Reinforcement Learning for Efficient Design and Control Co-optimisation of Energy Systems

Marine Cauz [1 2] **Adrien Bolland** [3] **Christophe Ballif** [1] **Nicolas Wyrsch** [1]

Abstract

The ongoing energy transition drives the development of decentralised renewable energy sources, which are heterogeneous and weather-dependent, complicating their integration into energy systems. This study tackles this issue by introducing a novel reinforcement learning (RL) framework tailored for the co-optimisation of design and control in energy systems. Traditionally, the integration of renewable sources in the energy sector has relied on complex mathematical modelling and sequential processes. By leveraging RL's model-free capabilities, the framework eliminates the need for explicit system modelling. By optimising both control and design policies jointly, the framework enhances the integration of renewable sources and improves system efficiency. This contribution paves the way for advanced RL applications in energy management, leading to more efficient and effective use of renewable energy sources.

1. Introduction

1.1. Background and motivation

Energy systems are undergoing significant transformations to meet increasing demands for sustainability and energy efficiency, particularly through the integration of decentralised and intermittent renewable energy sources. Traditionally, these systems are developed in two distinct phases: design, which determines the optimal size of components, and control, which focuses on their optimal operation. This sequential approach, as highlighted by (Dranka et al., 2021), often

leads to inefficiencies and missed opportunities for optimal performance. To address the increasing complexity driven by renewable integration, co-optimisation has emerged as a key approach, jointly handling design and control to enhance system reliability and affordability. Recent literature underscores the importance of co-optimising design and operation using techniques such as linear programming (Krishnan et al., 2016; Daadaa et al., 2021; Jayadev et al., 2020), stochastic models (Clack et al., 2015; Qiu et al., 2017), robust optimisation (Popovici & Winston, 2015; Khojasteh, 2020), and evolutionary algorithms (Li et al., 2018; Gjorgiev & Sansavini, 2018; Bao et al., 2019). Among these methods, Mixed-Integer Linear Programming (MILP) is the most commonly used but requires mathematical modelling of the system and its interactions. These methods aim to optimise performance comprehensively while addressing uncertainties and multi-objective challenges. Overall, these diverse approaches highlight both the technical challenges and the critical importance of co-optimisation in enhancing the efficiency and sustainability of energy systems (Sachio et al., 2022; Fazlollahi & Maréchal, 2013; Dranka et al., 2021).

Data-driven methods, such as reinforcement learning (RL), have shown significant potential in computing control policies across various applications, including energy, offering a promising alternative to traditional approaches (François-Lavet et al., 2018; Quest et al., 2022; Perera et al., 2020). However, standard RL methods typically focus solely on operational control without integrating system design, limiting insights into how design changes influence outcomes. Despite its potential, RL is not fully exploited in the energy field (Perera & Kamalaruban, 2021). Recent advancements in RL, particularly gradient-based optimisation techniques like actor-critic methods, facilitate learning control policies for complex problems, opening new opportunities.

Building on these advancements, researchers have proposed algorithms to efficiently tackle joint design and control challenges. In (Schaff et al., 2019), the authors introduced an RL framework that optimises both design and control by maintaining a distribution over designs, using the Proximal Policy Optimization (PPO) algorithm (Schulman et al., 2017) for policy training and the REINFORCE update rule for design adjustments (Williams, 1992). This approach has

---

[1]Photovoltaics and Thin-Film Electronics Laboratory (PV-lab), Institute of Electrical and Micro Engineering (IEM), École Polytechnique Fédérale de Lausanne (EPFL), Neuchâtel, Switzerland [2]Planair SA, Yverdon-les-Bains, Switzerland [3]Department of Electrical Engineering and Computer Science, University of Liège, Liège, Belgium. Correspondence to: Marine Cauz <marine.cauz@epfl.ch>.

been successfully applied in various robotic environments, outperforming other techniques (Bhatia et al., 2022; Ha, 2019). Alternatively, (Luck et al., 2020) enhances adaptability for joint design and control using Soft Actor-Critic (SAC) (Haarnoja et al., 2018), despite involving complex optimisation problems. The algorithm from (Bolland et al., 2022) refines this approach by combining policy gradients with model-based optimisation, It was applied to systems with photovoltaic (PV) panels and battery (Cauz et al., 2023), though it faces limitations due to finite time horizons and on-policy nature. Other approaches (Chen et al., 2020; Jackson et al., 2021) focus on learning system parameters directly, assuming the system dynamics are parameterised, but are restrictive when modelling complex energy systems where design decisions are directly related to explicit costs or rewards.

1.2. Contribution

Capitalising on these recent developments in policy gradient techniques, this study advances an integrated RL framework specifically tailored to address the co-optimisation challenges within energy systems. As introduced by (Schaff et al., 2019), the proposed framework employs a parametric design distribution, whose parametric nature is effective for modelling distributions over continuous supports and allows for using gradient based methods easily. This approach contrasts with most of the previous methods that employ a deterministic representation of the design variable (Chen et al., 2020; Jackson et al., 2021; Bolland et al., 2022), which can make model-free optimisation and efficient exploration challenging. Additionally, this framework distinguishes from (Schaff et al., 2019) by incorporating entropy regularisation, as in (Haarnoja et al., 2018), into the optimisation process to prevent convergence to local optima. Furthermore, this framework relies on a deterministic policy parameterisation, which is optimised using an off-policy actor-critic algorithm, namely Deep Deterministic Policy Gradient (DDPG) (Lillicrap et al., 2019). This allows for accommodating infinite time horizons, addressing a significant gap in methodologies (Bolland et al., 2022; Cauz et al., 2023). Unlike most existing studies, including (Schaff et al., 2019), the control policy training is off-policy, thereby enhancing sample efficiency by learning from a diverse range of past experiences stored in a replay buffer. Finally, this framework is also model-free, eliminating the need for a predefined mathematical model of the system, which simplifies implementation and broadens its applicability. None of the previously cited methods combine all these features.

By integrating these capabilities, this approach maximises the potential of RL to address the co-optimisation of design and operation within energy systems, a challenge often overlooked in RL research. This integrated framework bridges the gap between theoretical RL research and its practical application in energy systems, establishing a new benchmark for employing RL to tackle co-optimisation challenges in the energy sector.

The paper is structured as follows: Section 2 details the proposed RL method, covering both control and design aspects. Section 3 describes the energy system and experimental setup. Section 4 presents the findings, with Section 5 discussing their implications and potential impact. Finally, Section 6 summarizes the key insights and contributions of the research.

2. Method

This section outlines the conventional RL approach for system control and then details the adaptations made to enable learning system designs.

2.1. Control Policy

Formally, RL is conceptualised as an interplay between an agent and an environment. This environment is mathematically formalised as a Markov Decision Process (MDP) (Bellman, 1957), which is defined by its model $\mathcal{M} = (\mathcal{S}, \mathcal{A}, T, R, p_0, \gamma)$, where $\mathcal{S}$ denotes the state space, $\mathcal{A}$ denotes the action space, $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ denotes the transition function (i.e., $T(s_{t+1}|s_t, a_t)$ denotes the probability of reaching a state s_{t+1} when taking an action a_t from state s_t), $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ denotes the reward function (i.e., $R(s_t, a_t)$ is the immediate reward received by taking action a_t from state s_t), $p_0 : \mathcal{S} \to [0, 1]$ denotes the initial distribution, $\gamma \in [0, 1)$ denotes the discount factor (i.e., γ models the importance of future rewards, with a lower value placing more emphasis on immediate rewards). Within the MDP framework, the agent's objective is to find a policy, $\pi \in \Pi$, namely a conditional distribution over actions that can be used to take actions in each state by sampling. The optimal policy denoted π^* maximises the cumulative reward, called expected return of the policy: $\mathbb{E}[\mathcal{R}_t]$, such as $\mathcal{R}_t = \sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)$.

Actor-critic algorithms combine policy gradient and value-based methods for efficient policy learning and evaluation (François-Lavet et al., 2018). The actor proposes actions based on a policy π modelled by a neural network with parameters θ, while the critic evaluates these actions by estimating value functions. This mechanism allows for ongoing refinement of the policy based on the critic's feedback and updating the critic as the policy changes. Among the various actor-critic implementations, DDPG (Silver et al., 2014; Lillicrap et al., 2019) stands out due to its off-policy nature, meaning the policy can be improved using trajectories where actions are taken from another policy, and is suitable for environments with continuous action spaces. The critic approximates the state-action value function $Q^\theta(s, a)$, aiding

in policy update gradients. To ensure stable learning, DDPG employs target networks for temporal-difference learning benchmarks and adds Gaussian noise to policy outputs for sufficient exploration.

2.2. Design Policy

Conventional RL typically focuses on optimising a control policy π_θ^* for a fixed system design. Building on this primary objective, this study explores both the design space X and control strategies to identify an optimal system design x^* and its corresponding control policy $\pi_\theta^*(a_t|s_t, x^*)$. To each design $x \in X$ corresponds a different MDP, as defined in Subsection 2.1. The objective is to maximise the expected return over a design distribution, effectively co-optimising design and control to enhance overall system performance. The proposed RL framework extends the traditional control policy optimisation by incorporating a probability distribution $p_\phi(x)$ over potential designs $x \in X$. The learnable parameters ϕ represent the parameters of this design distribution. The ultimate goal is to find the optimal parameters ϕ^* and θ^* that jointly maximise the expected discounted reward:

$$\phi^*, \theta^* = \arg\max_{\phi,\theta} \mathbb{E}_{x \sim p_\phi(\cdot)} \left[\mathbb{E}_{\substack{s_0 \sim p_0(\cdot) \\ a_t \sim \pi_\theta(\cdot|s_t,x) \\ s_{t+1} \sim T(\cdot|s_t,a_t)}} [\mathcal{R}_t] \right] \quad (1)$$

The co-optimisation framework is designed to maximise the expected discounted reward by effectively integrating system design and control. It is compatible with any standard RL algorithm, however, this implementation specifically uses the DDPG algorithm. This algorithm adapts the control policy π_θ to maximise expected returns across a range of designs drawn from the design probability distribution p_ϕ. Each training iteration consists of two concurrent processes:

- The control policy π_θ is refined using gradient ascent to enhance reward expectations over the sampled designs.

- Simultaneously, the design distribution p_ϕ is updated to increase the likelihood of designs that yield higher performance under the current policy.

Algorithm 1 describes the co-optimisation procedure. It starts with the initialisation of the design distribution, fostering a wide-ranging exploration of designs. During training, the framework adjusts the policy parameters θ and the design parameters ϕ to gradually phase out less effective designs, allowing the policy to specialise and focus on a narrowing set of promising designs. As a result, the variance within the design distribution p_ϕ decreases, guiding the system towards the convergence on an optimal design

---

Algorithm 1 Co-optimisation of design and control
___
Initialise actor $\pi_\theta(s, x)$ and critic $Q_\theta(s, a, x)$
Initialise target networks and replay buffer (capacity N)
Initialise design distribution $p_\phi(x)$
repeat
  Sample designs $\{x_1, \ldots, x_d\}$ from $p_\phi(x)$
  Compute expected return R_i for each x_i with DDPG
  Update critic by minimising the loss:
  $L(\theta^Q) = \frac{1}{N} \sum_{n=0}^{N-1} (y_n - Q_\theta(s_n, a_n, x_n))^2$
  Update actor by one step of gradient descent:
    $\nabla_{\theta^\pi} J(\theta^Q, \theta^\pi) \approx$
  $\frac{1}{N} \sum_{n=0}^{N-1} \nabla_{\theta^\pi} Q_\theta(s_n, \pi_\theta(s_n, x_n), x_n)$
  Update target networks
  Compute the loss function for the design update:
  $\mathcal{L}(\phi) = -\frac{1}{d} \sum_{i=1}^{d} (\log p_\phi(x_i) \cdot R_{i,t} - \lambda \cdot \log p_\phi(x_i))$
  Update p_ϕ by minimising the loss with respect to ϕ
until End of training

---

x^* and associated policy π_θ^*, thereby maximising the overall system performance.

In comparison with the framework proposed by (Schaff et al., 2019), two notable modifications in the design distribution enhance its suitability for energy systems. Firstly, instead of using a Gaussian Mixture Model to parameterise the design distribution p_ϕ, which may require clipping to ensure physical feasibility, this framework employs a log-normal mixture model. This model inherently restricts the design space to $X = \mathbb{R}^+$, ensuring all design values remain within physically feasible limits for energy systems. The mixture model parameters, including the mean and variance of each log-normal component and their respective (unscaled) weights, are updated using stochastic gradient ascent based on the REINFORCE gradient estimates (Williams, 1992). The second modification introduces entropy regularisation to the design distribution to mitigate the risk of local optima, a common challenge in energy system optimisations as noted in (Cauz et al., 2023). Initially, the design distribution is set with random means and high variance to encourage diverse explorations. Additionally, an entropy term is added in the loss function (Ahmed et al., 2019), which gradually decreases to strategically reduce exploration over time. There is no straightforward computation of the entropy of a log-normal mixture model, hence the entropy is estimated by extending the return with the log probability of the design samples, bypassing the need for computationally intensive methods.

3. Experiments

This section describes the experimental set up to evaluate the proposed framework on a building-scale PV-battery system. The aim is to minimise total electricity costs by optimising

both the investment in system components, namely the design parameters, and the operational strategies for storage management, meaning the control policy. Operational costs are derived from grid interactions required to meet building energy demands. Performance is measured against the average expected return of the system's total cost, reflecting the economic impact of chosen design and control strategies. For comparative analysis, the RL co-optimisation is benchmarked against traditional approaches highlighted in Section 1. First, it is compared to a MILP approach for selecting the best design, followed by an RL technique for determining the optimal policy for this design. Second, it is compared to expert rule-based controllers.

3.1. Building-Scale System

The system is a building-scale energy system within an office setting, equipped with a PV installation and a stationary lithium-ion battery to satisfy its electricity requirements. The system also features a bidirectional EV (Electric Vehicle) charging point, whose usage is stochastically modelled based on typical patterns. Moreover, the building is connected to the electrical grid, subject to dynamically varying electricity prices. The main objective is to determine the optimal design for the PV installation and the battery capacity, while simultaneously developing an optimal control policy for battery and EV management. This aims to minimise the total cost of ownership, encompassing both capital and operational expenses, as well as grid costs. The design and control model of this system is formulated using an MDP, which is described in detailed in Appendix A.

The model is trained on a historical one-year dataset of normalised PV production and electrical consumption, divided into training and validation sets to capture seasonal fluctuations. This dataset is supplemented with synthetic data for dynamic grid tariffs and EV arrival times. Details of both datasets are provided in Appendix A. The MDP's time horizon is truncated after $T = 168$ hours (one week), with long-term dependencies captured via bootstrapping in the critic training. Ideally, the time horizon would cover an entire year or the system's lifecycle to capture seasonal production and consumption variations and potential equipment degradation. Performance is regularly evaluated in two ways, (i) across the full training dataset, corresponding to $T = 8088$ hours, to assess long-term effectiveness, and (ii) across the full validation dataset, corresponding to $T = 672$ hours, to avoid overfitting.

3.1.1. EXPERIMENT SETUP

The actor π_θ and critic Q_θ are both implemented using neural networks, each consisting of two hidden layers with 256 neurons and ReLU activation functions. For the actor network, a tanh activation function is applied in the final layer to map the output to the action space. The critic network concatenates the state and action at the input layer, with a linear activation function in the output layer. To facilitate integration with existing RL libraries, the design parameters x are appended to the state variables before they are inputted into the control network.

The design distribution, $p_\phi(x)$, is modelled as a log-normal mixture with three components, each parameterised with two design parameters. The means, variances and weights of each component are initialised randomly within the interval $[0, 1[$, set high and uniformly distributed, respectively, to ensure the distribution covers a large range of $\mathbb{R}^+$. The entropy weight linearly decreases throughout training and reaches zero during the last half of iterations.

The framework employs the DDPG algorithm to train the control policy π_θ. Each iteration consists of a batch of 32 episodes, each lasting $T = 168$ hours, i.e., one week. Additionally, at every iterations, a set of design parameters is sampled and evaluated with the current policy across the full training dataset to monitor performance over a duration close to one year, i.e., $T = 8088$ hours. To prevent gradient explosion during training, gradient clipping is implemented. Moreover, the performance are evaluated every iterations still with a batch of 32 episodes across the full validation dataset, i.e., $T = 672$, with the current design distribution and control policy. Finally, the medians and quartiles of the design parameter distribution are computed at the end of training, after 500 iterations. To ensure reliability and account for variability in initialisation, all experiments are conducted using 30 different seed values ranging from 0 to 30.

3.1.2. RULE-BASED BASELINE

A rule-based baseline is established as a fixed control policy, focusing solely on optimising the design. This setup allows for a direct comparison between joint optimisation using DDPG and simple design optimisation under a given expert policy. The rule-based discharges the stationary battery when consumption exceeds PV production and charges it when production is higher than consumption. The bidirectional EV's battery, when available, follows the same logic to augment the system's capacity. This rule-based controller operates within the same MDP environment but does not require a training phase, as it involves no trainable control parameters θ. Performance evaluations of the system's design under this controller are conducted over 500 iterations, focusing exclusively on updating the design parameters ϕ, given that the control policy is static and predetermined. This experiment is referred to as the **design-only scenario**, as only the design parameters are trained, without assuming perfect foresight.

3.1.3. MILP BASELINE

MILP is the most widely used tools for designing energy systems (Dranka et al., 2021; Perera & Kamalaruban, 2021). This approach requires mathematical modelling of the system and its interactions, assuming a perfect foresight approach. In this study, an environment formulated as a mathematical program consisting of constraints and objectives is developed, similar to the MDP presented in Appendix A. This formulation allows for computing the optimal design, which is then controlled by a policy learned using DDPG for this particular design. This methodology enables benchmarking the proposed co-optimisation framework against a two-step baseline where the design is initially computed using MILP over the full training dataset, and subsequently controlled with DDPG. This experiment is referred to as the **best two-step scenario** because it involves a co-optimisation using the best response two-step algorithm.

Moreover, a final scenario, referred to as the **fixed scenario**, is computed based on the rule-based control. The rule-based policy is implemented as constraints in the MILP (actions are constrained based on the state), allowing for the computation of the optimal design for this fixed control policy using MILP. This experiment provides a static performance, assuming fixed design and control. Since there are no training parameters, this scenario is computed to verify that the solutions provided exceed this fixed baseline.

4. Results

This section details the performance of the proposed framework in co-optimising the design and operation of a building-scale PV-battery system.

4.1. Training Dynamics

Figure 1 tracks the performances over 500 iterations during training of (i) the co-optimisation using DDPG (blue), (ii) the best two-step optimisation using DDPG for control with a fixed design derived from the MILP (orange), (iii) the design-only scenario using a rule-based control policy while optimising the design distribution (green), and (iv) the fixed scenario corresponding to the solution provided first by the MILP design computed with the rule-based constraints and then by applying the rule-based control policy (black). For all scenarios, 30 experiments are conducted with different seeds ranging from 0 to 30. Figure 1 reports the median and quartiles of the return during each learning procedure. In the best two-step scenario, the design parameters resulting from the MILP are fixed at 6 kWp for PV and 14 kWh for battery capacity. In the fixed scenario, the design parameters resulting from the MILP are fixed at 3 kWp for PV and 14 kWh for battery capacity, indicating that the integration of the rule-based constraint within the MILP constraints
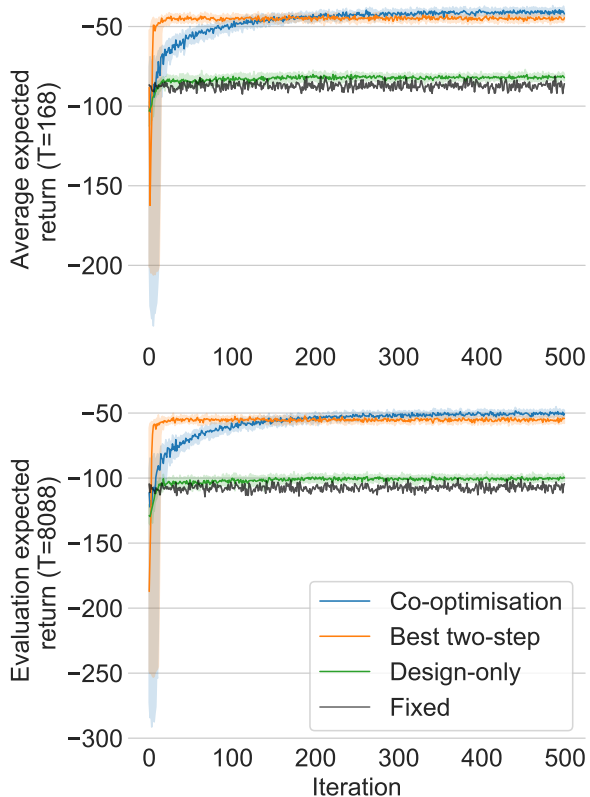

Figure 1. Training performances, over 500 iterations, for the co-optimisation (blue), best two-step (orange), and design-only (green) scenarios. Experiments were conducted using seed values ranging from 0 to 30, with the figure showing the median and quartiles. The top subplot illustrates the evolution of average expected returns on T=168, i.e., the effective training. The bottom subplot assesses the average expected return throughout the full training dataset on T=8088, i.e., the long-term performance.

reduces the optimal PV power by half.

The top subplot of Figure 1 illustrates the weekly average expected returns, computed from designs sampled from the current design distribution in the co-optimisation and design-only scenarios, across batches of 32 episodes, each lasting $T = 168$ hours. For these two scenarios, design parameters ϕ are updated during training and then the weekly average expected return stabilises by 500 iterations at mean values of -41.7 and -82.9 at the last iteration, respectively, as reported in Table 1. In the best two-step scenario, due to its static design, training converges faster, with results stabilising around -44.8. In the fixed scenario, since the design was previously computed using MILP and the control policy is predefined, there is no further optimisation, and it converges to -85.9. The variations are linked to the

samples in the initial states. The bottom subplot of Figure 1 evaluates long-term performance over a batch of 32 episodes, each with a duration of the entire training dataset, $T = 8088$ hours. The difference in performance between the co-optimisation and the best two-step scenario remains similar, with results converging to -49.8 and -54.8, respectively. In the design-only and fixed scenarios, the results slightly decrease compared to the weekly results, converging to -101.1 and -104.3, respectively. This assessment confirms that the co-optimisation maintains performance over extended operational periods, which is essential for the infinite horizon characteristic of energy systems.

Table 1. Average expected returns and standard deviation at the last iteration over the 30 seed experiments for training, long-term performance, and validation in the co-optimisation, best two-step, design-only, and fixed scenarios.

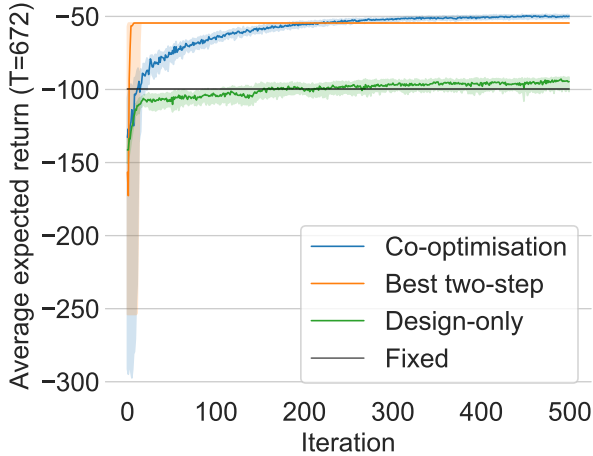| Scenario	Training T=168	Long-term T=8088	Validation T=672
Co-optim.	**-41.7$\pm$3.2**	**-49.8$\pm$4.9**	**-50.1$\pm$2.1**
Best 2-step	-44.8$\pm$4.4	-54.8$\pm$4.1	-54.5$\pm$0.0
Design-only	-82.9$\pm$6.5	-101.1$\pm$7.9	-97.1$\pm$7.4
Fixed	-85.9$\pm$0.0	-104.3$\pm$0.0	-99.6$\pm$0.0

4.2. Evaluation Process

Figure 2. Validation performances, over 500 iterations, for the co-optimisation (blue), best two-step (orange), design-only (green), and fixed (black) scenarios. Experiments were conducted using seed values ranging from 0 to 30. The figure shows the median and quartiles of the average expected return computed over the entire validation dataset, i.e., T=672.

The validation process, illustrated in Figure 2, involves com-

puting the average expected return of the current control policy and designs sampled from the current design distribution every iteration over the full validation dataset, i.e., $T = 672$ hours. The validation performance of the co-optimisation scenario (blue) is benchmarked against the best two-step scenario (orange), the design-only scenario (green), and the fixed scenario (black). The experiments are conducted using 30 different seed values, with the median and quartiles reported in Figure 2. All scenarios quickly converge to a unique solution for this specific validation episode over $T = 672$ hours. Interestingly, the difference in performance between the co-optimisation and best two-step scenarios is greater than during the training process. This might result from the perfect foresight approach in the MILP parameter selection, which allows for selection based on future information during learning that is unknown at the time of evaluation. As reported in Table 1, the scenarios converge to the following average values at the last iteration: -50.1 for co-optimisation, -54.5 for best two-step, -97.1 for design-only, and -99.6 for the fixed scenario.

Figure 3 represents the final distribution (estimated with 1000 samples) after 500 training iterations of one of the 30 seed experiments, for both the co-optimisation scenario and the design-only scenario. The median and quartiles are used to highlight the narrow confidence interval of the parameter distribution within the design space $X = \mathbb{R}^+$. This illustration effectively shows that both scenarios converge to similar optimal design parameter intervals. For co-optimisation with DDPG (blue), the interval between the first and third quartile is $[3, 6]$ kWp for PV and $[5, 10]$ kWh for battery capacity. For the design-only scenario with the rule-based control policy (green), the interval between the first and third quartile is $[2, 5]$ kWp for PV and $[0, 1]$ kWh for battery capacity. The mean design parameter over all 30 scenarios is reported in Table 2. Note that these design parameter values are consistent with the assumptions of the building-scale system environment. Additionally, they differ from those computed using MILP, which in the best two-step scenario are equivalent to 6 kWp for PV and 14 kWh for battery capacity, reflecting different optimisation dynamics.

Table 2. Mean of the design distribution at the last iteration, averaged over the 30 seed experiments in the co-optimisation and design-only scenarios. For the best two-step and fixed scenarios, the reported values are those computed using MILP.

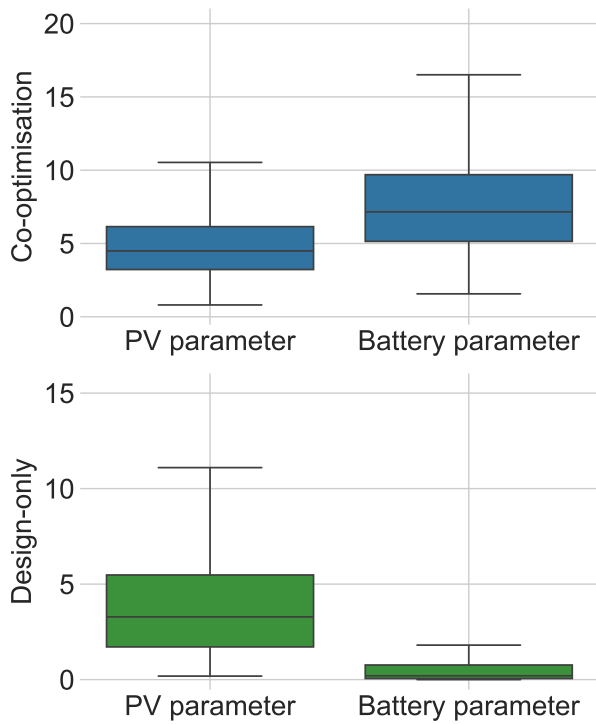| Design parameter	PV	Battery
Co-optim.	**6.6$\pm$1.4**	**4.5$\pm$0.9**
Best 2-step	6	14
Design-only	6.3$\pm$5.3	3.5$\pm$12.2
Fixed	3	14

Figure 3. Design parameter distribution after training for the co-optimisation (top, blue) and design-only (bottom, green) scenarios. The boxplots are computed based on a sample of 1000 designs drawn from the final design distribution of one of the 30 seed experiments.

5. Discussion

This study investigates the co-optimisation of design and operation in energy systems using a novel RL framework. The primary goal was to assess the feasibility and effectiveness of RL in developing integrated design strategies within a co-optimisation framework, aiming to enhance system performance by minimising total electricity costs. The results confirm that the framework successfully converges to high-performing design parameters while achieving superior control performance in both short and long-term periods. The co-optimisation scenario outperforms both the design-only and best two-step scenarios in training and validation performances, while converging to different design parameter values.

First, the two RL-based design optimisations, i.e., the co-optimisation and design-only scenarios, converged to different design parameters while achieving significantly different operational performances, underscoring the significance of co-optimisation. The convergence to optimal design parameters in both scenarios is evidenced by the narrowing of the boxplot charts, indicating a non-dispersed solution.

The optimised design parameters, although modest, align realistically with the environmental conditions and model assumptions. Additionally, they should be considered in relation to electricity requirements (i.e., 2.5 kWh on average per hour). These results highlight the framework's capacity to provide precise solutions.

The choice of DDPG among actor-critic algorithms was motivated by its off-policy nature and suitability for environments with continuous action spaces. These characteristics make DDPG significantly more sample-efficient, an advantage in the energy sector where system designs are typically based on data from a single year or, at best, a few years. The fast convergence of the control parameters further confirms the suitability of this algorithm for energy applications.

The main limitation of this proposed framework is the difficulty in guaranteeing an optimal design, in contrast to the one computed using MILP. The results can vary due to sensitivity to hyperparameters, necessitating a detailed analysis of the evolution to ensure convergence to an optimal solution. This becomes even more complex when different algorithms are used and converge to different solutions, owing to the sensitivity to hyperparameters that must be carefully studied. Additionally, this study examines two design parameters. Scaling up the method to include additional design parameters presents two main challenges. First, it increases the difficulty in sampling interesting design spaces for all parameters, likely requiring more iterations. Second, it results in higher variance in the gradient estimates. This is analogous to problems where the optimal control policy is learned.

Finally, two important advantages from the energy perspective are: first, the framework provides an interval of optimal design values rather than a unique solution, as MILP typically does, offering more flexibility and sensitivity information. Second, this framework offers better performance without assuming perfect foresight, likely explaining the superior validation performance in Figure 2, as the MILP did not have access to the validation dataset while computing the design parameters.

6. Conclusion

The primary achievement of this study has been leveraging theoretical advances in RL to bridge the gap with practical energy challenges, focusing on the co-optimisation of design and operation within energy systems. This work has harnessed recent developments in policy gradient techniques to introduce an integrated, off-policy, and model-free RL framework tailored to tackle the co-optimisation challenge in energy systems.

The successful demonstration of RL's feasibility and effectiveness in developing integrated design strategies within

a co-optimisation framework paves the way for future research and expands the capabilities of RL in the energy sector. This conclusion aligns with two notable reviews: (Dranka et al., 2021), which underscores the importance of addressing co-optimisation in energy and highlights the absence of integrated solutions, and (Perera & Kamalaruban, 2021), which notes that RL is not fully exploited in energy and suggests that using RL for design would be a promising new research area.

The outcomes validate the relevance of using RL to design energy systems, demonstrating how co-optimisation can effectively compute control and design policies jointly, and surpass traditional approaches. Additionally, this framework does not mandate a specific control algorithm or restrict to RL alone, instead, it requires the problem to be formulated as an MDP. Adherence to RL standards, i.e., Gymnasium library (Towers et al., 2023), is advised to ensure seamless integration with existing control algorithms, even though they have been developed from scratch in this case.

The practical application reveals the framework potential through a single year's data analysis. For greater accuracy and to evaluate long-term co-optimisation effects, it is advantageous to extend the dataset to encompass multiple years. Expanding the dataset would enhance the framework's ability to manage annual fluctuations in energy supply and demand. Further complexity could be introduced into the energy system model, like integrating multiple electric vehicles and accounting for non-linear heat pump dynamics. Additionally, incorporating more complex energy system dynamics such as real-time pricing or demand-response capabilities could improve the model's precision and relevance. The framework has also demonstrated promising outcomes that suggest the potential for generalisation to enhance sim-to-real transfer (Peng et al., 2018), a significant step towards ensuring that the insights and predictions generated can be effectively applied in realistic operational settings (Schaff et al., 2023). Future directions might also include integrating a critic architecture directly into the design learning process and extend the off-policy nature to the design part.

In conclusion, the findings and the comparison to traditional approaches, such as the design-only and best two-step scenarios, highlight that optimal design and optimal control are intrinsically linked. These insights affirm the value of integrated co-optimisation strategies over traditional, segregated approaches, especially in complex and dynamic settings like modern energy systems.

Acknowledgements

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Ahmed, Z., Roux, N. L., Norouzi, M., and Schuurmans, D. Understanding the impact of entropy on policy optimization, June 2019. URL http://arxiv.org/abs/1811.11214.

Bao, Z., Chen, D., Wu, L., and Guo, X. Optimal inter- and intra-hour scheduling of islanded integrated-energy system considering linepack of gas pipelines. *Energy*, 171:326–340, March 2019. ISSN 0360-5442. doi: 10.1016/j.energy.2019.01.016. URL https://www.sciencedirect.com/science/article/pii/S0360544219300180.

Bellman, R. A markovian decision process. *Journal of Mathematics and Mechanics*, 6(5):679–684, 1957. ISSN 0095-9057. URL https://www.jstor.org/stable/24900506. Publisher: Indiana University Mathematics Department.

Bhatia, J. S., Jackson, H., Tian, Y., Xu, J., and Matusik, W. Evolution Gym: A Large-Scale Benchmark for Evolving Soft Robots, January 2022. URL http://arxiv.org/abs/2201.09863.

Bolland, A., Boukas, I., Berger, M., and Ernst, D. Jointly learning environments and control policies with projected stochastic gradient ascent. *Journal of Artificial Intelligence Research*, 73:117–171, 2022. ISSN 1076-9757. doi: 10.1613/jair.1.13350. URL https://www.jair.org/index.php/jair/article/view/13350.

Cauz, M., Bolland, A., Miftari, B., Perret, L., Ballif, C., and Wyrsch, N. Reinforcement learning for joint design and control of battery-PV systems. *Proceedings of ECOS 2023*, 2023. doi: 10.52202/069564-0281. 36th International Conference on Efficiency, Cost, Optimization, Simulation and Environmental Impact of Energy Systems.

Chen, T., He, Z., and Ciocarlie, M. Hardware as Policy: Mechanical and Computational Co-Optimization using Deep Reinforcement Learning, November 2020. URL http://arxiv.org/abs/2008.04460.

Clack, C. T. M., Xie, Y., and MacDonald, A. E. Linear programming techniques for developing an optimal electrical system including high-voltage direct-current transmission and storage. *International Journal of Electrical Power & Energy Systems*, 68:103–114, June 2015. ISSN 0142-0615. doi: 10.1016/j.ijepes.2014.12.049. URL https://www.sciencedirect.com/science/article/pii/S0142061514007765.

Daadaa, M., Séguin, S., Demeester, K., and Anjos, M. F. An optimization model to maximize energy generation in short-term hydropower unit commitment using efficiency points. *International Journal of Electrical Power & Energy Systems*, 125:106419, February 2021. ISSN 0142-0615. doi: 10.1016/j.ijepes.2020.106419. URL https://www.sciencedirect.com/science/article/pii/S0142061519342218.

Dranka, G. G., Ferreira, P., and Vaz, A. I. F. A review of co-optimization approaches for operational and planning problems in the energy sector. *Applied Energy*, 304:117703, December 2021. ISSN 0306-2619. doi: 10.1016/j.apenergy.2021.117703. URL https://www.sciencedirect.com/science/article/pii/S0306261921010588.

Fazlollahi, S. and Maréchal, F. Multi-objective, multi-period optimization of biomass conversion technologies using evolutionary algorithms and mixed integer linear programming (MILP). *Applied Thermal Engineering*, 50(2):1504–1513, 2013. ISSN 1359-4311. doi: 10.1016/j.applthermaleng.2011.11.035. URL https://www.sciencedirect.com/science/article/pii/S1359431111006636.

François-Lavet, V., Henderson, P., Islam, R., Bellemare, M. G., and Pineau, J. An introduction to deep reinforcement learning. *Foundations and Trends in Machine Learning*, 11(3):219–354, 2018. ISSN 1935-8237, 1935-8245. doi: 10.1561/2200000071. URL https://www.nowpublishers.com/article/Details/MAL-071.

Gjorgiev, B. and Sansavini, G. Electrical power generation under policy constrained water-energy nexus. *Applied Energy*, 210:568–579, January 2018. ISSN 0306-2619. doi: 10.1016/j.apenergy.2017.09.011. URL https://www.sciencedirect.com/science/article/pii/S0306261917312977.

Ha, D. Reinforcement Learning for Improving Agent Design. *Artificial Life*, 25(4):352–365, November 2019. ISSN 1064-5462. doi: 10.1162/artl_a_00301. URL https://doi.org/10.1162/artl_a_00301.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 1861–1870. PMLR, July 2018. URL https://proceedings.mlr.press/v80/haarnoja18b.html. ISSN: 2640-3498.

Jackson, L., Walters, C., Eckersley, S., Senior, P., and Hadfield, S. Orchid: Optimisation of robotic control and hardware in design using reinforcement learning. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4911–4917, 2021. doi: 10.1109/IROS51168.2021.9635865.

Jayadev, G., Leibowicz, B. D., and Kutanoglu, E. U.S. electricity infrastructure of the future: Generation and transmission pathways through 2050. *Applied Energy*, 260:114267, February 2020. ISSN 0306-2619. doi: 10.1016/j.apenergy.2019.114267. URL https://www.sciencedirect.com/science/article/pii/S0306261919319543.

Khojasteh, M. A robust energy procurement strategy for micro-grid operator with hydrogen-based energy resources using game theory. *Sustainable Cities and Society*, 60:102260, September 2020. ISSN 2210-6707. doi: 10.1016/j.scs.2020.102260. URL https://www.sciencedirect.com/science/article/pii/S2210670720304819.

Krishnan, V., Ho, J., Hobbs, B. F., Liu, A. L., McCalley, J. D., Shahidehpour, M., and Zheng, Q. P. Co-optimization of electricity transmission and generation resources for planning and policy analysis: review of concepts and modeling approaches. *Energy Systems*, 7(2):297–332, May 2016. ISSN 1868-3975. doi: 10.1007/s12667-015-0158-4. URL https://doi.org/10.1007/s12667-015-0158-4.

Li, B., Roche, R., Paire, D., and Miraoui, A. Optimal sizing of distributed generation in gas/electricity/heat supply networks. *Energy*, 151:675–688, May 2018. ISSN 0360-5442. doi: 10.1016/j.energy.2018.03.080. URL https://www.sciencedirect.com/science/article/pii/S0360544218304894.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. Continuous control with deep reinforcement learning. *arxiv*, (arXiv:1509.02971), 2019. doi: 10.48550/arXiv.1509.02971. URL http://arxiv.org/abs/1509.02971.

Luck, K. S., Amor, H. B., and Calandra, R. Data-efficient Co-Adaptation of Morphology and Behaviour with Deep Reinforcement Learning. In *Proceedings of the Conference on Robot Learning*, pp. 854–869. PMLR, May

2020. URL https://proceedings.mlr.press/v100/luck20a.html. ISSN: 2640-3498.

Peng, X. B., Andrychowicz, M., Zaremba, W., and Abbeel, P. Sim-to-Real Transfer of Robotic Control with Dynamics Randomization. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3803–3810, May 2018. doi: 10.1109/ICRA.2018.8460528. URL http://arxiv.org/abs/1710.06537.

Perera, A. and Kamalaruban, P. Applications of reinforcement learning in energy systems. *Renewable and Sustainable Energy Reviews*, 137:110618, 2021. doi: 10.1016/j.rser.2020.110618.

Perera, A. T. D., Wickramasinghe, P. U., Nik, V. M., and Scartezzini, J.-L. Introducing reinforcement learning to the energy system design process. *Applied Energy*, 262:114580, 2020. ISSN 0306-2619. doi: 10.1016/j.apenergy.2020.114580. URL https://www.sciencedirect.com/science/article/pii/S0306261920300921.

Popovici, E. and Winston, E. A framework for co-optimization algorithm performance and its application to worst-case optimization. *Theoretical Computer Science*, 567:46–73, February 2015. ISSN 0304-3975. doi: 10.1016/j.tcs.2014.10.038. URL https://www.sciencedirect.com/science/article/pii/S0304397514008305.

Qiu, T., Xu, B., Wang, Y., Dvorkin, Y., and Kirschen, D. S. Stochastic Multistage Coplanning of Transmission Expansion and Energy Storage. *IEEE Transactions on Power Systems*, 32(1):643–651, January 2017. ISSN 1558-0679. doi: 10.1109/TPWRS.2016.2553678. URL https://ieeexplore.ieee.org/document/7454784. Conference Name: IEEE Transactions on Power Systems.

Quest, H., Cauz, M., Heymann, F., Rod, C., Perret, L., Ballif, C., Virtuani, A., and Wyrsch, N. A 3d indicator for guiding AI applications in the energy sector. *Energy and AI*, 9:100167, 2022. ISSN 2666-5468. doi: 10.1016/j.egyai.2022.100167. URL https://www.sciencedirect.com/science/article/pii/S2666546822000234.

Sachio, S., Mowbray, M., Papathanasiou, M. M., del Rio-Chanona, E. A., and Petsagkourakis, P. Integrating process design and control using reinforcement learning. *Chemical Engineering Research and Design*, 183:160–169, 2022. ISSN 0263-8762. doi: 10.1016/j.cherd.2021.10.032. URL https://www.sciencedirect.com/science/article/pii/S0263876221004421.

Schaff, C., Yunis, D., Chakrabarti, A., and Walter, M. R. Jointly learning to construct and control agents using deep reinforcement learning. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 9798–9805. IEEE Press, 2019. doi: 10.1109/ICRA.2019.8793537. URL 10.1109/ICRA.2019.8793537.

Schaff, C., Sedal, A., Ni, S., and Walter, M. R. Sim-to-real transfer of co-optimized soft robot crawlers. *Autonomous Robots*, 47(8):1195–1211, December 2023. ISSN 1573-7527. doi: 10.1007/s10514-023-10130-8. URL https://doi.org/10.1007/s10514-023-10130-8.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal Policy Optimization Algorithms, August 2017. URL http://arxiv.org/abs/1707.06347.

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., and Riedmiller, M. Deterministic policy gradient algorithms. In *Proceedings of the 31st International Conference on Machine Learning*, pp. 387–395. PMLR, 2014. URL https://proceedings.mlr.press/v32/silver14.html.

Towers, M., Terry, J. K., Kwiatkowski, A., Balis, J. U., Cola, G. d., Deleu, T., Goulão, M., Kallinteris, A., KG, A., Krimmel, M., Perez-Vicente, R., Pierré, A., Schulhoff, S., Tai, J. J., Shen, A. T. J., and Younis, O. G. Gymnasium, March 2023. URL https://zenodo.org/record/8127025.

Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3):229–256, 1992. ISSN 1573-0565. doi: 10.1007/BF00992696. URL 10.1007/BF00992696.

A. Appendix Building-Scale System – Environment definition

This Annex details the building-scale energy system used within an office setting, equipped with a PV (Photovoltaic) installation and a stationary lithium-ion battery to satisfy its electricity requirements. The system also features a bidirectional EV (Electric Vehicle) charging point, whose usage is stochastically modelled based on typical patterns. Moreover, the building is connected to the electrical grid, subject to dynamically varying electricity prices. The main objective is to determine the optimal design for the PV installation (P^{NOM}) and the battery capacity (B), while simultaneously developing an optimal control policy for battery and EV management. This aims to minimise the total cost of ownership, encompassing both capital and operational expenses, as well as grid costs. The environment is formulated below as an MDP and Table 3 gathers all parameters of this environment.

	Parameter	Value	Set	Unit	Description
GRID	P^{IMP}		$\mathbb{R}_+^T$	kW	imported power (from the grid)
	P^{EXP}		$\mathbb{R}_+^T$	kW	exported power (to the grid)
	$C_{\text{GRID}}^{\text{IMP}}$		$\mathbb{R}^T$	CHF/kWh	imported electricity price
	$C_{\text{GRID}}^{\text{EXP}}$		$\mathbb{R}^T$	CHF/kWh	exported electricity price
	C_{GRID}		$\mathbb{R}^T$	CHF	total electricity grid cost
PV	P^{NOM}		$\mathbb{R}_+$	kW$_p$	nominal power of the PV installation
	$P_{\text{MIN}}^{\text{NOM}}$	0	$\mathbb{R}_+$	kW$_p$	minimal nominal PV power
	$P_{\text{MAX}}^{\text{NOM}}$	∞	$\mathbb{R}_+$	kW$_p$	maximal nominal PV power
	P^{PROD}		$\mathbb{R}_+^T$	kW	generated PV power
	p^{PROD}		$\mathbb{R}_+^T$	kW	normalised PV power
	L^{PV}	20	$\mathbb{N}$	years	PV lifetime
	R_{PV}		$\mathbb{R}_+$	-	annuity factor
	$\text{OX}_{\text{PV}}^{\text{FIX}}$	0	$\mathbb{R}_+$	CHF	OPEX PV fixed cost
	$\text{OX}_{\text{PV}}^{\text{VAR}}$	100	$\mathbb{R}_+$	CHF/kW	OPEX PV variable cost
	$\text{CX}_{\text{PV}}^{\text{FIX}}$	100	$\mathbb{R}_+$	CHF	CAPEX PV fixed cost
	$\text{CX}_{\text{PV}}^{\text{VAR}}$	775	$\mathbb{R}_+$	CHF/kW	CAPEX PV variable cost
BATTERY	B		$\mathbb{R}_+$	kWh	nominal capacity of the battery
	SOC		$\mathbb{R}_+^T$	kWh	state of charge of the battery
	P^B		$\mathbb{R}^T$	kW	power exchanged with the battery
	B_{MIN}	0	$\mathbb{R}_+$	kWh	minimal nominal battery capacity
	B_{MAX}	∞	$\mathbb{R}_+$	kWh	maximal nominal battery capacity
	η^{B}	0.9	$]0,1]$	-	battery efficiency
	L^{B}	10	$\mathbb{N}$	years	battery lifetime
	R_{B}		$\mathbb{R}_+$	-	annuity factor
	$\text{OX}_{\text{B}}^{\text{FIX}}$	0	$\mathbb{R}_+$	CHF	OPEX Battery fixed cost
	$\text{OX}_{\text{B}}^{\text{VAR}}$	10	$\mathbb{R}_+$	CHF/kW	OPEX Battery variable cost
	$\text{CX}_{\text{B}}^{\text{FIX}}$	50	$\mathbb{R}_+$	CHF	CAPEX Battery fixed cost
	$\text{CX}_{\text{B}}^{\text{VAR}}$	300	$\mathbb{R}_+$	CHF/kW	CAPEX Battery variable cost
EV	b^{EV}		$\mathbb{R}_+^T$	-	binary indicator of EV presence
	B^{EV}	80	$\mathbb{R}_+$	kWh	maximal nominal EV battery capacity
	SOC^{EV}		$\mathbb{R}_+^T$	kWh	state of charge of the EV battery
	$\text{SOC}_{\text{MIN}}^{\text{EV}}$	32	$\mathbb{R}_+$	kWh	minimum state of charge of the EV battery
	P^{EV}		$\mathbb{R}_+^T$	kW	power exchange with the EV battery
	$P_{\text{MAX}}^{\text{EV}}$	5	$\mathbb{R}_+$	kW	maximal power exchange with the EV battery
	η^{EV}	1	$]0,1]$	-	EV battery efficiency
	$C_{\text{EV}}^{\text{IMP}}$	-1.5	$\mathbb{R}$	CHF/kWh	imported electricity price from the EV battery
	$C_{\text{EV}}^{\text{EXP}}$	1	$\mathbb{R}$	CHF/kWh	exported electricity price to the EV battery
SYSTEM	T		$\mathbb{N}$	-	time horizon
	Δt	1	$\mathbb{R}_+$	h	time steps
	r	0.05	$\mathbb{R}$	-	discount rate
	P^{LOAD}		$\mathbb{R}_+^T$	kW	uncontrollable electricity consumption

Table 3. Set of constants and parameters of the building-scale PV-battery system studied.

The **State Space** of the system can be fully described by

$$s_t = (h_t, d_t, \text{SOC}_t, P_t^{\text{PROD}}, P_t^{\text{LOAD}}, C_{\text{GRID},t}^{\text{IMP}}, C_{\text{GRID},t}^{\text{EXP}}, b_t^{\text{EV}}, \text{SOC}_t^{\text{EV}}) \in \mathcal{S} \tag{2}$$

- $h_t \in \{0, ..., 23\}$ denotes the hour of the day at time t.

- $d_t \in \{0, ..., 364\}$ denotes the day of the year at time t.

- $\text{SOC}_t \in [0, B]$ is the state of charge of the battery at time t, this value is upper bounded by the nominal capacity of the installed battery B.

- $P_t^{\text{PROD}} \in \mathbb{R}_+$ represents the expected PV power at time t. This value is obtained by scaling normalized historical data p_t^{PROD} with the design of PV power (P^{NOM}) and considering h_t and d_t values.

- $P_t^{\text{LOAD}} \in \mathbb{R}_+$ denotes the expected value of the electrical load at time t. The load profile is determined using historical data that corresponds to the same hour and day as the PV power.

- $C_{\text{GRID},t}^{\text{IMP}} \in \mathbb{R}$ represents the cost per unit of electricity imported from the grid at time t. This value is dynamically determined from a predefined dataset.

- $C_{\text{GRID},t}^{\text{EXP}} \in \mathbb{R}$ corresponds to the compensation received per unit of electricity exported to the grid at time t. Like the import costs, this value is derived from a dataset.

- $b_t^{\text{EV}} \in \{0, 1\}$ is a binary indicator indicating whether a bidirectional EV is present at the charging station at time t. This state affects the potential for energy storage or retrieval from the EV's battery, thereby influencing the overall energy management strategy. The value is updated according to usage patterns captured in the dataset.

- $\text{SOC}_t^{\text{EV}} \in [\text{SOC}_{\text{MIN}}^{\text{EV}}, B^{\text{EV}}]$ specifies the current charge level of the EV's battery, when present. This value ranges between 40% of B^{EV} and B^{EV} when the EV is connected, and is set to zero when no EV is present. The charge level is initialised randomly based on probable starting conditions and adjusted according to actual charging and discharging activities dictated by the control policy and EV usage scenarios from the dataset.

The **Action Space** comprises the power exchanged with the stationary battery and the EV's battery when present. Positive values indicate discharging, and negative values represent charging. The continuous action space is defined as:

$$a_t = (\widetilde{P}_t^B, \widetilde{P}_t^{\text{EV}}) \in \mathcal{A} = [-\frac{B}{\Delta t}, \frac{B}{\Delta t}] \times [-P_{\text{MAX}}^{\text{EV}}, P_{\text{MAX}}^{\text{EV}}] \tag{3}$$

The **Initial Distribution** set the initial state as follows. The hour h_0 is set to 0. During the training process, the initial day d_0 is randomly selected, whereas for the validation process, d_0 is set to the earliest date within the year. The initial SOC_t is randomly determined during training and set to half of the battery capacity B during validation. All other initial state values are derived from an predefined input dataset based on the corresponding initial hour and day.

The **Transition Probability** becomes a transition function, as there is no randomness involved. This function updates the system state at each hourly time step.

The hour of the day h_t increments each hour, and the day d_t increments every 24 hours:

$$h_{t+1} = (h_t + 1) \bmod 24 \tag{4}$$

$$d_{t+1} = \text{Int}(\frac{h_t + 1}{24}) \tag{5}$$

where the function Int takes the integer value of the expression.

The state of charge for both the stationary battery SOC_t and the EV's battery SOC_t^{EV} are updated based on the respective power actions $\widetilde{P}_t^B$ and $\widetilde{P}_t^{\text{EV}}$. These actions specify the power to be charged or discharged from the batteries over one hour ($\Delta t = 1h$). However, the actual power exchanged is constrained either by the battery capacity when charging it or by the energy stored in the battery when discharging it.

$$P_t^B = \begin{cases} \frac{\text{B}-\text{SOC}_t}{\Delta t} & \text{if } \widetilde{P}_t^B > \frac{\text{B}-\text{SOC}_t}{\Delta t} \\ \frac{\text{SOC}_t}{\Delta t} & \text{if } \widetilde{P}_t^B < -\frac{\text{SOC}_t}{\Delta t} \\ P_t^B & \text{otherwise} \end{cases} \tag{6}$$

Similarly for the EV's battery:

$$P_t^{\text{EV}} = \begin{cases} \frac{B^{\text{EV}} - \text{SOC}_t^{\text{EV}}}{\Delta t} & \text{if } \widetilde{P}_t^{\text{EV}} > \frac{B^{\text{EV}} - \text{SOC}_t^{\text{EV}}}{\Delta t} \\ \frac{\text{SOC}_t^{\text{EV}}}{\Delta t} & \text{if } \widetilde{P}_t^{\text{EV}} < -\frac{40\% \cdot B_t^{\text{EV}}}{\Delta t} \\ P_t^{\text{EV}} & \text{otherwise} \end{cases} \tag{7}$$

$$\text{with} - P_{\text{MAX}}^{\text{EV}} \leq \widetilde{P}_t^{\text{EV}} \leq P_{\text{MAX}}^{\text{EV}} \tag{8}$$

Using these power exchanges, the state of charge for the next time step is calculated as:

$$\text{SOC}_{t+1} = \text{SOC}_t + P_t^B \cdot \Delta t \cdot (\eta^{\text{B}} \text{ if } P_t^B \geq 0 \text{ else } \frac{1}{\eta^{\text{B}}}) \tag{9}$$

$$\text{SOC}_{t+1}^{\text{EV}} = \text{SOC}_t^{\text{EV}} + P_t^{\text{EV}} \cdot \Delta t \cdot (\eta^{\text{EV}} \text{ if } P_t^{\text{EV}} \geq 0 \text{ else } \frac{1}{\eta^{\text{EV}}}) \tag{10}$$

where η^{B} and η^{EV} are respectively the efficiency of the battery and EV's battery.

The **Reward Function** quantifies the system's performance by incorporating economic factors that include investment cost (CAPEX), operating cost (OPEX), and costs associated with the purchase and sale of electricity from the grid. The reward at each time step t is calculated as the negative total expenditure (TOTEX):

$$r_t = -\text{TOTEX}_t \tag{11}$$

$$= -(\text{CAPEX} + \text{OPEX} + C_{\text{GRID},t}) \tag{12}$$

$$= -(\text{CAPEX} + \text{OPEX} + P_t^{\text{IMP}} \cdot C_{\text{GRID},t}^{\text{IMP}} - P_t^{\text{EXP}} \cdot C_{\text{GRID},t}^{\text{EXP}}) \tag{13}$$

where $C_{\text{GRID},t}$ represents the net cost of electricity exchanged with the grid at time t.

The total cost (TOTEX) includes:

$$\text{TOTEX} = \text{OPEX} + \text{CAPEX} + C_{\text{GRID}} \tag{14}$$

Operating costs (OPEX) and capital expenditure (CAPEX) are defined for both PV and battery design parameters as:

$$\text{OPEX} = \text{OX}_{\text{PV}} + \text{OX}_{\text{B}} \tag{15}$$

$$\text{CAPEX} = \text{CX}_{\text{PV}} \cdot R_{pv} + \text{CX}_{\text{B}} \cdot R_B \tag{16}$$

where R_{PV} and R_B are annuity factors adjusting the CAPEX for the lifetime of the system's components, considering their financial amortisation over a finite period T.

The annuity factor R is derived as follows to prorate the CAPEX over the operational duration T, acknowledging T in hours and 8760 as the number of hours in a year:

$$R = \frac{r \cdot (1+r)^L}{(1+r)^L - 1} \cdot \frac{T}{8760} \tag{17}$$

This factor is calculated using the annual discount rate r and the expected lifetime L of the components, thereby aligning the investment costs proportionally to the duration T of the optimisation horizon.

The **Optimisation Horizon** refers to the period over which the system is optimised, corresponding to the duration of an episode. In this model, each time step of the MDP represents a single hour, and the horizon is truncated after $T = 168$ hours, equivalent to one week. Long-term dependencies are captured through the bootstrapping method used to train the critic. However, the time horizon ideally would span an entire year, or even the full lifecycle of the system to capture seasonal fluctuations in production and consumption, as well as potential equipment degradation. To assess performance over such a longer time horizon, the performances are regularly evaluated during the training phase across the full training dataset, corresponding to $T = 8088$.

The **Historical Datasets** of the system are detailed in Table 3. The historical data for the normalised PV production and the electrical consumption are derived from real monitoring of an office building in Switzerland in 2021, as shown in Figure 4. This dataset is divided into training and validation parts, each selected to represent the seasonal fluctuations. The dataset used for the electricity prices supplied to and from the grid, as well as the arrival times of the EV, are synthetically generated and summarised in Table 4. The grid export cost, $C_{\text{GRID}}^{\text{EXP}}$, is set to 0 at all times to discourage making money by reselling PV production and to maximise self-consumption. The duration of the EV's presence is randomly varied between 5 and 8 hours, and the initial state of charge (SoC) of the EV is randomly set between 40 % and 100 % of its battery capacity, B^{EV}.

Table 4. Synthetic dataset of electricity pricing and EV arrival time of the building-scale PV-battery system studied.

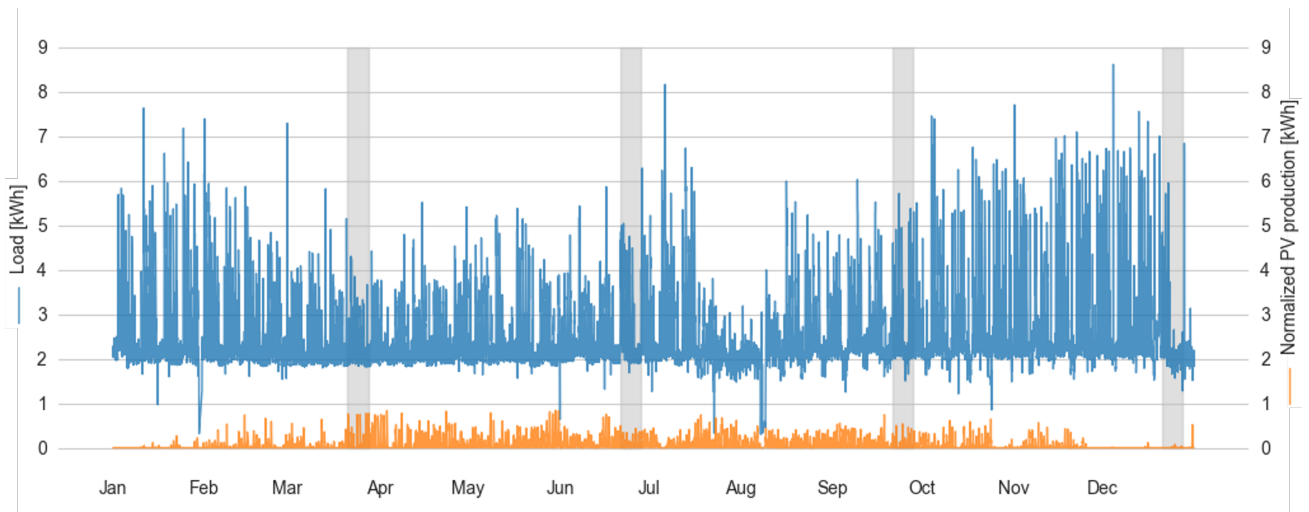| h [-]	C_{GRID}^{EXP} [CHF/kWh]	C_{GRID}^{IMP} [CHF/kWh]	Probability of EV arrival time [-]
0	0	-0.3	0
1	0	-0.3	0
2	0	-0.3	0
3	0	-0.3	0
4	0	-0.3	0
5	0	-0.3	0
6	0	-0.5	0
7	0	-0.5	0.75
8	0	-0.5	0.9
9	0	-0.5	0.9
10	0	-0.3	0.75
11	0	-0.3	0.1
12	0	-0.3	0.1
13	0	-0.3	0.1
14	0	-0.3	0
15	0	-0.3	0
16	0	-0.5	0
17	0	-0.5	0
18	0	-0.5	0
19	0	-0.5	0
20	0	-0.5	0
21	0	-0.5	0
22	0	-0.3	0
23	0	-0.3	0

Figure 4. Visualisation of the historical dataset covering a year of the building electricity consumption and its normalised PV production. The white background indicates the training set, while the grey background represents the validation dataset.