To evaluate the MPC and the baseline policies, we sample 64 exogenous time series through the procedure described in Appendix C.1, and we run the simulations with each of these time series independently. More pre-

Hyperparameter	Search space	RL policy	RL dense	RL retail	RL retail dense
$\lambda_{\text{GAE}}$	{0.90, 0.95, 0.99}	0.90	0.90	0.90	0.90
$\eta$	{ $5e-4$ , $5e-5$ , $5e-6$ }	$5e-5$	$5e-5$	$5e-5$	$5e-5$
$\gamma$	{0.9995, 0.95, 0.99}	0.99	0.99	0.99	0.99
$\beta$	{1, 2, 4}	2	2	1	1
$N_{\text{upd}}$	{5, 10}	10	10	10	10
$B$	{32, 64, 96}	64	64	64	64
$\lambda_{v_\phi}$	{1, 0.1, 0.01, 0.001}	1	0.01	1	1
$T_{\text{min}}$	{25, 50, 100}	50	50	50	50
$W$	{1, 10, 100}	1	1	1	1

Table 7: Values of hyperparameters related to PPO for RL policies in REC-2, with the search space in the second column; see Appendix A.2 for more details about the hyperparameters. Unspecified hyperparameters have been left to default values defined by the programming library RLLib [31].

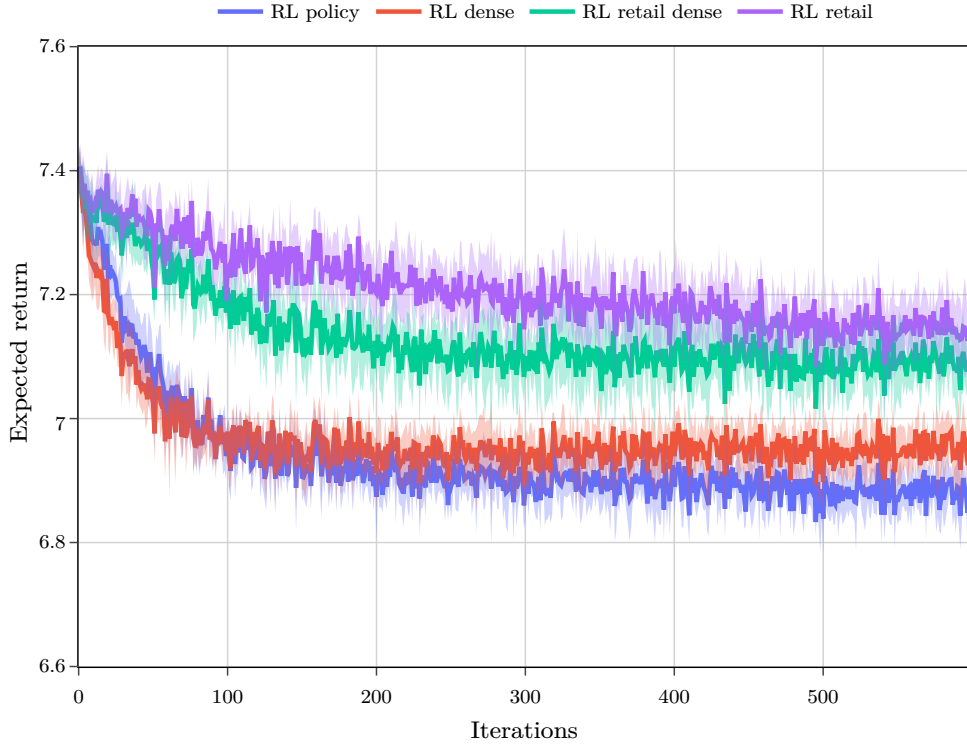


Figure 7: Mean expected return with standard error over training iterations of the RL policies in REC-2.

REC Member	Buying €/kWh	Selling €/kWh
M1	0.214907	0.075388
M2	0.208757	0.075152
M3	0.202735	0.076381
M4	0.20846	0.077213
M5	0.20846	0.078153
M6	0.206301	0.080649
M7	0.210234	0.081928

Table 8: Retail buying and selling prices for REC-7.

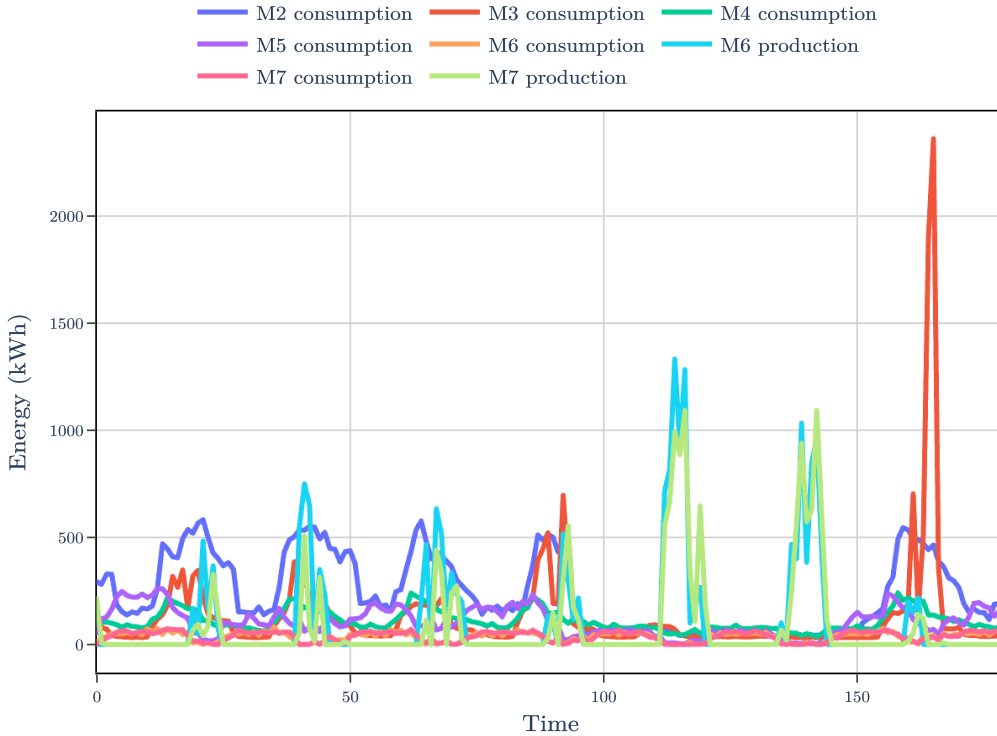


Figure 8: Consumption and production profiles for REC-7 (resampled for ease of readability). The members M2, M3, M4 and M5 do not produce electricity. The members M6 and M7 both consume and produce electricity. Electricity flows of the member M1 are only generated from its battery.

Specification	Value
Maximum capacity ( $S_{M2}^c$ )	5256 kWh
Maximum charge power ( $U^+$ )	525 kW
Maximum discharge power ( $U^-$ )	1051 kW
Charging efficiency ( $\nu^+$ )	0.88
Discharging efficiency ( $\nu^-$ )	0.71

Table 9: Specifications of the battery of member M1 in REC-7.

cisely, they are evaluated by averaging the sample expected returns resulting from these simulations through 16 distinct random seeds. We run the MPC policies with the values of  $\alpha$  in  $\{1.0, 0.95, 0.5\}$  and the values of  $K$  in  $\{1, \dots, 100, 132, 164, 196, 266, 330, 394, 458, 522, 586, 650, 721\}$ .

#### C.4.2 RL policies

Figure 9 shows the architectures of two independent deep neural networks. The first one computes the parameters of Gaussian distributions (to sample the next actions) and the second one is the critic value. Table 10 shows the hyperparameters (including the grid search space) used to train the RL policies. To evaluate the RL policies, we have run the PPO algorithm for 1000 iterations through 16 distinct random seeds. For each iteration, the current policy is run through 128 simulations in the environment. At the end of this procedure, 128 additional simulations are sampled to estimate the expected return. Figure 10 shows the evaluation of the RL policies at each iteration of the PPO algorithm.

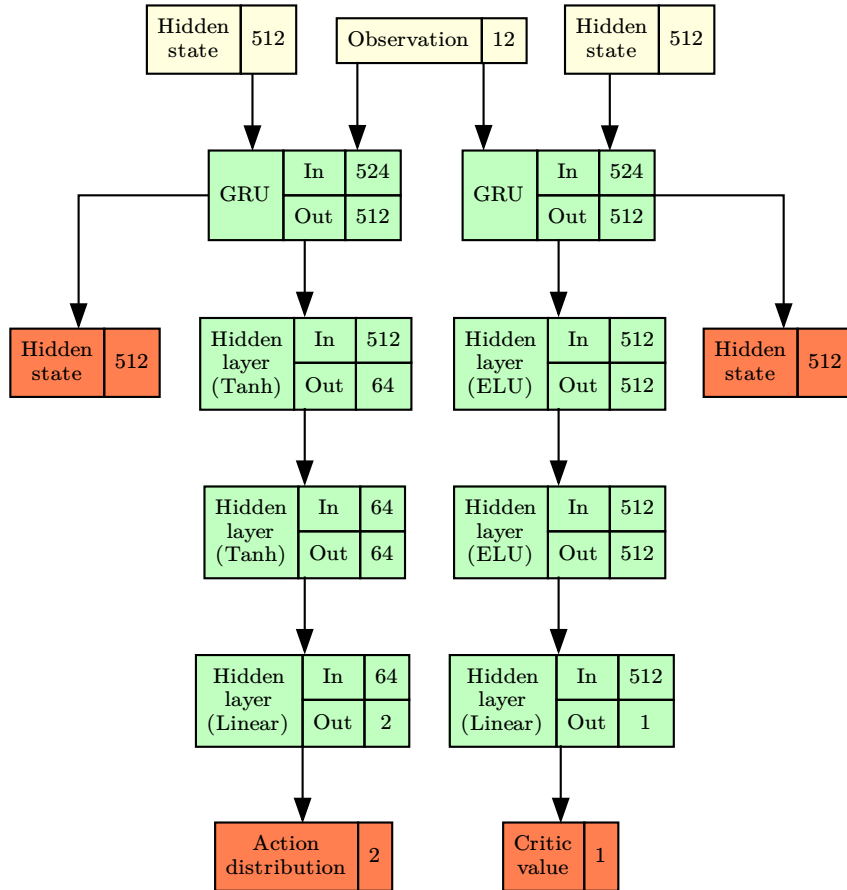


Figure 9: Forward view of underlying deep recurrent learning model of RL policies for REC-7 for a single observation. Yellow nodes are the inputs (including the RNN hidden states). Green nodes are hidden layers with activation functions. Orange nodes are output (including the next RNN hidden states).

Hyperparameter	Search space	RL policy	RL dense	RL retail	RL retail dense
$\lambda_{\text{GAE}}$	{0.9, 0.95, 0.99}	0.9	0.9	0.9	0.9
$\eta$	{ $5e-4$ , $5e-5$ , $5e-6$ }	$5e-5$	$5e-5$	$5e-5$	$5e-5$
$\gamma$	{0.99993, 0.95, 0.99}	0.99	0.99	0.99	0.99
$\beta$	{1, 2, 4}	2	2	1	1
$N_{\text{upd}}$	{5, 10}	5	5	5	5
$B$	{360, 720}	360	360	360	360
$\lambda_{v_\phi}$	{1, 0.1, 0.01, 0.001}	0.001	0.001	0.001	0.001
$T_{\text{rnn}}$	{180, 360}	360	360	360	360
$W$	{1, 10, 100}	100	100	100	100

Table 10: Values of hyperparameters related to PPO for RL policies in REC-7, with the search space in the second column; see Appendix A.2 for more details about the hyperparameters. Unspecified hyperparameters have been left to default values defined by the programming library RLlib [31].

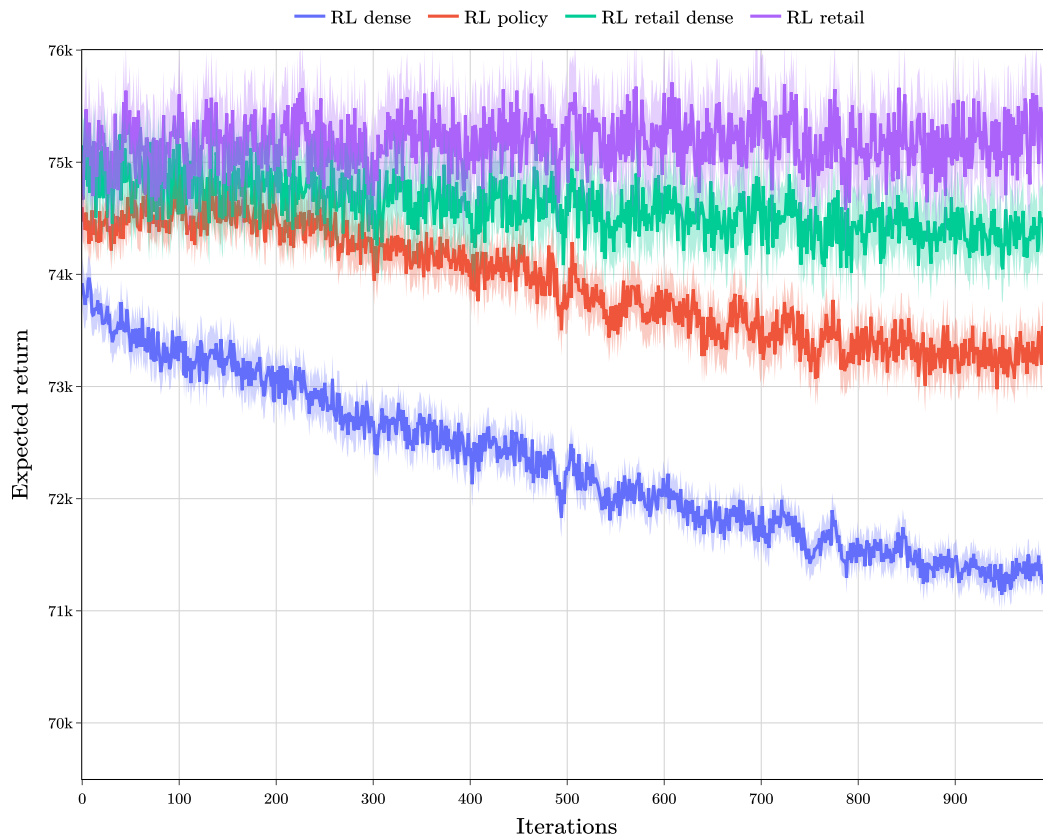


Figure 10: Mean expected return with standard error over training iterations of the RL policies in REC-7.