A Comparative Study of Multi-Stage Stochastic Optimization Approaches for an Energy Management System

Daniel Mimouni, Jiamin Zhu, Welington de Oliveira and Paul Malisani

***Abstract*—In this paper, we implement and compare four different control strategies for the Energy Management System (EMS) of a stationary battery using ground truth measurements of electricity consumption and production from a predominantly commercial building in France. This work is motivated by the fact that, on the one hand, classical approaches such as MPC, which deal with uncertainties by computing deterministic solutions in a receding horizon fashion, can lead suboptimal economic performance; on the other hand, risk-free multi-stage stochastic programming lacks robustness when reality deviates from forecast. To achieve a good trade-off between robustness and performance, we explore and compare other control models including robust stochastic optimization models and Reinforcement Learning strategies. Through numerical experiments, we evaluate these models in terms of cost efficiency, computational scalability, and out-of-sample robustness, offering a comprehensive comparison and insights into their practical interest for real-world EMS problems.**

***Index Terms*—Multistage Stochastic Optimization, Distributionally Robust Optimization, Reinforcement Learning, Stochastic Programming with Recourse, EMS.**

I. Introduction

ENERGY Management Systems (EMS) play a central role in optimizing electricity consumption, and reducing operational costs in modern power networks. However, managing energy efficiently requires making sequential decisions under uncertainty: the future is not fully known when decisions must be made. Traditionally, such problems are addressed using stochastic programming models, which aim to minimize expected costs by simulating a wide range of future scenarios. These models assume that the probability distribution of future data is known in advance. Although reasonable, this assumption is not always satisfied in practice, and the true distribution sometimes has to be estimated from limited historical data, which may affect the reliability of the results derived from it.

To overcome this limitation, practitioners often rely on Model Predictive Control (MPC) [11], [15], [21], a control model that solves a deterministic optimization problem at each stage using the most recent data and a single scenario for predicting the unknown variables. While MPC is simple to implement and widely adopted in industry, it does not explicitly account for future uncertainties and may result in suboptimal performance [10].

Another strategy is Robust Optimization (RO) [10], [27], which discards the need for probabilities altogether. Instead, it assumes the worst-case scenario among all the possible outcomes. This approach yields conservative solutions, which are often too costly or overly cautious for practical use [12].

A more balanced approach, known as Distributionally Robust Optimization (DRO), considers distributions lying within an ambiguity set – a set of plausible probability distributions constructed from a finite collection of scenarios. The goal is to optimize decisions to perform well against the worst-case distribution in this set. In fact, rather than relying on a single estimated probability measure, DRO acknowledges the uncertainty in this estimation and instead considers all distributions that lie within a certain neighborhood of the empirical measure. The quality of the DRO solution strongly depends on how this ambiguity set is defined as it captures the trade-off between robustness and reliance on a specific distribution. Among distributionally robust approaches, recent works have focused on the use of Wasserstein distance [9], [16] to define ambiguity sets. While the Wasserstein distance offers attractive mathematical properties, allowing for models to be both robust, its application to multistage problems raises nontrivial challenges [8], [25].

The recent work [19], situated within the field of risk-averse stochastic optimization, introduces a novel approach that integrates variance penalization directly into multistage SP models. By accounting for cost variability across scenarios, the approach effectively balances robustness and performance.

A key challenge shared by most of the optimization-based approaches discussed so far is their reliance on scenario generation. While increasing the number of scenarios can enhance solution accuracy, it also leads to substantial computational overhead. Although scenario selection and tree reduction techniques offer partial remedies [2], [13], [14], [20], they do not eliminate the fundamental dependency on pre-specified scenario sets.

An alternative to optimization-based models is Reinforcement Learning (RL), which learns decision policies directly from data through interaction with the environment [4], [7], [29]. RL avoids the need to specify probability distributions or generate accurate scenarios. Instead, it learns from repeated experience to make sequential decisions under uncertainty. Although promising, RL often requires a large amount of data and may struggle with constraints typically present in EMS problems [17].

This paper systematically compares several approaches,

Daniel Mimouni (corresponding author) is with the Centre de Mathématiques Appliquées (CMA), Mines Paris PSL, and the Applied Mathematics Department, IFP Energies nouvelles, 92852 Rueil-Malmaison, France (e-mail: daniel.mimouni1@gmail.com).

Jiamin Zhu is with the Control Signals and Systems Department, IFP Energies nouvelles, 92852 Rueil-Malmaison, France (e-mail: jiamin.zhu@ifpen.fr).

Welington de Oliveira is with the Centre de Mathématiques Appliquées (CMA), Mines Paris PSL, 06560 Sophia Antipolis, France (e-mail: welington.oliveira@minesparis.psl.eu).

Paul Malisani is with the Applied Mathematics Department, IFP Energies nouvelles, 92852 Rueil-Malmaison, France (e-mail: paul.malisani@ifpen.fr).

This work received no specific funding.

including MPC, (risk-free and risk-averse) stochastic programming models, robust optimization, distributionally robust models, and reinforcement learning—for a specific EMS application. We evaluate each approach in terms of robustness, computational efficiency, and out-of-sample performance, providing guidance on their practical use in real-world energy systems.

The remainder of this paper is structured as follows. Section II introduces the EMS problem we study and provides the mathematical background for the models discussed previously. Section III presents the computational methods, along with the specific algorithmic parametrizations tailored to address the EMS problem. Finally, Section IV reports numerical results, comparing the different approaches after performing cross-validation to ensure each method is tuned for the best performance on the EMS task.

NOTATIONS

Let X be a Hilbert space, we denote by $\langle .,. \rangle_{\mathrm{X}}$ its inner product, and by $\|.\|_X$ its corresponding norm. Let $\mathrm{E} \subset \mathrm{X}$ be convex, we denote $\mathrm{i}_{\mathrm{E}} : \mathrm{X} \mapsto \mathbb{R} \cup \{+\infty\}$ the indicator function of E, i.e. $\mathrm{i}_E(x) = 0$ if $x \in E$ and $\mathrm{i}_E(x) = +\infty$ otherwise. Additionally we denote $\mathbb{1}_E : X \mapsto \{0, 1\}$ the characteristic function of E, i.e. $\mathbb{1}_E(x) = 1$ if $x \in E$ and $\mathbb{1}_E(x) = 0$ otherwise . Given a Fréchet-differentiable function $f : \mathrm{X} \mapsto \mathbb{R}$ we denote $f' \in \mathrm{X}$ the Fréchet-derivative of f. Let $(\Omega, \mathcal{F}, P)$ be a probability space, we denote random variables from Ω to X using bold characters such as $\boldsymbol{\xi} : \Omega \mapsto \mathrm{X}$. If Ω is a discrete set, i.e. $\Omega = \{\omega_1, \dots, \omega_S\}$ we denote indifferently $\boldsymbol{\xi}(s)$ or ξ^s the outcome $\boldsymbol{\xi}(\omega_s)$ of the random variable. We denote with blackboard capital letters sets of random variable such as $\mathbb{X} := \{\boldsymbol{x} : \Omega \mapsto \mathrm{X}\}$. We denote $\mathbb{E}_P$ the mathematical expectation with respect to the measure P. The random Hilbert space $\mathbb{X}$ is endowed with the inner product $\langle .,. \rangle_{\mathbb{X}} := \mathbb{E}\left(\langle .,. \rangle_{\mathrm{X}}\right)$ and the corresponding norm $\|.\|_{\mathbb{X}} := \mathbb{E}(\|.\|_X)$. Given $p \in [1, +\infty]$, we denote $\mathrm{L}^p(A; B)$ (or L^p) the Lebesgue spaces of functions from A to B and we denote $\|.\|_{\mathrm{L}^p}$ the corresponding p-norm. For all $1 \leq p < +\infty$, we denote $\mathbb{L}^p$ the space of random variables $\boldsymbol{\xi} : \Omega \mapsto \mathrm{L}^p$ and we denote $\|\boldsymbol{\xi}\|_{\mathbb{L}^p} := \mathbb{E}(\|\boldsymbol{\xi}\|_{\mathrm{L}^p}^p)^{\frac{1}{p}}$. We denote $\mathbb{L}^\infty$ the space of random variables $\boldsymbol{\xi} : \Omega \mapsto \mathrm{L}^\infty$ and we denote $\|\boldsymbol{\xi}\|_{\mathbb{L}^\infty} := \inf\{y \in \mathbb{R} : \mu(\{\omega \in \Omega : \|\boldsymbol{\xi}(\omega)\|_{\mathrm{L}^\infty} > y\}) = 0\}$. Let X be a set, and let $x \in X$, we denote δ_x the Dirac delta function on x. Finally, let $x \in \mathbb{R}^n$, we denote x_k the k^{th} coordinate of x and let $M \in \mathbb{R}^{n \times m}$, we denote $M_{k,l}$ the value of the k^{th} row and l^{th} column of M.

II. PROBLEM STATEMENT

We consider a Stochastic Optimal Control Problem (SOCP) modeling the operation to a stationary battery connected downstream of a prosumer's electricity meter. A *prosumer* is an end-user which both consumes and produces electricity, typically via intermittent sources such as photovoltaic panels, as depicted in Figure 1. Often the goal is to minimize the expected electricity cost over a finite time horizon, while accounting for battery dynamics and the stochastic nature of both **consumption** and **production**.

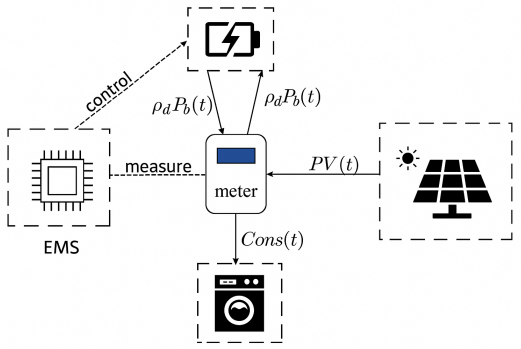


Fig. 1. Schematic diagram of a domestic system with a stationary battery controlled by an EMS [19]

A. The EMS battery strategy

Let P_b, $\mathbf{Cons}$, and $\mathbf{PV}$ be respectively the battery's charging power, the electric consumption and photovoltaic power both measured at the meter. The electricity bill to minimize, as a function of these variables, writes

$$J_{t_1:t_2}(P_b, \mathbf{Cons}, \mathbf{PV}) := \int_{t_1}^{t_2} p_r^c(t) \max\{\boldsymbol{P_m}(t), 0\} + p_r^d(t) \min\{\boldsymbol{P_m}(t), 0\} \mathrm{d}t \quad (1)$$

where $p_r^c(t)$ [€/kWh] (resp. $p_r^d(t)$) is the buying (resp. selling) prices of electricity at time t. They satisfy $0 \leq p_r^d(t) \leq p_r^c(t)$ at all times, therefore eq. (1) is convex. The power measured at the meter, $\boldsymbol{P_m}$ [kW] writes

$$\boldsymbol{P_m}(t) := \mathbf{Cons}(t) - \mathbf{PV}(t) + \frac{1}{\rho_c} \max\{P_b(t), 0\} + \rho_d \min\{P_b(t), 0\}, \quad (2)$$

The parameters $\rho_c = 0.97$, and $\rho_d = 0.97$ represent the battery charge and discharge efficiencies, respectively. Now, the battery's dynamics are governed by:

$$\dot{E}(t) = P_b(t), \quad (3)$$

and the operational constraints of the battery's charging / discharging process are

$$E \in \mathrm{L}^\infty([t_1, t_2]; [0, 13 \text{ kWh}]), \quad (4a)$$
$$P_b \in \mathrm{L}^2([t_1, t_2]; [-8, 8\rho_c \text{ kW}]), \quad (4b)$$
$$E(t_1) = E(t_2) = E^0. \quad (4c)$$

In this formulation, $E(t)$ [kWh] represents the battery's State of Energy (SoE) at time t, while E^0 denotes the initial and terminal state of energy, ensuring periodicity over the time horizon. The interactions between the sources of power are depicted in the diagram of Figure 1. For the sake of readability, we denote C the set of admissible deterministic charging powers defined as follows

$$\mathrm{C} := \left\{P_b \in \mathrm{L}^2 : \text{ eqs. (3) and (4) hold}\right\}. \quad (5)$$

III. Computing solutions for different models

In this work, we are interested in exploring different optimization models for our EMS problem taking in account uncertainties on consumption and (photovoltaic) production. We start by describing a standard model predictive control formulation.

A. Model predictive control

At each time step t, MPC algorithms [11], [15], [21] consist of predicting a realization of the random variables $\mathbf{Cons}$ and $\mathbf{PV}$, respectively denoted $\widehat{\mathrm{Cons}}$ and $\widehat{\mathrm{PV}}$, on a defined and finite horizon H, and solve the following deterministic optimization problem:

$$\min_{P_b \in \mathrm{C}} J_{t:t+H}(P_b, \widehat{\mathrm{Cons}}, \widehat{\mathrm{PV}}). \tag{MPC}$$

Note that, in the MPC framework, the charging battery power (P_b), the energy in the battery (E), are not random variables since they only depend on the predicted scenario $\widehat{\mathrm{Cons}}$, $\widehat{\mathrm{PV}}$ which is deterministic. This MPC strategy is a rolling horizon method of actualization horizon $h < H$. This method is easy to implement and relies on deterministic optimization algorithms. However, the solutions can be far from optimal depending on the accuracy of single predicted scenario ($\widehat{\mathrm{Cons}}$, $\widehat{\mathrm{PV}}$).

B. Scenario-based methods

1) Mathematical framework: The control methods presented in this section all rely on a finite set of scenarios drawn from a given probability distribution. In the following we denote $\Xi_{t_1:t2}$ the discrete set of scenarios defined as follows

$$\Xi^S_{t_1:t_2} := \left\{ \mathrm{Cons}^s - \mathrm{PV}^s \in \mathrm{L}^2([t_1, t_2]; \mathbb{R}) : s = 1, \dots, S \right\}. \tag{6}$$

Each scenario, $\mathrm{Cons}^s - \mathrm{PV}^s$, is associated to a probability $p_s > 0$, satisfying $\sum_{s=1}^S p_s = 1$. For notation simplicity, we denote $\xi^s := \mathrm{Cons}^s - \mathrm{PV}^s$. Scenario based methods consist in computing an optimal control for each scenario $\xi^s \in \Xi_{t_1:t_2}$, that is to say, S charging/discharging chronicles, denoted P_b^s. However, these S optimal controls are not independent as they must satisfy a Δ-non-anticipative constraint: If $s \neq s'$ are such that $\xi^s(t) = \xi^{s'}(t)$, for all $t \leq \tau$, then necessarily $P_b^s(t) = P_b^{s'}(t)$, for all $t \leq \tau + \Delta$. This Δ-non-anticipativity constraint allows for taking into account the measurement delays. Indeed, consumption at time t, $\mathrm{Cons}(t)$ is the mean power consumption on the time interval $[t, t+\Delta]$ and therefore is only measured at time $t + \Delta$. Thus, the battery power P_b on $[t, t + \Delta]$ must be computed without knowledge of the consumption nor the PV production on this time interval hence the Δ-non-anticipativity constraint. Throughout the paper, we denote $\mathcal{N}_\Delta$ the set of Δ-non-anticipative controls. Mathematical details on scenario-based methods can be found in [6], [23], [26].

Therefore, the corresponding set of admissible random charging powers, associated with set C and accounting for non-anticipativity, is

$$\mathbb{C} := \{ \boldsymbol{P_b} \in \mathcal{N}_\Delta : \boldsymbol{P_b} \in \mathrm{C} \text{ almost surely} \}. \tag{7}$$

2) Desterministic-control approach for SP: The deterministic-control approach to stochastic programming (DSP) model involves computing a single control policy that minimizes the expected electricity bill across all scenarios. This contrasts with standard SP models, which seek a scenario-dependent, non-anticipative control P_b^s, for $s = 1, \dots, S$. The deterministic control approach is workable due to the problem's structure: the plausible set does not depend on uncertainty.

Problem 1 (Desterministic-Control Approach for SP (DSP)): The DSP problem is defined as

$$\inf_{P_b \in \mathrm{C}} \left[\sum_{s=1}^S p_s J_{t_1:t_2}(P_b, \mathrm{Cons}^s, \mathrm{PV}^s) \right], \tag{DSP}$$

where $\mathrm{Cons}^s - \mathrm{PV}^s \in \Xi^S_{t_1:t_2}$ is the s-th scenario.
Problem 1 fits a standard setting of constrained optimal control problem and can be solved using numerical methods such as [18]. This model's main advantage is that it remains numerically tractable even when using a large number of scenarios. This method is highly robust but can be sub-optimal.

3) Stochastic programming with variance penalization: The stochastic programming with variance penalization consists of solving the following optimal control problem, parameterized with the parameter $\alpha \geq 0$

Problem 2 (SP with variance penalization): The SP model with variance penalization consists in solving for $\boldsymbol{P_b}$ the following optimal control problem

$$\inf_{\boldsymbol{P_b} \in \mathbb{C}} \Bigg[\sum_{s=1}^S p_s J_{t_1:t_2}(\boldsymbol{P_b}(s), \mathrm{Cons}^s, \mathrm{PV}^s) + \frac{\alpha}{2} \sum_{s=1}^S p_s \left\| \boldsymbol{P_b}(s) - \sum_{s'=1}^S p_{s'} \boldsymbol{P_b}(s') \right\|^2_{\mathrm{L}^2} \Bigg], \tag{VSP}$$

where $\boldsymbol{P_b}$ is the discrete random variable associating to each scenario s the optimal charging power $\boldsymbol{P_b}(s) := P_b^s$, and where $\mathrm{Cons}^s - \mathrm{PV}^s \in \Xi^S_{t_1:t_2}$.
For $\alpha = 0$, Problem 2 reduces to the standard risk-neutral optimal control problem [23]. For $\alpha = +\infty$, the problem consists of finding a deterministic battery's charging/discharging power that minimizes the electricity bill's expectation, i.e., (VSP) boils down to (DSP). For $\alpha \in (0, +\infty)$, Problem 2 balances the risk-neutral and Problem 1. The solving algorithm, presented in [19], for Problem 2 is displayed in Algorithm 1 below. We highlight that Step 1 of this algorithm is the only computationally intensive step and consists of solving S independent deterministic optimal control problems. Efficient algorithms such as [5], [18] are available for this task.

Let $\Lambda(s,t) := \left\{ j \in \{1, \dots, S\} \;\middle|\; \xi_l^j = \xi_l^s, \forall l \leq t - \Delta \right\}$ be the set of scenarios sharing the same history as scenario s up to stage t.

Then, for $\boldsymbol{z} \in \mathbb{L}^2$, the projection $\boldsymbol{x} = \mathrm{Proj}_{\mathcal{N}_\Delta}(\boldsymbol{z})$ is given component-wise by

$$\boldsymbol{x}_t(s) = \frac{1}{\sum_{j \in \Lambda(s,t)} p_j} \sum_{j \in \Lambda(s,t)} p_j \boldsymbol{z}_t(j), \quad t = 1, \dots, T, s = 1, \dots, S. \tag{11}$$

Algorithm 1 Regularized Progressive Hedging Algorithm (RPHA)

1: **Initialization:** Set $\boldsymbol{z}^0 \in \mathbb{L}^2$, $\boldsymbol{\lambda}^0 \in \mathcal{N}_\Delta^\perp$, $r > 0$, tolerance $\texttt{Tol} > 0$, $k \leftarrow 0$ and success $\leftarrow False$

2: **while** not success **do**

3: **Step 1:** Solve for each scenarios indexed by s:

$$\boldsymbol{P_b}^{k+1}(s) \leftarrow \arg\min_{P_b \in \mathrm{C}} \Big\{ J_{t_1:t_2}(P_b, \mathrm{Cons}^s, \mathrm{PV}^s) + \langle \boldsymbol{\lambda}^k(s), P_b \rangle_{\mathrm{L}^2} + \frac{r}{2} \left\| P_b - \boldsymbol{z}^k(s) \right\|_{\mathrm{L}^2}^2, \quad s = 1, \dots, S \Big\} \tag{8}$$

4: **Step 2:** Project onto the orthogonal non-anticipative space:

$$\boldsymbol{\lambda}^{k+1} \leftarrow \boldsymbol{\lambda}^k + r \mathrm{Proj}_{\mathcal{N}_\Delta^\perp}(\boldsymbol{P_b}^{k+1}) \tag{9}$$

5: **Step 3:** Update the dual variables:

$$\boldsymbol{z}^{k+1} \leftarrow \boldsymbol{z}^k - \boldsymbol{P_b}^{k+1} + \frac{\alpha}{r+\alpha} \sum_s p_s \left(2\boldsymbol{P_b}^{k+1}(s) - \boldsymbol{z}^k(s) \right) + \frac{r}{r+\alpha} \mathrm{Proj}_{\mathcal{N}_\Delta} \left(2\boldsymbol{P_b}^{k+1} - \boldsymbol{z}^k \right) \tag{10}$$

6: **Step 4:** Check convergence:

$$\text{success} \leftarrow \| \boldsymbol{P_b}^{k+1} - \boldsymbol{P_b}^k \|_{\mathbb{L}^2} < \texttt{Tol}$$

7: **end while**

In words, for each stage t and scenario s, $\boldsymbol{z}_t(s)$ is replaced by the average over all scenarios that share the same history up to t. And

$$\mathrm{Proj}_{\mathcal{N}_\Delta^\perp}(\boldsymbol{z}) = \boldsymbol{z} - \mathrm{Proj}_{\mathcal{N}_\Delta}(\boldsymbol{z}). \tag{12}$$

4) Distributionally robust optimization: In this setting, we use two fixed sets of scenarios, denoted Ξ^S and Ξ^L of respective size S and L with $L \geq S$. Each scenario $\xi^l := \mathrm{Cons}^l - \mathrm{PV}^l \in \Xi^L$ is associated with a probability p_l. The DRO setting consists of solving the following problem:

Problem 3 (Distributionnaly Robust Problem):

$$\inf_{\boldsymbol{P_b} \in \mathbb{C}} \sup_{q \in \mathcal{P}_\theta} \left[\sum_{s=1}^S q_s J(\boldsymbol{P_b}(s), \mathrm{Cons}^s, \mathrm{PV}^s) \right], \tag{DRO}$$

where $\mathcal{P}_\theta$ is the ambiguity set defined as follows

$$\mathcal{P}_\theta := \Big\{ q \in \mathbb{R}_+^S : \sum_s q_s = 1, \ \mathrm{W}_2 \left(\sum_{s=1}^S q_s \delta_{\xi^s}, \sum_{l=1}^L p_l \delta_{\xi^l} \right) \leq \theta \Big\}. \tag{13}$$

Here W_2 is the 2-Wasserstein distance between two discrete probability measures [24, Ch. 5]. The parameter θ controls the size of the ambiguity set. The larger θ is, the larger the ambiguity set becomes. Thus, large values of θ lead to a worst-case type behavior by assigning all probabilities q_s except one to zero, which is equivalent to selecting the worst scenario over Ξ^S as in RO. Conversely, when θ is small, the ambiguity set is also small, and the weights q_s are adjusted so that the probability defined over Ξ^S is close to that defined over Ξ^L, approaching a risk-neutral problem.

Before describing an algorithm to solve (DRO), we need to introduce two mathematical operations, the projection onto the ambiguity set $\mathcal{P}_\theta$ from eq. (13) and the projection onto the epigraph of the optimal control problem described in eqs. (1), (3) and (4).

Definition 1 (Projection onto the ambiguity set $\mathcal{P}_\theta$): Let $p \in \mathbb{R}^S$, let $\boldsymbol{\xi} \in \Xi^S$ and $\hat{\boldsymbol{\xi}} \in \Xi^L$, we denote the projection of p onto $\mathcal{P}_\theta$

$$\mathrm{Proj}_{\mathcal{P}_\theta}(p) := \bar{x} \tag{14}$$

where $\bar{x}$ is defined as follows

$$(\bar{x}, \bar{\eta}) \in \underset{x \in \mathbb{R}_+^S, \eta \in \mathbb{R}_+^{S \times L}}{\arg\min} \|x - p\|^2 \tag{15}$$

such that

$$\sum_{s=1}^S \sum_{l=1}^L \eta_{s,l} \left\| \xi^s - \hat{\xi}^l \right\|_{\mathrm{L}^2}^2 \leq \theta \tag{16a}$$

$$\sum_{s=1}^S \eta_{s,l} = p_l, \quad l = 1, \dots, L \tag{16b}$$

$$\sum_{l=1}^L \eta_{s,l} = x_s, \quad s = 1, \dots, S \tag{16c}$$

$$\sum_{s=1}^S \sum_{l=1}^L \eta_{s,l} = 1. \tag{16d}$$

Thus, the projection onto the ambiguity set is a quadratic problem of size $S(1+L)$, therefore, computationally intensive depending on S and L.

Definition 2 (Projection onto the epigraph): First, let us consider the following deterministic optimal control problem

$$\bar{J}_{t_1:t_2}^s(P_b) := J_{t_1:t_2}(P_b, \mathrm{Cons}^s, \mathrm{PV}^s) + \mathrm{i}_\mathrm{C}(P_b)$$

where C is defined in eq. (5). Let $(P_b, \rho) \in \mathrm{L}^2 \times \mathbb{R}$, we denote $\mathrm{Proj}_{\mathrm{epi}\, \bar{J}_{t_1:t_2}^s}(P_b, \rho)$ the corresponding projection onto the epigraph of $\bar{J}_{t_1:t_2}^s$ defined as the solution of the following optimal control problem

$$\mathrm{Proj}_{\mathrm{epi}\, \bar{J}_{t_1:t_2}^s}(P_b, \rho) \in \underset{P \in \mathrm{L}^2, \lambda \in \mathbb{R}}{\arg\min} \|P - P_b\|_{\mathrm{L}^2}^2 + (\rho - \lambda)^2 \tag{17}$$

under the following constraints

$$\dot{E}(t) = P(t) \tag{18a}$$

$$\dot{Q}(t) = p_r^c(t) \max\{P_m(t), 0\} + p_r^d(t) \min\{P_m(t), 0\} \tag{18b}$$

$$P_m(t) = \mathrm{Cons}^s(t) - \mathrm{PV}^s(t) + \frac{1}{\rho_c} \max\{P(t), 0\} + \rho_d \min\{P(t), 0\} \tag{18c}$$

$$Q(t_1) = 0 \tag{18d}$$

$$Q(t_2) - \lambda \leq 0 \tag{18e}$$

$$E(t_1) = E(t_2) = E^0. \tag{18f}$$

This is a standard deterministic optimal control problem and can be efficiently solved numerically.

We are now ready to detail the solving algorithm for problem (DRO) which is a direct adaptation of [8] to the problem at hand. The second step of Algorithm 2, i.e. the projection onto the ambiguity set, is the bottleneck of the algorithm. Indeed, the complexity of this task dramatically increases with the sizes of the scenarios sets. On the other hand, the projection onto the epigraphs performed at Step 3 is parallelizable with respect to the scenarios, and each individual task is a standard, easy-to-solve optimal control problem.

Algorithm 2 Scenario Decomposition with Alternating Projections - SDAP [8]

1: **Initialization:** Set $\boldsymbol{z}_{P_b}^0 \in \mathbb{L}^2$, $z_u^0 \in \mathbb{R}^S$, $z_v^0 \in \mathbb{R}^S$, $r > 0$, two tolerances $\text{Tol}_{P_b}, \text{Tol}_J > 0$, set $k \leftarrow 0$, and success $\leftarrow False$

2: **while** not success **do**

3: **Step 1:** Define

$$\boldsymbol{P_b}^k \leftarrow \text{Proj}_{\mathcal{N}_\Delta}(\boldsymbol{z}_{P_b}^k), \quad u^k \leftarrow \frac{z_u^k + z_v^k}{2}$$

4: **Step 2:** Projection on $\mathcal{P}_\theta$

$$\hat{v}^{k+1} \leftarrow (2u^k - z_v^k) - \frac{1}{r}\text{Proj}_{\mathcal{P}_\theta}\left(r(2u^k - z_v^k)\right)$$

5: **Step 3:** Projection onto the epigraphs

6: **for** $s = 1, \dots, S$ **do**

7:

$$(\hat{\boldsymbol{P}}_b^{k+1}(s), \hat{u}_s^{k+1}) \leftarrow \text{Proj}_{\text{epi}\, \bar{J}_{t_1:t_2}^s}\left(2\boldsymbol{P_b}^k(s) - \boldsymbol{z}_{P_b}^k(s), 2u_s^k - z_{us}^k\right)$$

8: **end for**

9: **Step 4:** Update:

$$\begin{aligned} \boldsymbol{z}_{P_b} &\leftarrow \boldsymbol{z}_x^k + \hat{\boldsymbol{P}}_b^{k+1} - \boldsymbol{P_b}^k \\ z_u^{k+1} &\leftarrow z_u^k + \hat{u}^{k+1} - u^k \\ z_v^{k+1} &\leftarrow z_v^k + \hat{v}^{k+1} - u^k \end{aligned}$$

10: **Step 5:** Check convergence:

$$\begin{aligned} \text{success} \leftarrow &\left\|\hat{\boldsymbol{P}}_b^{k+1} - \boldsymbol{P_b}^k\right\| < \text{Tol}_{P_b} \\ \text{and } &\left\|\hat{u}^{k+1} - u^k\right\| \leq \text{Tol}_J \text{ and } \left\|\hat{v}^{k+1} - u^k\right\| \leq \text{Tol}_J \end{aligned} \tag{19}$$

$k \leftarrow k + 1$

11: **end while**

C. Reinforcement learning

In reinforcement learning (RL), an agent learns an optimal behavior by interacting with an environment and receiving costs (or rewards) from these interactions. To solve the EMS problem, we will use a tabular Q-learning algorithm inspired by [29]. We define a finite horizon Markov Decision Process (MDP) as $(\mathcal{T}, \mathcal{S}, \mathcal{A}, \mathbb{P}, c)$, where $\mathcal{T} = \{t_1, t_1 + \Delta, \dots, t_2\}$ represents the finite time horizon, $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$, $\mathbb{P}(s'|s, a)$ is the transition probability of passing from state s to state s' given action a, and c is the cost function representing the cost received after taking action a in state s. We choose the state to be $s := (E, \bar{\xi}) \in \mathcal{S}$, where:

- $E \in \{0, 0.13, \cdots, 13 \text{ kWh}\}$ is the energy state of the battery and, where the discretization step of 0.13 kWh is taken from [29];
- $\bar{\xi} = \text{Cons} - \text{PV} \in \{-8\times 2, -16+1, \cdots, 16\rho_c \text{ kW}\}$ is the difference between the electricity demand and production, where the discretization step of 1 kW is taken from [29].

In the following we introduce the notation $s^\tau := (E(\tau), \bar{\xi}(\tau))$ to denote the state at time τ. We set the action to be the battery charging power and thus the action space $\mathcal{A}$ regroups the feasible battery powers $\mathcal{A} := \{-8, 0, 8\rho_c \text{ kW}\}$, where the discretization this choice for $\mathcal{A}$ is motivated by [29].

Finally, $c^\tau : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the cost function at time τ:

$$c^\tau(s, P_b) = p_r^c(\tau)\max\{P_m^\tau, 0\} + p_r^d(\tau)\min\{P_m^\tau, 0\}, \tag{20}$$

with

$$P_m^\tau = \bar{\xi}^\tau + \frac{1}{\rho_c}\max\{P_b, 0\} + \rho_d \min\{P_b, 0\}. \tag{21}$$

In this paper, we consider deterministic policies: in state s at time τ, we define π so that the action taken at time τ is entirely defined by $P_b^\tau = \pi(\tau, s) \in \mathcal{A}$. The goal is to find a policy π^* minimizing the expectation of its finite horizon return: $\mathbb{E}_\pi\left[\sum_{\tau=t}^{t_2} c^\tau(\boldsymbol{s}^\tau, \boldsymbol{P}_b^\tau)\right]$.

The Q-function associated with a policy π is defined as:

$$Q^\pi(t, s, P_b) = \mathbb{E}\left[\sum_{\tau=t}^{t_2} c^\tau(\boldsymbol{s}^\tau, \boldsymbol{P}_b^\tau) \mid \boldsymbol{s}^t = s, \boldsymbol{P}_b^t = P_b\right]. \tag{22}$$

It is the expected value of choosing action P_b in state s starting at time step t and following policy π afterwards. The optimal Q-function Q^* is defined as $\forall (t, s, P_b) \in \mathcal{T} \times \mathcal{S} \times \mathcal{A}$, $Q^*(t, s, P_b) = \min_\pi Q^\pi(t, s, P_b)$, and it satisfies the Bellman optimality equation:

$$Q^*(t, s, P_b) = \mathbb{E}_{s'}\left[c^t(s, P_b) + \min_{\bar{P}_b \in \mathcal{A}} Q^*(t + \Delta t, s', \bar{P}_b)\right], \tag{23}$$

with $s' \sim \mathbb{P}(\cdot|s, P_b)$. Then the optimal policy is simply $\pi^*(t, s) = \arg\min_{P_b \in \mathcal{A}} Q^*(t, s, P_b)$.

The controlled process $\boldsymbol{E}$ satisfies (3), hence we can avoid the exploration of the controlled part of the state space by parallel simulations for every pairs of $(E, P_b) \in \{0, dE, \cdots, 13 \text{ kWh}\} \times \mathcal{A}$. The exploration will be needed only for the unknown process $\bar{\boldsymbol{\xi}}$. Therefore, instead of standard Q-learning [28], our RL agent will learn from a dataset $\mathcal{D}$ consisting of previously collected transitions on the difference between consumption and production: $\mathcal{D}$ contains $|\mathcal{D}|$ days of historical data of **Cons** and **PV**. We denote $N(s)$ the number of times s is present in the dataset $\mathcal{D}$.

Given a dataset $\mathcal{D}$, the detailed training process is given in Algorithm 3. Then a policy π is derived as

$$\pi(t, s) = \arg\min_{P_b \in \mathcal{A}} Q(t, s, P_b), \quad \forall (t, s) \in \mathcal{T} \times \mathcal{S}. \tag{24}$$

The obtained policy π will be applied like a feedback controller on new days: at each timestep t, the state E and $\bar{\xi}$

Algorithm 3 Offline Q-Learning (Training)

Input: Given a training dataset $\mathcal{D}$, initialize $N(s) = 0$ for all $(t, s) \in \mathcal{T} \times \mathcal{S}$.
Initialize $Q(t, s, P_b) = 0$ for all $(t, s, P_b) \in \mathcal{T} \times \mathcal{S} \times \mathcal{A}$
for $t = t_2, t_2 - \Delta, \cdots, t_1$ **do**
  for each sample $((t, \bar{\xi}), (t + \Delta, \bar{\xi}')) \in \mathcal{D}$ **do**
    For all (E, P_b), collect state-action pairs $(s, P_b) = (E(t), \bar{\xi}(t), P_b)$ to be updated, and update the visit times of these state $N(s) \leftarrow N(s) + 1$
    Compute temporal difference error for all collected state-action pairs:

$$\mathrm{TD}(s, P_b) \leftarrow c(s, P_b) + \min_{\tilde{P}_b} Q(t + \Delta, s', \tilde{P}_b) - Q(t, s, P_b)$$

    Update Q-function for all collected state-action pairs:

$$Q(t, s, P_b) \leftarrow Q(t, s, P_b) + \frac{1}{N(s)} \cdot \mathrm{TD}(s, P_b)$$

  end for
end for
Return: Learned policy $\pi(t, s) = \arg\min_{P_b \in \mathcal{A}} Q(t, s, P_b)$

over the last intervale $[t - \Delta, t]$) is observed, then an action P_b is chosen according to the policy π, and the action $P_b = \pi(t, s)$ is applied to the system – see Algorithm 4.

Algorithm 4 Offline Q-learning (Application on new days)

Given a learned policy π.
for $t = t_1, \ldots, t_2$ **do**
  Measure $E(t)$ and $\bar{\xi}^t$ to define s^t
  Apply $P_b = \pi(t, s)$ to the battery
end for

Every days the $\mathcal{D}$ is enriched and therefore Algorithm 3 retrained before the newly learned policy is applied to the next day with Algorithm 4.

IV. Numerical example

The buying and selling prices of electricity are known: the selling price is set to 0, and the buying price is the real spot price collected over the 2-year period 2022-01-22 to 2024-04-22. Only the consumption and production are stochastic, and all the measures are historical data from a mainly commercial building in Solaize (France).

A. MPC's updating parameter setting

The MPC control strategy only requires setting the updating horizon h, i.e., the time interval at which the EMS re-computes a prediction and optimization over a 24-hour horizon. For this numerical example, we have set $h = 30$ minutes.

B. Scenario-based methods tuning

1) Scenario generation: To perform stochastic optimization, it is essential to generate a sufficient number of scenarios to capture daily variability. Leveraging historical data from the system we study, we apply the method proposed by [19] inspired by [1], to produce plausible scenarios that reflect the underlying distribution of the measurements.
In the following methods, we aim to strike a balance between accurately representing uncertainties and maintaining computational tractability. More specifically, this involves controlling the number of scenarios to avoid an exponential increase in complexity. A common approach consists in generating a large set of L equiprobable scenarios and then selecting a reduced subset of $S < L$ scenarios such that the Wasserstein distance to the original distribution is minimized. For this selection step, we use the Fast Forward Selection (FFS) algorithm (see Algorithm 2.1 in [13]). To further improve the quality of the reduced scenario tree, we refine this set by minimizing the nested distance [22] between the reduced and original trees. This refinement is achieved using a boosted version of Kovacevic and Pichler's reduction tree method (KP), as described in Algorithm 3 of [20], and originally introduced in [14].
In the following, we generate an initial set of $10\,000$ scenarios and reduce it to an optimized set of 225 scenarios for RPHA. The DSP model is simpler to solve and can directly handle $40\,000$ scenarios. Due to the limitations of our computational setup and the quadratic programming solver, the SDAP algorithm for solving the problem (DRO) struggles to handle a large number of scenarios, mainly because of the projection step onto the ambiguity set (see Definition 1). Therefore, we restricted our experiments to only $S = 100$ and $L = 400$ scenarios.

2) Rolling-horizon framework: The rolling-horizon we use is 24 hours, with $t_1 = 0{:}00$ and $t_2 = 24{:}00$, corresponding to a typical daily cycle of production and consumption, and thus of battery charging and discharging. Empirically, shorter horizons have been found to yield inferior results. The interval between two steps is $\Delta = 10$ minutes.

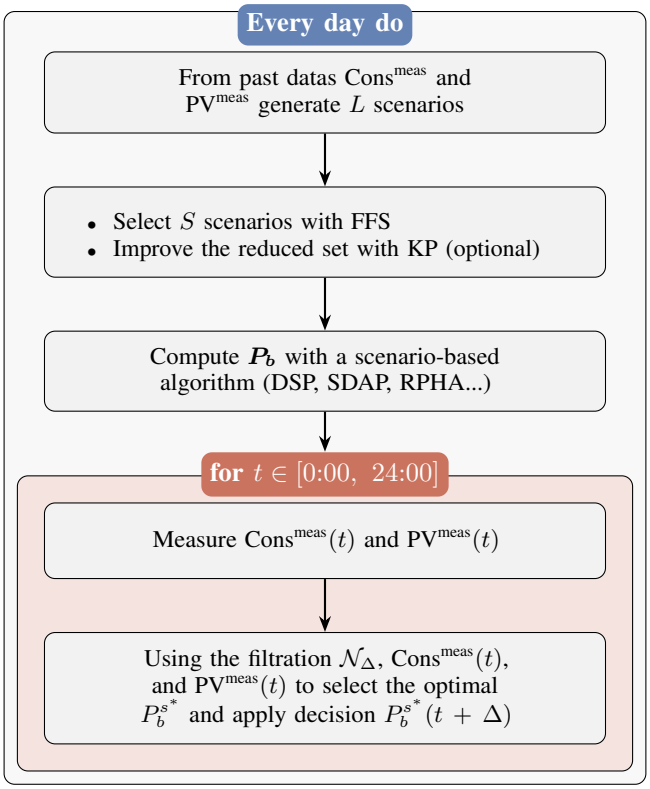
The control algorithm is described in the Process 1 above, where $\mathrm{Cons}^{\mathrm{meas}}, \mathrm{PV}^{\mathrm{meas}}$ denote, respectively, the electric consumption and photovoltaic production measured at the meter. These variables are deterministic in the sense that they represent a particular realization of a stochastic process.

3) Tuning hyperparameters: The optimization involves the tuning of hyperparameters: α for VSP and θ for DRO. They are selected through a cross-validation procedure in which Process 1 is run over a 60-day period, from 2024-05-04 to 2024-07-03, for various values of θ and α and with and without the optional step KP. The ground truth for electrical consumption and production is provided by the measured data $\mathrm{Cons}^{\mathrm{meas}}$ and $\mathrm{PV}^{\mathrm{meas}}$, respectively.

a) VSP: Figure 2 illustrates the impact of the combination of FFS and KP and identifies the best value of the hyperparameter α.

For VSP, it is preferable to apply only the FFS reduction method. Attempting to further align the reduced set with the original one using the KP algorithm does not improve performance and may even degrade it by constraining the diversity of the scenarios.

Process 1 Decision process for scenario-based methods



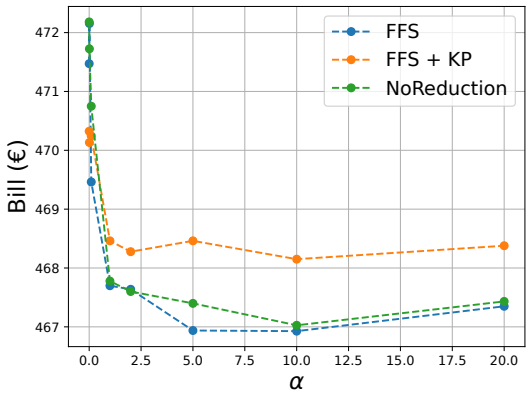


Fig. 2. Cross-validation VSP: Bill values over a 60 day period for different values of α, with or without the scenario reduction algorithm KP.

b) DRO: The results of this cross-validation are shown in Figure 3. As the radius of the ambiguity set varies, the level of robustness is adjusted, and there exists the best value of θ that minimizes the electricity bill. Figure 3 also illustrates that the combination of FFS and KP yields better performance compared to using FFS alone.

C. Out-of-sample tests

Finally, we evaluate and compare the performance of the various models—VSP (RPHA), SP, DSP, MPC, RL, and DRO (SDAP)—on the problem over a two-year period, from 2022-01-22 to 2024-01-22.

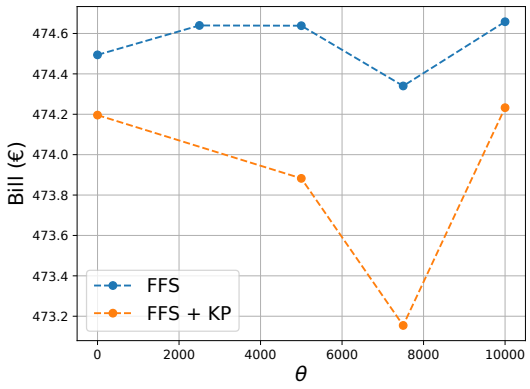


Fig. 3. Cross-validation DRO: Bill values over a 60 day period for different values of θ and with or without the scenario reduction algorithm KP.

Figure 4 presents the evolution of the performance ratio η, defined as:

$$\eta(\text{Day}) := 100 \times \frac{\rho - \text{Bill}}{\rho} \tag{25}$$

where ρ represents the battery-less electrical bill:

$$\rho := \int_{t_0}^{t_f} \mathrm{p}_r^c(t) \max\left(\text{Cons}^{\text{meas}}(t) - \text{PV}^{\text{meas}}(t), 0\right) \tag{26}$$

$$+ \mathrm{p}_r^d(t) \min\left(\text{Cons}^{\text{meas}}(t) - \text{PV}^{\text{meas}}(t), 0\right) dt. \tag{27}$$

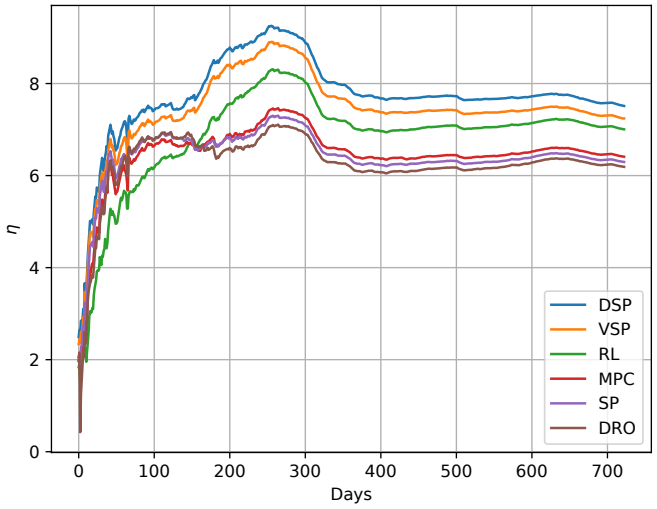


Fig. 4. Percentage reduction in electricity bill relative to storage-less baseline over two years for models: DSP, VSP (RPHA), SP, EMS, and DRO (SDAP).

Figure 4 highlights the strong performance of DSP, which outperforms all other models. The ideal performance that would be achieved with perfect foresight (solving a fully deterministic problem [3, Section 4] where both production and consumption are perfectly known) is a bill reduction of 12.5%. Thus, DSP reaches 62.5% of the perfect score. Its strong performance is likely due to its ability to leverage a very large number of scenarios, resulting in a highly robust solution. Besides it minimizes the expected cost over all the determinist solutions, making this model highly robust. Both the VSP and RL models

demonstrate superior results compared to the classical SP and MPC approaches. Notably, the RL model achieves competitive performance without requiring extensive fine-tuning or cross-validation, in contrast to other methods. At the begining RL does not train on a large dataset $\mathcal{D}$ this is why it displays poor results until a bit less than 200 days are trained on. The SP model outperforms the DRO approach in terms of bill reduction, this can be attributed to its implementation via PHA with a larger number of scenarios, leading to a more robust solution. In comparison, the SDAP-based DRO model is limited in that it only adjusts the scenario probabilities, leaving the scenario values unchanged. As a result, the model's ability to fine-tune robustness is constrained by the initial scenario set, potentially limiting its overall effectiveness.

An interesting observation is the contrasting efficiency trends during the summer of 2022: the performance of VSP and RL improve, while the standard SP's efficiency declines. This period coincided with unusually high spot electricity prices in France, driven by limited availability of nuclear power plants and elevated gas prices following the Russian invasion of Ukraine. Under such conditions, an effective control strategy must adopt a risk-averse stance to minimize costly electricity consumption. In this regard, the proposed robust approaches demonstrate greater risk aversion than the standard SP and yields better performance compared to the MPC strategy.

We define four criteria to help explain the differences in model performance, where $P_d := -\min(\rho_d P_b^{\text{meas}}, 0)$ and $P_c := \max(P_b^{\text{meas}}/\rho_c, 0)$:

- **Autoproduction gain ratio:** the amount of energy discharged from the battery that helps meet the demand. This gain is defined as

$$\texttt{PG} = 100 \times \frac{\int \texttt{min}\{P_d(t), \text{Cons}^{\text{meas}}(t) - \text{PV}^{\text{meas}}(t)\} \mathbb{1}_{\mathcal{C}_1}(t) \text{d}t}{\int \text{Cons}^{\text{meas}}(t) dt}. \tag{28}$$

  Where, $\mathcal{C}_1 : \text{PV}^{\text{meas}}(t) < \text{Cons}^{\text{meas}}(t)$.

- **Autoconsumption gain ratio:** the amount of surplus photovoltaic energy that is effectively stored in the battery. This gain is defined as

$$\texttt{CG} = 100 \times \frac{\int \texttt{max}\{P_c(t), \text{PV}^{\text{meas}}(t) - \text{Cons}\} \mathbb{1}_{\mathcal{C}_2}(t) dt}{\int \text{PV}^{\text{meas}}(t) dt}.$$

  Where, $\mathcal{C}_2 : \text{PV}^{\text{meas}}(t) > \text{Cons}^{\text{meas}}(t)$.

- **Discharging error ratio:** the amount of energy discharged from the battery that exceeds the energy need, and is somehow spoiled. This ratio is defined as

$$\texttt{DE} = 100 \times \frac{\int \left(P_d(t) - (\text{Cons}^{\text{meas}}(t) - \text{PV}^{\text{meas}}(t))\right) \mathbb{1}_{\mathcal{C}_3}(t) dt}{\int \text{Cons}^{\text{meas}}(t) dt}.$$

  Where, $\mathcal{C}_3 : \text{PV}^{\text{meas}}(t) < \text{Cons}^{\text{meas}}(t)$ and $\text{Cons}^{\text{meas}}(t) - \text{PV}^{\text{meas}}(t) < P_d(t)$.

- **Grid charging ratio:** the amount of energy charged to the battery directly from the grid. This ratio is defined as

$$\texttt{GC} = 100 \times \frac{\int (P_c(t) - (\text{PV}^{\text{meas}}(t) - \text{Cons}^{\text{meas}}(t))) \mathbb{1}_{\mathcal{C}_4}(t) dt}{\int \text{PV}^{\text{meas}}(t) dt}.$$

  Where, $\mathcal{C}_4 : \text{PV}^{\text{meas}}(t) > \text{Cons}^{\text{meas}}(t)$ and $\text{PV}^{\text{meas}}(t) - \text{Cons}^{\text{meas}}(t) < P_c(t)$.

These criteria are integrated over the two year period and these values are reported in Table I.

TABLE I
EVALUATION OF MODELS ACCORDING TO FOUR PERFORMANCE CRITERIA. VALUES IN PARENTHESES INDICATE THE PERCENTAGE DIFFERENCE RELATIVE TO MPC.

Method	CG	PG
DSP	1.2900 (49.7%)	1.4437 (12.2%)
VSP	0.9248 (7.3%)	1.3086 (1.7%)
RL	0.9189 (6.6%)	1.1824 (-8.1%)
MPC	0.8620 (0.0%)	1.2867 (0.0%)
DRO	0.8850 (2.7%)	1.2715 (-1.2%)
SP	0.8811 (2.1%)	1.2883 (0.1%)

Method	DE	GC
DSP	0.0378 (-65.3%)	2.1109 (-9.8%)
VSP	0.0783 (-28.1%)	2.2602 (-3.4%)
RL	0.0649 (-40.5%)	2.0080 (-14.2%)
MPC	0.1090 (0.0%)	2.3400 (0.0%)
DRO	0.1135 (4.1%)	2.2927 (-2.0%)
SP	0.1223 (12.2%)	2.3593 (+0.8%)

DSP achieves the highest autoproduction and autoconsumption gain ratios among all methods. This indicates that it stores surplus photovoltaic energy more effectively and makes better use of the battery to cover consumption when needed. Moreover, DSP has the lowest autoproduction loss ratio, meaning it makes fewer errors by discharging the battery when it is not necessary, compared to the other models. In addition, except for RL, DSP also has the smallest autoconsumption loss ratio, which suggests that it rarely charges the battery when there is insufficient solar production—therefore, it tends to purchase less electricity from the grid. According to Table I, the performance of scenario-based models is primarily driven by their ability to capture gains, whereas the performance of the RL model stems from its ability to limit losses. Indeed, RL exhibits low gain ratios, but also notably low loss ratios.

REFERENCES

[1] L. Amabile, D. Bresch-Pietri, G. El Hajje, S. Labbé, and N. Petit. Optimizing the self-consumption of residential photovoltaic energy and quantification of the impact of production forecast uncertainties. *Advances in Applied Energy*, 2:100020, 2021.

[2] F. Beltran, W. de Oliveira, and E. C. Finardi. Application of scenario tree reduction via quadratic process to medium-term hydrothermal scheduling problem. *IEEE Transactions on Power Systems*, 32(6):4351–4361, 2017.

[3] J. R. Birge and F. Louveaux. *Introduction to Stochastic Programming*. Springer Science & Business Media, 2011.

[4] D. Bousnina and G. Guerassimoff. Deep reinforcement learning for optimal energy management of multi-energy smart grids. In *International Conference on Machine Learning, Optimization, and Data Science*, pages 15–30. Springer, 2021.

[5] J.B. Caillau, R. Ferretti, E. Trélat, and H. Zidani. Chapter 15 - an algorithmic guide for finite-dimensional optimal control problems. In *Numerical Control: Part B*, volume 24 of *Handbook of Numerical Analysis*, pages 559–626. Elsevier, 2023.

[6] P. Carpentier, J.-P. Chancelier, G. Cohen, and M. De Lara. *Stochastic Multi-Stage Optimization: At the Crossroads between Discrete Time Stochastic Control and Stochastic Programming*, volume 75 of *Probability Theory and Stochastic Modelling*. Springer International Publishing, Cham, 2015.

[7] D. Cole, H. Sharma, and W. Wang. Contextual reinforcement learning for offshore wind farm bidding. *arXiv preprint arXiv:2312.10884*, 2023.

[8] W. de Oliveira. Risk-averse stochastic programming and distributionally robust optimization via operator splitting. *Set-Valued and Variational Analysis*, 29(4):861–891, 2021.

[9] C. Duan, W. Fang, L. Jiang, L. Yao, and J. Liu. Distributionally robust chance-constrained approximate ac-opf with Wasserstein metric. *IEEE Transactions on Power Systems*, 33(5):4924–4936, 2018.

[10] P. M. Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1):115–166, 2018.

[11] C. E. Garcia, D. M. Prett, and M. Morari. Model predictive control: Theory and practice—a survey. *Automatica*, 25(3):335–348, 1989.

[12] H. Golpira and S. A. R. Khan. A multi-objective risk-based robust optimization approach to energy management in smart residential buildings under combined demand and supply uncertainty. *Energy*, 170:1113–1129, 2019.

[13] H. Heitsch and W. Römisch. Scenario reduction algorithms in stochastic programming. *Computational Optimization and Applications*, 24:187–206, 2003.

[14] R. M. Kovacevic and A. Pichler. Tree approximation for discrete time stochastic processes: a process distance approach. *Annals of Operations Research*, 235(1):395–421, 2015.

[15] M. Y. Lamoudi. *Distributed model predictive control for energy management in buildings*. PhD thesis, Université de Grenoble, 2012.

[16] J. Li, Z. Xu, H. Liu, C. Wang, L. Wang, and C. Gu. A Wasserstein distributionally robust planning model for renewable sources and energy storage systems under multiple uncertainties. *IEEE Transactions on Sustainable Energy*, 14(3):1346–1356, 2022.

[17] Q. Li, X. Meng, F. Gao, G. Zhang, W. Chen, and K. Rajashekara. Reinforcement learning energy management for fuel cell hybrid systems: A review. *IEEE Industrial Electronics Magazine*, 17(4):45–54, 2022.

[18] P. Malisani. Interior point methods in optimal control. *ESAIM: Control, Optimisation and Calculus of Variations*, 30(59), 2024.

[19] P. Malisani, A. Spagnol, and V. Smis-Michel. Robust stochastic optimization via regularized PHA: Application to energy management systems. *arXiv preprint arXiv:2411.02015*, 2024.

[20] D. Mimouni, P. Malisani, J. Zhu, and W. de Oliveira. Scenario tree reduction via Wasserstein barycenters. *arXiv preprint arXiv:2411.14477*, 2024.

[21] F. Oldewurtel. *Stochastic model predictive control for energy efficient building climate control*. PhD thesis, ETH Zurich, 2011.

[22] G. C. Pflug and A. Pichler. A distance for multistage stochastic optimization models. *SIAM Journal on Optimization*, 22(1):1–23, 2012.

[23] R. T. Rockafellar and R. J.-B. Wets. Scenarios and policy aggregation in optimization under uncertainty. *Mathematics of Operations Research*, 16(1):119–147, 1991.

[24] F. Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs and Modeling*. Birkhäuser Cham, 2015.

[25] A. Shapiro. Distributionally robust modeling of optimal control. *Operations Research Letters*, 50(5):561–567, 2022.

[26] A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory*. Society for Industrial and Applied Mathematics, 2009.

[27] J. E. Smith and R. L. Winkler. The optimizer's curse: Skepticism and postdecision surprise in decision analysis. *Management Science*, 52(3):311–322, 2006.

[28] C. J. C. H. Watkins and P. Dayan. Q-learning. *Machine Learning*, 8:279–292, 1992.

[29] L. Weber, A. Bušić, and J. Zhu. Reinforcement learning based demand charge minimization using energy storage. In *Proceedings of the 62nd IEEE Conference on Decision and Control (CDC)*, pages 4351–4357. IEEE, 2023.

Daniel Mimouni received his engineering degree from Centrale Lyon and an MSc from Imperial College London. He is currently pursuing a PhD in Applied Mathematics at IFP Énergies nouvelles and Mines Paris (CMA). His research focuses on optimization under uncertainty for energy management, with interests in optimal transport, reinforcement learning, convex and stochastic optimization.

Jiamin Zhu received her Ph.D. degree in applied mathematics from Sorbonne University, France. She is currently a research scientist at IFP Énergies nouvelles. Her research interests include the development of reinforcement learning and distributed control algorithms, with a focus on applications in energy management systems.

Welington de Oliveira is an Associate Professor at the Centre de Mathématiques Appliquées, Mines Paris - PSL, France. He obtained his PhD in systems engineering and computer science from the Federal University of Rio de Janeiro, Brazil, and has a Habilitation in applied mathematics from Université Paris 1 Panthéon Sorbonne, France. Welington has extensive experience in nonsmooth optimization and stochastic programming, having published numerous research articles and served as an associate editor for several reputable journals in the field.

Paul Malisani received the Ph.D. degree in mathematics and control from Mines Paris, France. He is currently a research scientist at IFP Energies nouvelles, specializing in optimization for energy systems. His research interests include the development of optimization algorithms for energy management, with a focus on optimal control, stochastic programming, and non-convex optimization.