# Optimal Control Strategies for Seasonal Thermal Energy Storage Systems With Market Interaction

Jesus Lago, Gowri Suryanarayana, Ecem Sogancioglu, and Bart De Schutter, *Fellow, IEEE*

***Abstract*—Seasonal thermal energy storage systems (STESSs) can shift the delivery of renewable energy sources and mitigate their uncertainty problems. However, to maximize the operational profit of STESSs and ensure their long-term profitability, control strategies that allow them to trade on wholesale electricity markets are required. While control strategies for STESSs have been proposed before, none of them addressed electricity market interaction and trading. In particular, due to the seasonal nature of STESSs, accounting for the long-term uncertainty in electricity prices has been very challenging. In this article, we develop the first control algorithms to control STESSs when interacting with different wholesale electricity markets. As different control solutions have different merits, we propose solutions based on model predictive control and solutions based on reinforcement learning. We show that this is critical since different markets require different control strategies: MPC strategies are better for day-ahead markets due to the flexibility of MPC, whereas reinforcement learning (RL) strategies are better for real-time markets because of fast computation times and better risk modeling. To study the proposed algorithms in a real-life setup, we consider a real STESS interacting with the day-ahead and imbalance markets in The Netherlands and Belgium. Based on the obtained results, we show that: 1) the developed controllers successfully maximize the profits of STESSs due to market trading and 2) the developed control strategies make STESSs important players in the energy transition: by optimally controlling STESSs and reacting to imbalances, STESSs help to reduce grid imbalances.**

***Index Terms*—Demand response, electricity markets, model predictive control (MPC), optimal control, reinforcement learning (RL), seasonal storage systems.**

Manuscript received February 27, 2020; revised May 22, 2020; accepted August 5, 2020. Date of publication September 25, 2020; date of current version August 5, 2021. Manuscript received in final form August 10, 2020. This work was supported in part by the European Union's Horizon 2020 Research and Innovation Program under the Marie Skłodowska-Curie Grant Agreement 675318 (INCITE), in part by the Research Foundation—Flanders (FWO), and in part by the Flemish Government. Recommended by Associate Editor N. Quijano. *(Corresponding author: Jesus Lago.)*

Jesus Lago is with the Delft Center for Systems and Control, Delft University of Technology, 2628 CN Delft, The Netherlands, and also with the Department of Algorithms, Modeling and Optimization, VITO-Energyville, 3600 Genk, Belgium (e-mail: j.lagogarcia@tudelft.nl).

Gowri Suryanarayana is with the Department of Algorithms, Modeling and Optimization, VITO-Energyville, 3600 Genk, Belgium (e-mail: gowri.suryanarayana@vito.be).

Ecem Sogancioglu is with the Department of Radiology and Nuclear Medicine, Radboud University, 6525 GA Nijmegen, The Netherlands (e-mail: ecem.lago@radboudumc.nl).

Bart De Schutter is with the Delft Center for Systems and Control, Delft University of Technology, 2628 CN Delft, The Netherlands (e-mail: b.deschutter@tudelft.nl).

This article has supplementary downloadable material available at https://ieeexplore.ieee.org, provided by the authors.

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCST.2020.3016077

## I. Introduction

WHILE the energy transition [1] has the potential to highly improve our society, e.g., by mitigating climate change, it also poses some potential problems that need to be tackled [2]. Especially, due to the weather dependence of renewable sources, a large integration of renewables implies more uncertain energy generation. In the case of electricity, as generation and consumption have to be balanced at all times, the more renewable sources are integrated, the more imbalances between generation and consumption occur, and the more complex the control and balance of the electrical grid becomes [3]. In this context, energy storage systems offer a promising solution for uncertain generation by providing flexibility and ancillary services, leading to smooth and reliable grid operation [4].

### A. Energy Storage Systems

Depending on the type of technology, there are different energy storage solutions [4], [5], e.g., lithium-ion batteries, pumped hydrostorage, ultracapacitors, flywheels, molten-salt batteries, thermal storage systems, compressed air storage, or hydrogen storage. While most of these technologies can ensure efficient short- and medium-term energy storage, efficient long-term energy storage has traditionally been more difficult to achieve: although some of these technologies can store energy for long periods, they are not economically very efficient [4]. However, long-term energy storage is arguably one of the most important elements to ensure the success of the energy transition. Particularly, as the share of wind and solar energy by 2030 is expected to reach very high levels (70%–80% in some countries), and as the generation of renewables is seasonal dependent [5], seasonal energy storage solutions [5] that can store energy across several weeks or months are crucial in order to reduce seasonal fluctuations [4].

With regard to seasonal storage, there are primarily three solutions available that can provide electricity back to the grid: hydrogen storage, synthetic natural gas storage, and vanadium redox flow batteries [5], [6]. The first two approaches are power-to-gas technologies that make use of renewable sources to generate synthetic fuels, i.e., primarily hydrogen and methane [7]. The third belongs to the next generation of batteries that can potentially store electricity for long horizon [6], [8]. In this context, besides vanadium redox flow batteries, there is also undergoing research into the next generation of post-lithium-ion technologies with capabilities of long-term storage [9], [10]. Despite their potential, these technologies still have several problems that make them economically nonviable: first, they are expensive technologies and in an early stage of development [6]–[12]. Second, synthetic fuels have very low energy efficiency due to conversion losses [7]. Third, vanadium redox flow batteries and other postlithium-ion batteries are yet not profitable and face multiple challenges that difficult their commercial deployment [6], [9], [10], [13].

Another option for storing energy over long horizons are thermal energy storage (TES) systems [14]. While, in general, these systems cannot provide electricity back to the grid, they are a more mature technology, have the advantage of being significantly less expensive than electrical energy storage [4], and can be used to satisfy heating and cooling demands.

In the context of TES technologies, there are three main categories: sensible heat storage, latent heat storage, and chemical energy storage [14], [15]. While the last two have higher energy densities, they are both more expensive and less mature, i.e., sometimes at the laboratory testing stage and with no large-scale seasonal project completed [14]. By contrast, sensible energy storage is the simplest, cheapest, most widespread, and most mature technology [15]. As a result, sensible heat storage systems are the focus of this article. Note that, aligned with the literature [16]–[18], we use the name of seasonal TES systems (STESSs) to refer to TES systems based on sensible heat storage.

### B. Control of Nonseasonal Storage Systems

The problem of controlling storage systems is a developed area of research that contains many approaches that consider market interaction. However, within this context, all the research has usually focused on short-term storage systems, i.e., non-seasonal storage. The aim of this section is to provide a brief overview of the different families of approaches within the field, describe which markets the control algorithms are designed for, and which control horizons are usually considered. It is important to note that, since the number of contributions to this field is numerous, this will not be a thorough literature review but a brief summary of the research field.

Optimization-based approaches have been employed in numerous applications [19]–[27] and are arguably the most widely used family. In order to interact with different markets, these approaches are formulated as sequential multistage optimization problems. Another family of approaches is based on dynamic programming and Markov processes [28]–[30]. While these approaches often provide global optimal solutions, they do not scale for large systems [31]. The third family is rule-based approaches [23], [32] that derive a set of logical rules to control the storage systems. Finally, there are game-theoretical models [33] that are based on competition economic models.

In terms of markets, control approaches have been proposed in many different cases. The most common of them are trading in the day-ahead market together with the balancing market [19]–[21], [27], [33] or with the real-time market [24]–[26]. Other proposed strategies include frequency regulation coupled with energy arbitrage markets [29]; day-ahead market [30]; primary frequency response market [32]; real-time markets [22]; or day-ahead, intraday, and balancing markets [23], [28]. To the best of our knowledge, approaches that exploit the imbalance markets have not been proposed.

In terms of the horizon, the majority of the approaches perform price arbitrage between day-ahead and markets closer to real-time considering optimization horizons of one day [19]–[29]. In this context, no approaches provide solutions for trading energy over long horizons, e.g., months.

### C. Control of Seasonal Storage Systems

In the context of seasonal storage systems, several optimal control strategies have been also proposed. However, none of the proposed methods are designed for market interaction. In [17] and [34], model predictive control (MPC)-based strategies are proposed to control aquifer TES systems; however, while the controller is designed to satisfy physical constraints and stochastic heat demand, the STESS does not interact with electricity markets. Similarly, in [35], a dynamic programming approach is proposed to control borehole thermal storage systems; however, the controller assumes a constant market price and does not distinguish between different markets. In [18] and [36], two control algorithms are proposed to control solar communities with a borehole thermal storage system; however, similar to other studies, price and markets are not considered, and the controller is limited to satisfy the system constraints and the heat demand. In [16], a data-driven stochastic predictive control scheme to operate an energy hub with seasonal storage capabilities is proposed; the goal of the approach is to minimize the total energy consumption and be cost-efficient; however, here also, the algorithm does not consider real market prices nor market trading. Similarly, in [37], an optimal charging strategy for borehole thermal storage systems is proposed; however, the focus of the controller is to maximize the renewable energy use and to reduce $CO_2$ emissions, and also here, no prices nor market interaction are considered.

### D. Motivation of the Research

While the field of control for storage systems features several approaches, they are either limited to approaches for short-term storage with market interaction or seasonal storage without market interaction.

Generic methods for storage systems, while they model market interaction, cannot cope with long optimization horizons. Particularly, all the existing methods [19]–[21], [23], [27]–[29], [33] provide trading approaches where storage systems trade energy with daily/weekly horizons and use price differences to perform price arbitrage. This poses a challenge for seasonal storage systems, such as STESSs, where the optimization has to be performed over yearly horizons. The reason why the existing methods cannot be applied to STESSs is twofold:

1) STESSs require forecasts of electricity prices over yearly horizons. While there are several forecasting methods [3], [38], [39] for short-term horizons, i.e., days, there

are no reliable methods to forecasts for long-term horizons.

2) Because of the long optimization horizons, the number of variables in the optimization problems grows very large. In this context, the existing methods become computationally intractable, e.g., many of them are based on mixed-integer optimization.

In the context of control algorithms for seasonal storage, while long horizons are sometimes considered, none of the existing methods are able to model electricity market interaction. This interaction is of primary importance for several reasons:

1) To maximize the profit of STESSs, they should be allowed to interact with markets. In particular, while controlling STESSs to satisfy heat demand and/or to maximize renewable energy usage are important goals, they do not necessarily optimize the economic cost of STESSs. This is especially important to increase the number of storage systems in the electrical grid: if the time for return on investment of STESSs is too long, STESSs might become unattractive investments.
2) As we will show in this article, the profits of the STESSs are maximized when interacting with multiple markets. Therefore, controlling STESSs based on a single price or a single market is economically suboptimal.
3) To help reduce grid imbalances, STESSs need to be able to arbitrage in more than one market. In particular, to provide up-regulation in the imbalance markets, i.e., a real-time market, STESSs need to first buy that electricity in a market with an earlier gate closure time.

### *E. Contributions*

To fill the scientific gap described earlier, we present four contributions in this article:

- We propose and develop different control strategies for STESSs that can interact with multiple wholesale electricity markets. In particular, considering that there are several trading markets for STESSs, we propose control approaches for two cases: interaction with the day-ahead market alone and simultaneous interaction with the day-ahead and imbalance markets. In addition, as different control approaches have different merits, for each market interaction, we propose an MPC-based controller and an RL-based controller.
- We propose the first control algorithms for storage systems that consider long optimization horizons. Particularly, unlike the existing literature on seasonal storage systems, the proposed methods quantify the price variations and uncertainty over a horizon of a year and exploit these variations to maximize the profits of the storage system. In the case of the MPC approaches, this is obtained using a novel two-stage optimization problem, a forecasting method for long horizons, and a variable time grid formulation. In the case of the RL approaches, the solution involves a new simulation framework for long horizons and a collaborative RL strategy.
- We assess the merits of each control solution for the different markets and show that, while MPC-based methods are most suitable for day-ahead markets, RL-based methods perform better when trading in the imbalance market.
- Finally, we empirically demonstrate that STESSs can play an important role in the energy transition by helping grid operators to reduce grid imbalances. We show that the economic incentives of STESSs are aligned with the regulatory duties of the grid operators and that STESSs can help balancing the grid to allow further integration of renewable sources. To the best of our knowledge, this is the first time that trading on the imbalance market is explicitly evaluated from the perspective of balancing the grid and the regulatory duties of the TSO.

We also have two additional contributions: we propose a simple scenario generation method for generating long-term price scenarios and a novel method for imbalance price forecasting. This contribution refers specifically to forecasting imbalance prices and not real-time local marginal prices (LMPs). Although, for the latter, there are already forecasting methods [40], [41], imbalance prices have different properties than real-time LMPs and are much harder to predict.

### *F. Organization of This Article*

This article is organized as follows. Section III introduces and defines the framework of a general STESS interacting with electricity markets. Sections IV and V present, respectively, the proposed MPC and RL approaches. Finally, Section VI studies the performance of the proposed control approaches under several case studies and considering a real STESS. Appendix A in the Supplementary Material describes the proposed scenario generation method, Appendix B in the Supplementary Material explains the imbalance price forecasting method, Appendix C in the Supplementary Material introduces and defines wholesale electricity markets, and Appendix D in the Supplementary Material presents the theoretical basis of MPC and RL.

## II. Motivation for the Selected Methodology

Designing controllers for STESSs that trade in multiple electricity markets is a very challenging task as selecting the right control algorithms or right markets is not straightforward.

### *A. Control Algorithms for STESSs With Market Trading*

Considering the difficulty of market trading, state-of-the-art control approaches, e.g., MPC [42] or reinforcement learning (RL) [43], are highly desirable. However, in the case of MPC [42], several problems appear:

- MPC requires realistic forecasts and/or scenarios of electricity prices over yearly horizons. While there are several forecasting methods [3], [38], [39] and scenario generation methods [44]–[46] for short-term horizons, i.e., days, there are no reliable methods, to the best of our knowledge, to forecasts or generate scenarios for long-term horizons.

- In real-time electricity markets, e.g., imbalance markets, an action has to be taken within seconds. As the MPC works with a year horizon and the price resolution is typically 15 min, the number of variables in the optimization problem grows very large. As a result, MPC suffers from computational tractability problems to provide optimal action within the available time frame.

While data-driven and RL techniques can mitigate or solve these two issues, they also have problems of their own:

- While they do not require forecasts or scenarios of electricity prices, they need to generate artificial time series of electricity prices to simulate the market conditions. Thus, a method to generate realistic prices is still needed.
- As they are trained offline, they do not have the real-time computation issues of MPC. However, that comes at the cost of adaptability: if market conditions change or if the STESS suffers from a problem, e.g., a heat exchanger breaks, the controller has to be retrained again. As the training can take several days, this limits the adaptability of RL to changes in the environmental conditions. In contrast, as MPC computes the solution online, any change in the environment can be directly included as a change in the optimization problem or by reestimating the dynamical model with little impact on computation cost.
- The solutions of RL are at best a good approximation of the optimal solution, while MPC obtains an optimal solution by explicitly solving the given control problem.
- Unlike MPC, RL cannot explicitly model hard constraints (they can only be modeled as part of the reward). As such, RL cannot guarantee that the provided solutions do not violate constraints.

Based on these arguments, it becomes clear that the perfect method does not exist and considering RL or MPC involves several tradeoffs. As a result, for this research, we will propose different methods based on the two families and analyze the performance of each.

### B. Electricity Markets for Trading With STESSs

Another important point to consider is that not all electricity markets are the same. While, in theory, STESSs could trade in any electricity market, there are two trading strategies that are especially relevant: trading only in the day-ahead market and trading in both the day-ahead and the imbalance market. Trading only in the day-ahead market is arguably the safest trading strategy for STESSs as the day-ahead market is the electricity market with the largest volume of renewable energy trading, i.e., with low but volatile prices, and players incur no risks as they submit bidding curves.

While trading only in the day-ahead market is a low-risk and cost-effective trading strategy, it might still not be the most optimal economic strategy for STESSs. In particular, while, on average, prices in the imbalance market are larger than in the day-ahead market, since the imbalance prices are much more volatile, there are periods of time where imbalance prices are much lower (sometimes becoming even negative). In addition, by participating in the imbalance market, STESSs might be able to help reduce grid imbalances: as during periods of positive imbalances, i.e., generation larger than consumption, prices are low, STESSs could wait for these periods to buy their energy; by doing so, they would not only reduce grid imbalances but also increase their own profits. Similarly, as prices are high during periods of negative imbalances, STESSs can make use of their charging flexibility to first buy energy in the day-ahead market and then sell it in the imbalance market if imbalances are negative or use it if they are positive. By doing so, STESSs could potentially increase their profits while helping to reduce negative imbalances.

It is important to note that, despite all these potential benefits, trading strategies for the imbalance market have much higher risks: in the imbalance market, agents take an action for the next time interval without knowing the imbalance price. Particularly, as imbalance prices are based on the grid imbalances during a period of time, the price is only known after the period is over. Thus, trading strategies for the imbalance market heavily rely on price forecasters and have an associated risk.

In this article, we will explore both trading strategies, i.e., trading in just the day-ahead market and trading in both the day-ahead and imbalance markets, and study the benefits of each.

## III. Seasonal Storage System Framework

In order to introduce the control algorithms, we need to define the framework of a general STESS interacting with the electricity markets. For notational simplicity, concatenations of several vectors, e.g., $[x^\top, y^\top]^\top$, will be shortened as $(x, y)$.

### A. Dynamical Model

An STESS can be defined as a general dynamical system with an internal state $\mathbf{x}(t)$, controls $\mathbf{u}(t) = (\dot{\mathbf{Q}}^{\text{in}}(t), \dot{\mathbf{Q}}^{\text{out}}(t))$, $n_{\text{units}}$ storage units, and external disturbances $\mathbf{d}(t)$. The internal state $\mathbf{x}(t)$ represents the state of charge of the system. The controls $\dot{\mathbf{Q}}^{\text{in}}(t) \in \mathbb{R}^{n_{\text{in}}}$ and $\dot{\mathbf{Q}}^{\text{out}}(t) \in \mathbb{R}^{n_{\text{out}}}$, respectively, represent the rate at which energy is injected and extracted into/from the system. The disturbance represents any uncontrollable input, e.g., the external temperature.

The dynamics of the system are defined by a partial differential equation (PDE). For a sensible heat storage device with water stratification, the system can be divided into $n_{\text{units}}$ layers acting as individual storage units, and the dynamics of a layer $i$ are represented by the following PDE [47]:

$$\frac{\partial x_i}{\partial t} = a_1 \frac{\partial^2 x_i}{\partial z^2} + a_2(d - x_i) + a_3\big(\dot{Q}_i^{\text{in}} - \dot{Q}_i^{\text{out}}\big) \tag{1}$$

where $z$ represents the direction of stratification.

### B. Heat Demand and Purchased Energy

In general, an STESS is required to supply an uncertain heat demand $\dot{Q}^{\text{d}}(t)$. To do so, an STESS buys energy $\dot{Q}^{\text{m}}(t)$ from some market, stores it, and then delivers it to follow $\dot{Q}^{\text{d}}(t)$. To maximize the profits, it needs to consider the price of $\dot{Q}^{\mathrm{m}}(t)$, the storage efficiency, and an estimation of the future heat demand $\dot{Q}^{\mathrm{d}}(t)$. Therefore, the following holds:

$$\dot{Q}^{\mathrm{d}}(t) = \sum_{i=1}^{n_{\mathrm{out}}} \dot{Q}_i^{\mathrm{out}}(t), \qquad \dot{Q}^{\mathrm{m}}(t) = \sum_{i=1}^{n_{\mathrm{in}}} \dot{Q}_i^{\mathrm{in}}(t) \tag{2}$$

i.e., the heat demand should equal the sum of the energy extracted from the STESS, and the energy bought in the market should equal to sum of the energy introduced in the STESS.

### C. Trading in the Day-Ahead Market

Given a day-ahead market with unknown daily hourly prices $(\lambda_1^{\mathrm{dam}}, \ldots, \ldots, \lambda_{24}^{\mathrm{dam}})$, the goal of any control algorithm for an STESS is to build optimal bidding curves to maximize the profit. In particular, the aim is to, one day in advance, build 24 optimal bidding curves $\dot{Q}_1^{\mathrm{b}}(\cdot), \ldots, \dot{Q}_{24}^{\mathrm{b}}(\cdot)$ such that, while the STESS always has enough energy to satisfy the demand $\dot{Q}^{\mathrm{d}}(t)$, the cost of the purchased power $\dot{Q}^{\mathrm{dam}}(t)$ is minimized. In this market structure, the purchased power $\dot{Q}^{\mathrm{dam}}(t)$ at every hour $h$ is defined by

$$\dot{Q}^{\mathrm{dam}}(t) = \dot{Q}_h^{\mathrm{b}}\big(\lambda_h^{\mathrm{dam}}\big), \quad \forall\, t \in [h, h+1]. \tag{3}$$

### D. Trading in the Imbalance Market

For the imbalance market, the imbalance price $\lambda^{\mathrm{imb}}$ is always unknown when purchasing/selling power as the price $\lambda^{\mathrm{imb}}$ is determined in real time by the reserves activated by the TSO. In particular, at time step $k$, a market agent has to decide whether to sell, buy, or not trade without knowing the imbalance price $\lambda_k^{\mathrm{imb}}$ for the interval. As $\lambda_k^{\mathrm{imb}}$ is usually known immediately at the next interval, the agent can take the decision based on past imbalance prices $\lambda_{k-1}^{\mathrm{imb}}, \lambda_{k-2}^{\mathrm{imb}}, \ldots$ or any other information available at time step $k-1$.

Defining as $\dot{Q}^{\mathrm{imb}}(t)$ the energy traded in the imbalance market, with positive and negative values, respectively, representing energy that is bought and sold, it holds that

$$-\dot{Q}^{\mathrm{imb}}(t) \le \dot{Q}^{\mathrm{dam}}(t) \tag{4}$$

i.e., the energy sold in the imbalance market by an STESS is limited by the energy purchased on any previous market (the day-ahead market in the case of the proposed control algorithms). Particularly, because the STESS cannot effectively convert heat back to electricity, any energy sold is limited by the energy bought within the same day in other markets, and the STESS cannot sell any energy that was previously stored. Similarly, it holds that

$$\dot{Q}^{\mathrm{m}}(t) = \dot{Q}^{\mathrm{dam}}(t) + \dot{Q}^{\mathrm{imb}}(t) \tag{5}$$

i.e., the total energy purchased for the STESS is the sum of the energy purchased in the day-ahead and imbalance markets.

Considering these definitions, a control algorithm for the imbalance market has to select the value of $\dot{Q}^{\mathrm{imb}}(t)$ for each time step $k$ so that, while the STESS has enough energy to to satisfy the demand $\dot{Q}^{\mathrm{d}}(t)$, the total cost of trading $\dot{Q}^{\mathrm{dam}}(t)$ and $\dot{Q}^{\mathrm{imb}}(t)$ is minimized. To do so, the control algorithm receives as an input the energy $\dot{Q}^{\mathrm{dam}}(t)$ purchased in the day-ahead and selects the value of $\dot{Q}^{\mathrm{imb}}(t)$.

## IV. MPC Approaches

In this section, we derive and explain the two proposed MPC approaches: one to trade exclusively on the day-ahead market and the other one to trade on both the day-ahead and the imbalance market.

### A. Bidding Functions

In the case of the day-ahead electricity market, the goal of the MPC is to provide the 24 optimal bidding functions $\dot{Q}_h^{\mathrm{b}}(\cdot)$, for $h = 1, 2, \ldots, 24$. Since standard MPC can only provide the optimal market power $\dot{Q}_{\tilde{\lambda}}^{\mathrm{dam}}$ for a fixed price $\tilde{\lambda}$, an additional step is needed. For each hour $h$, we propose the following approach:

1) Predefine $n_{\mathrm{p}}$ discrete prices $\{\lambda^1, \lambda^2, \ldots, \lambda^{n_{\mathrm{p}}}\}$ for the price $\lambda^{\mathrm{dam}}$ at hour $h$.
2) Fix the remaining 23 day-ahead prices using their expected value, e.g., a forecast.
3) Solve the MPC for each of these $n_{\mathrm{p}}$ prices and obtain the associated optimal market powers $\{\dot{Q}_{\lambda^1}^{\mathrm{dam}}, \dot{Q}_{\lambda^2}^{\mathrm{dam}}, \ldots, \dot{Q}_{\lambda^{n_{\mathrm{p}}}}^{\mathrm{dam}}\}$ at hour $h$.
4) Build the bidding function as a piecewise constant function based on the obtained solutions

$$\dot{Q}_h^{\mathrm{b}}(\lambda^{\mathrm{dam}}) = \begin{cases} \dot{Q}_{\lambda^1}^{\mathrm{dam}}, & \lambda^{\mathrm{dam}} \le \lambda^1 \\ \dot{Q}_{\lambda^2}^{\mathrm{dam}}, & \lambda^1 < \lambda^{\mathrm{dam}} \le \lambda^2 \\ \vdots & \\ \dot{Q}_{\lambda^{n_{\mathrm{p}}}}^{\mathrm{dam}}, & \lambda^{n_{\mathrm{p}}-1} < \lambda^{\mathrm{dam}} \le \lambda^{n_{\mathrm{p}}} \\ 0, & \lambda^{n_{\mathrm{p}}} < \lambda^{\mathrm{dam}}. \end{cases} \tag{6}$$

This approach for building bidding functions is obviously only possible as long as the bidding functions within one day are independent of each other. However, since STESSs are very large storage devices, their internal state does not vary much within one day. As a result, the choice of one bidding function does not affect much the others, and the assumption of independent bidding functions holds in practice.

Moreover, due to the market structure and the long optimization horizons of STESS, the 24 bidding functions are very similar. In detail, as the 24 daily bids are submitted at the same time, all the bids are built based on the same information, e.g., the STESS state. Moreover, as the STESS is flexible, it does not matter at which hour of the day it buys energy: because of the large storage size of the STESS, the state of the STESS barely changes with the action taken in a given hour. As such, the STESS states between consecutive days never differ too much and, as the optimal bidding functions only depend on the STESS state, it follows that the optimal bidding functions for every hour of a given day are similar. As a result, in a given day, the difference in price distribution between hours is not important, and the STESS reacts almost equally to a market price independently of the hour, i.e.,

$$\dot{Q}_1^{\mathrm{b}}(\lambda^{\mathrm{dam}}) \approx \dot{Q}_2^{\mathrm{b}}(\lambda^{\mathrm{dam}}) \approx \cdots \approx \dot{Q}_{24}^{\mathrm{b}}(\lambda^{\mathrm{dam}}), \quad \forall\, \lambda^{\mathrm{dam}}. \tag{7}$$

Thus, to build the 24 bidding functions, it is only needed to obtain the bidding function $\dot{Q}_1^{\mathrm{b}}(\cdot)$ for the first hour.

### *B. MPC for Day-Ahead Trading*

As motivated in Section IV-A, we only need to estimate the bidding function $\dot{Q}_1^{\mathrm{b}}(\cdot)$ for the first hour of the day. However, instead of solving a single OCP like in standard MPC, we need to discretize the first price $\lambda_1^{\mathrm{dam}}$ into a discrete set of prices $\{\lambda^1, \lambda^2, \ldots, \lambda^{n_\mathrm{p}}\}$ and, for each of these prices, solve the relevant OCP.

For the sake of simplicity, in this section, we will assume that each OCP is optimized using a discrete time grid $t_1, t_2, \ldots, t_{N+1}$, i.e., using an optimization horizon equal to $t_{N+1} - t_1$; the details of how the time grid is defined will be covered in Section IV-D. Similarly, we will assume that the expected day-ahead prices $\{\bar{\lambda}_k^{\mathrm{dam}}\}_{k=1}^N$, the expected heat demand values $\{\bar{\dot{Q}}_k^{\mathrm{d}}\}_{k=1}^N$, and the expected disturbances $\{\bar{\mathbf{d}}_k\}_{k=1}^N$ are also provided; the method to obtain these values is explained in Appendix A in the Supplementary Material.

Considering the previous definitions, at every day and for each price $\lambda^j$, the MPC approach solves the following OCP.

OCP($\lambda^j$):

$$\min_{\substack{\mathbf{x}_1,\dot{\mathbf{Q}}_1^{\mathrm{in}},\dot{\mathbf{Q}}_1^{\mathrm{out}},\dot{Q}_1^{\mathrm{dam}},\mathbf{x}_2,\ldots,\\ \dot{\mathbf{Q}}_N^{\mathrm{in}},\dot{\mathbf{Q}}_N^{\mathrm{out}},\dot{Q}_N^{\mathrm{dam}},\mathbf{x}_{N+1}}} \lambda^j\, \dot{Q}_1^{\mathrm{dam}} + \sum_{k=2}^{N} \bar{\lambda}_k^{\mathrm{dam}}\, \dot{Q}_k^{\mathrm{dam}} \tag{8a}$$

s.t.

$$\mathbf{x}_1 = \tilde{\mathbf{x}}_1 \tag{8b}$$

$$\mathbf{x}_{k+1} = f\left(\mathbf{x}_k, \dot{\mathbf{Q}}_k^{\mathrm{in}}, \dot{\mathbf{Q}}_k^{\mathrm{out}}, \bar{\mathbf{d}}_k\right), \qquad \text{for } k = 1, \ldots, N \tag{8c}$$

$$\dot{Q}_k^{\mathrm{dam}} \leq \dot{Q}_{\max}^{\mathrm{m}}, \qquad \text{for } k = 1, \ldots, N \tag{8d}$$

$$\sum_{i=1}^{n_{\mathrm{in}}} \dot{Q}_{k,i}^{\mathrm{in}} = \dot{Q}_k^{\mathrm{dam}}, \qquad \text{for } k = 1, \ldots, N \tag{8e}$$

$$\sum_{i=1}^{n_{\mathrm{out}}} \dot{Q}_{k,i}^{\mathrm{out}} = \bar{\dot{Q}}_k^{\mathrm{d}}, \qquad \text{for } k = 1, \ldots, N \tag{8f}$$

$$0 \leq \dot{\mathbf{Q}}_k^{\mathrm{in}} \leq g_{\mathrm{in}}(\mathbf{x}_k), \qquad \text{for } k = 1, \ldots, N \tag{8g}$$

$$0 \leq \dot{\mathbf{Q}}_k^{\mathrm{out}} \leq g_{\mathrm{out}}(\mathbf{x}_k), \qquad \text{for } k = 1, \ldots, N \tag{8h}$$

$$\mathbf{x}_{\min} \leq \mathbf{x}_k \leq \mathbf{x}_{\max}, \qquad \text{for } k = 1, \ldots, N \tag{8i}$$

$$\mathbf{x}_{N+1} = \tilde{\mathbf{x}}_1 \tag{8j}$$

where the following holds:

- The objective function represents the cost of purchasing energy considering that the first price is fixed and given by $\lambda^j$ and that the remaining prices on the horizon are the expected prices in the market.
- Equation (8b) fixes the initial state, which is assumed to be known and given by $\tilde{\mathbf{x}}_1$.
- Equation (8c) ensures that the dynamics of the system are ensured at every time step. To discretize the continuous PDE, i.e., (1), we consider an explicit Euler integration scheme [47] as it provides a good tradeoff between speed and accuracy for long optimization horizons
- To model the discrete dynamics, a multiple shooting [48] scheme is used. Unlike single shooting, multiple shooting explicitly includes the state $x$ in the optimization problem. This is done to obtain a sparse Hessian and an easier to optimize problem.
- The maximum power purchased from the market is limited by (8d).
- Equation (8e) ensures that the input power equals the power purchased from the market.
- Through (8f), it is ensured that the heat demand is met.
- Equations (8g) and (8h) ensure the individual charging and discharging limits of each individual storage device. The upper limit is usually a function of the state as the maximum power that can be charged/discharged usually depends on the state of charge.
- The limits on the STESS state are defined by (8i).
- The OCP should avoid depleting the STESS at the end of the horizon. To do so, as the optimization horizon is usually a seasonal periodic cycle (see Section IV-D for details), (8j) constrains the STESS to have the same state of charge at the beginning and at the end.
- The objective function is simplified to leave out some costs, e.g., maintenance costs, startup costs, or utility costs. Simplifying the objective function to only include the market cost is a design choice motivated by two reasons: first, some of these costs are orders of magnitude lower than the market cost.[1] Second, some costs simply offset the profitability by a constant or a scaling factor and are not relevant for the control algorithm.

After solving an OCP for each discrete price $\lambda^{\mathrm{j}}$, the optimal bidding function $\dot{Q}_1^{\mathrm{b}}(\cdot)$ can be estimated using (6), where the optimal market power $\dot{Q}_{\lambda^{\mathrm{j}}}^{\mathrm{dam}}$ equals $\dot{Q}_1^{\mathrm{dam}}$.

### *C. MPC for Day-Ahead and Imbalance Trading*

The MPC-based approach to trade in both the day-ahead and the imbalance market consists of two separate MPC algorithms that run one after the other:

- A first MPC algorithm that trades in the day-ahead market, but, unlike the MPC algorithm defined in Section IV-B, it considers that there is also a possible future interaction with the imbalance market.
- A second MPC algorithm that trades in the imbalance market and that considers that there are also possible future interactions with the day-ahead market. However, unlike the MPC algorithm for the day-ahead market, it runs in real time, and it does not build bidding functions. Instead, at time step $k-1$, it considers a forecast $\hat{\lambda}_k^{\mathrm{imb}}$ of the next imbalance price and then solves a single OCP to obtain the optimal power $\dot{Q}_k^{\mathrm{imb}}$ to trade in the imbalance market.

It is important to note that, as with the day-ahead market, both algorithms are based on deterministic MPC. Given the uncertainty in electricity prices, one could argue that a more optimal approach would be to employ stochastic MPC. However, due to the long horizons involved, the computation time required for stochastic MPC makes the approach infeasible for real-time applications (especially for the imbalance market). Particularly, for trading in the imbalance market, the MPC approach already requires (in the deterministic setting) to approximate the one-year horizon to one month, i.e., the

[1] This information was obtained from the case study site company.

obtained optimal solution is approximated and no longer optimal with respect to the yearly seasonal period. A stochastic setting would only make this approximation worse. While larger computation capabilities could perhaps mitigate the issue, there are is another problem: as MPC solves a nonconvex problem, there is no guarantee on the maximum computation time, and more computational power might not help much.

### *C.1 MPC for the Day-Ahead Market*

To define the OCP of the first MPC algorithm, we will again consider that the discrete time grid $t_1, t_2, \ldots, t_{N+1}$, the expected day-ahead prices $\{\bar{\lambda}_k^{\mathrm{dam}}\}_{k=1}^N$, imbalance prices $\{\bar{\lambda}_k^{\mathrm{imb}}\}_{k=1}^N$, heat demand values $\{\bar{\dot{Q}}_k^{\mathrm{d}}\}_{k=1}^N$, and disturbances $\{\bar{\mathbf{d}}_k\}_{k=1}^N$ are given. In addition, we will simplify the vector of input controls by $\mathbf{u}_k = (\dot{\mathbf{Q}}_k^{\mathrm{in}}, \dot{\mathbf{Q}}_k^{\mathrm{out}}, \dot{Q}_k^{\mathrm{dam}}, \dot{Q}_k^{\mathrm{imb}})$. Then, at every day and for each discrete price in $\{\lambda^1, \lambda^2, \ldots, \lambda^{n_{\mathrm{p}}}\}$, the MPC solves the following OCP.

OCP($\lambda^j$):

$$\min_{\substack{\mathbf{x}_1,\mathbf{u}_1,\mathbf{x}_2,\ldots,\\ \mathbf{u}_N,\mathbf{x}_{N+1}}} \lambda^j\, \dot{Q}_1^{\mathrm{dam}} + \sum_{k=2}^{N} \bar{\lambda}_k^{\mathrm{dam}}\, \dot{Q}_k^{\mathrm{dam}} + \sum_{k=1}^{N} \bar{\lambda}_k^{\mathrm{imb}}\, \dot{Q}_k^{\mathrm{imb}} \tag{9a}$$

s.t.

$$\mathbf{x}_1 = \tilde{\mathbf{x}}_1 \tag{9b}$$

$$\mathbf{x}_{k+1} = f\big(\mathbf{x}_k, \dot{\mathbf{Q}}_k^{\mathrm{in}}, \dot{\mathbf{Q}}_k^{\mathrm{out}}, \bar{\mathbf{d}}_k\big), \qquad \text{for } k = 1, \ldots, N \tag{9c}$$

$$\dot{Q}_k^{\mathrm{dam}} + \dot{Q}_k^{\mathrm{imb}} \leq \dot{Q}_{\max}^{\mathrm{m}}, \qquad \text{for } k = 1, \ldots, N \tag{9d}$$

$$\sum_{i=1}^{n_{\mathrm{in}}} \dot{Q}_{k,i}^{\mathrm{in}} = \dot{Q}_k^{\mathrm{dam}} + \dot{Q}_k^{\mathrm{imb}}, \qquad \text{for } k = 1, \ldots, N \tag{9e}$$

$$\sum_{i=1}^{n_{\mathrm{out}}} \dot{Q}_{k,i}^{\mathrm{out}} = \bar{\dot{Q}}_k^{\mathrm{d}}, \qquad \text{for } k = 1, \ldots, N \tag{9f}$$

$$0 \leq \dot{\mathbf{Q}}_k^{\mathrm{in}} \leq g_{\mathrm{in}}(\mathbf{x}_k), \qquad \text{for } k = 1, \ldots, N \tag{9g}$$

$$0 \leq \dot{\mathbf{Q}}_k^{\mathrm{out}} \leq g_{\mathrm{out}}(\mathbf{x}_k), \qquad \text{for } k = 1, \ldots, N \tag{9h}$$

$$\mathbf{x}_{\min} \leq \mathbf{x}_k \leq \mathbf{x}_{\max}, \qquad \text{for } k = 1, \ldots, N \tag{9i}$$

$$\dot{Q}_k^{\mathrm{dam}} \geq 0, \qquad \text{for } k = 1, \ldots, N \tag{9j}$$

$$-\dot{Q}_k^{\mathrm{dam}} \leq \dot{Q}_k^{\mathrm{imb}}, \qquad \text{for } k = 1, \ldots, N \tag{9k}$$

$$\mathbf{x}_{N+1} = \tilde{\mathbf{x}}_1. \tag{9l}$$

While the main structure is very similar to (8a), there are some important differences:

- The algorithm minimizes the cost of purchasing energy as in (8a) but includes future transactions in the imbalance market.
- The constraints that contain the power purchased from the market, i.e., (9d) and (9e), consider now the sum of the power purchased in both markets.
- Unlike the case of trading only in the day-ahead market, the STESS can now sell energy on the imbalance market if it has previously bought it in the day-ahead market. This is modeled by (9j) and (9k), which respectively guarantee that in the day-ahead market energy can only be bought and that the energy sold in the imbalance market is limited to the energy bought in the day-ahead market.
- The amount of energy traded is not limited by the system demand. In particular, the total energy traded is limited by $\dot{Q}_{\max}^{\mathrm{m}}$, which represents a safety upper bound that can be much larger than the heat demand $\dot{Q}^{\mathrm{d}}$ and that simply models how risk-averse the STESS is to price arbitration.

### *C.2 MPC for the Imbalance Market*

To define the second MPC algorithm, let us first make the following assumptions and definitions:

- The MPC algorithm for the imbalance market considers a new time grid $t'_1, t'_2, \ldots, t'_{N_1+1}$ with $t'_{N_1+1} \leq t_{N+1}$, i.e., a shorter horizon and a different discretization. The details on this discretization are provided in Section IV-D.
- The expected day-ahead prices $\{\bar{\lambda}_k^{\mathrm{dam}}\}_{k=1}^{N_1}$, imbalance prices $\{\bar{\lambda}_k^{\mathrm{imb}}\}_{k=1}^{N_1}$, heat demand values $\{\bar{\dot{Q}}_k^{\mathrm{d}}\}_{k=1}^{N_1}$, and disturbances $\{\bar{\mathbf{d}}_k\}_{k=1}^{N_1}$ are again provided (see Appendix A in the Supplementary Material for details).
- The optimal state at time $t_{N_1+1}$ is defined by $\mathbf{x}_{N_1+1}^{\star}$ and obtained from the solution of the MPC for the day-ahead market. In particular, this value can be obtained from the optimal solution of any of the $n_{\mathrm{p}}$ OCPs solved in the latest day-ahead market.
- An accurate forecast $\hat{\lambda}_1^{\mathrm{imb}}$ of the next price in the imbalance market is available. The details of this forecast are explained in Appendix B in the Supplementary Material.

Based on these definitions, before each imbalance market clearance, MPC solves the following OCP and trades the optimal solution $\dot{Q}_1^{\mathrm{imb}}$ in the imbalance market:

$$\min_{\substack{\mathbf{x}_1,\mathbf{u}_1,\mathbf{x}_2,\ldots,\\ \mathbf{u}_{N_1},\mathbf{x}_{N_1+1}}} \hat{\lambda}_1^{\mathrm{imb}}\, \dot{Q}_1^{\mathrm{imb}} + \sum_{k=1}^{N_1} \bar{\lambda}_k^{\mathrm{dam}}\, \dot{Q}_k^{\mathrm{dam}} + \sum_{k=2}^{N_1} \bar{\lambda}_k^{\mathrm{imb}}\, \dot{Q}_k^{\mathrm{imb}} \tag{10a}$$

s.t.

$$(9\mathrm{b}) - (9\mathrm{k}) \tag{10b}$$

$$\mathbf{x}_{N_1+1} = \mathbf{x}_{N_1+1}^{\star}. \tag{10c}$$

The new MPC scheme is very similar to the previous MPC for the day-ahead market but with some differences:

- As a bidding function is not needed, instead of solving the OCP multiple times for different possible prices, this MPC algorithm solves a single OCP considering the most likely imbalance price $\hat{\lambda}_1^{\mathrm{imb}}$ in the next market clearance. Then, it trades directly the optimal solution $\dot{Q}_1^{\mathrm{imb}}$ in the imbalance market.
- A distinction is made between the future expected imbalance prices $\{\bar{\lambda}_k^{\mathrm{imb}}\}_{k=2}^{N_1}$ and the forecast price $\hat{\lambda}_1^{\mathrm{imb}}$ in the next time step. This distinction is made because the accuracy of the forecast is better than that of the method used to generate the expected future values.
- As this MPC algorithm runs in real time, the computation time should be as small as possible. To reduce the computation time, a smaller horizon $t'_{N_1+1} < t_{N+1}$ is considered.
- As the optimization horizon $t'_{N_1+1}$ is now smaller than a periodical seasonal cycle, it is suboptimal to constrain the final state to be equal to the initial state. However, not constraining the final state leads to an OCP that does not account for what happens after $t'_{N_1+1}$. To solve this problem, (10c) constrains the final state to be equal to the optimal state $\mathbf{x}_{N_1+1}^{\star}$ at time $t'_{N_1+1}$, which is obtained from the solution of the latest day-ahead MPC algorithm.

### D. Time Grid and Optimization Horizon

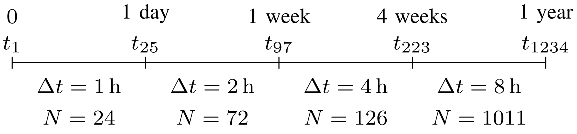
In Sections IV-B and C, we assumed that the discrete time grids where the OCPs were defined were given. In this section, we explain the methodology to define these time grids.

In general, to define a discrete time grid, we also need to define the optimization horizon $T$ and the discrete time step $\Delta t$. Then, based on $T$ and $\Delta t$, the number of time intervals $N$ is also defined. For an STESS, $T$ represents its seasonal horizon, which is typically a year. While most applications consider a constant $\Delta t$ along the optimization horizon, we argue that, for an STESS, this is not necessary and should in fact be avoided:

- As day-ahead markets have a different price every hour, the largest possible time step at the beginning of the horizon is $\Delta t = 1\,\mathrm{h}$. However, due to the long optimization horizons, it is not possible to accurately estimate with hourly resolutions the price and demand distributions at the end of the horizon. Instead, it is better to estimate the distributions over larger intervals, e.g., several hours, where due to noise averaging the uncertainty can be better quantified.
- Another reason to consider a variable $\Delta t$ is the computational cost: by increasing $\Delta t$ toward the end of the horizon we reduce the number of optimization points $N$ and the computational complexity of the OCP.
- As MPC only needs the optimal control at the first time point, it can be argued that lowering the time resolution at the end of the horizon has little impact on the first optimal control.

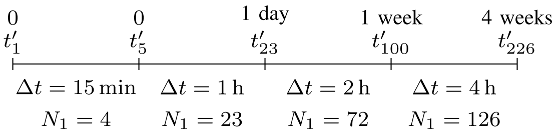
#### D.1 Day-Head Market

Considering that the day-ahead electricity market is cleared every day, the hourly resolution should only be needed for the first day. Based on this and the arguments above, for the day-ahead market MPC, we consider a time grid $t_1, t_2, \ldots, t_{N+1}$ with a year horizon, using four different $\Delta t$, and containing $N = 1233$ time intervals:


0
1 day
1 week
4 weeks
1 year
$t_1$
$t_{25}$
$t_{97}$
$t_{223}$
$t_{1234}$
$\Delta t = 1\,\mathrm{h}$
$N = 24$
$\Delta t = 2\,\mathrm{h}$
$N = 72$
$\Delta t = 4\,\mathrm{h}$
$N = 126$
$\Delta t = 8\,\mathrm{h}$
$N = 1011$

#### D.2 Imbalance Market

For the case of the imbalance market, the minimum $\Delta t$ is 15 min. Moreover, considering the large uncertainty in imbalances prices, we argue that the 15-min resolution is only needed for the first hour. Finally, as the MPC algorithm for the imbalance market runs in real time, the computation time should be as small as possible. Based on these arguments, we consider a time grid $t'_1, t'_2, \ldots, t'_{N_1+1}$ for the imbalance market with a horizon of four weeks, using four different $\Delta t$, and containing $N_1 = 225$ time intervals:


0
0
1 day
1 week
4 weeks
$t'_1$
$t'_5$
$t'_{23}$
$t'_{100}$
$t'_{226}$
$\Delta t = 15\,\mathrm{min}$
$N_1 = 4$
$\Delta t = 1\,\mathrm{h}$
$N_1 = 23$
$\Delta t = 2\,\mathrm{h}$
$N_1 = 72$
$\Delta t = 4\,\mathrm{h}$
$N_1 = 126$

It could be argued that considering a horizon of four weeks instead of a year (the standard seasonal cycle) leads to suboptimal solutions, i.e., the MPC cannot account for what happens during a full seasonal cycle. However, as explained in Section IV-C, MPC avoids this by constraining the state at the end of the four weeks to be equal to the optimal state $\mathbf{x}^\star_{226}$ at that time point.

## V. RL Approaches

In this section, we present the two proposed RL approaches: one to trade in the day-ahead market and the other one to trade in both the day-ahead and the imbalance markets.

### A. RL for Day-Ahead Trading

As with MPC, any RL control algorithm for the day-ahead market needs to estimate the bidding functions $\dot{Q}^{\mathrm{b}}_h(\cdot)$ for $h = 1, 2, \ldots, 24$. While in the case of MPC that required discretizing prices and solving multiple OCPs, for RL, the bidding functions can be directly obtained from the optimal policy $\pi^\star(\mathbf{s}_k)$. In detail, if the RL agent is set up, the following holds:

- The reward represents the cost of purchasing energy.
- The RL state $\mathbf{s}$ contains the day-ahead price $\lambda^{\mathrm{dam}}$.
- The action $\mathbf{u}$ includes the power $\dot{Q}^{\mathrm{dam}}$ purchased from the market.

Then, by definition, the bidding function $\dot{Q}^{\mathrm{b}}(\lambda^{\mathrm{dam}})$ is implicitly defined by the optimal policy $\mathbf{u}^\star = \pi^\star(\mathbf{s}) = \pi^\star(\lambda^{\mathrm{dam}}, \ldots)$. In the following, we provide further details on the proposed RL algorithm.

#### A.1 State and Control Spaces

The first step to define the RL algorithm is to define its state and control spaces. For the proposed algorithm, the state $\mathbf{s} = (\mathbf{x}, \tau, \lambda^{\mathrm{dam}})$ is defined by three different features:

1) the state $\mathbf{x}$ of the STESS.
2) the time position $\tau$ within the periodic seasonal cycle, e.g., the day of the year.
3) the market price $\lambda^{\mathrm{dam}}$.

The reason for selecting these three features is twofold:

- The optimal action $\mathbf{u}^\star = \pi^\star(\mathbf{s})$ can be selected based on both the state of the STESS and the environment.
- As we will show in Section V-A, given a fixed time point $\tilde{\tau}$ and STESS state $\tilde{\mathbf{x}}$, the bidding function $\dot{Q}^{\mathrm{b}}(\lambda^{\mathrm{dam}})$ is by definition given by the optimal policy $\tilde{\pi}^\star(\tilde{\mathbf{x}}, \tilde{\tau}, \lambda^{\mathrm{dam}})$.

To define the action space $\mathbb{U}$, we consider that a single action $u \in \mathbb{R}^{n_{\mathrm{in}}+1}$ has the following format:

$$\mathbf{u} = (u_1, u_2, \ldots, u_{n_{\mathrm{in}}}, j). \tag{11}$$

In detail, we consider that each input control $u_i$ can take $n_{\mathrm{dis}} + 1$ discrete values uniformly separated between 0 and 1 and that the real power $\dot{Q}^{\mathrm{in}}_i$ into the storage $i$ is obtained by multiplying $u_i$ by the maximum power $\dot{Q}^{\max}_i$, i.e., $\dot{Q}^{\mathrm{in}}_i = u_i\,\dot{Q}^{\max}_i$. This scaling is done because $\dot{Q}^{\max}_i$ might depend on the system state and can change throughout time. For the output control, a single storage unit $j$ is selected to provide the demand $\dot{Q}^{\mathrm{d}}$, i.e., $\dot{Q}^{\mathrm{d}} = \dot{Q}^{\mathrm{out}}_j$. The action space is then defined by the possible combinations of all these values.

### *A.2 Reward Function*

The reward $r_k$ at time step $k$ is defined as the negative of the cost of the energy purchased. Thus, assuming that the agent is at state $\mathbf{s}_k = (\mathbf{x}_k, \tau_k, \lambda_k^{\mathrm{dam}})$ and takes an action $\mathbf{u}_k = (u_{1,k}, \ldots, u_{n_{\mathrm{in}},k}, j)$, $r_k$ is defined as $-\lambda_k^{\mathrm{dam}} \sum_{i=1}^{n_{\mathrm{in}}} (u_{i,k} \cdot \dot{Q}_{i,k}^{\max})$. In addition, if the agent depletes the system and the demand $\dot{Q}_k^{\mathrm{d}}$ cannot be satisfied, the reward penalizes this situation with a cost of ten times larger than the cost of instantaneously buying $\dot{Q}_k^{\mathrm{d}}$ in the market.[2] Finally, as with standard RL algorithms, the reward at the last point in an episode is 0

$$r_k = \begin{cases} 0, & \text{If } k = T_{\mathrm{e}} \\ -\lambda_k^{\mathrm{dam}} \left( \sum_{i=1}^{n_{\mathrm{in}}} (u_{i,k}\, \dot{Q}_{i,k}^{\max}) + 10\, \dot{Q}_k^{\mathrm{d}} \right), & \text{If the system} \\ & \text{is depleted} \\ -\lambda_k^{\mathrm{dam}} \sum_{i=1}^{n_{\mathrm{in}}} (u_{i,k} \cdot \dot{Q}_{i,k}^{\max}), & \text{Otherwise.} \end{cases} \tag{12}$$

### *A.3 Episode Length and Time Grid*

Another critical point when designing an RL algorithm is to select the episode length $T_{\mathrm{e}}$. For STESS, it can be argued that, to avoid optimal policies that deplete the STESS, the minimum $T_{\mathrm{e}}$ should be two seasonal periodic cycles. In particular, if the episode length equals the cycle length, the agent would know the time position within an episode as the agent knows the time position $\tau$ within a seasonal cycle. Using that information, the agent could potentially deplete the STESS at the end of the episode/cycle to reduce the cost. This behavior would be undesirable as the STESS needs to provide energy for more than a seasonal periodical cycle.

For the size of the discrete time grid, we consider that a time transition $k \to k+1$ spans a day. In particular, as with MPC, it is assumed that the state of charge does not change dramatically from one day to another and that the optimal bidding curves within a day are very similar. It is important to note that selecting this time step size is just a design choice and that it is equally possible to consider time steps of 1 h at the expense of increasing the computation load.

### *A.4 Simulation Environment*

To train an RL agent to control STESSs, we use a simulation environment that recreates the world an STESS lives in. In detail, this environment consists of two modules:

- *STESS Simulator:* A simulator of the dynamical model of the STESS: $\mathbf{x}_{k+1}= f(\mathbf{x}_k, \dot{\mathbf{Q}}_k^{\mathrm{in}}, \dot{\mathbf{Q}}_k^{\mathrm{out}}, \bar{\mathbf{d}}_k)$.
- *Environment Simulator:* A simulator that produces realistic day-ahead market prices $\lambda^{\mathrm{dam}}$, heat demand values $\dot{Q}^{\mathrm{d}}$, and disturbances $\mathbf{d}$. To obtain a simulator that generates realistic time series, the method for scenario generation explained in Appendix A in the Supplementary Material is considered.

### *A.5 Training Algorithm*

The last step before training the agent is to select the specific RL algorithm to estimate the optimal policy $\pi^\star(\mathbf{s})$. For the case of STESSs, we propose using fitted-Q-iteration [43], [49] with boosting trees [50]. The reason for selecting this algorithm is that we empirically observed (using the real system presented in Section VI) that this algorithm performed as good as more advanced RL algorithms but without the additional computational complexity. Unlike the deterministic MPC approach, price uncertainty is implicitly included in this approach as the RL agent is trained with a probabilistic reward. Therefore, the RL agent can learn some notion of risk that quantifies the distribution of a reward for a given state.

### *A.6 Building Bidding Functions*

After the RL agent is trained, the optimal bidding functions $\dot{Q}^{\mathrm{b}}(\cdot)$ are directly obtained. In particular, given a fixed time point $\tilde{\tau}$ and STESS state $\tilde{\mathbf{x}}$, we have an optimal policy $\mathbf{u}^\star = \pi^\star(\tilde{\mathbf{x}}, \tilde{\tau}, \lambda^{\mathrm{dam}}) = \tilde{\pi}^\star(\lambda^{\mathrm{dam}})$ that selects the power purchased from the market as function of the market prices; thus, by definition, $\dot{Q}^{\mathrm{b}}(\lambda^{\mathrm{dam}})$ is directly defined by $\tilde{\pi}^\star(\lambda^{\mathrm{dam}})$.

## *B. RL for Day-Ahead and Imbalance Trading*

As with MPC, the RL-based approach to trade in both markets consists of two separate RL algorithms:

- The first RL algorithm that trades with the day-ahead market. This is the algorithm proposed in Section V-A, and it is agnostic of what happens in the imbalance market.
- The second RL algorithm that trades in the imbalance market and that considers the interaction with the day-ahead market. This algorithm runs in real time and it does not build bidding functions.

### *B.1 Training Multiple RL Agents*

As each electricity market has its own rules and working principles, it is clear that a different RL agent for each market is needed. As an example, an RL agent for the imbalance market has a different state $\mathbf{s}$ as it knows more information than the agent for the day-ahead market, e.g., it knows the prices and allocations of the day-ahead market.

Based on this premise, when using RL to trade in two electricity markets, the problem becomes a multiagent RL problem [51]. More specifically, as both agents are trying to minimize the economic cost, it becomes a collaborative multiagent RL problem [52], [53].

While the literature has several methods for collaborative RL, e.g., join-action learners [52], we argue that the available methods might not be very suitable for the case of STESS. In particular, when training several agents at the same time, the environment becomes nonstationary [51], i.e., as each agent improves and changes its own policy the environment that the other agents perceive changes as well. This nonstationary condition invalidates the convergence properties of most single-agent RL algorithms [51]. While there are methods that address this by allowing every agent to observe the state and actions of the other agents, these are not applicable to STESSs. In particular, due to the sequential decision-making nature of electricity markets, while the imbalance agent can know the state of the day-ahead agent, the opposite is not true, i.e., the information of the imbalance market is unknown at the time when bids need to be submitted to the day-ahead market.

Based on the previous argument, we propose an RL approach for trading in the two markets where agents are not trained simultaneously. Instead, the day-ahead agent is trained first using the algorithm proposed in Section V-A, and

[2] Selecting a factor of 10 is a design choice. The agent just needs a large penalty cost whenever it depletes the STESS.

the imbalance agent is trained afterward including in its state information from the day-ahead market. This scheme has two benefits:

- *Convergence:* As the two RL agents are independently trained in two stationary environments, standard RL algorithms have guarantees of convergence.
- *Flexibility:* As the imbalance market is highly volatile, STESSs owners could potentially want to stop trading in the imbalance market during periods of high volatility. As the agent for the day-ahead market is independent, STESSs could simply use the controls of this agent and be optimal in the more stable day-ahead market.

### *B.2 RL for the Imbalance Market*

As the RL agent for the day-ahead market is the same as the one described in Section V-A, we only need to define the RL agent that uses the information from the day-ahead and trades in the imbalance market. For the state space, besides the three values included in the state of the day-ahead agent, the new state includes past imbalance prices, past imbalance volumes, and the day-ahead price and energy allocation. In detail, at step $k$

$$\mathbf{s}_k = \left(\mathbf{x}_k, \tau_k, \lambda_k^{\mathrm{dam}}, \dot{Q}_k^{\mathrm{dam}}, \lambda_{k-1}^{\mathrm{imb}}, V_{k-1}^{\mathrm{imb}}, \ldots, \lambda_{k-n_{\mathrm{hrl}}}^{\mathrm{imb}}, V_{k-n_{\mathrm{hrl}}}^{\mathrm{imb}}\right) \quad (13)$$

where $V^{\mathrm{imb}}$ represents the overall grid imbalance and the number of historical past values $n_{\mathrm{hrl}}$ is defined by the last lag uncorrelated to the imbalance price $\lambda_k^{\mathrm{imb}}$. As an example, for The Netherlands, we observed $n_{\mathrm{hrl}} = 3$ to be a good choice.

To define the action space $\mathbb{U}$, a single action $u \in \mathbb{R}^{n_{\mathrm{in}}}$ has a similar format as before

$$\mathbf{u} = (u_1, u_2, \ldots, u_{n_{\mathrm{in}}}). \quad (14)$$

In detail, we consider that each input control $u_i$ can take $n_{\mathrm{dis}} + 1$ discrete values uniformly separated between $-1$ and 1. In particular, defining by $\dot{Q}_i^{\mathrm{in,dam}}$ the energy purchased for storage device $i$ in the day-ahead market, a value of $u_i = -1$ represents selling all the energy $\dot{Q}_i^{\mathrm{in,dam}}$ in the imbalance market, i.e., $\dot{Q}_i^{\mathrm{in}} = 0$. By contrast, a value of $u_i = 1$ represents buying all the energy that is still possible, i.e., $\dot{Q}_i^{\mathrm{max}} - \dot{Q}_i^{\mathrm{in,dam}}$, for storage device $i$, i.e., $\dot{Q}_i^{\mathrm{in}} = \dot{Q}_i^{\mathrm{max}}$. The selection of the output power is not considered as it is already selected by the day-ahead agent.

Besides the reward $r$ that also includes now the cost/income obtained in the imbalance market and the simulation environment that also generates imbalance prices, the other parts of the RL agent remain the same.

### *B.3 Market Interaction*

In terms of the interaction with the agent for the day-ahead market, the STESS is controlled with both agents acting sequentially. First, one day-ahead, the day-ahead agent builds the bidding functions for the next day's day-ahead market. Next, the day-ahead market is cleared, and the energy is allocated. Then, in real time, the imbalance agent uses the existing information of the day-ahead and imbalance markets to select the optimal power to buy/sell.

Unlike the agent for the day-ahead market, the imbalance agent does not build bidding functions as the imbalance market requires direct selection of the power $\dot{Q}^{\mathrm{imb}}$ to buy/sell. As a result, the optimal policy $\pi^\star(\mathbf{s}_k)$ at time $k$ directly selects the power to be traded based on available data $\mathbf{s}_k$ at that time step $k$ but not on the imbalance market price $\lambda_k^{\mathrm{imb}}$.

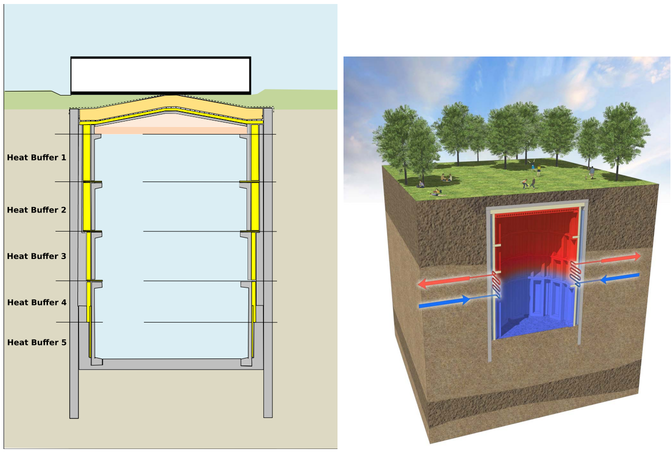


Fig. 1. Schematic representation of the STESS. Left: technical scheme representing the five heat buffers in the real system. Right: scheme representing the underground installation of the STESS.

## VI. Case Study

To study the quality of the proposed control strategies and in order to analyze the merits and disadvantages of each one of them, we consider the Ecovat vessel [54], a real SSTES. The system will be evaluated in eight case studies. First, the STESS will need to satisfy an uncertain heat demand for one year while minimizing the cost through the day-ahead market. Second, the STESS will need to supply the same heat demand but interacting with both the day-ahead and the imbalance market. For each of the two scenarios, we will consider two different heat demand profiles and two different countries.

### *A. Real STESS*

The considered STESS is a large subterranean thermal stratified storage vessel with the ability to store heat for seasonal periods and to supply heat demand to a cluster of buildings. The system is divided into different segments or heat buffers that can be charged and discharged separately; the system has five thermal buffers with the top four buffers (see Fig. 1) being able to be charged and discharged independently. Fig. 1 shows a schematic of the vessel, and Fig. 2 illustrates the real system when it was under construction. For further details on the system, we refer to [47].

### *B. System Dynamics*

The state of the STESS at time step $k$ is defined by $\mathbf{x}_k = (T_{1,k}, T_{2,k}, T_{3,k}, T_{4,k}, T_{5,k})$, i.e., by the temperature stored in each of the five buffers as it is proportional to the stored energy. Similarly, as the top four buffers can be charged and discharged independently, the input and output power are, respectively, defined by $\dot{\mathbf{Q}}_k^{\mathrm{in}} = (\dot{Q}_{1,k}^{\mathrm{in}}, \ldots, \dot{Q}_{4,k}^{\mathrm{in}})$ and $\dot{\mathbf{Q}}_k^{\mathrm{out}} = (\dot{Q}_{1,k}^{\mathrm{out}}, \ldots, \dot{Q}_{4,k}^{\mathrm{out}})$. Finally, using the dynamical model

Fig. 2. Construction of the STESS. Left: installation of the last heat buffer. Right: STESS almost completely sealed.

for thermal stratified vessels proposed in [47], the dynamics of each heat buffer $i$ at time $k$ can be defined by

$$\begin{aligned} T_{i,k+1} = \; & T_{i,k} + a_1\,(T_{i+1,k} + T_{i-1,k} - 2\,T_{i,k}) \\ & + a_2\,(T_\infty - T_{i,k}) + a_3\big(\dot{Q}^{\text{in}}_{i,k} - \dot{Q}^{\text{out}}_{i,k}\big) \end{aligned} \quad (15)$$

where $T_\infty$ represents the ambient temperature and is the only disturbance $\mathbf{d}$. For further details on the model, we refer to [47]. Note that this is the dynamical model used for the RL simulator and for defining the dynamics constraint in the MPC.

### C. Data

To set up the study, we consider the day-ahead and imbalance prices between 2015–2017 in The Netherlands,[3] and the heat demand of a cluster of five buildings with a yearly average heat demand of 220 MWh during the same time period.[4] As a second case study, we consider the day-ahead and the imbalance markets in Belgium during the same time period and the same heat demand.

The data of 2015 and 2016 is used as training data for the RL agents and as the historical data for generating scenarios. The data of 2017 are used as out-of-sample data to evaluate the performance of the different algorithms.

### D. Experimental Setup

To compare and study the control approaches, we evaluate their performance in terms of the economic cost that they incur when controlling the STESS for the full 2017 year in both The Netherlands and Belgium. As a baseline, we consider the economic cost of directly buying the instantaneous heat demand $\dot{Q}^{\text{d}}$ at the day-ahead market price. This baseline serves us to establish whether a control approach learns to trade energy, i.e., to study whether a control approach can use the STESS to reduce the energy cost. Moreover, to compare the algorithm in different conditions, the demand data are multiplied by 2 and used to evaluate the algorithms in the case of having ten buildings, i.e., a yearly average demand of 440 MWh.

The MPC algorithm is modeled using `Casadi` [55] and `python` and then solved using `Ipopt` [56]. For the RL approach, the fitted-Q-iteration algorithm is implemented in `python` using the `Xgboost` [50] library. The forecaster of imbalance prices is also done via the `Xgboost` library.

It is important to note that although both methods are based on completely different concepts, i.e., RL largely depends on the training data while MPC on the underlying optimization problem, the comparison between the methods is fair as the available data and dynamical model for both methods is exactly the same. In particular, MPC uses historical data to build price forecasts, and RL uses the same historical data to build the simulation framework. Moreover, both methods consider the same dynamical model: MPC does it explicitly in the optimization problems, while RL uses it in the simulation framework. While their solvers are different, this is the standard scenario in any comparison as different approaches have tailored solvers to the specific optimization problem, e.g., when comparing convex and nonconvex models, the convex models are estimated using a convex solver even though the nonconvex models cannot make use of it.

### E. MPC Approaches

To use the MPC approaches proposed in Section IV, a discrete set of prices has to be defined to build the bidding functions. To do so, we selected 15 discrete prices equally spaced between 0 and 70 €/MWh. This selection was done based on the price distribution in 2015–2016 and considering the computation time of solving a single OCP; however, a coarser or finer discretization could be used to, respectively, decrease the computation time or to increase the accuracy of the bidding functions. For prices above 70 €/MWh, the bidding function was set to 0 considering the seldom occurrence of prices above this threshold. For negative prices, the bidding function was defined as the solution at 0 €/MWh.

The OCPs are defined by (8)–(10), where:

- The dynamical constraint is represented by (15).
- The maximum power $\dot{\mathbf{Q}}^{\text{in}}_{\text{max}}$ to be traded in the market is defined by the electrical installation to charge the STESS. In our case, $\dot{\mathbf{Q}}^{\text{in}}_{\text{max}} = 300$ MW.
- The individual upper limits of charging and discharging, i.e., $g_{\text{in}}(\mathbf{x}_k)$ and $g_{\text{out}}(\mathbf{x}_k)$, are defined by the maximum heat transfer of the heat exchangers, which, in turn, is proportional to the temperature difference between the tank temperature and the temperature of the fluid in the heat exchangers.
- The limits on the STESS state are given by $\mathbf{x}_{\text{max}} = 286$ K and $\mathbf{x}_{\text{min}} = 263$ K, where the lower limit is defined by the outer soil temperature and the upper limit by the safety margin to prevent water boiling in the tank.

### F. RL Approaches

The RL control algorithms proposed in Section V can be directly applied to the current case study:

- The time position $\tau$ is simply the day of the year.
- As the STESS has a seasonal cycle of a year, an RL episode length is defined as two years.
- The time-dependent constraints on the maximum power are implicitly enforced within the action space as the actions are normalized with respect to the maximum power.

[3] Collected from https://transparency.entsoe.eu/

[4] Obtained from one of our research partners.

TABLE I

MPC and RL Comparison in Terms of Their Economic Cost When Only Trading in the Day-Ahead Market. The Savings Are Computed With Respect to The Cost of Not Having an STESS. For Each Case Study, the Best Method Is Indicated in Bold

| | | **The Netherlands** | | **Belgium** | |
|---|---|---|---|---|---|
| | | 10 buildings | 5 bldgs. | 10 bldgs. | 5 bldgs. |
| **Cost [€]** | No STESS | 19384 | 9692 | 23490 | 11744 |
| | MPC | **15206** | **6825** | **16826** | 7033 |
| | RL | 15942 | 7465 | 17636 | **7027** |
| **Savings** | MPC | **21.6%** | **29.6%** | **28.4%** | 40.1% |
| | RL | 17.8% | 23.0% | 24.9% | **40.2%** |

TABLE II

MPC and RL Comparison in Terms of Their Computation Time When Trading in the Day-Ahead Market. The Comparison Is Done in Terms of Online and Offline Computation Times

| | **Offline** | **Online** |
|---|---|---|
| **MPC** | 0 | 10–15 minutes |
| **RL** | 1–2 days | <1 second |

## G. Day-Ahead Market Trading

The main results of the first study, i.e., the comparison of MPC and RL when only trading in the day-ahead market, are listed in Tables I and II. Table I displays the yearly economic cost when using both algorithms and the cost of not having an STESS, i.e., the cost of buying directly the heat demand in the day-ahead market; it also lists the economic savings of both algorithms with respect to the case of not having an STESS. Table II lists the offline costs, i.e., one-time computations, and online costs, i.e., real-time computations, of both algorithms.

Independently of the country or heat demand level considered, the following observations can be made:

- Both algorithms can trade energy and make use of the STESS to reduce the economic cost. In particular, using the STESS and trading optimally, the algorithms can reduce the economic cost by 20%–40%.
- The performance of both algorithms is similar, but MPC can obtain slightly lower costs and larger profits.
- While RL requires a long offline computation time, its cost online is almost negligible. In particular, as the optimal bidding functions are estimated offline, the computation cost in real time is almost 0.
- By contrast, while MPC does not require offline computations, it needs 10–15 min in real time to build the bidding functions. However, as the bidding functions are submitted once per day and one day in advance, this large real-time computation cost does not represent a real problem/disadvantage.

Finally, to illustrate the generated bidding curves of both methods, Fig. 3 displays the generated bidding curves the first day of the five-building case study for the day-ahead market in The Netherlands. As it could be expected based on the results in Table I, both bidding curves are very similar.

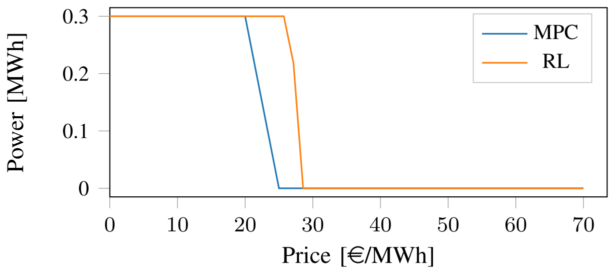


Fig. 3. Generated bidding curves by the MPC and RL algorithms on January 1, 2017, in The Netherlands when supplying heat for five buildings.

TABLE III

MPC and RL Comparison in Terms of Their Economic Cost When Trading in the Day-Ahead and Imbalance Markets. The Savings Are Computed With Respect to The Cost of Not Having an STESS. For Each Case Study, the Best Method Is Indicated in Bold

| | | **The Netherlands** | | **Belgium** | |
|---|---|---|---|---|---|
| | | 10 buildings | 5 bldgs. | 10 bldgs. | 5 bldgs. |
| **Cost [€]** | No STESS | 19384 | 9692 | 23490 | 11744 |
| | MPC | **9227** | 3544 | **10569** | 4401 |
| | RL | 11176 | **3437** | 11468 | **3872** |
| **Savings** | MPC | **52.4%** | 63.4% | **55.0%** | 62.5% |
| | RL | 42.3% | **64.5%** | 51.8% | **67.0%** |

TABLE IV

Computation Cost of the MPC and RL Approaches When Trading in the Imbalance Market. The Comparison Is Done in Terms of Online and Offline Computation Times

| | **Offline** | **Online** |
|---|---|---|
| **MPC** | 0 | 30–45 seconds |
| **RL** | 1–2 days | <1 second |

TABLE V

MPC and RL Comparison in Terms of % of Times That They Correctly Upregulate or Downregulate the Grid, i.e., % of Times That They Sell/Buy Energy in the Imbalance Market When the TSO Upregulates/Downregulates. For Each Case Study, the Best Method Is Indicated in Bold

| | | **The Netherlands** | | **Belgium** | |
|---|---|---|---|---|---|
| | | 10 buildings | 5 bldgs. | 10 bldgs. | 5 bldgs. |
| **Up-regul.** | MPC | **51%** | 47% | 44% | 47% |
| | RL | 49% | 46% | **50%** | **52%** |
| **Down-reg.** | MPC | 70% | 66% | 68% | 55% |
| | RL | **81%** | **81%** | **81%** | **80%** |

## H. Day-Ahead and Imbalance Market Trading

The main results of the second study, i.e., the comparison between MPC and RL when trading in both the day-ahead and imbalance markets, are listed in Tables III–V. Table III displays the yearly economical cost and economic savings of both algorithms. Table IV lists their offline and online computation costs when trading in the imbalance market (the computation cost for trading in the day-ahead is the same as in Table II). As an extra comparison, Table V summarizes the percentage of times that each algorithm correctly upregulates and downregulates the grid, i.e., the percentage of times that the algorithm sells (buys) energy in the imbalance market, while the TSO tries to up-regulate and down-regulate the system.

As before, independently of the case study considered, the following observations can be made:

- As for day-ahead trading, both algorithms perform very similarly to each other. However, unlike in the case of only day-ahead trading, MPC no longer performs slightly better. Instead, RL performs slightly better for lower heat demand profiles (five buildings), and MPC performs better for higher heat demand profiles (ten buildings).
- Trading in both markets is much more beneficial than trading only in the day-ahead market as the costs are halved with respect to day-ahead trading. In particular, while day-ahead trading reduces the economic cost by 20%–40%, trading in the two markets reduces the cost up to 60%–70%.
- As before, RL requires large offline computation costs but negligible online computation costs. By contrast, MPC has no offline computation costs but requires 30–45 s to obtain the optimal trading strategy for the imbalance market. Since the imbalance market is cleared every 15 min and optimal decisions are made within seconds, it can be argued that the online computation cost of MPC might now represent a problem.
- When buying energy in the imbalance market, the RL algorithm helps the TSO to down-regulate the grid. In particular, approximately 80% of the times the RL algorithm buys energy, the TSO simultaneously tries to reduce the grid generation or to increase the grid consumption. While the MPC algorithm also helps, this contribution is worse as it only helps to down-regulate 55%–70% of the time.
- By contrast, when selling energy in the imbalance market, none of the algorithms help much to up-regulate: only 45%–55% of the times, an algorithm sells energy the TSO is simultaneously trying to up-regulate.

## VII. Discussion

In this section, based on the obtained results, we discuss the merits and disadvantages of the proposed control approaches, the benefits of using STESSs for energy trading, how to optimally operate STESSs to maximize their profits, and the generality and optimality of the proposed methods.

### A. Merits of Each Control Approach

We start the discussion by analyzing the merits of the different proposed approaches in the two trading contexts.

#### A.1 Day-Ahead Trading

When trading only in the day-ahead market, both approaches can trade energy with a similar performance despite their underlying differences. Therefore, while MPC obtains slightly lower economic costs than RL, it is necessary to consider other metrics in order to make a meaningful comparison.

When considering the online computation time, both algorithms are feasible for real-life applications. Thus, the largest difference between both approaches is the offline computation time. While this metric does not play a role most of the time, i.e., it usually represents one-time computation costs, it might be important when the system regularly goes under maintenance, something breaks down, or the market has a big change. In particular, if any of these events happen, MPC can easily adapt itself by a change in the OCP or by reestimating the dynamical model (which does not take more than some minutes). By contrast, RL requires one to two days to reestimate the optimal policy under the new conditions, which hinders the day-ahead trading. Thus, MPC has, in general, better adaptability to environmental conditions.

Based on this analysis, it becomes clear that MPC is a better approach when trading only in the day-ahead market. Particularly, slightly better optimal solutions together with better adaptability to environmental changes make the proposed MPC approach a better solution in this case.

#### A.2 Day-Ahead and Imbalance Trading

Similar to the case of only day-ahead trading, when trading in the day-ahead and imbalance market, the two proposed approaches obtain good solutions. In particular, while RL performs slightly better for lower heat demand profiles (five buildings), and MPC performs better for higher heat demand profiles (ten buildings), these differences are not very large, and as before, other metrics need to be considered.

While the online computation time for day-ahead trading was not an issue, for the case of imbalance trading, it becomes one. In detail, due to the real-time nature of the imbalance market, optimal decisions should be made in seconds. As the proposed MPC approach requires 30–45 s to compute an optimal solution, it can potentially fail to provide an optimal trading strategy.

As a result, while the proposed MPC approach still has better adaptability to environmental changes, one could argue that it is a less appropriate control strategy than the proposed RL approach. The latter, with its negligible real-time computation cost, equal quality solutions, and better regulatory capabilities, is a better choice when it comes to trading in the imbalance market.

### B. Importance of Market Trading for STESSs

Based on the obtained results, it is clear that optimal control approaches, either MPC or RL, are key to maximize the profits of STESSs and ensure their widespread use as optimal control strategies and can reduce the energy cost by 60%–70%. In this context, the largest profits are obtained when the STESS trades in multiple markets. In particular, while a traditional STESS would restrict its trading to the day-ahead market to avoid unnecessary risks, in this article, we show that STESSs can dramatically reduce their costs by using optimal control strategies and trading also in the imbalance market.

### C. STESSs as Regulation Tools

Looking at the results of Table V, it can be argued that the economic goal of STESSs is (partially) aligned with the regulatory duties of the TSO. In particular, in the case of RL, 80% of the times the STESS buys energy in the imbalance market, it helps the TSO to down-regulate the system. This behavior is seen for the various case studies considered, which included different imbalance markets and different heat demands. In the case of MPC, this effect is not so pronounced; nevertheless, it still helps the TSO 55%–70% of the times.

While the same cannot be said about up-regulation, i.e., only 50% of the times the STESS sells energy in the imbalance market it is actually helping upregulate the grid, it can be argued that wrongly up-regulating is less critical than wrongly down-regulating. In particular, if the STESS wrongly sells energy in the imbalance market, the TSO can always request somebody to reduce their generation, i.e., down-regulate. However, if the STESS wrongly buys energy in the imbalance market, the TSO has to request somebody to increase their generation; as the generation is limited, there might not be an available agent that can provide that service.

As an additional remark, to further improve the regulatory services of STESSs, communication between the TSO and the STESS could be established. In particular, in the current setup, the STESS simply optimizes its profit without considering the TSO. Thus, to improve this, the TSO could simply indicate the STESS whether it is allowed to buy or sell energy, i.e., whether the TSO plans to down or up-regulate, and the STESS could take its optimal action if it helps the TSO and its own profit.

### D. Generality of the Methods

While the case study focused on a specific STESS, i.e., latent heat storage via water stratification, the proposed methods are general and can be applied to any STESS. Indeed, with the proposed methods, the several challenges that prevent the development of efficient control solutions for STESS trading can be tackled, namely: scenario generation and quantification of price uncertainty for long horizons, small computation costs for real-time control, and adaptability to market changes.

### E. Optimality of the Methods

The optimality property of the proposed methods is affected by the following elements: 1) the optimization problems are nonconvex; 2) the quality of the solutions depends on the accuracy of the forecasting method; and 3) in a multistage optimization problem, the decision taken at the first stage will have an effect in future stages. In this context, it is important to remark that the methods are nonetheless optimal from the perspective that they take a local optimal solution at every state with the information that is known.

1) The first optimization problem takes an optimal decision, considering that, at the moment of the decision, it only knows a forecast of future prices.
2) The second optimization problem takes an optimal decision with updated information and considering that market conditions have been changed. While this decision may differ from the first optimal solution, the solution is, nonetheless, a local optimum at the time when the decision is made.

Within the same context of optimality, to evaluate the proposed methods, the obtained solutions should ideally be compared with the real optimal solutions considering perfect knowledge of the future. However, this is neither possible nor fair for two reasons:

1) The optimization problem that provides the optimal solution is nonconvex. Therefore, such an analysis would involve comparing two local minima, and it would not involve a real optimal baseline.
2) The proposed approaches need to rely on forecasting methods, while the baseline solution has perfect knowledge of the future. In this context, the quality of the proposed methods depends on an external factor (forecasts) that the baseline solution does not.

## VIII. Conclusion

We have proposed several optimal control strategies for seasonal thermal storage systems (STESSs) when interacting with electricity markets. Particularly, while in the literature there are control strategies for STESSs and there are optimal trading strategies for traditional storage systems, the former does not allow STESSs to trade in the markets and the latter is not suitable for STESSs. To fill that gap, we have proposed a MPC and a RL approach for the case of having an STESS trading in the day-ahead electricity market. In addition, we argued that trading in one market is not optimal and proposed another MPC and another RL approach for the case of having an STESS trading in both the day-ahead and the imbalance markets.

Based on a case study involving a real STESS, it was shown that, despite the similarity in the optimal solutions of the proposed algorithms, MPC is a better trading strategy for the day-ahead market due to its larger adaptability. In contrast, for trading in the imbalance market, the proposed RL approach is a more suitable control strategy as it has negligible real-time computation costs, leads to similar economic costs as MPC, and has better regulatory capabilities.

It was also shown that STESSs are potential tools for grid regulation and that the economic incentive of STESSs is aligned with the regulatory duties of TSOs. Similarly, it was demonstrated that optimal control strategies are needed to optimize the profit of STESSs and to ensure their widespread use.

In future research, we intend to further explore the use of STESSs as regulation devices. Moreover, as stochastic approaches can further improve the performance of the control algorithms in the context of long horizons, we will analyze the advantages of using stochastic MPC approaches for seasonal storage systems. Finally, we will also study the tradeoffs between MPC and RL to obtain a set of generalizable tradeoffs that are independent of the case study.

## Acknowledgment

The authors would also like to thank Dr. Fjo De Ridder for the fruitful discussions and the people at Ecovat for granting access to their seasonal thermal energy storage system. The computational resources used in this work were provided by the VSC (Flemish Supercomputer Center).

## List of Symbols

| **Type** | **Symbol** | **Definition** |
|---|---|---|
| Indices | $t$ | Continuous time index |
| | $k$ | Discrete time index |
| | $h$ | Discrete hourly time index |
| Dynamic Systems | $x$ | State of a general dynamic system |
| | $u$ | Controls of a general dynamic system |
| | $d$ | Disturbances of a general dynamic system |
| | $\dot{Q}^{\mathrm{in}}$ | Input power |
| | $\dot{Q}^{\mathrm{out}}$ | Output power |
| | $\dot{Q}^{\mathrm{d}}$ | Heat demand |
| | $\dot{Q}^{\mathrm{max}}$ | Maximum input power |
| | $n_{\mathrm{in}}$ | Number of inputs |
| | $n_{\mathrm{out}}$ | Number of outputs |
| | $n_{\mathrm{units}}$ | Number of individual storage units |
| Electricity Markets | $\dot{Q}^{\mathrm{m}}$ | Allocated power from a general market |
| | $\dot{Q}^{\mathrm{dam}}$ | Allocated power from the day-ahead market |
| | $\dot{Q}_i^{\mathrm{in,dam}}$ | Power for device $i$ from the day-ahead market |
| | $\dot{Q}^{\mathrm{imb}}$ | Allocated power from the imbalance market |
| | $\dot{Q}^{\mathrm{b}}(\cdot)$ | Biding function for electricity market |
| | $\lambda$ | General price |
| | $\lambda^{\mathrm{dam}}$ | Price in the day-ahead market |
| | $\lambda^{\mathrm{imb}}$ | Price in the imbalance market |
| | $V^{\mathrm{imb}}$ | Volume of the imbalance |
| Forecast | $\bar{\dot{Q}}^{\mathrm{d}}$ | Generated scenarios of heat demand |
| | $\bar{\lambda}^{\mathrm{dam}}$ | Generated scenarios of day-ahead prices |
| | $\bar{\lambda}^{\mathrm{imb}}$ | Generated scenarios of imbalance prices |
| | $\hat{\lambda}^{\mathrm{imb}}$ | Forecast of imbalance market prices |
| MPC | $N$ | Number of discrete time intervals |
| | $T$ | Optimization horizon |
| | $\lambda^{\mathrm{b}}$ | Discrete price for building bidding functions |
| | $n_{\mathrm{p}}$ | Number of discrete price in bidding functions |
| RL | $s$ | State of agent |
| | $u$ | Action taken by agent |
| | $\mathbb{U}$ | Discrete set of possible actions |
| | $r$ | Reward obtained when taking action |
| | $\pi$ | Agent policy used to take actions |
| | $T_{\mathrm{e}}$ | Episode length |
| | $\tau$ | Seasonal time index in the agent state |
| | $n_{\mathrm{hrl}}$ | Number of past lags in the RL state |
| | $n_{\mathrm{dis}}$ | Number of discretized inputs |

## References

[1] B. K. Sovacool, "How long will it take? Conceptualizing the temporal dynamics of energy transitions," *Energy Res. Social Sci.*, vol. 13, pp. 202–215, Mar. 2016.

[2] U. Bardi, "The grand challenge of the energy transition," *Frontiers Energy Res.*, vol. 1, p. 2, 2013.

[3] J. Lago, F. De Ridder, and B. De Schutter, "Forecasting spot electricity prices: Deep learning approaches and empirical comparison of traditional algorithms," *Appl. Energy*, vol. 221, pp. 386–405, Jul. 2018.

[4] *Electricity Storage and Renewables: Costs and Markets to 2030*, Int. Renew. Energy Agency, Abu Dhabi, United Arab Emirates, 2017.

[5] S. Ugarte *et al.*, "Energy storage: Which market designs and regulatory incentives are needed?" Policy Dept. A, Econ. Sci. Policy, Eur. Parliament, Brussels, Belgium, Tech. Rep. PE 563.469, 2015. [Online]. Available: https://www.europarl.europa.eu/thinktank/en/document.html?reference=IPOL_STU(2015)563469

[6] O. Schmidt, S. Melchior, A. Hawkes, and I. Staffell, "Projecting the future levelized cost of electricity storage technologies," *Joule*, vol. 3, no. 1, pp. 81–100, Jan. 2019.

[7] A. Verkehrswende, "The future cost of electricity-based synthetic fuels," Agora Energiewende, Frontier Econ., Berlin, Germany, 2018. [Online]. Available: https://www.agora-energiewende.de/en/publications/the-future-cost-of-electricity-based-synthetic-fuels-1/

[8] A. Lucas and S. Chondrogiannis, "Smart grid energy storage controller for frequency regulation and peak shaving, using a vanadium redox flow battery," *Int. J. Electr. Power Energy Syst.*, vol. 80, pp. 26–36, Sep. 2016.

[9] M. Walter, M. V. Kovalenko, and K. V. Kravchyk, "Challenges and benefits of post-lithium-ion batteries," *New J. Chem.*, vol. 44, no. 5, pp. 1677–1683, 2020.

[10] J. W. Choi and D. Aurbach, "Promise and reality of post-lithium-ion batteries with high energy densities," *Nature Rev. Mater.*, vol. 1, no. 4, Apr. 2016.

[11] J. O. Abe, A. P. I. Popoola, E. Ajenifuja, and O. M. Popoola, "Hydrogen energy, economy and storage: Review and recommendation," *Int. J. Hydrogen Energy*, vol. 44, no. 29, pp. 15072–15086, Jun. 2019.

[12] O. Pesonen and T. Alakunnas, "Energy storage a missing piece of the puzzle for the self-sufficient living," Lapland Univ. Appl. Sci., Rovaniemi, Finland, Tech. Rep., 2017. [Online]. Available: https://www.theseus.fi/handle/10024/136086

[13] B. Battke and T. S. Schmidt, "Cost-efficient demand-pull policies for multi-purpose technologies the case of stationary electricity storage," *Appl. Energy*, vol. 155, pp. 334–348, 2015.

[14] J. Xu, R. Z. Wang, and Y. Li, "A review of available technologies for seasonal thermal energy storage," *Sol. Energy*, vol. 103, pp. 610–638, May 2014.

[15] I. Sarbu and C. Sebarchievici, "A comprehensive review of thermal energy storage," *Sustainability*, vol. 10, no. 2, p. 191, Jan. 2018.

[16] G. Darivianakis, A. Eichler, R. S. Smith, and J. Lygeros, "A data-driven stochastic optimization approach to the seasonal storage energy management," *IEEE Control Syst. Lett.*, vol. 1, no. 2, pp. 394–399, Oct. 2017.

[17] V. Rostampour and T. Keviczky, "Probabilistic energy management for building climate comfort in smart thermal grids with seasonal storage systems," *IEEE Trans. Smart Grid*, vol. 10, no. 4, pp. 3687–3697, Jul. 2019.

[18] E. Saloux and J. A. Candanedo, "Control-oriented model of a solar community with seasonal thermal energy storage: Development, calibration and validation," *J. Building Perform. Simul.*, pp. 1–23, 2018.

[19] J. Arteaga and H. Zareipour, "A price-maker/price-taker model for the operation of battery storage systems in electricity markets," *IEEE Trans. Smart Grid*, vol. 10, no. 6, pp. 6912–6920, Nov. 2019.

[20] N. Yu and B. Foggo, "Stochastic valuation of energy storage in wholesale power markets," *Energy Econ.*, vol. 64, pp. 177–185, May 2017.

[21] M. Kazemi, H. Zareipour, N. Amjady, W. D. Rosehart, and M. Ehsan, "Operation scheduling of battery storage systems in joint energy and ancillary services markets," *IEEE Trans. Sustain. Energy*, vol. 8, no. 4, pp. 1726–1735, Oct. 2017.

[22] H. Khani, R. K. Varma, M. R. D. Zadeh, and A. H. Hajimiragha, "A real-time multistep optimization-based model for scheduling of storage-based large-scale electricity consumers in a wholesale market," *IEEE Trans. Sustain. Energy*, vol. 8, no. 2, pp. 836–845, Apr. 2017.

[23] A. Gonzalez-Garrido, A. Saez-de-Ibarra, H. Gaztanaga, A. Milo, and P. Eguia, "Annual optimized bidding and operation strategy in energy and secondary reserve markets for solar plants with storage systems," *IEEE Trans. Power Syst.*, vol. 34, no. 6, pp. 5115–5124, Nov. 2019.

[24] S. Nojavan, A. Akbari-Dibavar, and K. Zare, "Optimal energy management of compressed air energy storage in day-ahead and real-time energy markets," *IET Gener., Transmiss. Distrib.*, vol. 13, no. 16, pp. 3673–3679, Aug. 2019.

[25] A. Akbari-Dibavar, K. Zare, and S. Nojavan, "A hybrid stochastic-robust optimization approach for energy storage arbitrage in day-ahead and real-time markets," *Sustain. Cities Soc.*, vol. 49, Aug. 2019, Art. no. 101600.

[26] X. Fang, B.-M. Hodge, L. Bai, H. Cui, and F. Li, "Mean-variance optimization-based energy storage scheduling considering day-ahead and real-time LMP uncertainties," *IEEE Trans. Power Syst.*, vol. 33, no. 6, pp. 7292–7295, Nov. 2018.

[27] E. Nasrolahpour, J. Kazempour, H. Zareipour, and W. D. Rosehart, "A bilevel model for participation of a storage system in energy and reserve markets," *IEEE Trans. Sustain. Energy*, vol. 9, no. 2, pp. 582–598, Apr. 2018.

[28] S. Karhinen and H. Huuki, "Private and social benefits of a pumped hydro energy storage with increasing amount of wind power," *Energy Econ.*, vol. 81, pp. 942–959, Jun. 2019.

[29] B. Cheng and W. Powell, "Co-optimizing battery storage for the frequency regulation and energy arbitrage using multi-scale dynamic programming," *IEEE Trans. Smart Grid*, vol. 9, no. 3, pp. 1997–2005, May 2018.

[30] S. Grillo, A. Pievatolo, and E. Tironi, "Optimal storage scheduling using Markov decision processes," *IEEE Trans. Sustain. Energy*, vol. 7, no. 2, pp. 755–764, Apr. 2016.

[31] R. Bellman, *Dynamic Programming*. Princeton, NJ, USA: Princeton Univ. Press, 1957.

[32] A. Oudalov, D. Chartouni, and C. Ohler, "Optimizing a battery energy storage system for primary frequency control," *IEEE Trans. Power Syst.*, vol. 22, no. 3, pp. 1259–1266, Aug. 2007.

[33] P. Zou, Q. Chen, Q. Xia, G. He, and C. Kang, "Evaluating the contribution of energy storages to support large-scale renewable generation in joint energy and ancillary service markets," *IEEE Trans. Sustain. Energy*, vol. 7, no. 2, pp. 808–818, Apr. 2016.

[34] V. Rostampour, M. Jaxa-Rozen, M. Bloemendal, and T. Keviczky, "Building climate energy management in smart thermal grids via aquifer thermal energy storage systems," *Energy Procedia*, vol. 97, pp. 59–66, Nov. 2016.

[35] F. De Ridder, M. Diehl, G. Mulder, J. Desmedt, and J. Van Bael, "An optimal control algorithm for borehole thermal energy storage systems," *Energy Buildings*, vol. 43, no. 10, pp. 2918–2925, Oct. 2011.

[36] Q. Xu and S. Dubljevic, "Model predictive control of solar thermal system with borehole seasonal storage," *Comput. Chem. Eng.*, vol. 101, pp. 59–72, Jun. 2017.

[37] W. Wei, C. Gu, D. Huo, S. Le Blond, and X. Yan, "Optimal borehole energy storage charging strategy in a low carbon space heat system," *IEEE Access*, vol. 6, pp. 76176–76186, 2018.

[38] J. Lago, F. De Ridder, P. Vrancx, and B. De Schutter, "Forecasting day-ahead electricity prices in Europe: The importance of considering market integration," *Appl. Energy*, vol. 211, pp. 890–903, Feb. 2018.

[39] J. Lago, K. De Brabandere, F. De Ridder, and B. De Schutter, "Short-term forecasting of solar irradiance without local telemetry: A generalized model using satellite data," *Sol. Energy*, vol. 173, pp. 566–577, Oct. 2018.

[40] F. Feijoo, W. Silva, and T. K. Das, "A computationally efficient electricity price forecasting model for real time energy markets," *Energy Convers. Manage.*, vol. 113, pp. 27–35, Apr. 2016.

[41] Y. Ji, R. J. Thomas, and L. Tong, "Probabilistic forecast of real-time LMP via multiparametric programming," in *Proc. 48th Hawaii Int. Conf. Syst. Sci.*, Jan. 2015, pp. 2549–2556.

[42] J. B. Rawlings, D. Q. Mayne, and M. M. Diehl, *Model Predictive Control: Theory, Computation, and Design*. Madison, WI, USA: Nob Hill, 2017.

[43] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.

[44] P. Pinson, H. Madsen, H. A. Nielsen, G. Papaefthymiou, and B. Klöckl, "From probabilistic forecasts to statistical scenarios of short-term wind power production," *Wind Energy*, vol. 12, no. 1, pp. 51–62, Jan. 2009.

[45] H. Louie, "Evaluation of bivariate archimedean and elliptical copulas to model wind power dependency structures," *Wind Energy*, vol. 17, no. 2, pp. 225–240, Feb. 2014.

[46] F. Golestaneh, H. B. Gooi, and P. Pinson, "Generation and evaluation of space–time trajectories of photovoltaic power," *Appl. Energy*, vol. 176, pp. 80–91, Aug. 2016.

[47] J. Lago, F. De Ridder, W. Mazairac, and B. De Schutter, "A 1-dimensional continuous and smooth model for thermally stratified storage tanks including mixing and buoyancy," *Appl. Energy*, vol. 248, pp. 640–655, Aug. 2019.

[48] H. G. Bock and K.-J. Plitt, "A multiple shooting algorithm for direct solution of optimal control problems," *IFAC Proc. Volumes*, vol. 17, no. 2, pp. 1603–1608, Jul. 1984.

[49] D. Ernst, P. Geurts, and L. Wehenkel, "Tree-based batch mode reinforcement learning," *J. Mach. Learn. Res.*, vol. 6, pp. 503–556, Apr. 2005.

[50] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.

[51] L. Busoniu, R. Babuška, and B. De Schutter, "A comprehensive survey of multiagent reinforcement learning," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 38, no. 2, pp. 156–172, Mar. 2008.

[52] C. Claus and C. Boutilier, "The dynamics of reinforcement learning in cooperative multiagent systems," in *Proc. 15th Nat./10th Conf. Artif. Intell./Innov. Appl. Artif. Intell.* Menlo Park, CA, USA: American Association for Artificial Intelligence, 1998, pp. 746–752.

[53] L. Matignon, G. J. Laurent, and N. Le Fort-Piat, "Independent reinforcement learners in cooperative Markov games: A survey regarding coordination problems," *Knowl. Eng. Rev.*, vol. 27, no. 1, pp. 1–31, Feb. 2012.

[54] Ecovat. (2018). *Ecovat Renewable Energy Technologies BV*. [Online]. Available: http://www.ecovat.eu

[55] J. A. E. Andersson, J. Gillis, G. Horn, J. B. Rawlings, and M. Diehl, "CasADi: A software framework for nonlinear optimization and optimal control," *Math. Program. Comput.*, vol. 11, no. 1, pp. 1–36, Mar. 2019.

[56] A. Wächter and L. T. Biegler, "On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming," *Math. Program.*, vol. 106, no. 1, pp. 25–57, Mar. 2006.

**Jesus Lago** received the B.Sc. degree from the University of Vigo, Vigo, Spain, in 2013, and the M.Sc. degree from the University of Freiburg, Freiburg im Breisgau, Germany, in 2016. He is currently pursuing the Ph.D. degree with the Delft Center for Systems and Control, Delft University of Technology, Delft, The Netherlands.

He is currently a Researcher with the Flemish Institute for Technological Research (VITO), Mol, Belgium, and with Energyville, Genk, Belgium. His research areas are forecasting algorithms and control techniques for smart energy systems that interact with electricity markets.

**Gowri Suryanarayana** received the M.Sc. degree in applied mathematics and the Ph.D. degree in numerical analysis from Katholieke Universiteit Leuven (K.U. Leuven), Leuven, Belgium, in 2001 and 2016, respectively.

She specialized in numerical integration and approximation of high-dimensional functions at K.U. Leuven. She is currently working at the Flemish Institute for Technological Research (VITO), Mol, Belgium, and at Energyville, Genk, Belgium, on research projects in the energy domain related to modeling, optimization, and optimal control of smart energy systems.

**Ecem Sogancioglu** received the bachelor's degree in computer science from Hacettepe University, Ankara, Turkey, in 2013, and the master's degree in computer science from the University of Freiburg, Freiburg im Breisgau, Germany, in March 2017, with a focus on machine learning. She is currently pursuing the Ph.D. degree with the Diagnostic Image Analysis Group, Radboud University, Nijmegen, The Netherlands.

Her research is focused on deep learning algorithms with application to medical images.

**Bart De Schutter** (Fellow, IEEE) received the Ph.D. degree from Katholieke Universiteit Leuven (K.U. Leuven), Leuven, Belgium, in 1996.

He is currently a Full Professor and the Department Head of the Delft Center for Systems and Control, Delft University of Technology, Delft, The Netherlands. He is the (co)author of three books, including *Reinforcement Learning and Dynamic Programming Using Function Approximators*.

Dr. De Schutter is also an Associate Editor of the IEEE TRANSACTIONS ON AUTOMATIC CONTROL and a Senior Editor of the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS.