mathematics

MDPI

Article

Safe Optimal Control of Dynamic Systems: Learning from Experts and Safely Exploring New Policies

Antonio Candelieri [1,*], Andrea Ponti [1], Elisabetta Fersini [2], Enza Messina [2] and Francesco Archetti [2]

[1] Department of Economics Management and Statistics, University of Milano-Bicocca, 20126 Milan, Italy; andrea.ponti@unimib.it
[2] Department of Computer Science Systems and Communication, University of Milano-Bicocca, 20126 Milan, Italy; elisabetta.fersini@unimib.it (E.F.); enza.messina@unimib.it (E.M.); francesco.archetti@unimib.it (F.A.)
* Correspondence: antonio.candelieri@unimib.it

Abstract: Many real-life systems are usually controlled through policies replicating experts' knowledge, typically favouring "safety" at the expense of optimality. Indeed, these control policies are usually aimed at avoiding a system's disruptions or deviations from a target behaviour, leading to suboptimal performances. This paper proposes a statistical learning approach to exploit the historical safe experience—collected through the application of a safe control policy based on experts' knowledge— to "safely explore" new and more efficient policies. The basic idea is that performances can be improved by facing a reasonable and quantifiable risk in terms of safety. The proposed approach relies on Gaussian Process regression to obtain a probabilistic model of both a system's dynamics and performances, depending on the historical safe experience. The new policy consists of solving a constrained optimization problem, with two Gaussian Processes modelling, respectively, the safety constraints and the performance metric (i.e., objective function). As a probabilistic model, Gaussian Process regression provides an estimate of the target variable and the associated uncertainty; this property is crucial for dealing with uncertainty while new policies are safely explored. Another important benefit is that the proposed approach does not require any implementation of an expensive digital twin of the original system. Results on two real-life systems are presented, empirically proving the ability of the approach to improve performances with respect to the initial safe policy without significantly affecting safety.

Keywords: optimal control; safe exploration; Gaussian Processes

MSC: 90-08

Citation: Candelieri, A.; Ponti, A.; Fersini, E.; Messina, E.; Archetti, F. Safe Optimal Control of Dynamic Systems: Learning from Experts and Safely Exploring New Policies. *Mathematics* **2023**, *11*, 4347. https://doi.org/10.3390/math11204347

Academic Editors: Rosita Guido and Sara Ceschia

Received: 18 September 2023
Revised: 15 October 2023
Accepted: 17 October 2023
Published: 19 October 2023

1. Introduction

The control of real-life dynamic systems often relies on policies biased towards *safety*, at the cost of suboptimal performances. Usually, these policies are designed by experts and based on their knowledge about the target system. From now on, we refer to these kind of policies with the term *safe-by-design* policies. *Safety* is a generic term related to preventing physical disruptions of the system itself (e.g., manufacturing plants, robotic systems, etc.) as well as avoiding poor Quality of Service (e.g., water/energy/gas supply networks, smart home systems, such as smart cooling/heating or lighting devices).

Nowadays, there could be the chance to work with a *digital twin* of the target system, which is a software program replicating the system's behaviour (i.e., by numerically solving all the relevant equations) depending on different setups and inputs. The main challenge is to deal with the *uncertainty* characterizing the real-life settings. If a model of the uncertainty is available, for instance, obtained from historical data, then *simulation–optimization* methods can be used to search for optimal control policies according to a large number of plausible *scenarios* sampled from the uncertainty model and simulated on the digital twin [1–3].

However, the availability of a digital twin is not so common, especially because designing and implementing it could be an expensive task depending on the complexity of the target system. More recently, research has focused on exploiting the *safe experience*; that is, the historical data collected over time as a result of the implementation of a *safe-by-design* policy. Thus, instead of using historical data just for estimating the uncertain components affecting the system's output, the control *actions* (also known as *decisions*) implied by the safe-by-design policy are also exploited. These data offer just a partial view—knowledge—of the overall system's behaviour; thus, Machine Learning (ML) algorithms are used to generalize for unseen data, specifically new control actions, while dealing with two different sources of uncertainty: the prediction error inherent in the resulting ML model (also known as *epistemic* uncertainty) and the stochastic components acting on the system (also known as *aleatoric* uncertainty).

Safety and *optimality* are strictly intertwined, as largely reported in the scientific literature about optimal learning and optimal control. The most relevant research studies, for this paper, regard the combination of Safe Active Learning (SAL) and Bayesian Optimization (BO) [4–7]. For instance, Ref. [8] address the optimal calibration of a PID (proportional–integrative–derivative) controller for a high-pressure fuel supply system of an engine through a SAL-BO method. The same problem—also known as safe optimal tuning of a PID controller's hyperparameters—has been largely investigated in [9–13], empirically proving the benefits offered by the combination of BO and SAL. In robotics, optimizing a control policy is largely a more complicated task than the optimal tuning of a PID controller, but similar approaches have also been proposed within the Reinforcement Learning (RL) framework, such as [14–16]. Interestingly, most of the quoted approaches use Gaussian Process (GP) regression to obtain a probabilistic model of the performance metric to be optimized under safety constraints. A different approach, based on Lipschitz optimization, has been proposed in the case that the safety constraint is related to a given threshold on the performance metric and uncertainty is a bounded noise whose maximum effect is known [17].

However, almost all of these studies massively use digital twins of the target systems to learn or tune a safe optimal control policy, instead of exclusively analysing historical data collected by having operated a safe-by-design policy.

The main contributions of this paper can be summarized as follows:

- Using data collected by having operated a safe-by-design policy, namely, *safe experience*, to obtain a probabilistic model of the the target system, both in terms of performance and safety. Although it is a partial representation of the system, this model allows the expensive design and implementation of a digital twin to be avoided;
- Using two separate GPs to obtain the probabilistic models of performance metric and safety, respectively. This allows different uncertainty components to be dealt with that could affect, separately as well as jointly, performances and safety;
- Generating new safe—and more effective—policies by solving a constrained optimization problem involving the two GP models mentioned before;
- Validating the proposed approach on two case studies inspired by real-life systems and quite commonly considered in the optimal control literature: the control of a house heating system [18,19] and the control of a water tank [20–22]. Safe-by-design and new safely explored policies are compared both in terms of performances and incurred risk (i.e., safety violations).

2. Safe Control of a Dynamic System

In this section, we briefly introduce the generalities about safe optimal control and present the two case studies.

2.1. System Control: Generalities

Denote with $\mathbb{S}$ the target dynamic system to be controlled, and with π the control policy deciding the control action to apply at time t, namely a_t, depending on the current

state of the system, s_t. According to its internal behaviour, the system will turn its state into $s_{t+\Delta_t}$ depending on both the control action a_t and some uncontrollable stochastic input, ξ_t. Specifically, ξ_t is unobservable or observable only after the control action a_t is operated. Along with the new state, $s_{t+\Delta_t}$, the system can provide further information, specifically some performance metric(s), denoted with $p_{t+\Delta_t}$, and constraints satisfaction, denoted with $g_{t+\Delta_t}$. It is important to remark that constraints can refer to different aspects of the system behaviour; in this paper, we are specifically interested in those related to the safety, namely *safety constraints*. The subscript $t + \Delta_t$ is used to denote the fact that state transition, performance metric(s) and constraint satisfaction are not revealed at the same time that a_t is operated; instead, some processing time (i.e., Δ_t) is needed.

Figure 1 depicts a schematic representation of a generic control loop for a dynamic system. The two case studies, detailed in the following two sections, can be resampled to this general schema.

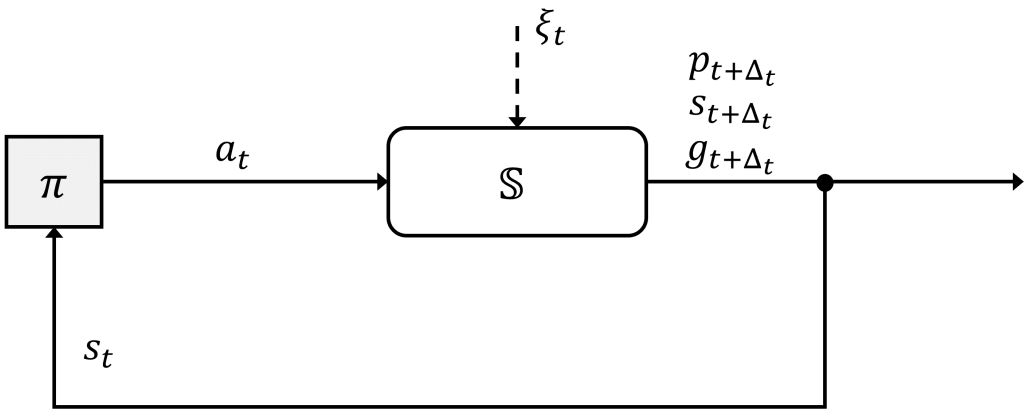


Figure 1. A schematic representation of the control of a dynamic system.

2.2. Controlling a House Heating System

The first case study considered in this paper is the control of a house heating system, schematically represented in Figure 2. The uncertain component acting on the system is the outside temperature, which varies during the day: the higher the difference between the in-house and outside temperature, the higher the heat dissipation is (depending on the physical properties of the house's walls). The control policy, π, receives the current inner temperature, s_t, as input, and sets the heater temperature, a_t, for the next Δ_t time.

The most common formulation of this control problem aims at reaching a target in-house temperature within a given interval of time (after the heater and the controller are switched on) while avoiding excessive sovraelongation and keeping the in-house temperature as stable as possible over the day. These constraints can be considered as *safety constraints*, where *safety*—in this case study—refers to avoiding discomfort to the household.

All the equations of this case study were adapted from the "Model A House Heating System" example available at the MathWorks website (https://www.mathworks.com/help/simulink/ug/model-a-house-heating-system.html (accessed on 18 September 2023)).

System transition equation. The system transition is regulated by the following *in-house changing temperature* equation:

$$s_{t+\Delta_t} = s_t + \frac{\Delta_t}{\kappa}\left[M\kappa(a_t - s_t) + \frac{s_t - \xi_t}{R}\right] \tag{1}$$

where $M\kappa(a_t - s_t) = \frac{dQ_G}{dt}$ is the *rate of heat gain* and $\frac{s_t - \xi_t}{R}$ is the *rate of heat loss*. All the others symbols are listed as follows (along with the values used in the case study):

- M is the mass of air of the heater ($M = 3600$ kg·h);
- κ is the heater capacity ($\kappa = 1500.4$ Joule/°C·kg);
- m is the mass of air in the house ($m = 1470$ kg);

- R is the thermal resistance ($R = 4.329 \cdot 10^{-7}$ °C·h/Joule).

Finally, for the outside temperature, $\xi(t)$, we used the *Yosemite temperature dataset* (https://github.com/facebook/prophet/blob/main/examples/example_yosemite_temps.csv (accessed on 18 September 2023). It stores the daily temperature measurements, sampled every minute, referring to 60 different days. All the temperature values, originally expressed as Fahrenheit degrees, were first converted into Celsius degrees.

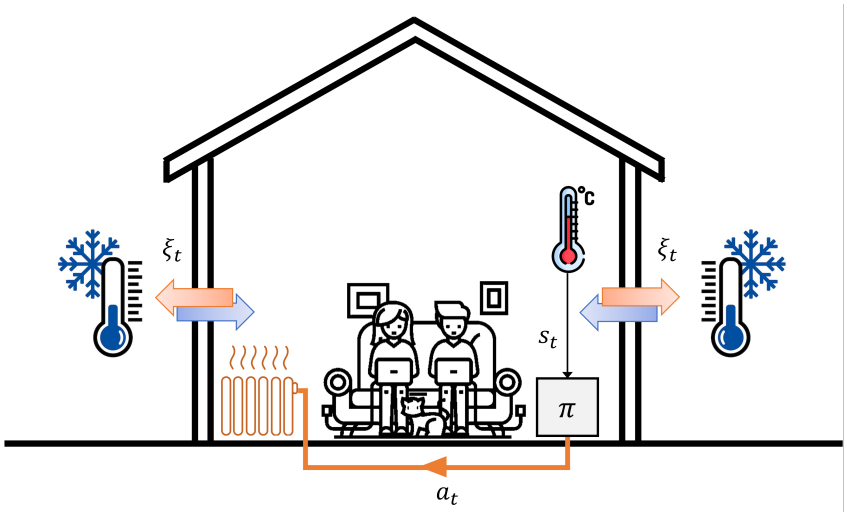


Figure 2. An illustrative representation of a house heating system.

Safety (and operating) constrains. As previously mentioned, in this case study, the safety constraints refer to:

- Bringing the in-house temperature close to a target one (i.e, $\hat{s} = 18$ °C) within 30 min from the heater and controller switch-on;
- Not exceeding a sovraelongation of 1 °C within the first 30 min (i.e., $s_t \leq 19$ °C, with $t \leq 30$ min);
- Keeping the in-house temperature close to the target, which is $s_t \in [\hat{s} - 0.5; \hat{s} + 0.5]$ °C, with $t > 30$ min.

Finally, there is only one operating constraint to consider, which is the maximum heat the heater can provide. In this study, we considered $0 \leq a_t \leq 70$ °C.

Safe-by-design policy. The safe-by-design policy for this case study is implemented through the following PID (proportional–integrative–derivative) control:

$$a_t = \min \left\{ K_p \varepsilon_t + K_i \sum_{j=0}^{t} \varepsilon_j + K_d \frac{\epsilon_t - \varepsilon_{t-1}}{\Delta t}, \quad a_{max} \right\} \tag{2}$$

where $\varepsilon_t = \hat{s} - s_t$, $a_{max} = 70$ °C, and the PID's parameters are manually set to $K_p = 43$, $K_i = 0.17$, and $K_d = 0$ to meet all the safety constraints.

Figure 3 shows the implementation of the safety-by-design policy over a certain day. On the top, the in-house temperature is reported, along with the safety constraints; on the bottom, the heat provided by the heater, according to the control policy, is depicted.

Figure 4, instead, shows (a) the in-house temperature, (b) the heat provided by the PID controller, and (c) the outside temperature, all over time for each one of the 60 days into the Yosemite temperature dataset. Specifically, it is easy to notice that the control policy implemented by the PID controller is always safe (i.e., Figure 4a).

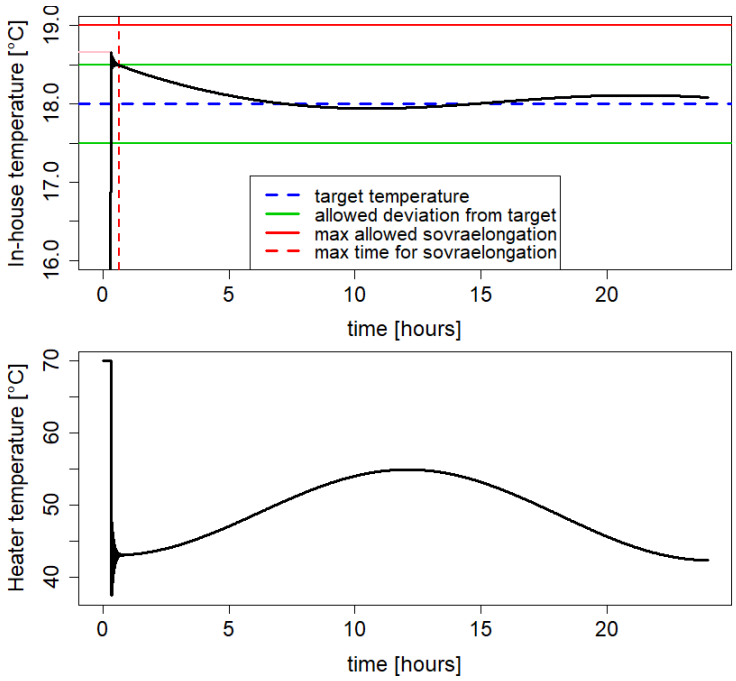
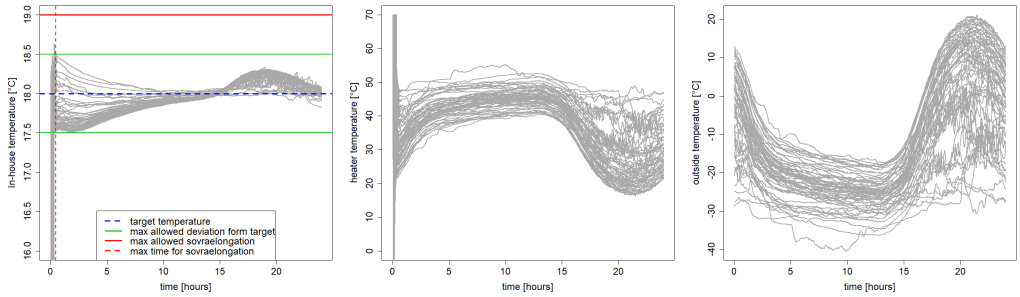


Figure 3. Safe-by-design policy for the house heating system case study: (top) in-house temperature over 24 h for a certain day, along with safety constraints; (bottom) amount of heat provided by the heater over the day.



(**a**) In-house temperature. (**b**) Heating by the PID. (**c**) Outside temperature.

Figure 4. A graphical representation of all the data for the case study, for each one of the 60 days stored in the Yosemite temperature dataset: (**a**) the in-house temperature along with the safety constraints, (**b**) the amount of heat provided by the safe-by-design control policy operated by the PID controller, and (**c**) the outside temperature (i.e., the stochastic component acting on the system).

Performance metric. In the original formulation of this control problem there is not any *economical* performance metric to optimize: the goal is just to keep the in-house temperature as close as possible to the target one. On the other hand, as far as a real-life setting is considered, a quite natural optimization the household could be interested in is the minimization of heating-related energy costs, under a Time-of-Use (ToU) energy tariff. Specifically, we considered the following ToU for the energy price:

$$price_t = \begin{cases} 1 & €/°C \quad \text{if } t \in [0;7] \cup [21;24] \\ 10 & €/°C \quad \text{if } t \in (7;10] \cup (17;21] \\ 5 & €/°C \quad \text{if } t \in (10;17] \end{cases}$$

Thus, the safe optimization problem consists of

$$\min_{a_t \in [0,\, 70]°C} \sum_{t=0}^{T} a_t \cdot price_t$$

subject to all the safety constraints previously defined.

2.3. Controlling a Water Tank

The second case study is the safe and optimal control of a water tank. The system is schematically depicted in Figure 5. The stochastic component affecting the system's output is the water demand, ξ_t, varying over time.

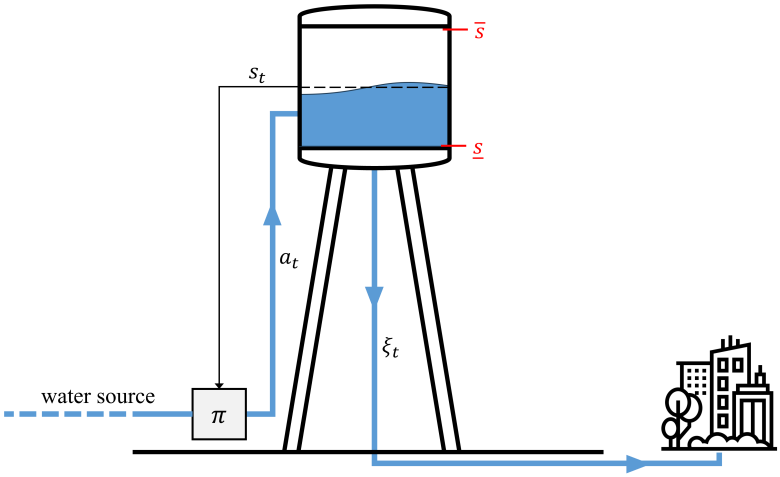


Figure 5. A schematic representation of the water tank control case study.

Contrary to the previous case, where a target temperature to reach is given, here there is not any target water level. The aim is to minimize the energy costs to pump water into the tank, under a ToU energy tariff, while matching stochastic demand.

Specifically, we considered the following ToU for the energy price:

$$price_t = \begin{cases} 1\,€/\text{kWh} & \text{if } t \in [1;6] \cup [23;24] \\ 2\,€/\text{kWh} & \text{if } t \in [9;18] \cup [21;23] \\ 3\,€/\text{kWh} & \text{if } t \in [7;8] \cup [19;20] \end{cases} \tag{3}$$

Then, the cost associated to a control action a_t also depends on the pump efficiency η. We assumed the following simple non-linear relation between action, pump efficiency, and energy-related costs:

$$cost_t = \frac{price_t \cdot a_t^3}{\eta}$$

As far as safety is concerned, the control policy π must guarantee that the level of water in the thank is always within a prefixed interval, formally $\underline{s} \leq s_t \leq \bar{s}$, with $t = 0, \ldots, T$.

The safe-by-design policy—typically operated in the real-life setting—simply consists of refilling the water tank depending on its current level s_t and the maximum amount of water that can be pumped in the unite of time, namely $\bar{a}$. Formally,

$$\pi : a_t = \min\{\bar{a}, \bar{s} - s_t\}$$

Specifically, $0 \leq a_t \leq \bar{a}$ is the operating constraint characterizing this case study.

Clearly, the reported safe-by-design policy is suboptimal—indeed, it does completely ignore energy-related costs—while our aim is to solve the following problem

$$\min_{a_t \in [0,\, \bar{a}]} \sum_{t=0}^{T} \frac{price_t \cdot a_t^3}{\eta}$$

subject to the mentioned safety (and operating) constraints.

As far as all the other technical details are concerned, we considered:

- $\underline{s} = 5\ \mathrm{m}^3$;
- $\bar{s} = 50\ \mathrm{m}^3$;
- $\bar{a} = 10\ \mathrm{m}^3$.

Finally, a dataset of 365 daily water demand time series (i.e., hourly data) was generated by sampling from typical real-life patterns. Figure 6 shows the water demand data.

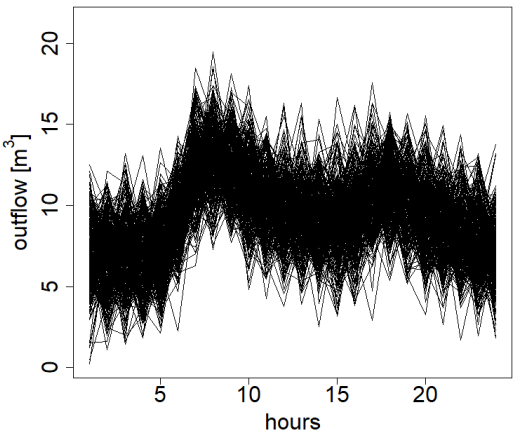


Figure 6. The dataset of 365 generated water demand time series, one for each day with hourly time step.

3. Safely Exploring More Efficient Control Policies

This section describes the proposed approach aimed at exploring new, more efficient, still safe policies, starting from the knowledge collected by having operated a safe-by-design policy. First, we describe the general methodology to model safety constraints and system performances from previous knowledge. Then, we provide all the details to specialize the general method on the two case studies.

3.1. Modelling Performance and Safety from Safe Experience

The safe experience collected by having operated a safe-by-design policy can be simply represented as a set of N tuples, each codified as:

$$< t, s_t, a_t, s_{t+\Delta_t}, p_{t+\Delta_t}, g_{t+\Delta_t} > \tag{4}$$

where symbols refer to the concepts described in Section 2.1. It is important to remark that, even if the stochastic components ξ_t was observable after a_t was operated, our approach does not require to explicitly store it into the tuple.

The first step consists of organizing the information within the set of tuples into two datasets, respectively, denoted with $\mathcal{D} = \left\{ \left(\mathbf{x}^{(i)}, z^{(i)} \right) \right\}_{1:N}$ and $\mathcal{P} = \left\{ \left(\mathbf{x}^{(i)}, p^{(i)} \right) \right\}_{1:N}$, where:

- $\mathbf{x}^{(i)} \in \mathbb{R}^d$ is an input vector of observable information, whose definition—and dimensionality d—is strictly problem-specific (as demonstrated in the next with respect to the two case studies). Anyway, it must contain at least s_t and a_t;
- $z^{(i)} \in \mathbb{R}$ is the next system's state, $s_{t+\Delta_t}$, observed depending on $\mathbf{x}^{(i)}$;
- $p^{(i)} \in \mathbb{R}$ is the performance value, $p_{t+\Delta_t}$, associated to $\mathbf{x}^{(i)}$.

Therefore, it is quite intuitive that the datasets $\mathcal{D}$ and $\mathcal{P}$ can be used to model, respectively, the system's dynamics and the system's performance, in response to an input vector $\mathbf{x}$ whose components must include the current state s_t and the control action a_t. The choice of the modelling strategy is crucial, especially to deal with the uncertainty due to the stochastic component acting on the target system.

In line with the more recent literature, we decided to adopt Gaussian Process (GP) regression as the modelling strategy. From an ML perspective, a GP is a *probabilistic*

regression model, where *probabilistic* means that the model provides not only the prediction of the output variable but also the associated *predictive uncertainty*. More importantly, a GP is able to deal with *noisy* observations (i.e., different values of the target variable for the same input). Indeed, the GP's predictive uncertainty consists of two different components: *epistemic*, increasing with the distance from observations (i.e., data) that the GP is fitted on, and *aleatoric*, which is the noise in the output variable due to inherently random effects (i.e., the effect of the stochastic uncontrollable component on the target system, in our setting).

For a comprehensive study of GP regression, the reader could consider [23,24]. Here, we report just the equations for the GP's predictive mean $\mu(\mathbf{x})$ (i.e., the predicted output) and the predictive variance $\sigma^2(\mathbf{x})$ (i.e., the square of the predictive uncertainty).

$$\mu(\mathbf{x}) = m_0(\mathbf{x}) + \mathbf{k}(\mathbf{x}, \mathbf{X}_N)\left[\mathbf{K}_N + \lambda^2\mathbf{I}\right]^{-1}(\mathbf{y}_N - m_0(\mathbf{X}_N)) \tag{5}$$

$$\sigma^2(\mathbf{x}) = k(\mathbf{x}, \mathbf{x}) - \mathbf{k}(\mathbf{x}, \mathbf{X}_N)\left[\mathbf{K}_N + \lambda^2\mathbf{I}\right]^{-1}\mathbf{k}(\mathbf{X}_N, \mathbf{x}) \tag{6}$$

where $m_0(\mathbf{x})$ is the *prior mean* (usually set constant and equal to 0 without loss of generality), and (X_N, y_N) is the training dataset, such that $X_N = \left\{\mathbf{x}^{(i)}\right\}_{i=1:N}$ are the input data and $\mathbf{y}_N = \left\{y^{(i)}\right\}$ are the associated output values. Here, the output values are assumed to be noisy observations of the unknown target function $f(\mathbf{x})$, that is $y^{(i)} = f(\mathbf{x}^{(i)}) + \varepsilon^{(i)}$, with $\varepsilon^{(i)} \sim \mathcal{N}(0, \lambda^2)$. Finally, $k(\mathbf{x}, \mathbf{x})$ is a *kernel* (also known as covariance) function establishing a prior on structural properties, specifically the *smoothness*, of the GP regression model. It follows that $\mathbf{k}(\mathbf{x}, \mathbf{X})$—as well as its transpose $\mathbf{k}(\mathbf{X}, \mathbf{x})$—is a vector with components $k(\mathbf{x}, \mathbf{x}^{(i)})$, and $\mathbf{K}_N$ is an $N \times N$ matrix with entries $K_{ij} = k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$.

The kernel function can be chosen among many possibilities [23,24]. For the purposes of this study, the Squared Exponential (SE) kernel was considered a reasonable choice, that is:

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \cdot e^{-\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{2\ell^2}}$$

with σ_f and ℓ being two kernel's hyperparameters regulating the vertical and horizontal span of the GP's output. Learning a GP regression model given a dataset $(\mathbf{X}_N, \mathbf{y}_N)$ means tuning the kernel hyperparameters to fit the data. The common choice—also adopted in this paper—consists of searching for the values of the kernel's hyperparameters, maximizing the marginal log-likelihood estimation (MLE).

As far as the goal of modelling knowledge from previously operated safe-by-design policies is concerned, we propose to fit two separate GP regression models: one learned on the dataset $\mathcal{D}$ for predicting the next state of the system, and one learned on $\mathcal{P}$ for predicting the system performances. As a result, we obtain the predictive means, $\mu_{\mathcal{D}}(\mathbf{x})$ and $\mu_{\mathcal{P}}(\mathbf{x})$, and the predictive uncertainties, $\sigma_{\mathcal{D}}(\mathbf{x})$ and $\sigma_{\mathcal{P}}(\mathbf{x})$.

3.2. New Safe Policy via Constrained Optimization

Given the two GP regression models introduced in the previous section, the safe exploration of more efficient and safe control policies is performed as a constrained optimization problem, generically formalized as follows:

$$\begin{aligned} a_t^* = \quad & \arg\min_{a_t \in \mathcal{A}} \mu_{\mathcal{P}}(\mathbf{x}) - \beta_{\mathcal{P}}\, \sigma_{\mathcal{P}}(\mathbf{x}) \\ & \text{s.t.} \quad \gamma(\mathbf{x}, \mu_{\mathcal{D}}(\mathbf{x}), \sigma_{\mathcal{D}}(\mathbf{x}), \beta_{\mathcal{D}}) \approx g_{t+\Delta_t} \geq 0 \end{aligned} \tag{7}$$

where $\mathbf{x}$ is an input vector whose components include, necessarily, the current s_t and the control action a_t to be chosen. It is important to remark that the only decision variable is a_t, while all the other components of $\mathbf{x}$, including s_t, are unchangeable observed values (they are denoted in blue in the following).

The optimal control action a_t^* must be searched within the space $\mathcal{A}$ that is specified by the operating constraints of the target system.

Without any loss of generality, we consider a minimization problem (i.e., we want to minimize costs in both the two case studies). Specifically, the lower confidence bound of the GP modelling the performance is considered, with the parameter $\beta_{\mathcal{P}}$ specifying the confidence interval to consider.

Finally, with the function $\gamma(.)$ we denote the approximation of the safety constraints $g_{t+\Delta_t} \geq 0$. This function depends on the input vector $\mathbf{x}$, the GP approximating the system's dynamics—specifically, $\mu_{\mathcal{D}}(\mathbf{x})$ and $\sigma_{\mathcal{D}}(\mathbf{x})$—and, similarly to the objective function, a coefficient $\beta_{\mathcal{D}}$ dealing with the confidence interval of the prediction.

The general formulation (7) is specifically declined for the two case studies (i.e., following optimization problems (8) and (9)). The same software framework is adopted, which is the R package **nloptr**. All the code is developed in R and it is available for free, along with data and results, as detailed in the "Data Availability Statement" at the end of the paper.

3.3. Detailing the Approach for the House Heating System Case Study

In the house heating system case study, the input vector is defined as $\mathbf{x} = (a_t, s_t, a_{t-\Delta_t})$, where the blue components are observed and unchangeable values.

At a generic time $t = 0, \ldots, T$, all the tuples referring to that specific time are retrieved and the two datasets $\mathcal{D}^{[t]}$ and $\mathcal{P}^{[t]}$ are built. Here, the superscript $[t]$ remarks that all the observations are related to the time step t, but it is omitted in the following to keep the notation as simple as possible. For $t = 0$, we assume $a_{t-\Delta_t} = 0$.

The two GP models are separately fitted on the two datasets and, according to the definition of the use case, the constrained optimization problem becomes:

$$\begin{aligned} a_t^* = \quad & \underset{a_t \in [0;\, 70]^{\circ}\mathrm{C}}{\arg\min} \; \mu_{\mathcal{P}}(\mathbf{x}) - \beta_{\mathcal{P}}\, \sigma_{\mathcal{P}}(\mathbf{x}) \\ \text{s.t.} \quad & 19\ ^{\circ}\mathrm{C} - \left[\mu_{\mathcal{D}}(\mathbf{x}) - \beta_{\mathcal{D}}\, \sigma_{\mathcal{D}}(\mathbf{x})\right] \geq 0 \quad \forall t \leq 30' \\ & 18.5\ ^{\circ}\mathrm{C} - \left[\mu_{\mathcal{D}}(\mathbf{x}) - \beta_{\mathcal{D}}\, \sigma_{\mathcal{D}}(\mathbf{x})\right] \geq 0 \quad \forall t \geq 30' \\ & \left[\mu_{\mathcal{D}}(\mathbf{x}) - \beta_{\mathcal{D}}\, \sigma_{\mathcal{D}}(\mathbf{x})\right] - 17.5\ ^{\circ}\mathrm{C} \geq 0 \quad \forall t \geq 30' \end{aligned} \tag{8}$$

The first constraint refers to the maximum allowed sovraelongation within 30′, while the other two refer to the allowed deviation from the target temperature (i.e., 18 °C).

With regard to the parameters considered, we chose to set $\beta_{\mathcal{P}} = 1$ and $\beta_{\mathcal{D}} = 6$. Indeed, we decided to be precautionary in terms of safety estimation. These values were obtained from preliminary experiments (not reported in the paper).

3.4. Detailing the Approach for the Water Tank Case Study

In the second case study, the optimal and safe control of a water tank, the input vector is defined as $\mathbf{x} = (a_t, s_t)$. Again, in blue is the component that is observed and unchangeable, specifically, the current amount of water in the tank. As for the previous case study, after retrieving the observations related to the current time step t, and after the two GPs are learned, the optimal and safe action a_t^* is obtained by solving the following constrained optimization problem:

$$\begin{aligned} a_t^* = \quad & \underset{a_t \in [0;\, 10]\mathrm{m}^3}{\arg\min} \; \mu_{\mathcal{P}}(\mathbf{x}) - \beta_{\mathcal{P}}\, \sigma_{\mathcal{P}}(\mathbf{x}) \\ \text{s.t.} \quad & 50\ \mathrm{m}^3 - \left[\mu_{\mathcal{D}}(\mathbf{x}) + \beta_{\mathcal{D}}\, \sigma_{\mathcal{D}}(\mathbf{x})\right] \geq 0\ \mathrm{m}^3 \\ & \left[\mu_{\mathcal{D}}(\mathbf{x}) - \beta_{\mathcal{D}}\, \sigma_{\mathcal{D}}(\mathbf{x})\right] - 5\ \mathrm{m}^3 \geq 0\ \mathrm{m}^3 \end{aligned} \tag{9}$$

For this specific case study, and according to preliminary experiments, we empirically defined the values of the parameters as follows: $\beta_{\mathcal{P}} = 1$ and $\beta_{\mathcal{D}} = 4$.

According to the definition of the input vector $\mathbf{x}$, it is easy to provide a graphical representation of how the approach works. Figure 7 provides an illustrative example for a specific day (i.e, the 354th in the dataset) and a specific hour of the day (i.e., $t = 7$),

in which s_t is observed and unchangeable and, therefore, a one-dimensional representation can be easily obtained. The top of the figure shows the GP modelling the cost associated to any possible action: the solid blue line is the GP's predictive mean, $\mu_{\mathcal{P}}(\mathbf{x})$, while the dashed line represents the lower confidence bound (i.e., the objective function), that is $\mu_{\mathcal{P}}(\mathbf{x}) - \beta_{\mathcal{P}}\ \sigma_{\mathcal{P}}(\mathbf{x})$. The shaded area represents the confidence interval; it is clear that the value of $\beta_{\mathcal{P}}$ does not affect the minimizer. The bottom of the figure depicts the GP devoted to predicting $s_{t+\Delta_t}$, depending on the possible actions (while s_t is always observed and kept fixed). The solid curve is the GP's predictive mean, $\mu_{\mathcal{D}}(\mathbf{x})$, and the dashed lines are the lower and the upper confidence bound, respectively, $\mu_{\mathcal{D}}(\mathbf{x}) - \beta_{\mathcal{D}}\ \sigma_{\mathcal{D}}(\mathbf{x})$ and $\mu_{\mathcal{D}}(\mathbf{x}) + \beta_{\mathcal{D}}\ \sigma_{\mathcal{D}}(\mathbf{x})$. It is clear that the higher $\beta_{\mathcal{D}}$ is, the smaller the estimated feasible (i.e., safety) region, directly affecting the location of the constrained minimizer.

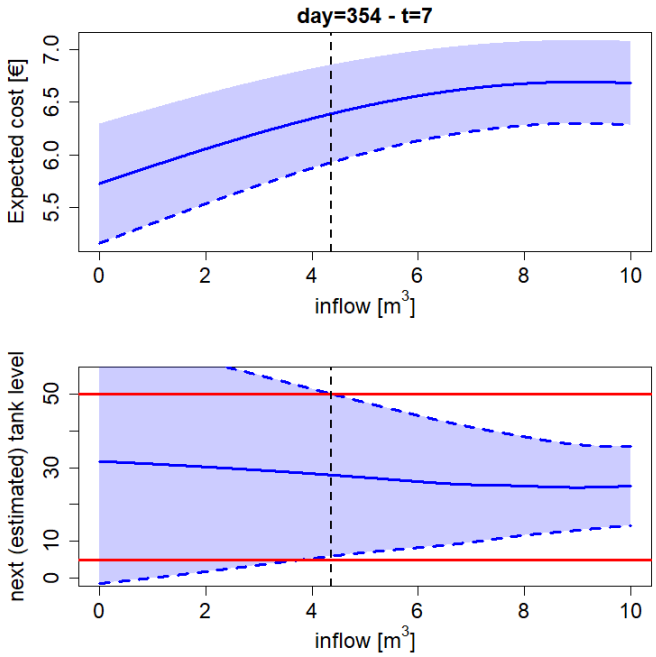


Figure 7. An illustrative example of how the proposed method works, for a specific day and hour of the day. (**top**) The GP regression model and the associated objective function $\mu_{\mathcal{P}}(\mathbf{x}) - \beta_{\mathcal{P}}(\mathbf{x})\ \sigma_{\mathcal{P}}(\mathbf{x})$ (dashed blue curve) to minimize; (**bottom**) GP regression model predicting the next water level in the tank: prediction and uncertainty are used to evaluate safety with respect to the max and min allowed level. Finally, the vertical dashed line is the optimal and safe solution a_t^* of the constrained optimization problem.

4. Results

In this section, the results obtained on the two case studies are reported, separately. The approach are evaluated in terms of efficiency improvement (i.e., cost reduction) and safety guarantees of the explored policies against the safe-by-design ones.

The validation schema is inspired by the well-known *leave-one-out validation* procedure largely adopted in ML. One day is left apart as a test, while all the others are used to train the GP models and to consequently apply our approach on the test day.

The energy costs obtained are compared with the historical data related to the safe-by-design policy. Moreover, the possible safety violations of the approach are also reported.

It is important to remark that *safety* has two different meanings in the two case studies: it is just related to *discomfort for the household* in the house heating system case study, while it refers to *service interruption and/or system disruption* in the water tank control case study.

4.1. Results on the House Heating System Case Study

As expected, operating the proposed approach leads to a reduction in the energy-related costs when compared with the safe-by-design policy. The overall cost saving is

approximately 2.5%. Figure 8 shows the box plot of the daily costs incurred by operating the two different policies, highlighting an overall shift towards lower costs over all the 60 days when the proposed approach is operated.

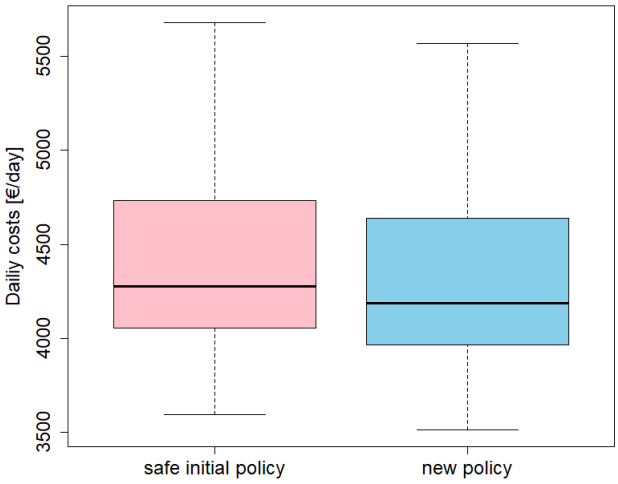


Figure 8. Comparison between the daily costs, over 60 days, incurred by operating the safe-by-design policy and the proposed approach, respectively denoted with safe initial policy and new policy in the chart.

On the other hand, the observed cost saving implies some safety violations. Specifically, the sovraelongation constraint (i.e., the in-house temperature must be lower than 19 °C within the first 30 min) is violated in 14 days out of 60. However, the entity of the violations is really small: the violating temperature is 19.2 °C, for both the average and median, with a standard deviation of 0.1 °C. Similarly, the constraints related to the allowed deviation for the target temperature are violated in 35 days, but, also in this case, the entities of the violations are affordable: the temperature violating the allowed deviation from below (i.e., lower than 17.5 °C) is 17.3 °C for the average and median, with a standard deviation of 0.3 °C, while the temperature violating the allowed deviation from above (i.e., higher than 18.5 °C) is 18.9 °C on average and 18.8 °C for the median, with a standard deviation of 0.3 °C.

Figure 9 compares the in-house temperatures under, respectively, the safe-by-design policy and the proposed approach, over a certain day. The goal of cost reduction, underlying the proposed approach, is clear: the in-house temperature is significantly lower, oscillating around the lower bound of the desired target range. Temperatures under the lower bound are due to erroneous modelling of the safety (and/or too small values of $\beta_{\mathcal{D}}$), which, in turn, is due to the fact that cost minimization leads the system to work in unseen conditions for which it is difficult to provide an accurate prediction.

Although a fine-tuning of the GP model and the $\beta_{\mathcal{D}}$ parameter could lead to an increase in terms of safety—but to a consequently worsening in terms of costs—we concluded that the resulting discomfort (i.e., safety violation) could be considered acceptable by the household when compared against the economical gain.

Finally, we are aware that the oscillating behaviour of the in-house temperature, implied by the proposed approach, is definitely far from the smooth one observed for the safe-by-design policy. Thus, we want to remark that our goal was not to replicate the actions operated by the PID controller, but to completely explore new policies starting from the safe experience collected so far. Therefore, the different observed behaviour has to be considered, in our opinion, as a positive result: our approach is actually able to discover new policies instead of replicating the original one. It is important to remark that, in any case, the oscillations could be removed/decreased through two possible strategies: adding a further constraint relative to smoothness or operating the control action less frequently (e.g., every 5 or 10 min, or every time the in-house temperature becomes too close to the

upper or lower bounds of the desired target interval). These strategies are not considered in this paper and will be investigated in future works.

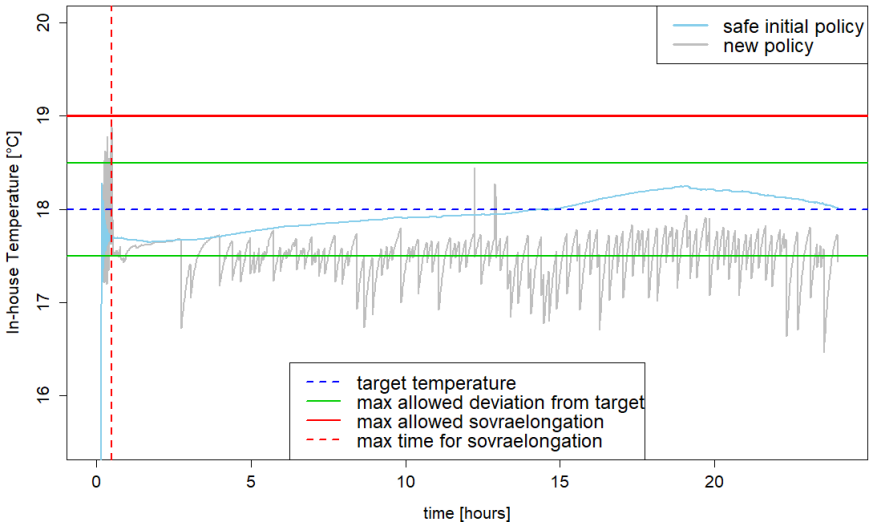
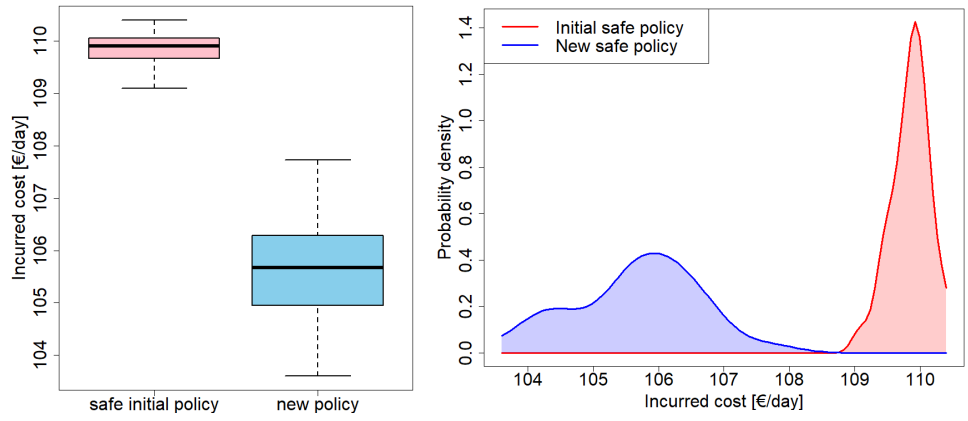


Figure 9. In-house temperature under the safe-by-design policy (light blue) and the proposed approach (gray) over a given day.

4.2. Results on the Water Tank Case Study

The results obtained on the water tank case study are exciting. According to the leave-out-procedure, the proposed approach was always able to drastically reduce energy-related costs **with no safety violations over all the 365 days**. This is crucial because, contrary to the previous case study, here a safety violation refers to a system disruption or a service interruption, two situations that must be definitely avoided.

Figure 10 shows the statistically significant difference between the daily costs incurred by operating the safe-by-design policy ("safe initial policy", in the chart) and those incurred by adopting the proposed approach ("new policy", in the chart). The reduction in terms of daily costs is clear in both the two reported charts: a box plot (on the left) and two probability density functions (on the right).



(**a**) Box plot. (**b**) Probability density function.

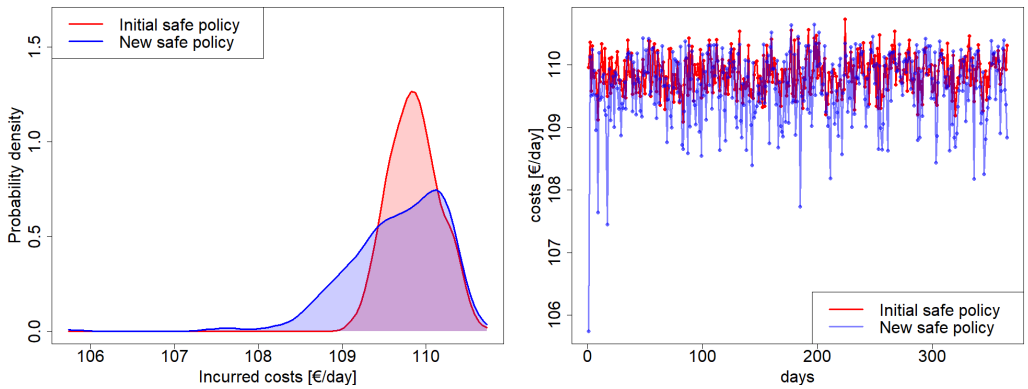
Figure 10. Comparison between daily energy-related costs incurred by using the safe-by-design ("safe initial policy") and the proposed approach ("new policy"). Daily energy-related costs of the proposed approach were computed through leave-one-out validation. The statistically significant difference in terms of daily costs is clear both in (**a**) the box plot and (**b**) the charts of the two probability density functions.

Long-run validation. An important consideration is that the water level in the tank, at the end of the day, is completely different when the proposed approach is used instead of the safe-by-design policy. Moreover, the resulting water level at the end of the day is significantly lower than the water level at the beginning of any other day in the dataset. However, the control of a real-life water tank must be operated in the long-run, from one day to the next without any interruption. This makes this case study more complicated than the previous one (where, instead, it is quite simple to keep the in-house temperature stable as soon as the target temperature is reached, due to the time horizon of the control). Thus, we decided to perform a long-run test by operating, in parallel and independently, the safe-by-design policy and our approach. Both start from the same initial water level in the tank and are operated for 365 days, subject to the same water demand.

As a result, the proposed approach again provides a relevant reduction in the energy-related costs. However, in order to avoid any possible safety violation, it was necessary to increase $\beta_{\mathcal{D}}$ from 4 to 5. This is reasonable because, in the long-run, the proposed approach leads the system to work in very unexplored conditions; thus, assuming a more precautionary attitude is more than advisable.

The cost reduction offered by the proposed approach can be clearly observed in Figure 11, as a comparison between both the two probability density functions and the cost time series. While the first chart shows that, overall, a large amount of high daily costs are shifted towards lower ones, the second demonstrates that the cost reduction occurs over the entire long-run period, not just at the beginning.

Finally, we report in Figure 12 the daily amount of water pumped into the tank, namely *inflow*, for the safe-by-design policy and our approach. Clearly, there is not any difference because this quantity is strictly linked to the water to be supplied to match the water demand (this is even more obvious in the long run). Indeed, the few outliers observed for the proposed approach refer just to the first day of the control, according to a lower water level in the tank, than the safe-by-design policy. From this consideration stems the most relevant result for the case study: since the daily inflows are significantly similar, the unique reason behind the cost reduction is a clever strategy for pumping and storing water into the tank. More simply, the proposed approach is able to pump water into the tank when this operation is more economically convenient, and, at the same time, it keeps the water level in the tank within a safety range.



(**a**) Probability density function. (**b**) Time series.

Figure 11. Comparing safe-by-design policy and proposed approach in the long run. Differences between daily costs as (**a**) probability density functions and (**b**) time series.

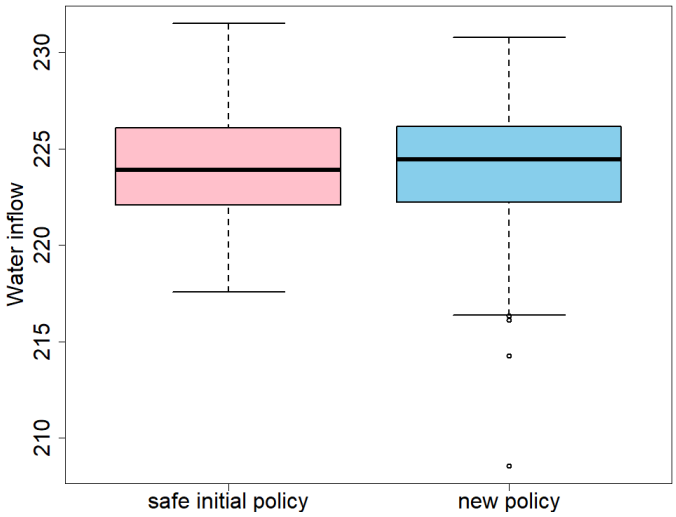


Figure 12. Comparison between the daily inflows operated by the safe-by-design policy and the proposed approach.

5. Discussion and Conclusions

We presented a novel approach to explore more efficient and safe control policies starting from data collected by having operated a safe-by-design one. The proposed approach does not require the expensive design and implementation of any digital twin of the target system, and deeply exploits the knowledge coded into a set of historical safe data (also known as safe experience).

As demonstrated in the paper, a more efficient policy can only be obtained by incurring some risk in terms of safety. Thus, uncertainty quantification becomes crucial to estimate such a risk, and GP modelling resulted in being a well-suited methodology for this purpose.

The safe exploration of new policies is obtained by solving a constrained optimization problem in which both the objective function and the safety constraints are predicted along with the associated uncertainty. Empirical results on two case studies, inspired from real-life systems, proved the effectiveness of the proposed approach, providing a significant cost reduction—with respect to the initial safe-by-design policy—without any relevant safety violation.

Indeed, the two case studies imply two different meanings of safety violation: a discomfort for the household in the house heating system case study and a system disruption or a service interruption in the water tank control case study. While relatively small violations can be considered acceptable in the first case, they must definitely be avoided in the second. The approach proved to work in the two different settings.

A limitation of the proposed approach can be found in the identification of the most suitable confidence interval for the safety estimation, namely the parameter $\beta_{\mathcal{D}}$ in the two case studies. It is difficult to choose a suitable value a priori: too large a value can lead to a poor performance improvement, while too small a value increases the risk for safety violations. Our suggestion is to adopt validation procedures, such as the leave-one-out validation used in our paper, to estimate a suitable value of the parameter and, in case, to slightly increase it before operating on the real-life system.

Ongoing work aims at (i) evaluating safety not only at time t but also looking ahead—even if this entails an additional computational cost—and (ii) augmenting the initial safe experience with all the new observations collected by operating the new approach. Indeed, as empirically demonstrated, our approach leads the system to work in previously unseen—but still safe—conditions and new observations can increase the knowledge about the overall behaviour of the target system. This would also help to address the previously reported limitation, by allowing us to adaptively refine the value of $\beta_{\mathcal{D}}$ over time as quickly as new knowledge is collected.

Author Contributions: Conceptualization, A.C. and F.A.; methodology, all the authors; software, A.P. and A.C.; validation, all the authors; investigation, E.M. and E.F.; writing—original draft preparation, A.C.; writing—review and editing, all the authors. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by the grant: ENERGIDRICA—Efficienza energetica nelle reti idriche (CUP B42F20000390006) Programma PON "Ricerca e Innovazione" 2014–2020—Azione II—OS 1.b.

Data Availability Statement: Code and data are available in our publicly accessible Github repository that does not issue DOIs: https://github.com/acandelieri/SafeExploringControlPolicies.git (accessed on 18 September 2023).

Acknowledgments: We greatly acknowledge the DEMS Data Science Lab, Department of Economics Management and Statistics (DEMS), University of Milano-Bicocca, for supporting this work by providing computational resources.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Rachih, H.; Mhada, F.; Chiheb, R. Simulation optimization of an inventory control model for a reverse logistics system. *Decis. Sci. Lett.* **2022**, *11*, 43–54. [CrossRef]
2. Chakraei, I.; Safavi, H.R.; Dandy, G.C.; Golmohammadi, M.H. Integrated simulation-optimization framework for water allocation based on sustainability of surface water and groundwater resources. *J. Water Resour. Plan. Manag.* **2021**, *147*, 05021001. [CrossRef]
3. Tordecilla, R.D.; Juan, A.A.; Montoya-Torres, J.R.; Quintero-Araujo, C.L.; Panadero, J. Simulation-optimization methods for designing and assessing resilient supply chain networks under uncertainty scenarios: A review. *Simul. Model. Pract. Theory* **2021**, *106*, 102166. [CrossRef] [PubMed]
4. Frazier, P.I. Bayesian optimization. In *Recent Advances in Optimization and Modeling of Contemporary Problems*; Informs: Catonsville, MD, USA, 2018; pp. 255–278.
5. Archetti, F.; Candelieri, A. *Bayesian Optimization and Data Science*; Springer: Cham, Switzerland, 2019.
6. Candelieri, A. A gentle introduction to bayesian optimization. In Proceedings of the 2021 Winter Simulation Conference (WSC), Phoenix, AZ, USA, 12–15 December 2021; pp. 1–16.
7. Garnett, R. *Bayesian Optimization*; Cambridge University Press: Cambridge, UK, 2023.
8. Schillinger, M.; Hartmann, B.; Skalecki, P.; Meister, M.; Nguyen-Tuong, D.; Nelles, O. Safe active learning and safe Bayesian optimization for tuning a PI-controller. *IFAC-PapersOnLine* **2017**, *50*, 5967–5972. [CrossRef]
9. Sui, Y.; Zhuang, V.; Burdick, J.; Yue, Y. Stagewise safe bayesian optimization with gaussian processes. In Proceedings of the International Conference on Machine Learning, PMLR, Stockholm, Sweden, 10–15 July 2018; pp. 4781–4789.
10. Kirschner, J.; Mutny, M.; Hiller, N.; Ischebeck, R.; Krause, A. Adaptive and safe Bayesian optimization in high dimensions via one-dimensional subspaces. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 10–15 June 2019; pp. 3429–3438.
11. Fiducioso, M.; Curi, S.; Schumacher, B.; Gwerder, M.; Krause, A. Safe contextual Bayesian optimization for sustainable room temperature PID control tuning. *arXiv* **2019**, arXiv:1906.12086.
12. Berkenkamp, F.; Krause, A.; Schoellig, A.P. Bayesian optimization with safety constraints: Safe and automatic parameter tuning in robotics. *Mach. Learn.* **2023**, *112*, 3713–3747. [CrossRef] [PubMed]
13. König, C.; Turchetta, M.; Lygeros, J.; Rupenyan, A.; Krause, A. Safe and efficient model-free adaptive control via bayesian optimization. In Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 30 May–5 June 2021; pp. 9782–9788.
14. Deisenroth, M.P.; Fox, D.; Rasmussen, C.E. Gaussian processes for data-efficient learning in robotics and control. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *37*, 408–423. [CrossRef] [PubMed]
15. Bischoff, B.; Nguyen-Tuong, D.; van Hoof, H.; McHutchon, A.; Rasmussen, C.E.; Knoll, A.; Peters, J.; Deisenroth, M.P. Policy search for learning robot control using sparse data. In Proceedings of the 2014 IEEE International Conference on Robotics and Automation (ICRA), Hong Kong, China, 31 May–7 June 2014; pp. 3882–3887.
16. Kamthe, S.; Deisenroth, M. Data-efficient reinforcement learning with probabilistic model predictive control. In Proceedings of the International Conference on Artificial Intelligence and Statistics, PMLR, Playa Blanca, Spain, 9–11 April 2018; pp. 1701–1710.
17. Sergeyev, Y.D.; Candelieri, A.; Kvasov, D.E.; Perego, R. Safe global optimization of expensive noisy black-box functions in the δ-Lipschitz framework. *Soft Comput.* **2020**, *24*, 17715–17735. [CrossRef]
18. Nägele, F.; Kasper, T.; Girod, B. Turning up the heat on obsolete thermostats: A simulation-based comparison of intelligent control approaches for residential heating systems. *Renew. Sustain. Energy Rev.* **2017**, *75*, 1254–1268. [CrossRef]
19. Akimov, V.I.; Polyakov, S.I.; Polukazakov, A.V. Design and Development of Cascade Heating Control for a «Smart» Residential Housing. In Proceedings of the 2020 International Russian Automation Conference (RusAutoCon), Sochi, Russia, 6–12 September 2020; pp. 42–48.

20. Ali, A.; Rahman, M.; Siddique, M.; Galib, S.; Nashiry, A. Design of an Automatic Rooftop Water Tank Filling System and Measurement of Consumed Water for Home Appliance. *Int. J. Autom. Smart Technol.* **2023**, *13*, 2371.
21. Xu, W.; Zhang, X.; Wang, H. A Water Tank Level Control System with Time Lag Using CGSA and Nonlinear Switch Decoration. *Appl. Syst. Innov.* **2023**, *6*, 12. [CrossRef]
22. Sun, C.; Puig, V.; Cembrano, G. Real-time control of urban water cycle under cyber-physical systems framework. *Water* **2020**, *12*, 406. [CrossRef]
23. Williams, C.K.; Rasmussen, C.E. *Gaussian Processes for Machine Learning*; MIT Press: Cambridge, MA, USA, 2006.
24. Gramacy, R.B. *Surrogates: Gaussian Process Modeling, Design, and Optimization for the Applied Sciences*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2020.