The Proof of the Pudding is in the Heating: A Field Experiment on Household Engagement with Heat Pump Flexibility

Baptiste Rigaux[1], Sam Hamels[2], and Marten Ovaere[3]

Department of Economics, Ghent University, Tweekerkenstraat 2, 9000 Gent (Belgium)
[1]baptiste.rigaux@ugent.be (Corresponding author), [2]sam.hamels@ugent.be, [3]marten.ovaere@ugent.be

December 6, 2024

Abstract

As renewable energy grows, flexible electricity demand becomes essential. We conducted a field experiment with nine heat pumps in well-insulated homes near Ghent, Belgium. During 287 flexibility interventions, we remotely deactivated heating until indoor temperatures reached predefined thresholds or households manually overruled the intervention. After initiating a flexibility event, the heat pump power is initially lowered by 250 W on average per unit in the fleet. As some heat pumps in the fleet reactivate, they consume more power to restore their threshold temperatures, triggering a rebound effect that gradually reduces net power savings achieved. On average, net power savings become zero after 18 hours, followed by a rebound period. Overall heat pump consumption was reduced by 1 kWh per event, stabilizing 36 hours after the event start. If flexibility activation is timed strategically, up to €1.1 can be saved through price arbitrage, assuming energy-crisis-level wholesale prices. Colder weather significantly influences savings, by increasing heat pump power available for flexibility but also amplifying rebound effects. This flexibility came with moderate comfort impacts: on average, indoor temperatures were 0.38°C lower during interventions. However, 19% of interventions were manually overruled when larger temperature drops occurred, with households citing discomfort, illness, or occupancy as factors on an online dashboard. These findings suggest that flexible residential heating can support renewable energy integration with moderate comfort impacts.

Keywords: Electricity Demand; Flexibility; Direct Load Control; Field Experiment; Household; Heat Pump; Thermal Comfort

JEL codes: Q40, Q41, D12

This research was supported by the "FlexSys" (A Flexible electricity System contributing to security of supply) project funded by the Energy Transition Fund of the Belgian federal government, managed by the FPS Economy, SMEs, Self-employed and Energy. M. Ovaere was funded by Research Foundation - Flanders (FWO) (mandate no. 12B7822N). B. Rigaux was funded by Research Foundation - Flanders (FWO) (mandate no. 11Q4224N). The authors are grateful to Joannes Laveyne and Nicolas Van Damme for helpful discussions during this research, and for the technical support in setting up and operating the field experiment.

1 Introduction

Electricity demand flexibility is crucial for the energy transition, especially with the growing variability of renewable sources like wind and solar. Achieving targets such as renewable energy accounting for 32% of gross final energy consumption in the EU by 2030 (Council of the European Union, 2018) requires changes not only in electricity supply but also in demand, with households and industries shifting their electricity consumption to periods when the sun shines or the wind blows. As the transition to renewables accelerates, projections show that the EU's flexibility resources for managing daily electricity consumption peaks will need to double between 2021 and 2030, reaching up to 362 TWh per year (European Environment Agency, 2023). Various economic incentives exist to support households in this transition, typically falling into two categories: implicit and explicit flexibility. Implicit flexibility occurs through time-varying prices and tariffs that encourage households to consume more when renewable supply is high and prices are low (due to the low marginal cost of renewable generation) and less when renewable supply is low and prices rise. Explicit flexibility programs, on the other hand, offer direct compensation from third parties, such as aggregators or Distribution System Operators (DSOs), for adjusting consumption to meet system or grid needs (Nouicer et al., 2020). Implicit and explicit flexibility benefit households, but they also produce wider societal benefits. Indeed, reduced system consumption peaks lower the need for costly, carbon-intensive peak power plants and fuel imports, while also decreasing the demand for reserve power plants and transmission capacity (Fischer & Madani, 2017). This, in turn, contributes to grid stability and minimizes the risk of power outages in a system increasingly based on renewable supply.

The adoption of low-carbon technologies, like electric vehicles (EVs) and heat pumps (HPs), is viewed as a critical step toward increasing residential demand flexibility. Heat pumps, in particular, hold significant potential for flexibility, as space and water heating —representing 78.3% of European households' final energy consumption (Eurostat, 2022)— can be adjusted by a few hours to better match electricity supply. This potential was recently demonstrated in practice by a study in the UK, where Bernard et al. (2024) found that HPs reduce gas use by 90% and CO_2 emissions by 36% in households where they are installed.

However, despite economic incentives like time-varying prices, many households may fail to fully exploit this flexibility due to several well-documented barriers. The behavioral economics literature on flexibility highlights for instance the status quo bias —where people tend to keep current habits even when incentives are offered— and risk and loss aversions, where perceived risks and potential losses are too high to trigger changes (see, e.g. (Darby & McKenna, 2012; Frederiks et al., 2015; Good, 2019; Good et al., 2017; Hobman et al., 2016)). Additionally, bounded rationality is a recognized barrier when the complexity or amount of available information about incentives overwhelms households, often leading to satisficing behavior, where simple heuristics for adjusting electricity consumption patterns result in sub-optimal gains (see (Frederiks et al., 2015; Good, 2019; Good et al., 2017; Gyamfi et al., 2013; Hobman et al., 2016; Kim & Shcherbakova, 2011)). This tendency is amplified when the incentives to shift electricity consumption trigger frequent and repeated adjustments, leading to what is referred to as response fatigue (Kim & Shcherbakova, 2011). Moreover, the delayed financial gains from adjusting consumption contribute to time-discounting behavior (see (Frederiks et al., 2015; Good, 2019; Hobman et al., 2016)). In addition to their timing, the magnitude of these gains may also play a role (Fabra et al., 2021): as noted by Bailey et al. (2024), they may sometimes be too small to offset the households' opportunity cost of the time they invest in adjusting their consumption.

Nowadays, manufacturers are addressing these challenges by developing tools such as smartphone apps and automation systems, which simplify control and reduce the effort needed to adjust electricity consumption. This raises two critical questions, central to our study: how willing are households to give up control over their assets, and how do they respond to potential discomfort, such as thermal discomfort? Accurately estimating the flexibility potential of households requires a better understanding of these behavioral barriers.

This paper studies the flexibility potential of residential HPs through a field experiment. Our treatment interventions involve temporarily switching off HPs until one of three predefined stopping scenarios is triggered: the indoor temperature drops below a certain threshold, the domestic hot water (DHW) tank

temperature falls below 40 °C, or the household overrides the intervention via an online platform, where they must provide a reason for doing so. In cases where interventions were notified a day in advance, households could also override preemptively. As such, this setup mimics what is commonly referred to as 'direct load control', where "a utility has the right to control the customers' appliances, usually based on a contract which is useful for peak demand reduction or emergency situation handling" (Kostková et al., 2013). By implementing a default and decentralized automation of the HPs, our experiment simplifies the typically complex decision-making process required for households to respond to flexibility incentives, and helps alleviate some of the behavioral barriers discussed above.

We conducted 287 flexibility interventions over two winter seasons (2022-2023 and 2023-2024). Before the first season, a survey revealed that the participants were generally richer, more environmentally conscious, and living in better-insulated homes compared to the Belgian average. This is expected, as less than 5% of Belgian households are equipped with a heat pump (Rosenow et al., 2022), and these are typically installed in homes with above-average insulation. The survey also explored how households formed expectations about flexibility and their comfort temperature preferences during interventions.

We find that the flexibility interventions unfold in two distinct phases. In the first phase, power consumption decreases to almost zero as the HP is temporarily switched off. In the second phase, power consumption increases beyond the usual levels as the HP compensates to return to the setpoint temperature (i.e. the predefined temperature chosen by the household on their thermostat). This phenomenon, known as a rebound peak, has been shown to occur when many automated flexible assets resume normal operation at the same time after periods with high day-ahead prices or high time-of-use tariffs (Dewangan et al., 2022; Ludwig & Winzer, 2022; Muratori et al., 2014).

To quantify HP flexibility during interventions, we measure five key variables: the duration of how long the HP remains turned off (hours), the power reduction over the course of the intervention (kW), the total decrease in electricity consumption during the intervention (kWh), the increase in electricity consumption after the intervention (kWh), and the financial savings resulting from the intervention (€). We conduct our analysis at two levels: the individual HP level, focusing on periods when HPs are blocked (referred to as 'interventions'), and the fleet level, capturing the aggregated response of both blocked HPs and those that have already got unblocked after an intervention is initiated in a fleet of HPs (referred to as 'flexibility events').

First, we analyze the interventions at the individual HP level. We find that interventions lasted an average of 12.8 hours before meeting the criteria for one of the stopping scenarios. A regression analysis of intervention duration shows that the strongest predictor of duration is the indoor temperature at the start of the intervention, with a one-degree increase extending the duration by 1.9 hours on average over the whole sample. Similarly, a higher initial DHW tank temperature and higher outdoor temperature positively influence intervention length. We did not find any evidence that households altered their behavior by raising the indoor temperature prior to pre-notified interventions. This is consistent with a post-experiment survey, where households indicated that they did not place high importance on being notified of upcoming interventions. Interestingly, the time of day when the intervention started had no measurable effect on its duration. During the interventions, HP power consumption dropped to around 50 W, as some electricity is still required to operate the HP's electrical boards and circulatory pumps. Using the average HP power profile outside intervention and rebound periods as a counterfactual, we estimate that HP power consumption is reduced by 84% on average during an intervention, translating to an average reduction in electricity consumption of 3.5 kWh over the duration of an intervention. However, this flexibility comes with significant post-intervention rebound effects, requiring an average additional 612 W of power consumption in the first post-intervention hour, or 2.5 kWh of electricity consumption over 16 hours to restore the indoor and DHW temperatures back to the user setpoint. A regression analysis of rebound energy consumption within 16 hours after an intervention stop shows that it is primarily driven by the difference between the household's setpoint temperature at the end of the intervention and the actual indoor temperature. Each one-degree increase in this difference leads to an average increase of 0.9 kWh in rebound energy consumption. Additionally, for each

degree increase in the average outdoor temperature during the 16-hour period, rebound energy consumption decreases by 0.7 kWh. Similar to intervention duration, the time of day when the intervention stops has no significant effect on rebound consumption.

Second, we broaden the scope of our analysis to study fleet-level behavior, aggregating all HPs into a single profile to quantify flexibility over time relative to the start of the intervention. We refer to this as flexibility events. Unlike the analysis of individual interventions, which focuses only on HPs during their blocked periods, flexibility events capture the aggregated response of the entire fleet after heating is temporarily deactivated. During a flexibility event, some HPs remain off, while others gradually get unblocked at different times—depending on building and household characteristics—and experience rebound effects. On average, the power reduction at the start of a flexibility event is around 250 W per HP, gradually decreasing to 0 W after 18 hours. Beyond this point, the fleet-level rebound peak occurs, during which the fleet consumes more electricity than it would have without the intervention. Flexibility events last longer than individual interventions because individual HPs reactivate at different points in time, resulting ina lower average fleet-level rebound than at the individual HP level, thereby allowing still-blocked HPs to offset the unblocked HPs' rebound impact and extend the period of the fleet-level net power reduction. We find a clear relationship with weather conditions. When average temperatures within the first 18 hours of an event are below 3°C, the initial power reduction reaches 600 W per HP, with energy consumption reduced by 4.5 kWh in the first phase. However, the rebound phase causes total net savings to drop to only 1 kWh after 36 hours. In contrast, when average temperatures are above 9°C, consumption decreases by 1.5 kWh, with no rebound observed, resulting in sustained savings. Note that all houses in our sample have an energy performance label A, indicating a high degree of thermal insulation and impacting both the observed power reductions and intervention durations.

To monetize flexibility events, we simulate financial gains using real day-ahead prices in Belgium. We calculate gains by assuming a flexibility event is initiated at each hour of the winter seasons 2022-2023 and 2023-2024. Using outdoor temperature data, each event is assigned to one of four average power reduction profiles—specific to defined temperature ranges—and matched with day-ahead electricity prices to estimate net savings. On average, net savings at 36 hours after the event start amount to €0.13 per event per HP. However, targeting periods of high and volatile electricity prices can increase net savings to €1.1 per event per HP. In practice, flexibility aggregators are likely to adopt such targeted strategies, attempting to maximize financial gains through day-ahead price volatility and other value streams.

Finally, we assess the discomfort experienced by households during the interventions. Our findings indicate that most interventions resulted in a modest temperature decrease of 0.38°C within a household, representing the average change across the entire intervention period. However, in cases where interventions were manually overruled, the temperature drop was significantly larger, reaching 1.06°C by the end of the intervention (i.e., at the moment of overrule). This suggests that households' observed overruling behavior aligns with their subjective perceptions of discomfort, though this is not consistent across all interventions, as the correlation between temperature drop and manual overrule is weak. This is further supported by the reasons provided by households upon overruling, which can be categorized into three main categories: "Too low temperature" (accounting for 65% of manual overrides); followed by "Sickness/health concerns", when illness triggers a desire for thermal comfort (20.5%), and "Presence at home", when an individual, usually working or studying at home, requires higher temperatures (16.5%). In a post-experiment survey, households reported experiencing low to moderate discomfort, with an average score of 2.4 on a 1–5 Likert scale (*No discomfort - Very large discomfort*), and confirmed a strong preference for being offered the option to overrule interventions, averaging 3.9 on a 1–5 scale (*Absolutely not important - Extremely important*).

Related literature and contributions

Our study contributes to the growing literature that studies the responses of households to incentives to make their electricity consumption more flexible. Research on time-varying prices and other forms of remuneration as strategies to shift consumption away from peak periods often shows weak or limited reac-

tions. For instance, Herter and Wayland (2010) find a 5% reduction in residential electricity consumption under a critical peak pricing scheme. More recently, Fabra et al. (2021) show that Spanish households do not significantly react to a real-time pricing scheme, while Enrich et al. (2024) found that consumption during peak periods only reduced by an average of 9% when the aforementioned scheme was later replaced by a time-of-use tariff. These relatively low reactions may be partly explained by a lack of awareness about incentives among the households involved (Enrich et al., 2024; Fabra et al., 2021). This suggests a need for technologies that make price incentives more effective and salient, such as in-home displays or luminous orbs, which can lead to stronger household responses compared to the absence of such technologies, as shown by (Allcott, 2011; Jessoe & Rapson, 2014). Additionally, numerous studies identify automation as a key enabling technology, reducing or even eliminating the need for households to manually adjust their electricity consumption to incentives (Bollinger & Hartmann, 2015, 2019; Faruqui & Sergici, 2010; Harding & Lamarche, 2016; Harding & Sexton, 2017; Herter et al., 2007).

In the behavioral economics literature on flexibility, a consensus is emerging in favor of automation, which is seen as a way to reduce the cognitive load imposed on households and to help eliminate well-documented barriers outlined above such as status quo bias, bounded rationality, and response fatigue (Darby & McKenna, 2012; Frederiks et al., 2015; Good et al., 2017). Acknowledging advancements in broader energy economics research, including studies on the behavioral aspects of residential flexibility, our paper positions itself within an emerging literature focusing on field trials of automated heating flexibility. In their study, Bailey et al. (2024) compare two automation treatments under time-varying price conditions. They find that households with access to an app to manage appliances (thermostats, hot water heaters, EV chargers) reduced peak consumption by about 5%, similar to those adjusting manually. However, the reduction jumps to 26% for households whose appliances were automatically controlled by the utility as part of the treatment. Similarly, Blonz et al. (2021) show that automating thermostats with time-of-use pricing reduces the electricity consumption of air conditioning units by 63% during peak periods in summer, resulting in savings of CA$0.21 per household per day of activation, with most participants experiencing no discomfort. Extending this, Fu et al. (2024) study the implementation of a time-of-use tariff in California, where electricity prices doubled during summer peak hours. Analyzing the response of households with smart thermostats, they show that this led to an average increase in air conditioner thermostat setpoints of 1.04°F (0.6°C), reducing runtime and lowering the utility's electricity load by 5% on the hottest summer days. Finally, Kane et al. (2024) conducted a recent field experiment using Wi-Fi-controlled switches to automatically manage A/C units during flexibility interventions, resulting in an 8.5% reduction in household electricity demand and a 2.3% reduction in CO_2 emissions during events.

The benefits of automation have also been demonstrated specifically in the context of HP flexibility. For example, Jensen et al. (2018) test the automatic and continuous adjustment of HPs of eight households based on electricity prices, weather forecasts and user-defined temperature preferences. They found that such automation reduced household energy costs by 6.8% to 16.9%, while avoiding significant discomfort. Similarly, Bernard et al. (2024) study a large-scale implementation of a specific time-of-use tariff for HPs on 6,631 households. Although automation was not a specific feature of the experiment, most participants reported using smart thermostats to automate their HPs, leading to a 50% reduction in electricity consumption during peak (expensive) periods and a significant shift to off-peak periods. This resulted in savings of up to 18% on households' yearly bill. In another pilot study by the same research center, 30 four-hour flexibility interventions were tested on 43 HPs (Centre for Net Zero & Nesta, 2024). Participants were allowed to set their own minimum and maximum temperature thresholds using smart thermostats, enabling homes to be preheated to the maximum during the first two hours of an intervention and then cooled down to the minimum during the blocking period. The study found a 74% reduction in HP consumption during the blocking period, with no significant rebound consumption afterward. Indoor temperatures were on average 0.85°C higher during preheating. Only 9% of events were overruled by participants in advance, mainly to avoid unnecessary preheating when they were away, and just 2% were overruled during events due to discomfort.

Despite this growing body of evidence on automated flexibility, there are three well-identified gaps

which make up our contributions to the literature. First, to the best of our knowledge, our paper, along with Bernard et al. (2024), is one of the first to provide quantitative estimates of the potential for HPs to deliver flexibility in real-world settings. This includes both load shedding and monetary savings, while integrating actual human behavior and preferences. Existing studies on residential flexibility are often constrained by the theoretical nature of these programs and rely heavily on methods like choice experiments for investigating stated preferences (Harold et al., 2021; Richter & Pollitt, 2018; Ruokamo et al., 2019; Yilmaz et al., 2022; Yilmaz et al., 2021) or on modeling approaches. For example, modeling in Georges et al. (2017) showed that a fleet of 100 HPs can provide an average flexibility potential of 510 W per unit during "downward modulation" events. However, Good (2019) states that existing works, including models, often assume "well-informed, rational actors". As a result, they may overlook real-world human behavior, especially when households make trade-offs involving thermal comfort. As Bernard et al. (2024) note, "there is currently no causal evidence on how heat pump interventions, particularly time-of-use tariffs, affect actual energy demand in practice.". Our study expands on this by providing further empirical evidence of HP flexibility.

Second, unlike most other studies, our treatment design does not explicitly rely on user-defined temperature setpoints or standards for comfort temperatures to guide interventions[1]. Instead, we set lower temperature thresholds exogenously, allowing us to explore household reactions in a manner agnostic to specific user-defined preferences, aiming to cover a broader range of the comfort spectrum. While future flexibility schemes may indeed enable households to specify comfort boundaries ex-ante or may include algorithms to minimize comfort impact, Zhang et al. (2016) highlight that standards for comfort temperatures can be overly conservative, while Aghniaey and Lawrence (2018) suggest that households can develop habits and tolerance for wider temperature ranges over time.

Third, our setup is specifically designed to test whether participants behave consistently with how their comfort is affected by the interventions. As such, we contribute to studies on households' thermal comfort. Only recently has research on residential flexibility begun to focus on the role of thermal comfort, including in determining households' economic gains (Da Fonseca et al., 2021). Although some of the works mentioned above acknowledge the role of comfort —often indicating that flexibility has no significant impact on this front— this is not true for all studies (e.g., (Bernard et al., 2024)). Furthermore, even when thermal comfort is discussed, it is rarely used to provide insights into the behavior observed during the experiment. To our knowledge, no existing research has directly compared households' subjective perceptions of discomfort with objective discomfort metrics, as we do in our study through the use of a proxy metric for discomfort.

The remainder of this paper is structured as follows: Section 2 details the experimental design and data collection. Section 3 presents the methods used to construct and graphically represent counterfactuals of key HP consumption variables during intervention periods, while Section 4 presents the results of the empirical analyses at both perspectives of flexibility interventions and flexibility events, quantifying power and energy consumption over time and discussing the impact on household comfort. Finally, Section 5 concludes and formulates policy recommendations for shaping the future of residential flexibility.

2 Experimental setting and data

2.1 Experimental setting

The experiment was carried out in collaboration with Energent, a Ghent-based local energy cooperative with around 2,000 members, developing renewable energy projects[2]. The cooperative selected nine participating households based on previous related and successful projects and coordinated the installation of hardware devices making it possible to remotely monitor and steer their HPs. All HPs in the sample are of the air-to-water type. Our research's target period is in the winter months, when HPs are used to heat homes. Therefore,

[1] See (Da Fonseca et al., 2021) for a review of various standards and metrics used to evaluate thermal discomfort in flexibility studies, and for instance, the standard established by the American Society of Heating, Refrigerating and Air-conditioning Engineers (ASHRAE), encountered in studies like (Kaspar et al., 2024).

[2] See https://energent.be/.

the experiment spans two heating seasons (HS): HS1 spans November 21, 2022 to April 15, 2023, while HS2 spans October 30, 2023 to March 24, 2024. The outdoor temperatures during both heating seasons were slightly milder than average, with a mean of 7.5 °C in HS1 and 8.3 °C in HS2, compared to the historic average of 5.4 °C for the same period in Belgium (see Appendix A.6).

Temporarily deactivating HPs can help balance supply and demand in an electricity system increasingly based on renewable generation. To explore how this affects electricity consumption and household behavior, our experiment examined the impact of flexibility interventions designed to replicate future flexibility management schemes of HPs. Specifically, the flexibility interventions involved momentarily turning off HPs in our sample, until they are reactivated when one of three predefined scenarios occurs. These scenarios are:

- The participants manually overrule the intervention via an online dashboard, where they are prompted to briefly state their reasons. This overrule interrupts the ongoing intervention and any intervention scheduled within the next 24 hours.
- The indoor temperature reaches a scheduled threshold value (see below).
- The temperature of the domestic hot water tank (DHW) inside the HP falls below 40 °C. This was done to avoid any impact on the ability of households to take a hot shower or bath whenever they wanted. In our sample, three out of the nine HPs, referred to as 'decoupled' HPs, were technically capable of turning off space heating separately from sanitary hot water provision. Therefore, this scenario was only applied to the 'non-decoupled' HPs, where space heating and DHW could only be controlled together.

When one of these scenarios is met, the HP resumes normal operation, i.e. restoring the space heating setpoint to the value selected by the household (e.g. 21°C) or going into standby if specified in a clock setting by the user.[3]

For each of the two heating seasons, a common schedule of 32 interventions was designed in advance, with interventions sent simultaneously to all HPs. Across both heating seasons, this resulted in up to 64 scheduled interventions per HP, depending on when it was added to the sample. The interventions vary by:

- The intervention's start time (2 a.m., 8 a.m., 2 p.m. or 8 p.m.)[4].
- The intervention's day of the week.
- The intervention's indoor temperature threshold value (16 °C, 17 °C, 18 °C or 19 °C).

The schedule was randomized across these dimensions and the interventions were distributed throughout the heating seasons, across both week- and weekend days. In HS1, households were informed of all 32 interventions one day in advance via text message. In HS2, notifications were sent for only a randomly selected half (16) of the 32 interventions, delivered via e-mail (presented in Appendix E). This randomization avoided correlations with other intervention characteristics, allowing us to test the hypothesis that households strategically reacted to receiving notification e-mails.

Appendix A.7 presents an overview of the sample composition for each heating season. In terms of sample attrition, Household 5 withdrew from the experiment near the end of January 2024, citing diminished motivation as a key reason. This was attributed to energy prices during HS2 having returned to relatively normal, in contrast to the most intense phase of the energy crisis experienced during HS1. Additionally, Household 4 did not participate in HS2 due to technical issues, Household 7 was equipped with the proper hardware right before the start of HS2 and Household 8 joined only at the end of January 2023 (during HS2).

[3]Note that our experiment used Smart Grid Ready (SGR) to communicate with the HPs, which is a standardized communication interface to either request or force HPs to turn on or off. It includes four possible commands, of which we only used "force off", instructing the HP to cease its operations. Due to limitations in the SGR standard –notably, the lack of control over the setpoint temperature to which rooms are heated–, it was impossible in our experimental setup to perform interventions where buildings are pre-heated to a higher setpoint before a HP is turned off, as done in (Centre for Net Zero & Nesta, 2023, 2024)

[4]Variation in the intervention start time ensures that we test households' responses to interventions at different times of the day, i.e. different levels of home occupation, including when the entire household is likely to be present.

Overall, there were eight HPs in each heating season, resulting in a final sample of 287 interventions across both seasons (168 in HS1 and 119 in HS2) —- somewhat lower than initially planned. This shortfall is due to several factors. First, connection losses between the HPs and the online platform prevented some interventions from taking place correctly. Also, for some HP brands, data collection relied on the manufacturer's API, which was occasionally interrupted by updates or limitations on the number of API calls allowed within a short period. Interventions with substantial missing data on core variables, such as indoor temperature or HP power consumption, were excluded from the final sample.[5] Furthermore, households turning off their HP for extended periods, such as during holidays, also contributed to the reduction in the number of completed interventions. However, the final dataset of 287 interventions is still random across the starting hour of the day, day of the week, and indoor temperature threshold, as shown in Appendix C.

2.2 Data

2.2.1 Survey data

Before the start of the first heating season, all nine participating households were invited to complete a detailed survey. The survey explained the upcoming flexibility interventions to their HPs and gathered data on their socio-demographic and housing characteristics. It also investigated attitudinal and behavioral factors, including stated preferences for comfort temperatures during and outside the flexibility interventions. One household did not complete the entire survey.

Table 1 presents the key characteristics of the participating households.[6] The findings indicate that our sample is not representative of the wider population in terms of socio-demographic, housing and likely pro-environmental characteristics. The participants can be viewed as generally larger, more educated and better-off households than the Belgium average. They live in bigger, newer and more energy-efficient[7] homes, which they also predominantly own rather than rent. They demonstrate a good understanding and awareness of the concepts related to the experiment, and can be considered as environmentally conscious.

This selection bias is anticipated since HP ownership itself is still not widespread in Belgium (Rosenow et al., 2022), likely drawing a demographic that differs from the general population.

2.2.2 Heat pump data

The HP sample includes two brands: Daikin and Viessmann. Each HP has been equipped for the experiment with a piece of hardware that provides us with a rich panel dataset at the 5-minute level on the power consumption of the HP, the indoor temperature inside the house (along with the setpoint temperature selected by the household), the outdoor temperature (measured by a sensor placed outside, away from direct exposure to the Sun), the temperature of the water in the DHW tank, the amount of electricity generated by solar panels, and an indicator showing whether the HP is currently blocked by an ongoing intervention.

Participants were offered the opportunity to monitor these variables (except for the intervention indicator) on an online dashboard developed by the project partner EnergieID, a Belgian company specialized in developing tools to monitor energy consumption [8].

3 Methods

To estimate the effect of interventions on various outcomes of interest, our analysis compares how a HP operates during an intervention to how it would operate in the absence of an intervention. Since our sample

[5] Long interventions, especially on decoupled HPs, are likely underrepresented in the final sample as they are more prone to connection losses or missing data, resulting in their exclusion from the final dataset.

[6] In addition, the survey also revealed that all participants are equipped with solar panels, and the yearly electricity consumption reported by seven respondents averages 5270 kWh.

[7] In Belgium's energy performance certificate rating system, a rating of A means a primary energy use of less than 100 kWh/m^2 per year. The rating ranges from A (best) to F (worst).

[8] See https://www.energyid.eu/en.

Table 1: Participants' characteristics

	Total respondents	Sample statistics	National average	
Household characteristics				
Mean household size (persons)	8	3.38	2.25	[a]
Mean number of children < 6 years old	8	.63	—	
Share of respondents and/or partner employed full time	8	100%	76.5%	[b]
Share of respondents and/or partner holding a university degree	8	100%	50.0%	[c]
Share of households with total monthly income > € 5,000	8	62.5%	—	[d]
Participants' housing characteristics				
Share residing in urban or suburban environment	8	87.5%	85.5%	[e]
Share residing in a semi-detached house	8	37.5%	42.1%	[f]
Share residing in an apartment	8	0%	22.9%	[g]
Share residing in a home surface > 150 m^2	8	50%	—	[h]
Share residing in a home built after 2006	8	37.5%	12.5%	[i]
Share home has been energy-retrofitted	8	75%	—	
Share energy performance certificate rated A	6	100%	—	[j]
Behavioral metrics				
Understanding of flexibility-related concepts [*]	9	3.39	—	
Pro-environmental behavior [**]	8	4.59	—	
Frequency of engagement in electricity-savings practices [***]	8	4.06	—	

[a] (Statbel, 2024b); [b] Of the population aged over 18 (Eurostat, 2024b).; [c] Of the 25-34 years old (Statbel, 2024a).; [d] The average household disposable income is $47,446 a year, i.e. about € 3,760/month in 2022 (OECD, 2024).
[e,f,g] (Eurostat, 2024c); [h] The average home size is 145.5 m^2 (Eurostat, 2024a).; [i] (Statbel, 2024c); [j] The average energy performance in non-retrofitted homes built after 2006 is between 160 (apartments) and 190 (single-family houses) kWh/m^2.year in primary energy use. Both correspond to a rating of B (See: https://www.epcwaarde.be/epc-score/, 21 August 2024).
[*] Scale 1-4 (*Never heard of it - I know a lot about it*) based on (Li et al., 2017); [**] Scale 1-5 (*Strongly disagree - Strongly agree*) and items based on (Bauwens & Devine-Wright, 2018); [***] Scale 1-5 (*Never - Always*) and items based on (Herabadi et al., 2021). Appendix D presents the full list of statements.

does not include a separate control group, we exploit the randomness of the intervention schedules of the two heating seasons, assuming that the average operation of the HP during non-intervention periods can serve as the control. We adopt two different approaches, depending on whether the outcomes (e.g., HP power consumption or indoor temperature) are analyzed with respect to a time variable expressed in absolute (e.g., time of day) or in relative terms (e.g., time elapsed since intervention start or stop). Specifically, absolute time variables allow direct comparisons of outcomes, while relative time variables require constructing counterfactuals

3.1 Direct comparison of daily profiles

When outcomes are analyzed with respect to the time of day, we directly compare the sample average of observations during and outside intervention periods. This comparison is assumed to estimate the treatment effect, given the randomized intervention schedules (see Appendix C) which support the assumption that the intervention treatment is independent of the potential outcomes (the independence assumption). Formally, for each outcome of interest y, we calculate the average outcome within the time-of-day bin t at a 5-min resolution during intervention periods (yielding $y^1(t)$, where the superscript 1 indicates the treatment period, i.e., the intervention) and outside intervention periods ($y^0(t)$), across the entire sample:

$$\hat{y}^z(t) = \frac{1}{n(S^z_{y,t})} \sum_{\tilde{y} \in S^z_{y,t}} \tilde{y} \tag{1}$$

where $z \in \{0, 1\}$ represents the treatment status, $S^z_{y,t}$ is the set of all observations of outcome y in treatment status z within the same time-of-day bin t and $n(S^0_{y,t})$ is the number of such observations. To minimize bias

in this comparison, we exclude observations from $y^0(t)$ if they occur less than 20 minutes prior to the start of an intervention (to account for any irregularities in HPs processing the intervention signal[9]) or within 16 hours after the end (to account for rebound effects on HP operation). These observations are therefore not included in the sets $S^0_{y,t}$. In addition, the sets $S^0_{y,t}$ include observations spanning one week prior to the first intervention of each heating season and one week after the last, defining the analysis window.

This direct comparison of averages reflects the imbalance in the data differently for $z = 0$ and $z = 1$: y^1 is more influenced by households with more observations in intervention periods (i.e., with more or longer interventions), while y^0 is more influenced by households with more observations in non-intervention periods (i.e., with fewer or shorter interventions). Therefore, a more robust estimation of the average treatment effect (ATE) of interventions on the outcome y, is given by the following fixed-effects (FE) regression model:

$$y_{h\tau} = \beta_1 \cdot I_{h\tau} + \alpha_h + \lambda_\tau + \varepsilon_{h\tau} \tag{2}$$

where $I_{h\tau}$ is an indicator for intervention periods on household h and at datetime τ across the experimental period, α_h denotes household fixed effects to account for characteristics that are constant across interventions, and λ_τ captures hour-of-day fixed effects. The coefficient β_1 identifies the ATE, representing the average effect of interventions on y, within households, controlling for household-specific characteristics. Given the small number of clusters (nine heat pumps), we estimate the model using a wild cluster bootstrap (with 100,000 repetitions) at the household-level to obtain cluster-robust standard errors (Cameron et al., 2008; MacKinnon et al., 2023).

3.2 Counterfactual construction with relative-to-absolute time variable mapping

For outcomes analyzed with respect to relative time variables, such as time elapsed since intervention start or stop, we use a three-step process. First, we construct household-specific counterfactual profiles by calculating the average outcome for each time of day during non-intervention periods for each household:

$$\hat{y}^0_h(t) = \frac{1}{n(S^0_{y,h,t})} \sum_{\tilde{y} \in S^0_{y,h,t}} \tilde{y} \tag{3}$$

where $S^0_{y,h,t}$ is the set of all observations of the outcome y for household h in time-of-day bin t during non-intervention periods and $n(S^0_{y,h,t})$ is the number of such observations. Observations occurring less than 20 minutes prior to the start or within 16 hours after the end of an intervention are excluded from the counterfactual calculation to minimize bias.

Second, for each intervention, we map every value of the relative time variable[10] t_{rel} to its corresponding time-of-day bin t. This mapping is unique to each intervention and constructs a dataset $S^0_{y,t_{rel}}$ at each t_{rel}, containing all the assigned counterfactual values across households.

Finally, we calculate the average counterfactual at each value of t_{rel} by taking the mean of the assigned counterfactual values over $S^0_{y,t_{rel}}$:

$$\hat{y}^0(t_{rel}) = \frac{1}{n(S^0_{y,t_{rel}})} \sum_{\tilde{y} \in S^0_{y,t_{rel}}} \tilde{y} \tag{4}$$

Where households with more interventions contribute more to both the counterfactual through $S^0_{y,t_{rel}}$, as well as the observed average $y(t_{rel})$, which ensures consistent handling of data imbalance in counterfactual and observed outcomes.

Given the limited numbers of clusters in the data, bootstrapped standard errors that fully account for both intra-household and inter-household variability would lack practical interpretability. To address the data constraints, we adopt a tractable approach to capture part of the variability in the mean counterfactual

[9] Heat pumps typically respond to the blocking signal within a few seconds, but this margin ensures minimum bias.
[10] Binned in 5-minutes intervals, to match data resolution.

at any value of t_{rel} by assuming independence among observations in $S^0_{y,t_{rel}}$ and estimating the standard error using within-bin variability:

$$SE(\hat{y}^0(t_{rel})) = \frac{SD(S^0_{y,t_{rel}})}{\sqrt{n(S^0_{y,t_{rel}})}} \tag{5}$$

where $SD(S^0_{y,t_{rel}})$ is the standard deviation of the observations in $S^0_{y,t_{rel}}$. The standard errors represent the variability in average counterfactual power across interventions, pooling observations from different households, at each value of t_{rel}. This method likely results in an underestimation of the true variability.

3.3 Graphical representation

To facilitate interpretation of the figures, the results of the empirical analysis, including the mean estimates across the sample and the confidence intervals constructed using the approach described above, are smoothed in the figures using local polynomial regressions. With the polynomial degree set to 0 (identified as the optimal value), this approach corresponds to a Nadayara-Watson regression (see (Nadaraya, 1964; Watson, 1964)). The bandwidths are set to their optimal cross-validated values, and the weights are determined by the Epanechnikov kernel. This approach produces a locally weighted smoothing of the sample averages, where the kernel assigns weights within the bandwidth.

4 Results

4.1 Effects of flexibility interventions on individual heat pumps

In this section, we analyze the key factors summarizing the impact of flexibility interventions at the individual HP level. We examine why and when interventions ended, quantify their impact on both ambient temperature and HP power usage, and characterize the post-intervention increase in electricity usage.

4.1.1 Duration of interventions

Table 2 shows a breakdown of the duration d of the 287 studied interventions (in hours) for each of the three reasons for stopping an intervention. The most common scenario was the automatic stop triggered by the DHW temperature threshold amongst non-decoupled HPs, accounting for 201 of all stops. Fifty-four interventions were manually stopped by households, including eight that were preemptively overruled. Thirty-two interventions ended when the indoor temperature threshold was reached. Specifically, twenty-four stopped at the 19 °C threshold, six at at the 18 °C threshold, and only two at the 17 °C threshold. Notably, the 16 °C threshold was never reached. This indicates that households avoided letting their homes cool down too much, as shown by the histogram of indoor temperatures during non-intervention periods presented in Appendix A.1.

Table 2: Intervention durations d (in hours) for each reason for stopping an intervention

Reason for intervention stop	N	Percent	Average d	95% CI	Min	Median	Max
DHW temperature threshold	201	70.0%	12.68	11.21, 14.15	0	11.51	82.08
Manual overrule	54	18.8%	15.51	9.82, 21.20	0	11.49	135.80
Indoor temperature threshold	32	11.2%	9.30	5.35, 13.26	0	4.71	34.67
Full sample	287	100%	12.84	11.30, 14.37	0	11.17	135.80

Table 2 shows that interventions in the full sample lasted 12.84 hours on average. Interventions ending by a manual overrule last the longest, while interventions ending because of the indoor temperature threshold last the shortest. However, three t-tests with unequal variances show that the differences in average intervention duration across the stopping scenarios are not significant at the 5% level.

These averages conceal significant variation among individual interventions, as shown by the histogram in Fig. 1. Approximately 15.3% of interventions lasted less than one hour and 12.5% of interventions over 24 hours. This substantial variation suggests that several factors, both endogenous (e.g., household comfort preferences) and exogenous (e.g., weather conditions, time of day the intervention is initiated), may influence the duration of flexibility interventions.

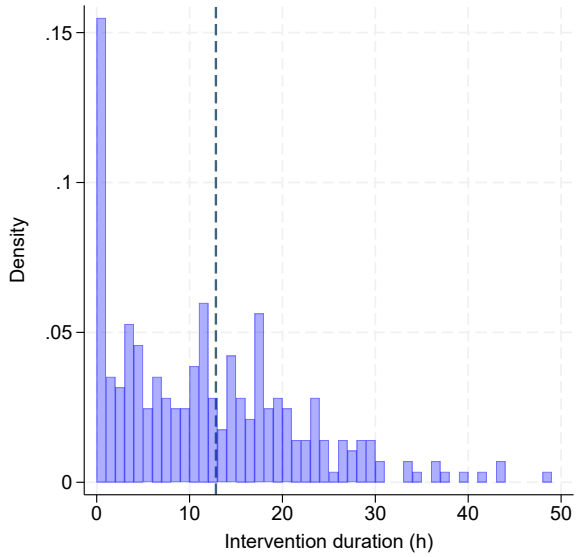


Figure 1: Histogram of the intervention durations d (in hours) of flexibility interventions ($N = 287$). The dotted line indicates the mean duration. The histogram is truncated at $d = 50$ hours to exclude outliers.

To determine which external factors influence the duration of an intervention i on household h (d_{ih}, in hours), we estimate the following regression specification with households FE:

$$\begin{aligned} d_{ih} =& \beta_1 \cdot D_{\text{notif},ih} + \beta_2 \cdot T^0_{\text{in},ih} + \beta_3 \cdot \min(T^{\leq 5\text{h}}_{\text{out},ih}) \\ &+ \beta_4 \cdot \text{TOD}_{\text{2AM},ih} + \beta_5 \cdot \text{TOD}_{\text{8AM},ih} + \beta_6 \cdot \text{TOD}_{\text{8PM},ih} \\ &+ \mathbb{1}(\text{FE} = 0) \cdot \left(\beta_7 \cdot T^0_{\text{DHW},ih} \cdot \delta_{\text{decoupled}} + \beta_8 \cdot T^0_{\text{DHW},ih} \cdot \left(1 - \delta_{\text{decoupled}}\right) \cdot \delta_{\left\{T^0_{\text{DHW},ih} \geq 40°\text{C}\right\}} \right) \\ &+ \mathbb{1}(\text{FE} = 1) \cdot \beta_9 \cdot T^0_{\text{DHW},ih} + \mathbb{1}(\text{FE} = 1) \cdot \alpha_h + \mathbb{1}(\text{FE} = 0) \cdot \beta_0 + \varepsilon_{ih} \end{aligned} \tag{6}$$

Where the initial conditions of the intervention are represented by the following parameters: $D_{\text{notif},ih}$, a dummy variable indicating whether the intervention i on household h was pre-notified; $T^0_{\text{in},ih}$ and $T^0_{\text{DHW},ih}$, the indoor and DHW temperatures at the start of the intervention[11], and three time-of-day dummies for start of the intervention ($\text{TOD}_{\text{2AM},ih}$, $\text{TOD}_{\text{8AM},ih}$, $\text{TOD}_{\text{8PM},ih}$ with 2 p.m. as the base category). The dynamic component of the intervention is captured by $\min(T^{\leq 5\text{h}}_{\text{out},ih})$, the minimum outdoor temperature observed within five hours after the start of the intervention. This variable has been found to lead to slightly higher adjusted R^2 values than other parametrizations (e.g. the average outdoor temperature during the intervention).

We estimate the model using linear regression, controlling for household FE (α_h in eq. (6)) to capture within-household variation by accounting for household characteristics that are invariant across interventions, such as insulation levels and comfort temperature preferences. As the decoupled dummy ($\delta_{\text{decoupled}}$) is constant across observations for a given household, the DHW temperature at the start of the intervention enters the FE model via a single parameter $T^0_{\text{DHW},ih}$, interpreted differently. To address the small number of clusters (nine HPs), we use wild cluster bootstrap (100,000 repetitions) at the HP level to compute cluster-robust standard errors and report the p-values derived from the empirical distribution of the bootstrapped

[11] The parameter $T^0_{\text{DHW},ih}$ is split into two terms to distinguish between decoupled and non-decoupled HPs.

estimates. Table 3 presents the results for four models, from a parsimonious specification to the full model with household FE in eq. (6).

Table 3: Linear regression results for intervention duration (in hours)

	(1)	(2)	(3)	(4)
D_{notif}		-2.766	-1.645	-1.936
		(0.682)	(0.764)	(0.730)
T^0_{in}	1.741*	1.863*	2.653	2.854+
	(0.019)	(0.012)	(0.116)	(0.082)
$\min(T^{\leq 5\text{h}}_{\text{out}})$	0.562*	0.585*	0.604*	0.664*
	(0.036)	(0.047)	(0.023)	(0.016)
$T^0_{\text{DHW}} \cdot \delta_{\text{decoupled}}$	0.340*	0.329*		
	(0.015)	(0.020)		
$T^0_{\text{DHW}} \cdot (1-\delta_{\text{decoupled}}) \cdot \delta_{\{T^0_{\text{DHW}} \geq 40°\text{C}\}}$	0.197+	0.186		
	(0.086)	(0.129)		
T^0_{DHW}			0.573+	0.567*
			(0.069)	(0.046)
TOD_{2AM}				-0.678
				(0.536)
TOD_{8AM}				-0.756
				(0.641)
TOD_{8PM}				-2.819+
				(0.090)
Constant	-36.732**	-36.659**		
	(0.008)	(0.007)		
Household-FE	No	No	Yes	Yes
Adj. R-Square	0.137	0.141	0.191	0.189
N observations	286	286	286	286

Linear regression estimates. Models (3) and (4) include household fixed effects. The reference category for TOD is 2 p.m. P-values (in parentheses) are derived from wild cluster bootstrapped standard errors (100,000 repetitions) clustered at the household level (nine clusters for all models). $^{+}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$.

The results reveal consistent general trends across models. Although the significance and magnitude of the estimates vary across specifications, they show expected signs. In particular, intervention duration increases with the indoor temperature at the start, the DHW temperature at the start (both reflecting the thermal reservoir available for the intervention) and the minimum temperature within five hours after the start (which affects the rate at which it depletes). Another key general trend is that the dummy variable for day-ahead notification of the intervention is insignificant in all models where it is included, suggesting that preemptive overrules by households in response to notifications were minimal, or that households did not consistently increase their thermostat temperature in preparation for an intervention. As a result, this variable is not relevant and excluded from further analyses.

Estimates remain robust within the same type of models. Among the non-FE models, Model (2) achieves the highest adjusted R^2. It shows that the indoor temperature at the start has the largest marginal effect: a one-degree rise in T^0_{in} increases the intervention by 1.9 hours on average. Outdoor temperature also has a positive effect, with each additional degree in $\min(T^{\leq 5\text{h}}_{\text{out}})$ increasing the duration by about 35 minutes. The effect of a one-degree increase in DHW temperature at the start is not significant for non-decoupled HPs[12] but contributes an average 20-minute increase in intervention duration on decoupled HPs.

[12]It should be noted that the variation in non-zero observations for $T^0_{\text{DHW}} \cdot (1-\delta_{\text{decoupled}}) \cdot \delta_{\{T^0_{\text{DHW}} \geq 40°\text{C}\}}$ is limited, where the 95% CI ranges narrowly from 46.97 to 47.83 °C (213 observations). The 95% CI for $T^0_{\text{DHW}} \cdot \delta_{\text{decoupled}}$ spans 39.94 to 42.97 °C (50 observations).

Among the FE models, our preferred specification is Model (4) which adds the TOD dummies, with only a 1% drop in adjusted R^2 compared to Model (3). The marginal effect of the minimum outdoor temperature is similar to the non-FE models: a one-degree increase in $\min(T_{\text{out}}^{\leq 5\text{h}})$ extends the intervention by approximately 40 minutes on average. However, adding household FE further changes the estimates for parameters typically affected by household HP usage patterns. For example, a one-degree increase in the initial DHW temperature extends the intervention by 34 minutes on average within a household. The estimate for the indoor temperature at the start is borderline significant ($p = 0.08$), likely because there is little variation in indoor temperature within households, as most avoid letting their homes get too cold, as noted earlier. Still, within a household, a one-degree increase in T_{in}^{0} extends the intervention by nearly 2.9 hours at the 10% level. Interestingly, the time of day when the intervention starts does not significantly affect its duration at the 5% level. However, Model (4) shows evidence at the 10% level that interventions starting at 8 p.m. are shorter than those starting at 2 p.m. by an average 2.8 hours. This may be because interventions starting in the evening do not benefit from cooking activities or solar irradiance (even after controlling for outdoor temperature), which contribute to reheating the home during interventions initiated over the afternoon, thereby extending their duration.

4.1.2 Reduction in indoor temperature during interventions

Fig. 2 (left panel) shows the average indoor temperature daily profiles during and outside intervention periods, across all HPs. During non-intervention periods, indoor temperature remains stable throughout the day, rising from about 20.4°C in the morning to 20.8°C by late afternoon, averaging 20.59°C over the day. Interventions reduce this daily average to 20.43°C, i.e. a small but statistically significant difference of 0.16°C at the 1% level, as shown by a t-test for equality of means with unequal variances.

However, this result does not account for household-specific characteristics or data imbalance. To address this, we estimate the household FE model in eq. (2), which provides an ATE of -0.38°C (95% CI: -0.66, -0.10 °C; $p = 0.02$ derived from wild cluster robust standard errors at the household level). This ATE represents the average within-household reduction in indoor temperature caused by interventions over their entire duration, controlling for household characteristics invariant across interventions.

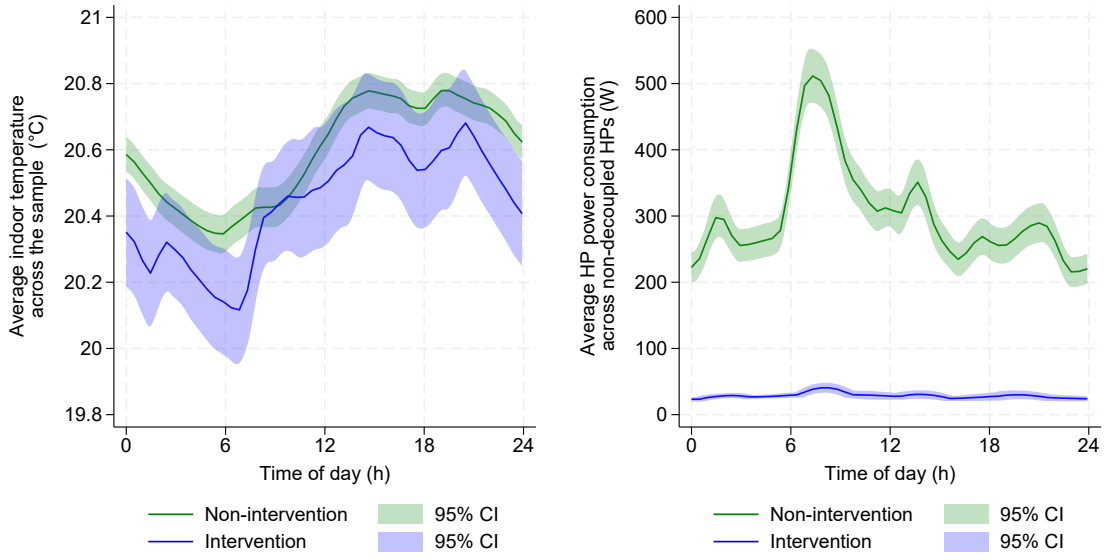


Figure 2: Daily profiles of average indoor temperature (left panel) and heat pump power (right panel, decoupled units only) during and outside intervention periods, averaged across heat pumps and heating seasons. The profiles are smoothed using local polynomial regression of degree 0 for the mean and confidence intervals. Standard errors reflect the variability of the mean in 5-min-of-day bins, assuming independence among observations.

4.1.3 Reduction in power consumption during interventions

Fig. 2 (right panel) shows the average reduction in HP power consumption during intervention periods, a key outcome of flexibility programs, for non-decoupled units only[13]. On non-intervention days, households typically begin heating their homes around 6 a.m., resulting in a peak power of around 500 W at 7 a.m., due to the demand for both space and domestic water heating. Throughout the day, HP power consumption averages 338 W.

During interventions, HP power does not become zero, but is reduced by an average 285 W (statistically significant at the 1% level, based on a two-sample t-test for equality of means with unequal variances), resulting in an average consumption of 53 W over the entire HP sample, i.e., a reduction of 84%. This residual power consumption supports the HP's essential functions, including maintaining electrical circuits, circulatory pumps, and its connection to the online control platform. The household FE regression model in eq. (2) provides an ATE estimate of -292 W (95% CI: -390, -200 W; $p < 0.01$ derived from wild cluster robust standard errors at the household level). This accounts for data imbalance across households and represents the average within-household reduction in power due to interventions, averaged over their duration.

Appendix A.3 shows that HP power consumption is negatively correlated with outdoor temperature, as one expects, with a Pearson correlation of -0.25 ($p < 0.01$) estimated across the entire sample. The profiles for different outdoor temperature ranges indicate that, as outdoor temperatures increase, power peaks shorten, and overall power consumption decreases, while the overall shape of the daily profile (in terms of overall peak timing and pattern) remains consistent.

4.1.4 Reduction in electricity consumption during interventions

The amount of electricity consumption reduction during an intervention is calculated as:

$$\Delta E_{ih}^{during} = \int_{t_0}^{t_f} \left(P_{cf,ih}(t) - P_{obs,ih}(t) \right) \mathrm{d}t \tag{7}$$

Where ΔE_{ih}^{during} represents the decrease in electricity consumption (in kWh) observed in household h during intervention i, from its start at t_0 to its end at t_f; $P_{cf,ih}(t)$ is the household-specific counterfactual power consumption (as defined in eq. (3)) during intervention i if no intervention had occurred, and $P_{obs,ih}(t)$ is the observed (residual) consumption. The histogram of ΔE_{ih}^{during} for all interventions is shown in Fig. 3. On average, 3.49 kWh (95% CI[14]: 3.10, 3.88 kWh) of electricity was saved during the interventions. As expected, it is strongly and positively correlated with the intervention duration d, with a Pearson coefficient of 0.73 (significant at the 1% level).

4.1.5 Increase in electricity consumption after the interventions

The estimate ΔE_{ih}^{during} reflects the reduction in energy consumption during an intervention only. However, when the intervention ends, the HP resumes operation from a lower indoor and/or DHW temperature (depending on whether the unit is decoupled) than the household's typical thermostat setpoint for that time of day. This leads to increased power consumption during the post-intervention period, as the HP works to restore the home and/or the DHW tank temperatures to the setpoint. This phenomenon, often referred to as the rebound peak in the literature (Dewangan et al., 2022; Ludwig & Winzer, 2022; Muratori et al., 2014), is calculated for each intervention as a function of t, the time relative to intervention stop:

$$\Delta P_{ih}(t) = P_{obs,ih}(t) - P_{cf,ih}(t) \tag{8}$$

[13] Appendix A.2 shows power profiles during interventions for the entire sample, including both decoupled and non-decoupled HPs. Decoupled units occasionally exhibit multi-kW power spikes to heat water for DHW usage or to above 60°C to prevent the spread of Legionella bacteria, but these do not contribute to heating the ambient space.

[14] These confidence intervals reflect variability across interventions but likely underestimate the true variability, as they do not account for intra-household correlation or variability around the mean counterfactual power level at each time point.

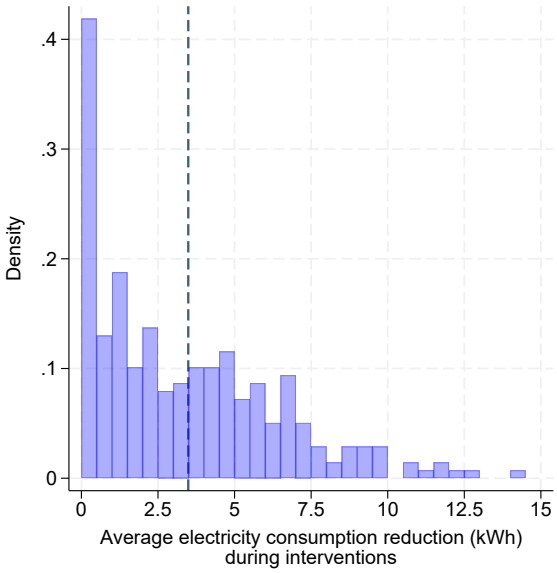


Figure 3: Histogram of the average decrease in electricity consumption (in kWh) during each intervention, computed using the average daily heat pump consumption profile as the counterfactual (defined in eq. (3)). The dotted line indicates the mean energy saved. The histogram is truncated at 15 kWh to exclude outliers.

A positive value for $\Delta P_{ih}(t)$ indicates that the HP consumption (in W) is higher than it would have been, had no intervention occurred. Additionally, this extra power consumption can be integrated over time to quantify the additional electricity consumption at a period Δt after the intervention stop :

$$\Delta E_{ih}^{after}(\Delta t) = \int_{t_f}^{t_f+\Delta t} \left(P_{obs,ih}(t) - P_{cf,ih}(t)\right) \mathrm{d}t \tag{9}$$

Fig. 4 illustrates the rebound peak in terms of additional power consumption (left panel, in W) estimated via eq. (8) and energy consumption (right panel, in kWh) estimated via eq. (9) with $\Delta t = 16h$ after intervention stop. Most of the rebound effect occurs within the first eight hours after the intervention, with approximately one hour required for the rebound to settle in (as shown in the left panel). On average, the excess power consumption jumps to around 700 W just after the intervention stops, with an average of 612 W (95% CI[15]: 576, 647 W) in the first post-intervention hour, but quickly drops, with an average of 275 W (95% CI: 265,285 W) within the first eight hours. The average electricity consumption rebound within the first 16 hours is 2.46 kWh (95% CI: 1.86, 3.06 kWh).

We conclude that an intervention consists of two phases. During the first phase, consumption decreases by an average of about 3.5 kWh. In the second phase, the rebound phase following the intervention stop, the consumption increases by approximately 2.5 kWh on average. Overall, the average net effect of an intervention, accounting for the rebound over the 16 hours post-stop, is a reduction of approximately 1 kWh in HP electricity consumption.

To study the factors influencing rebound energy consumption, Appendix F presents a regression analysis of ΔE_{ih}^{after} at 16 hours post-intervention, with outdoor conditions during the rebound period and indoor temperature at intervention stop as regressors. The model is estimated with wild cluster bootstrap to address the small number of clusters. Results show that rebound consumption depends not on the indoor temperature at intervention stop itself, but on the difference between the indoor temperature and the setpoint temperature[16]. In Model (3) with household FE, each 1°C increase in this difference leads to an additional 0.88 kWh of rebound consumption on average and within a household at 16 hours post-intervention. Additionally, a 1 °C rise in the average outdoor temperature over the 16 hours post-stop period reduces rebound by 0.72 kWh.

[15]These confidence intervals reflect variability across interventions but likely underestimate the true variability, as they do not account for intra-household correlation, autocorrelation structures in errors over time, or variability around the mean counterfactual power level at each time point.

[16]The temperature set on the thermostat by the user, which would have been reached without the intervention.

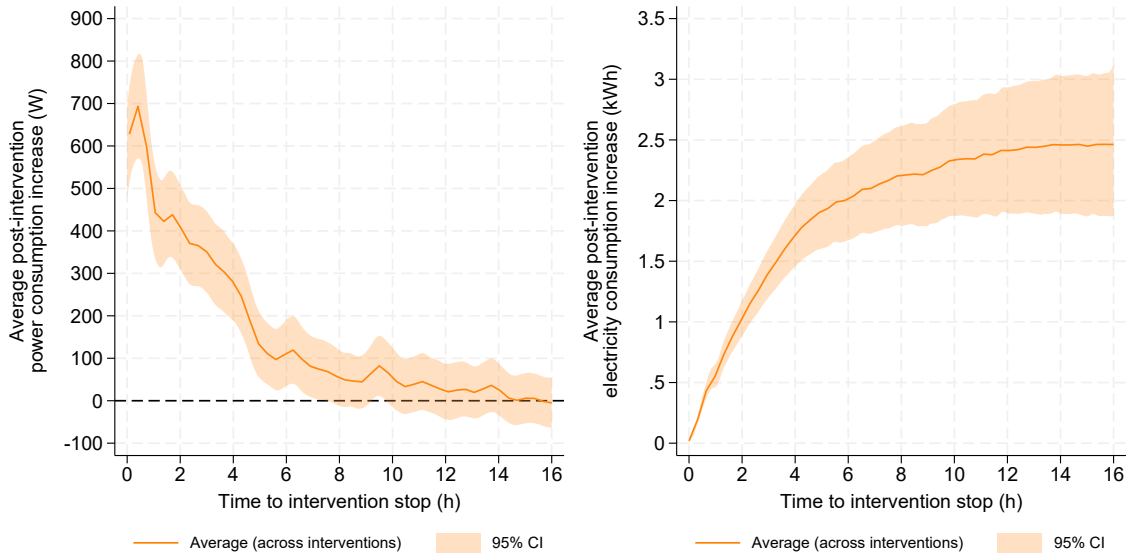


Figure 4: Average increase in heat pump power consumption during the post-intervention rebound period (left panel, in W) and the corresponding energy consumption (right panel, in kWh) across all interventions. The averages are computed using the average daily heat pump consumption profile as the counterfactual (defined in eq. (3)). Standard errors reflect the variability of the mean in 5-min-to-intervention-stop bins, assuming independence among observations. The means and confidence intervals are smoothed using local polynomial regression of degree 0.

Notably, the time of day when the intervention ends has no significant effect once these factors are controlled for.

4.2 Effects of flexibility events on a fleet of heat pumps

For a flexibility program operator, understanding individual HP responses to interventions is secondary to assessing the behavior of an entire fleet of flexible HPs. Once an intervention is initiated fleet-wide, the operator needs reliable estimates of aggregate power reductions, their evolution over time, and the resulting financial savings from reduced consumption during peak hours. This shift in focus acknowledges that, at any point after an intervention begins, some HPs may have already resumed operation, thereby consuming more power than the counterfactual during their rebound period each, while the fleet as a whole may still achieve a net power reduction. This section analyzes these fleet-level dynamics during what we refer to as 'flexibility events', combining interventions and HPs into a single profile relative to time to intervention start.

4.2.1 Fleet-level power consumption profiles during flexibility events

The left panel of Fig. 5 shows the aggregated power consumption per HP in the fleet, averaged over the entire intervention sample and across both heating seasons, relative to the time to intervention start. The figure compares the average fleet power consumption per HP to the control HP power level. The fleet-level control consumption, as defined in eq. (4), is derived by aligning all HP-specific counterfactual consumption values during interventions by time to intervention start. As a result, the fleet-level control shows periodic large and smaller peaks, corresponding to the two main –morning and midday– consumption peaks observed in the right panel of Fig. 2. Additionally, the right panel of Fig. 5 shows the resulting average net power reduction, calculated as the difference between the intervention and control power levels at each value of time to intervention start.

In the pre-flexibility event period, the power consumption of soon-to-be-blocked HPs aligns closely with the control power (green curve, left panel). Once the event begins and interventions are initiated on all HPs in the fleet, the average power consumption per HP quickly drops to just under 100 W (left), resulting in a power reduction of around 250 W at the start of the intervention (right). As the event progresses, some

HPs resume normal operation, while others remain blocked by the still-ongoing intervention, leading to a gradual increase in the average fleet power consumption (blue curve, left). This gradually reduces the fleet-level power reduction over time (right) until it eventually turns negative. Around 18 hours after the event starts, the fleet's average power consumption (left) exceeds the control consumption. At this point, the flexibility event no longer reduces electricity consumption, and the fleet begins consuming more electricity than usual. This may be viewed as the fleet-level equivalent of the rebound period in the period following the end of flexibility interventions on individual HPs. Similarly, the average power reduction drops to 0 W after about 18 hours (right). Over the unsmoothed intervention data, the net power reduction averages 251 W during the first hour of a flexibility event (95% CI: 234, 268 W), decreasing to 110 W on average over the first 18 hours (95% CI: 101, 119 W). Between 18 and 36 hours, the fleet-level rebound is limited to just 47 W on average (95% CI: 42, 53 W), and up to only 132 W. These values align closely with the trends depicted in the locally smoothed plots.

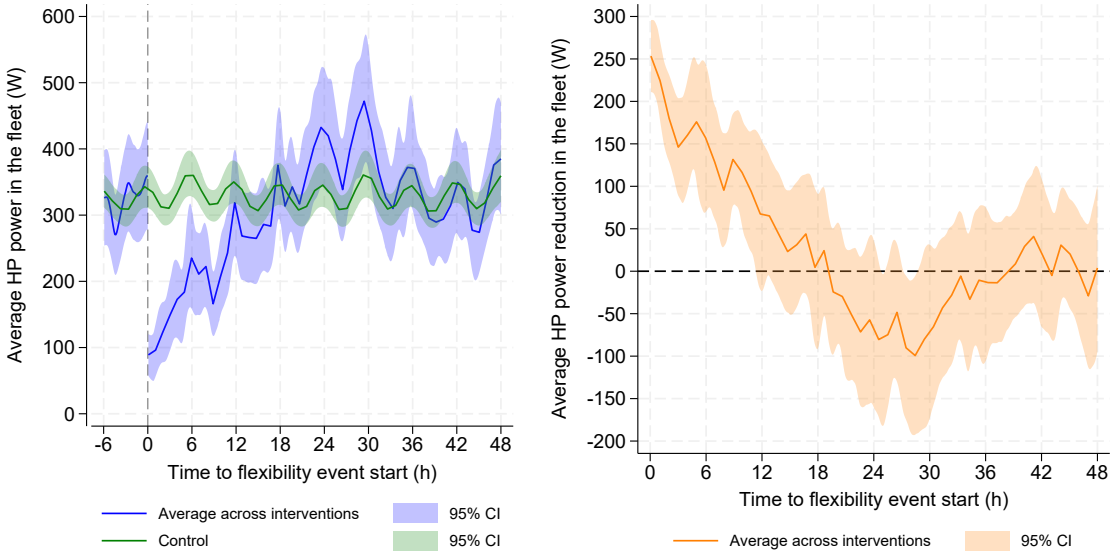


Figure 5: Average heat pump power consumption per unit in the fleet relative to the time of intervention start (left panel) and the resulting net power reduction (right panel), averaged across all interventions. The control curve is computed using the average daily heat pump consumption profile as the counterfactual, aligned to the time of intervention start (eq. (4)). Standard errors reflect the variability of the means in 5-min-to-intervention-start bins, assuming independence among observations (eq. (5)). The means and confidence intervals are smoothed using a local polynomial of degree 0, with the bandwidth of the intervention curve (left panel) set to the optimal value over the entire plotted period.

These quantitative results differ from those in Sect. 4.1.3. Specifically, in the left panel, the minimum power consumption per HP immediately after the intervention starts is about 100 W, higher than the average residual consumption of 53 W observed at the individual HP level. Similarly, in the right panel, the maximum power reduction of around 250 W is lower than the ATE of 292 W (see Sect. 4.1.3), which reflects the effect of interventions on HP power within a household, controlling for household-invariant characteristics. These differences arise because the fleet-level analysis of flexibility events includes data for interventions that stop immediately[17] or shortly after being initiated. Additionally, some HPs take longer to adjust to reduced operation during interventions, particularly when starting from a high-power state (e.g., heating the DHW tank just before an intervention begins). As a result, the fleet's power reduction at the start of the intervention is lower than at the level of flexibility interventions on individual HPs in Sect. 4.1.3.

A further difference between flexibility interventions and events is their duration: flexibility events reduce power consumption for 18 hours on average, significantly longer than interventions. This is because the fleet-level reflects the balance between HPs that remain blocked and those that become unblocked during the event. Blocked HPs reduce their consumption by a constant amount (292 W each on average), while un-

[17] Because the intervention's initial conditions already meet the conditions for one of the termination scenarios.

blocked ones experience individual rebound effects. However, as HPs become unblocked at different times (e.g., due to differences in insulation levels), a staggering occurs at the fleet level. This leads to an average excess consumption from unblocked HPs in the fleet (shown in Appendix A.4) of only 122 W over the first 18 hours, enabling still-blocked HPs to compensate for the unblocked ones over a longer period. The natural staggering in the return to normal operation observed in our setup aligns with the approach studied in (Müller & Jansen, 2019) to mitigate rebound effects in flexible HPs.

The two distinct phases of flexibility interventions at the individual HP level (see Sect. 4.1.5) are also reflected in the fleet-level analysis of flexibility events, as shown in Fig. 6. This figure presents the average reduction in electricity consumption (kWh) per event and per HP in the fleet relative to the time of event start. Flexibility events reduce electricity consumption within the first 18 hours by up to approximately 2 kWh compared to the counterfactual. Following this, post-event rebound effects gradually increase consumption, reducing the net savings, which eventually stabilize at around 1 kWh after 36 hours. This net gain of 1 kWh aligns with the findings for flexibility interventions at the individual HP level.

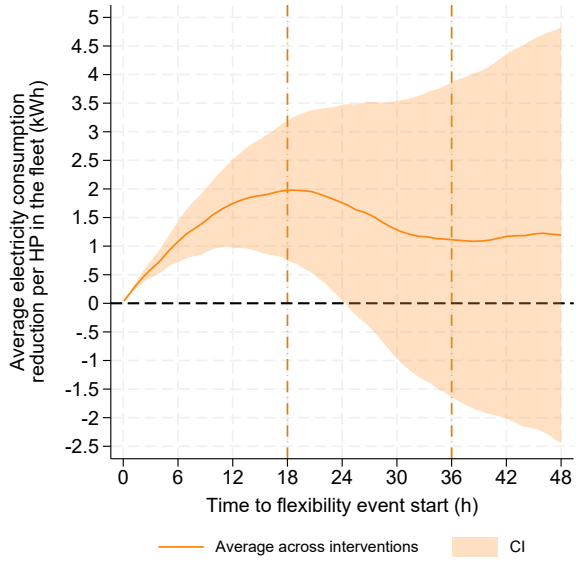


Figure 6: Average electricity consumption reduction (in kWh) per heat pump in the fleet across all interventions. Calculated using the average heat pump daily consumption profile as the counterfactual ((3)). Vertical dotted lines mark the transition from the first phase (reduced consumption) to the rebound period and the start of the stabilizing phase. CIs are derived from the upper and lower bounds (UL and LL) of the power reduction estimates. Mean and CIs are smoothed using a local polynomial of degree 0.

As HPs operate at higher power levels when outdoor temperatures are lower, there is more electricity consumption available to reduce during the first phase of flexibility events. To account for weather variability, we proceed in three steps. First, we assign each intervention a counterfactual power consumption profile, following a similar approach as in eq. (3). These counterfactuals are household-specific and constructed for four categories of daily average temperature: below 3 °C, 3 to 6 °C, 6 to 9 °C, and over 9 °C. For each intervention, we select the counterfactual profile that matches the household, the time of day, and the category of daily average outdoor temperature on the day the intervention is initiated, ensuring this temperature falls within the same category used to construct the counterfactual.

Second, we categorize flexibility events into the same four groups ($<$ 3 °C, 3 - 6 °C, 6 - 9 °C, $>$ 9 °C) now using the average outdoor temperature within the first 18 hours of the event. Each individual intervention is then assigned to one of these event categories by matching the average temperature within 18 hours after the intervention starts to the corresponding temperature bin for the event.

Third, we construct the fleet-level power profile of flexibility events for each temperature category by aligning both the observed power profiles and the household- and temperature-specific counterfactuals relative to the time to flexibility event start, as in eq. (4). Appendix A.5 and Appendix A.6 present the figures

analogous to the left and right panels of Fig. Appendix A.5 shows that, as outdoor temperatures get colder, the initial power reduction in the fleet increases, reaching up to 600 W under 3 °C.

Following this three-step approach, Fig. 7 shows the average[18] electricity consumption reduction per HP in a fleet relative to the time to flexibility event start, across four outdoor temperature ranges. Flexibility events occurring in very cold conditions (average outdoor temperature within 18 hours after the start below 3 °C) achieve the highest reductions in the first phase of the event, reaching 4.5 kWh reduction on average at around 23 hours after the event start. However, the post-event rebound considerably reduces these gains in the second phase, stabilizing at just 1 kWh reduction after 36 hours.

The post-event rebound appears to diminish consumption reductions less as outdoor temperatures increase, aligning with the regression results at the individual HP level discussed in Sect. 4.1.5. Notably, for events between 6 and 9 °C, the rebound effect seems to fully offset the consumption reductions, leaving virtually no net reductions 36 hours after the event start. This may result from outdoor temperatures being high enough for HPs to operate a higher power levels, thereby limiting kWh reductions in the first phase, while they are still low enough for the rebound effect to reduce these small gains during the second phase. Above 9 °C, no rebound effect is observed, suggesting that the rebound's effect at the flexibility event level is not continuous across temperature ranges. At these mild temperatures, the fleet does not require significant additional electricity to restore temperature setpoints, as heat loss to the outside environment is limited.

At 36 hours after the start, all events except those between 6 and 9 °C reduce net consumption by 1 to 1.5 kWh on average, with slightly higher net average reductions for events between 3 and 6 °C compared to those below 3 °C. Finally, while net reductions stabilize at 36 hours after the event start on average across the sample (see Fig. 6), the exact point at which they stabilize for low-temperature events remains unclear.

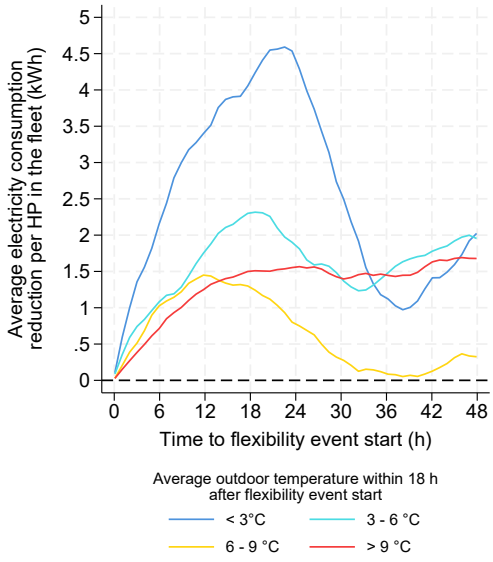


Figure 7: Average electricity consumption reduction (in kWh) per heat pump in the fleet, categorized by four outdoor temperature ranges. Reductions are calculated using the average heat pump daily consumption profile as the counterfactual, based on outdoor temperatures within the first 18 hours after event start: below 3 °C, 3-6 °C, 6-9 °C, and above 9 °C. The means are smoothed using a local polynomial of degree 0.

4.2.2 Monetary valuation of a flexibility event in a fleet of heat pumps

We estimate the monetary impact of flexibility events across a fleet of HPs using day-ahead electricity prices. We calculate the monetary value of simulated events initiated at each hour throughout the 2022-2023 and 2023-2024 heating seasons (a total of 7,622 hours, meaning almost as many events, as last hours of HS2 lack sufficient data for 36-hour events), using actual data on day-ahead electricity prices and outdoor temperatures

[18] Due to the scarcity of very cold days in the sample, the confidence intervals for these estimates are not shown as they are wider and harder to interpret. We restrict our discussion to the averages.

to capture the relationship between outdoor temperature and its effect on HP power reduction at a fleet-level. As day-ahead electricity prices are negatively correlated with outdoor temperature, with a correlation coefficient of -0.46 ($p < 0.01$) throughout the 2022-2023 and 2023-2024 heating seasons, lower temperatures not only influence energy consumption reductions during events but also frequently coincide with higher electricity prices.

The calculation involves two steps. First, using historical outdoor temperature data for Belgium (Royal Meteorological Institute of Belgium, 2024), we compute the average temperature over the 18 hours following each event start to assign it to an appropriate electricity consumption reduction profile from Fig. 7. Second, we estimate the cost savings by matching the dataset of simulated events with actual day-ahead electricity prices for Belgium (European Network of Transmission System Operators for Electricity, 2024). To explicitly account for post-intervention increases in electricity consumption, we calculate the average cost savings at two time points: 18 hours and 36 hours after each event start.

On average, we find that flexibility events result in savings of €0.280 (95% CI[19]: €0.273,0.289) after 18 hours per HP and event, which decrease to €0.125 (95% CI: €0.122, 128) at 36 hours. However, aggregators can achieve higher savings by targeting specific high-price periods. This is reflected in the left panel of Fig. 8, which shows the empirical cumulative distribution of the average savings at these two time points as a function of the percentile of events, i.e., relative to the share of time in HS1 and HS2. In the top 5% of events ($n = 380$), net savings at 36 hours increase significantly, averaging €0.58 per HP and event, and averaging €0.76 in the top 1 % of hours ($n = 76$). The maximum observed savings are €1.10 per HP and event. Remarkably, cited savings above the 95th percentile occurred almost entirely in November and December 2022, at the peak of the electricity crisis.

Around 10% of events ($n = 781$) resulted in negative monetary savings at 36 hours, averaging -€0.03 and reaching as low as -€0.40. The majority of such events occurred when the average temperature over the first 18 hours of the event was between 6 and 9 °C, where consumption reductions are entirely offset by rebound consumption (see Fig. 7). Besides, during these events, the average day-ahead price over the 18-hour reduction period was slightly lower than the average price over the full 36-hour event, making rebound consumption more costly than the reductions. Such price dynamics may happen more frequently in this temperature category as it is the one that shows the weakest correlation between outdoor temperatures and day-ahead prices, estimated at just -0.08 (although significant at the 1% level). Overall, this reinforces the importance of targeting the events in a smart way, so that aggregators avoid scenarios where flexibility events result in extra costs rather than generate savings.

The right panel of Fig. 8 shows that, as average day-ahead prices increase during the first phase, the gap between savings at 18 hours and 36 hours after the intervention widens. This aligns with the findings from Fig. 7, where very cold temperatures (associated to high prices) amplify this gap due to a larger rebound consumption. However, 36-hour savings still increase with higher prices, as expected. The slight deviation at very high prices is due to these bins containing more events occurring below 3 °C, where net savings at 36 hours on average are slightly lower than those achieved by events in the 3 to 6 °C range.

Finally, the calculation of savings of up to €1.10 per flexibility event is based solely on the volatility of day-ahead prices. In systems with capacity markets[20], an additional value stream is the capacity value during peak events. At a capacity value of $700 per kW (Bollinger & Hartmann, 2015, 2019), an average flexible HP in our sample could save up to $175 per year in investment in additional peak generation capacity, provided that post-intervention increases in consumption are appropriately managed. Additionally, aggregators may also exploit other value streams, such as participation in ancillary services markets, including balancing services.

[19]The confidence intervals in this section reflect the variability of mean savings across events but likely underestimate the true variability, as they do not account for variability around the mean counterfactual power level at each time point during the simulated events.

[20]Which is not the case in Belgium.

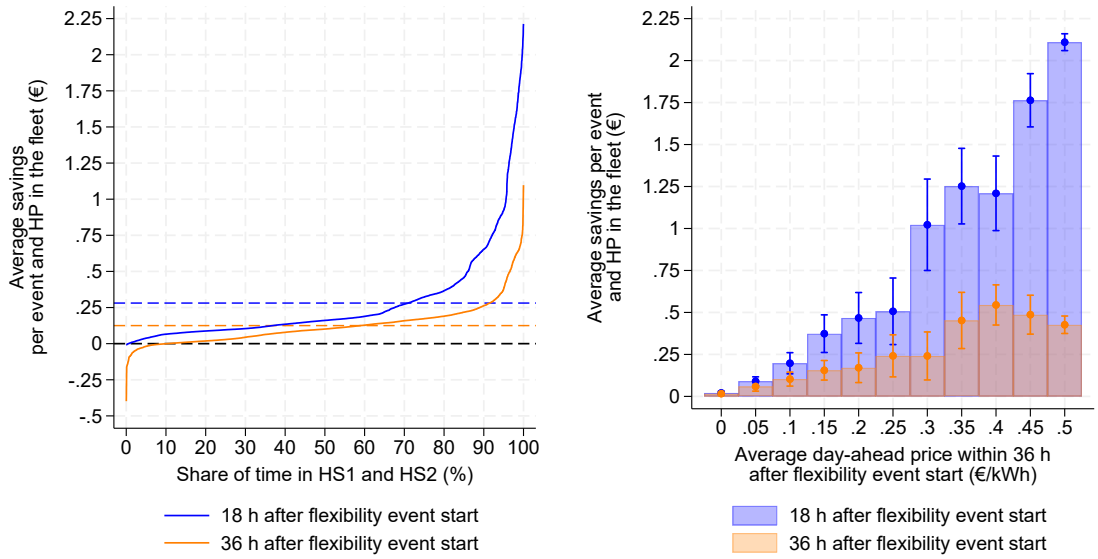


Figure 8: Average savings per event and per HP in the fleet at 18 and 36 hours after the flexibility event start for flexibility events initiated at each hour throughout heating seasons 1 and 2, using the temperature-specific average consumption reduction profiles from Fig. 7. Left: cumulative distribution of savings by share of time across heating seasons 1 and 2, smoothed using a local polynomial of degree 0. Horizontal dotted lines indicate the sample averages. Right: histogram of savings binned by the average day-ahead electricity price within the first 18 hours after the event starts. Confidence intervals at the 95% level, represented by error bars, capture the variability in mean savings within each bin, excluding variability around the heat pump counterfactual power level.

4.3 Comfort impact and household responses

In this section, we analyze households' overriding pattern during ongoing interventions, reflecting their subjective perception of discomfort. A thematic analysis of comments left at the time of overruling provides qualitative insights. We also examine whether subjective discomfort aligns with a quantitative proxy for discomfort. Finally, we present post-experiment survey results to understand participants' responses to flexibility interventions and the features they considered important when scaling such programs.

4.3.1 Thematic analysis of reasons for manual overrides

We analyze the reasons households provided on the online dashboard when overriding flexibility interventions, listed in Appendix E. To classify them, a thematic analysis was conducted with the assistance of a large language model [21] (OpenAI, 2024). The analysis identified eight (possibly overlapping) categories of manual overrules. The three main categories are: "Too low indoor temperature" ($N = 35$; 65%), "Sickness/health concerns" when users request the interruption of the intervention or the cancellation of any upcoming intervention due to individuals being temporarily vulnerable to low temperatures ($N = 11$; 20.5%) –this category contains half of the preemptive overrules–, "Presence at home" with individuals present at home requiring comfort to work or study ($N = 9$; 16.5%). Five further categories are identified, although with fewer occurrences each: "Need for comfort during the weekend" ($N = 3$; 5.5%), "Low water temperature" ($N = 3$; 5.5%), "Returning home" when participants return to a colder home after being away ($N = 2$; 3.5%), "Visitors" when households expect guests ($N = 2$; 3.5%), and "Technical issue" to report errors with the installation ($N = 1$; 2.0%). Households' responses to drops in DHW temperature (in non-decoupled HPs) are further evidenced by the morning overrides of Household 5, likely driven by increased hot water demand at that time of the day. This supports the automatic stop implemented when the DHW temperature falls below 40 °C. The morning period from 7 to 12 a.m. accounts for 41% of all manual overrules (excluding preemptive ones), making it the most predominant time of day for overruling.

[21] The prompt and output are available with the replication codes and data.

Interestingly, ongoing interventions were overruled at relatively mild indoor temperatures, averaging 19.4 °C (95% CI: 19.1, 19.7 °C) over all overrules. This can be compared to the pre-survey results, where households were asked to report their minimum comfortable indoor temperature during interventions[22]. The average reported temperature was 18.5°C across the nine participating households but varied significantly: from 14 °C (Household 9, who overruled at 18.7 °C on average over non-preemptive overrules) to 20 °C (Households 2, 3, 5 and 6, who overruled at 20.1 °C on average over non-preemptive overrules).

While this highlights discrepancies between stated and revealed preferences, the indoor temperature at the moment of overruling does not fully capture household discomfort at that stage of the intervention. The temperature drop, i.e., the difference between the initial indoor temperature at the intervention start and the final temperature must also be considered. Moreover, the placement of thermal sensors —typically in a central room— may not always reflect conditions if households activities occur in colder rooms. The location of the thermostat is indeed a key factor affecting the potential of flexibility events, as noted by (Centre for Net Zero & Nesta, 2023).

4.3.2 Quantitative proxy for household discomfort and manual overruling patterns

Beyond the average temperature reduction during an intervention or the temperature at the time of overrule, a more representative proxy for discomfort is the temperature drop attributable to each intervention. This is calculated as the difference between the initial indoor temperature at the start of the intervention, $T_{in}(t^0)$, and at its end, $T_{in}(t^f)$, for each intervention i on household h:

$$\Delta T_{ih}^{drop} = T_{in,ih}(t^0) - T_{in,ih}(t^f) \tag{10}$$

We find that the average temperature drop across the entire intervention sample equals 0.69 °C (95% CI: 0.59, 0.78 °C). As expected, this value, which reflects the total impact of the intervention, is higher than the average temperature reduction of 0.16 °C observed throughout the duration of interventions (see Sect. 4.1.2). Additionally, in the subsample of interventions that were automatically stopped (by either the T_{in} or T_{DHW} thresholds), the temperature drop shows a moderate correlation with the indoor temperature at the start of the intervention ($r = 0.40$, $p < 0.01$) and with the intervention duration ($r = 0.56$, $p < 0.01$); the negative correlation with the average outdoor temperature during the intervention is close to significance ($r = -0.12$, $p = 0.06$).

To analyze whether household overruling patterns align with the proxy for discomfort in eq. (10), Fig. 9 presents a histogram of the temperature drop for each intervention, distinguishing between those that were automatically stopped and those that were manually overruled (excluding eight preemptive overrules, as they do not result from discomfort induced by an intervention).

Most interventions resulted in a positive temperature drop from the initial value, indicating that households experienced actual discomfort. A few outliers, including one automatic stop with a drop of -4.3 °C, likely occurred when interventions were followed by sunny weather or high home occupancy with activities like cooking, both of which contribute to reheating the home, resulting in a negative ΔT^{drop}. The peak observed at a zero temperature drop reflects the interventions that were too short in duration to cause significant deviations from the initial temperature (see Sect. 4.1.1).

The average temperature drop across automatically stopped interventions is 0.62 °C (95% CI: 0.51, 0.72 °C), but was considerably higher across manually stopped ones, at 1.06 °C on average (95% CI: 0.83, 1.30 °C). The difference between the two means is statistically significant at the 1% level. This indicates that the discomfort households experience at the time of manual overrule is substantially higher than the discomfort built up during automatically stopped interventions.[23]. This aligns with the thematic analysis of comments

[22] This temperature was not intended to be used as a feature in the experiment, and households were informed of this.

[23] This finding is consistent with results from a regression analysis including household FE, and over households that overruled interventions at least once. The coefficient for the manual overrule dummy (excluding preemptive overrules) is 0.24 °C, $p = 0.03$ (wild cluster bootstrap 95% CI: 0.03, 0.53 °C), i.e., the average within-household temperature drop was 0.24 °C higher for manually overruled interventions (controlling for household characteristics invariant across interventions).

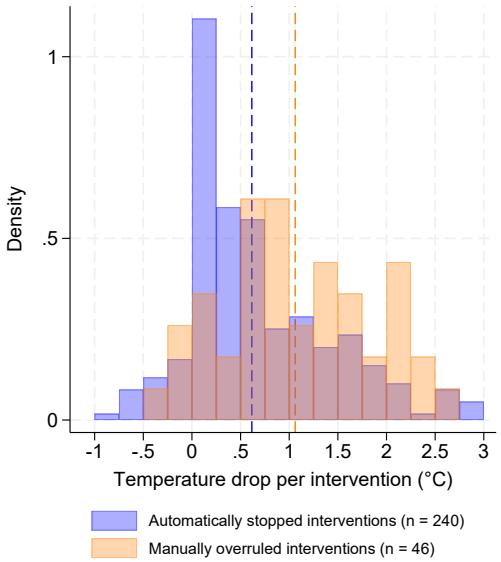


Figure 9: Histogram of the temperature drop induced by interventions, by automatically stopped and manually overruled interventions. The drop is calculated using eq. (10). Preemptive overrules are excluded from the manually overruled subsample. The vertical dotted lines indicate the mean temperature drop for each subsample. The histogram is truncated at -1 °C to exclude one outlier.

left at overrule, which showed that most manual overrules are motivated by thermal discomfort.

Although this might suggest rational behavior from households—overruling interventions when discomfort is greater— it is important to note that the correlation between a dummy variable for manually overruled interventions and the temperature drop is weak, at only 0.20 ($p < 0.01$). Households did not consistently overrule interventions with larger temperature drops, which could be partly explained by their absence from home during some of these interventions.[24]

4.3.3 Post-experiment participants' feedback

In May 2024, a few weeks after the last intervention of HS2, a short survey was sent to participants to gather feedback on their experience. Eight of the nine households that participated in HS1 or HS2 responded.

Households were asked to rate how interventions impacted their indoor temperature comfort on a 1-5 scale (*No discomfort - Extreme discomfort*), with an average score of 2.4 indicating only slight to moderate discomfort. This is consistent with the finding that interventions did not cause a large temperature drop on average across the entire intervention sample. Discomfort related to reduced DHW availability was minimal, with an average response of 1.1, suggesting that the automatic trigger at 40 °C was effective.

Regarding the features of flexibility programs tested in the experiment, households did not assign much importance to being notified in advance about upcoming interventions, giving it an average rating of 2.5 on a 1-5 scale (*Not important at all - Extremely important*), indicating it was considered slightly to moderately important. However, they rated the ability to overrule interventions as very important, giving it an average score of 3.9. Households also valued real-time communication with experiment organizers and coordinators via the dashboard, as shown by messages left on the platform, which were often used to ask questions or report technical issues without the intention to override an intervention. In some cases, households left messages requesting the overruling of potential upcoming interventions within the next 24 hours, even when none were scheduled.

Finally, when asked about strategies to reduce discomfort during interventions, two households reported opting for warmer clothing instead. This aligns with the regression results of intervention duration presented

[24] Moreover, descriptive evidence over a given heating season suggests that households are more likely to overrule interventions they experience first compared to those they experience later. While temperature drops do not significantly differ between first and later interventions, this pattern could suggest potential habituation or learning effects, although the evidence remains suggestive.

in Sect. 4.1.1, which show no evidence that households increased the thermostat before a notified event.

5 Conclusions and policy implications

This field experiment is one of the first practical implementations of a residential electricity flexibility scheme, with a specific focus on flexible heating. The experimental design is suited for studying the interaction between the technical capabilities of heat pumps and the comfort boundaries of the households involved, which together determine the real-world flexibility potential. Our analysis builds on two complementary perspectives: individual heat pump flexibility and aggregated fleet-level flexibility.

First, our experiment was designed so that interventions could either stop automatically or be manually overruled by the user. This mixed approach provides participants with some degree of control, which is likely to become a standard feature of future residential flexibility schemes. We observed that on average, individual interventions lasted approximately 13 hours, with the primary constraint being the demand for domestic hot water. Further, our findings indicate that intervention duration is significantly influenced by the initial heat pump and outdoor temperature conditions, but is not affected by the specific time of day the intervention is initiated. Although interventions reduce electricity consumption, they also lead to rebound effects of up to 700 W as the intervention ends and the heat pump resumes operation. On average, interventions save 1 kWh of electricity on a net basis at 16 hours after the intervention stops.

Second, we present the fleet-level dynamics, focusing on the average contribution of each heat pump during "flexibility events", i.e., coordinated actions across a fleet of assets. These results are particularly valuable, as they offer more realistic insights into the flexibility potential of heat pumps, considering that some units return to normal operation earlier than others or have their interventions manually overruled, which reduces the maximum achievable power reduction. This perspective reveals that the power consumption of a fleet of heat pumps can be reduced by up to 250 W per heat pump, gradually decreasing to 0 after 18 hours on average, before the fleet consumes more electricity than usual to restore temperature setpoints. This is considerably longer than individual interventions, due to the natural staggering of heat pumps returning to normal operation. This staggering smooths rebound effects, which peak at just 130 W and average 50 W over the 18-hour rebound period.

While it could be argued that large impacts on household comfort would result in a low acceptability of heat pump flexibility, creating a barrier to its large-scale adoption in the future, our findings suggest otherwise. In fact, heat pumps can be turned off for several hours without a noticeable impact on indoor temperatures, as demonstrated by the vast majority of the hundreds of interventions performed in our experiment. On average, the temperature decreased by just 0.69 °C by the end of an intervention, although larger drops were observed in cases of manual overrides. These overrides, along with feedback gathered after the experiment, indicate that households value and actively use their ability to intervene and restore comfort levels. They suggest that such a feature is an important prerequisite for encouraging greater participation of households in heat pump flexibility schemes. It reflects that households are not homogeneous actors who predictably allow or reject flexibility but that their decisions are instead driven by multiple factors, including sick children and dinner parties, which all influence the time-dependent and context-specific acceptability of flexibility.

The key driver expected to motivate heat pump owners to operate them flexibly is the potential for financial savings. These savings depend on the different stakeholders and their abilities to access value streams like price volatility or ancillary services; our study focuses on savings from day-ahead price volatility only. We show that colder outdoor temperatures, frequently coupled with higher day-ahead prices, provide more power reduction potential as heat pumps operate at higher levels, but also result in greater heat loss during rebound periods. Accounting for this trade-off, net savings amount to €0.13 on average per heat pump and flexibility event, occasionally reaching much higher savings up to €1.10 during periods of extreme price volatility. Importantly, because these interventions have minimal impact on household comfort, it becomes

feasible to implement them daily throughout the heating season. Automating such interventions allows many small individual savings to add up to substantial annual reductions in heat pump running costs, encouraging more households to participate in flexibility. This is particularly true as automation minimizes the cognitive burden on users by reducing the effort required, as evidenced by the many interventions we conducted with only a few manual overrules.

Our findings lead to several key policy recommendations. First, policymakers should support and invest in commercial and academic efforts to advance the understanding, development, and large-scale adoption of heat pump flexibility. This flexibility can address different challenges in a decarbonized electricity system, such as managing national peaks in electricity consumption, and addressing period of low wind and/or solar production, both of which typically only last a few hours. However, our research shows that heat pumps are not a solution to extended periods of low renewable generation, such as a two-week 'Dunkelflaute'. Interestingly, households in our study did not strategically adapt to day-ahead notifications of flexibility interventions and reported minimal importance for receiving such notifications in the post-experiment survey. This suggests the potential for unannounced, shorter-term flexibility to respond to rapid changes in electricity system conditions. Overall, heat pump flexibility is a relatively inexpensive way to increase the energy system efficiency, as it builds on assets that households are already adopting, without requiring costly new infrastructures for electricity production and storage. Policymakers can further facilitate this gain in efficiency by adjustments in regulations to improve interoperability, implement standardized communication protocols, and ensure that heat pumps can separate space from domestic hot water heating out of the box.

Second, practitioners, such as energy suppliers, aggregators, heat pump manufacturers, and system and grid operators benefit from progress in research as well. Heat pump flexibility is shaped not only by technical capabilities of assets, but also on in-situ factors like building characteristics, which affect event duration, temperature drop rates, and heat loss driving rebound consumption. User behavior, such as setpoint values choice, influences how much of the reduction in consumption is sustained after the rebound period. Additionally, indoor and outdoor temperatures also play a key role in determining flexibility potential and event duration. Practitioners must consider all these factors to accurately assess the net impact of fleet-wide heat pump flexibility events, particularly in accounting for rebound effects, as explored in our study. Fleet-level results show that a staggered return to normal operation after flexibility interventions can significantly enhance the benefits of heat pump flexibility. While this staggering occurred naturally in our setup, practitioners may need to develop dedicated algorithms to achieve this at scale. Furthermore, developments in software are also necessary to enable the separation of control between space and domestic hot water heating.

There is ample room for further work on this topic. For instance, one straightforward way to address the question of acceptability of flexible heating involves studying interventions that pre-heat homes (e.g., by 1-2°C above user setpoint) before interrupting heating during high-price periods. Such interventions could not be tested within the constraints of heat pumps in our sample but hold potential for better understanding comfort thresholds. Similarly, future research could focus on interventions targeting heat pumps with decoupled space and water heating, once these become more common. Besides, as heat pump adoption scales up, it will be essential to study how our findings generalize across different building and household characteristics. In particular, less-well-insulated homes may represent a significant portion of future flexibility potential if high-temperature heat pumps gain popularity as replacements for fossil-fueled systems. In such homes, heat pump flexibility is likely to remain viable without substantially impacting comfort, though the dynamics will differ, with higher power reductions but presumably shorter intervention durations. Exploring these in future research could provide additional insight to policymakers and practitioners on the design of flexibility schemes that make the electricity demand of heat pump users more price-responsive, with minimal impact on comfort.

Data availability

All data and code necessary to replicate the results are available on GitHub, along with the prompts and outputs used for the AI-assisted thematic analysis of comments left by participants during manual overrule: https://github.com/RigauxBaptiste/Proof_of_the_Pudding_Heating.git.

Author contributions

B.R.: Conceptualization, data collection, software, formal analysis, writing – original draft, review and editing, visualization, data curation.
S.H.: Funding acquisition, conceptualization, data collection, coordination, supervision, writing – original draft, review and editing.
M.O.: Funding acquisition, conceptualization, coordination, supervision, writing – original draft, review and editing.

Declaration of generative AI and AI-assisted technologies

During the preparation of this work, the authors used OpenAI's ChatGPT to assist with: the thematic analysis of comments left at overrule by participating households overall, the optimization of the analysis code, and the improvement of the readability and language of the manuscript. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

Aghniaey, S., & Lawrence, T. M. (2018). The impact of increased cooling setpoint temperature during demand response events on occupant thermal comfort in commercial buildings: A review. *Energy and Buildings*, *173*, 19–27. https://doi.org/10.1016/j.enbuild.2018.04.068

Allcott, H. (2011). Rethinking real-time electricity pricing. *Resource and Energy Economics*, *33*(4), 820–842. https://doi.org/10.1016/j.reseneeco.2011.06.003

Bailey, M., Brown, D. P., Wolak, F. A., & Shaffer, B. (2024). Centralized vs Decentralized Demand Response: Evidence from a Field Experiment.

Bauwens, T., & Devine-Wright, P. (2018). Positive energies? An empirical study of community energy participation and attitudes to renewable energy. *Energy Policy*, *118*, 612–625. https://doi.org/10.1016/j.enpol.2018.03.062

Bernard, L., Hackett, A., Metcalfe, R., & Schein, A. (2024). *Decarbonizing Heat: The Impact of Heat Pumps and a Time-of-Use Heat Pump Tariff on Energy Demand* (tech. rep. w33036). National Bureau of Economic Research. Cambridge, MA. https://doi.org/10.3386/w33036

Blonz, J. A., Palmer, K. L., Wichman, C. J., & Wietelman, D. C. (2021). Smart Thermostats, Automation, and Time-Varying Prices. https://media.rff.org/documents/WP_21-20_April_2023_Update_May_1_VkrzXOU.pdf

Bollinger, B. K., & Hartmann, W. R. (2015). Welfare Effects of Home Automation Technology with Dynamic Pricing [Working Paper No. 3274]. https://www.gsb.stanford.edu/faculty-research/working-papers

Bollinger, B. K., & Hartmann, W. R. (2019). Information vs. Automation and Implications for Dynamic Pricing. *Management Science*, *66*(1), 290–314. https://doi.org/10.1287/mnsc.2018.3225

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-Based Improvements for Inference with Clustered Errors. *The Review of Economics and Statistics*, *90*(3), 414–427. https://doi.org/10.1162/rest.90.3.414

Centre for Net Zero & Nesta. (2023). Automating Heat Pump Flexibility: Results From a Pilot: Heatflex uk - a collaboration between nesta and centre for net zero. https://www.centrefornetzero.org/papers/automating-heat-pump-flexibility-results-from-a-pilot

Centre for Net Zero & Nesta. (2024). HeatFlex: the untapped potential of heat pump flexibility: Heatflex uk - a collaboration between nesta and centre for net zero. https://www.centrefornetzero.org/papers/heatflex-the-untapped-potential-of-automated-heat-pump-flexibility

Council of the European Union. (2018). Directive (EU) 2018/2001 of the European Parliament and of the Council: of 11 December 2018 on the promotion of the use of energy from renewable sources. http://data.europa.eu/eli/dir/2018/2001/oj

Da Fonseca, A. L., Chvatal, K. M., & Fernandes, R. A. (2021). Thermal comfort maintenance in demand response programs: A critical review. *Renewable and Sustainable Energy Reviews*, *141*, 110847. https://doi.org/10.1016/j.rser.2021.110847

Darby, S. J., & McKenna, E. (2012). Social implications of residential demand response in cool temperate climates. *Energy Policy*, *49*, 759–769. https://doi.org/10.1016/J.ENPOL.2012.07.026

Dewangan, C. L., Singh, S., Chakrabarti, S., & Singh, K. (2022). Peak-to-average ratio incentive scheme to tackle the peak-rebound challenge in TOU pricing. *Electric Power Systems Research*, *210*, 108048. https://doi.org/10.1016/j.epsr.2022.108048

Enrich, J., Li, R., Mizrahi, A., & Reguant, M. (2024). Measuring the impact of time-of-use pricing on electricity consumption: Evidence from Spain. *Journal of Environmental Economics and Management*, *123*, 102901. https://doi.org/10.1016/j.jeem.2023.102901

European Environment Agency. (2023). *Flexibility solutions to support a decarbonised and secure EU electricity system* (tech. rep.). European Environment Agency. Publications Office of the European Union. https://doi.org/10.2800/104041

European Network of Transmission System Operators for Electricity. (2024). Day-Ahead Prices Transparency [Accessed: 29 July 2024]. https://transparency.entsoe.eu/transmission-domain/r2/dayAheadPrices/show

Eurostat. (2022). Final energy consumption in households by type of end-use - quantities [Data extract for year 2022. Accessed: 16 October 2024]. https://ec.europa.eu/eurostat/databrowser/view/nrg_d_hhq/default/table?lang=en

Eurostat. (2024a). Average size of dwelling by household composition and degree of urbanisation [Data coverage: 2023 to 2023. Accessed: 21 August 2024]. https://ec.europa.eu/eurostat/databrowser/view/ilc_lvho31__custom_12632371/default/table?lang=en

Eurostat. (2024b). Distribution of population aged 18 and over by part-time or full-time employment, income group and sex - EU-SILC survey [Data coverage: 2003 to 2023. Accessed: 21 August 2024]. https://ec.europa.eu/eurostat/databrowser/view/ilc_lvhl04__custom_12638500/default/table?lang=en

Eurostat. (2024c). Distribution of population by degree of urbanisation, dwelling type and income group - EU-SILC survey [Data coverage: 2003 to 2023. Accessed: 21 August 2024]. https://ec.europa.eu/eurostat/databrowser/view/ilc_lvho01__custom_12637832/default/table?lang=en

Fabra, N., Rapson, D., Reguant, M., & Wang, J. (2021). Estimating the Elasticity to Real-Time Pricing: Evidence from the Spanish Electricity Market. *AEA Papers and Proceedings*, *111*, 425–429. https://doi.org/10.1257/pandp.20211007

Faruqui, A., & Sergici, S. (2010). Household response to dynamic pricing of electricity: A survey of 15 experiments. *Journal of Regulatory Economics*, *38*(2), 193–225. https://doi.org/10.1007/s11149-010-9127-y

Fischer, D., & Madani, H. (2017). On heat pumps in smart grids: A review. *Renewable and Sustainable Energy Reviews*, *70*, 342–357. https://doi.org/10.1016/j.rser.2016.11.182

Frederiks, E. R., Stenner, K., & Hobman, E. V. (2015). Household energy use: Applying behavioural economics to understand consumer decision-making and behaviour. *Renewable and Sustainable Energy Reviews*, *41*, 1385–1394. https://doi.org/10.1016/j.rser.2014.09.026

Fu, Z., Novan, K., & Smith, A. (2024). Do time-of-use prices deliver energy savings at the right time? *Journal of Environmental Economics and Management*, *128*, 103054. https://doi.org/10.1016/j.jeem.2024.103054

Georges, E., Cornélusse, B., Ernst, D., Lemort, V., & Mathieu, S. (2017). Residential heat pump as flexible load for direct control service with parametrized duration and rebound effect. *Applied Energy*, *187*, 140–153. https://doi.org/10.1016/j.apenergy.2016.11.012

Good, N. (2019). Using behavioural economic theory in modelling of demand response. *Applied Energy*, *239*, 107–116. https://doi.org/10.1016/j.apenergy.2019.01.158

Good, N., Ellis, K. A., & Mancarella, P. (2017). Review and classification of barriers and enablers of demand response in the smart grid. *Renewable and Sustainable Energy Reviews*, *72*, 57–72. https://doi.org/10.1016/j.rser.2017.01.043

Gyamfi, S., Krumdieck, S., & Urmee, T. (2013). Residential peak electricity demand response—Highlights of some behavioural issues. *Renewable and Sustainable Energy Reviews*, *25*, 71–77. https://doi.org/10.1016/j.rser.2013.04.006

Harding, M., & Lamarche, C. (2016). Empowering consumers through data and smart technology: Experimental evidence on the consequences of time-of-use electricity pricing policies. *Journal of Policy Analysis and Management*, *35*(4), 906–931. https://EconPapers.repec.org/RePEc:wly:jpamgt:v:35:y:2016:i:4:p:906-931

Harding, M., & Sexton, S. (2017). Household response to time-varying electricity prices. *Annual Review of Resource Economics*, *9*(1), 337–359. https://doi.org/10.1146/annurev-resource-100516-053437

Harold, J., Bertsch, V., & Fell, H. (2021). Preferences for curtailable electricity contracts: Can curtailment benefit consumers and the electricity system? *Energy Economics*, *102*, 105454. https://doi.org/https://doi.org/10.1016/j.eneco.2021.105454

Herabadi, A. G., Kadarusman, Y. B., & Yachinta, C. (2021). Effect of environmental optimism on responsible electricity consumption with price concern as a moderator. *Psychological Research on Urban Society*, *4*. https://doi.org/DOI:10.7454/proust.v4i2.128

Herter, K., McAuliffe, P., & Rosenfeld, A. (2007). An exploratory analysis of California residential customer response to critical peak pricing of electricity. *Energy*, *32*(1), 25–34. https://doi.org/10.1016/j.energy.2006.01.014

Herter, K., & Wayland, S. (2010). Residential response to critical-peak pricing of electricity: California evidence [Demand Response Resources: the US and International Experience]. *Energy*, *35*(4), 1561–1567. https://doi.org/https://doi.org/10.1016/j.energy.2009.07.022

Hobman, E. V., Frederiks, E. R., Stenner, K., & Meikle, S. (2016). Uptake and usage of cost-reflective electricity pricing: Insights from psychology and behavioural economics. *Renewable and Sustainable Energy Reviews*, *57*, 455–467. https://doi.org/10.1016/j.rser.2015.12.144

Jensen, R. H., Kjeldskov, J., & Skov, M. B. (2018). Assisted Shifting of Electricity Use: A Long-Term Study of Managing Residential Heating. *ACM Transactions on Computer-Human Interaction*, *25*(5), 25:1–25:33. https://doi.org/10.1145/3210310

Jessoe, K., & Rapson, D. (2014). Knowledge is (Less) Power: Experimental Evidence from Residential Energy Use. *American Economic Review*, *104*(4), 1417–1438. https://doi.org/10.1257/aer.104.4.1417

Kane, N., Khanna, S., Martin, R., Muûls, M., Sinha, P., & Saha, S. K. (2024). *Leveraging automation and incentives to enhance power demand flexibility* (tech. rep.). Imperial College London. https://www.imperial.ac.uk/media/imperial-college/research-centres-and-groups/hitachi-decarbonisation/Tata_Powbal_Imperial-Report.pdf

Kaspar, K., Nweye, K., Buscemi, G., Capozzoli, A., Nagy, Z., Pinto, G., Eicker, U., & Ouf, M. M. (2024). Effects of occupant thermostat preferences and override behavior on residential demand response in CityLearn. *Energy and Buildings*, 114830. https://doi.org/10.1016/j.enbuild.2024.114830

Kim, J.-H., & Shcherbakova, A. (2011). Common failures of demand response. *Energy*, *36*(2), 873–880. https://doi.org/10.1016/j.energy.2010.12.027

Kostková, K., Omelina, Ľ., Kyčina, P., & Jamrich, P. (2013). An introduction to load management. *Electric Power Systems Research*, *95*, 184–191. https://doi.org/10.1016/j.epsr.2012.09.006

Li, R., Dane, G., Finck, C., & Zeiler, W. (2017). Are building users prepared for energy flexible buildings?—A large-scale survey in the Netherlands. *Applied Energy*, *203*, 623–634. https://doi.org/https://doi.org/10.1016/j.apenergy.2017.06.067

Ludwig, P., & Winzer, C. (2022). Tariff Menus to Avoid Rebound Peaks: Results from a Discrete Choice Experiment with Swiss Customers. *Energies*, *15*(17), 6354. https://doi.org/10.3390/en15176354

MacKinnon, J. G., Nielsen, M. Ø., & Webb, M. D. (2023). Cluster-robust inference: A guide to empirical practice. *Journal of Econometrics*, *232*(2), 272–299. https://doi.org/10.1016/j.jeconom.2022.04.001

Müller, F., & Jansen, B. (2019). Large-scale demonstration of precise demand response provided by residential heat pumps. *Applied Energy*, *239*, 836–845. https://doi.org/10.1016/j.apenergy.2019.01.202

Muratori, M., Schuelke-Leech, B.-A., & Rizzoni, G. (2014). Role of residential demand response in modern electricity markets. *Renewable and Sustainable Energy Reviews*, *33*, 546–553. https://doi.org/10.1016/j.rser.2014.02.027

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications*, *9*(1), 141–142. https://doi.org/10.1137/1109020

Nouicer, A., Meeus, L., & Delarue, E. (2020). *The economics of explicit demand-side flexibility in distribution grids: The case of mandatory curtailment for a fixed level of compensation* (RSCAS Working Paper No. 2020/45). European University Institute, Florence School of Regulation. https://cadmus.eui.eu/bitstream/handle/1814/67762/RSCAS%202020_45.pdf?sequence=1&isAllowed=y

OECD. (2024). Household Disposable Income [OECD Data Archive. Indicator: Household Disposable Income (2022). Accessed: 21 August 2024]. https://data-explorer.oecd.org/vis?lc=en&df[ds]=DisseminateArchiveDMZ&df[id]=DF_DP_LIVE&df[ag]=OECD&df[vs]=&av=true&pd=2022%2C2022&dq=BEL%2BOECD%2BOAVG.HHDI...A&to[TIME_PERIOD]=false&vw=tb

OpenAI. (2024). ChatGPT (version o1 preview) [AI language model] [Accessed: 16 October 2024. The prompt and output are available with the replication codes and data.]. https://chat.openai.com/

Richter, L.-L., & Pollitt, M. G. (2018). Which smart electricity service contracts will consumers accept? the demand for compensation in a platform market. *Energy Economics*, *72*, 436–450. https://doi.org/https://doi.org/10.1016/j.eneco.2018.04.004

Rosenow, J., Gibb, D., Nowak, T., & Lowes, R. (2022). Heating up the global heat pump market. *Nature Energy*, *7*(10), 901–904. https://doi.org/10.1038/s41560-022-01104-8

Royal Meteorological Institute of Belgium. (2024). Open Data - Royal Meteorological Institute of Belgium [Accessed: 29 July 2024]. https://opendata.meteo.be/

Ruokamo, E., Kopsakangas-Savolainen, M., Meriläinen, T., & Svento, R. (2019). Towards flexible energy demand – preferences for dynamic contracts, services and emissions reductions. *Energy Economics*, *84*, 104522. https://doi.org/https://doi.org/10.1016/j.eneco.2019.104522

Statbel. (2024a). Level of Education [Published: 28 March 2024. Accessed: 21 August 2024]. https://statbel.fgov.be/en/themes/work-training/training-and-education/level-education#news

Statbel. (2024b). Structure of the population - Households [Published: 5 June 2024. Accessed: 21 August 2024]. https://statbel.fgov.be/en/themes/population/structure-population/households

Statbel. (2024c). T04_21_BE_POC_CL - Nombre de logements classiques selon l'époque de construction, Nombre de logements classiques au 01-01-2021 - CENSUS - 2021 [Last updated: 19 April 2024. Accessed 21 August 2024 via: https://statbel.fgov.be/fr/themes/census/logement/epoque-de-construction.]. https://statbel.fgov.be/sites/default/files/files/documents/Census2021/T04_POC_BE_FR.XLSX

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, *26*(4), 359–372. Retrieved November 24, 2024, from http://www.jstor.org/stable/25049340

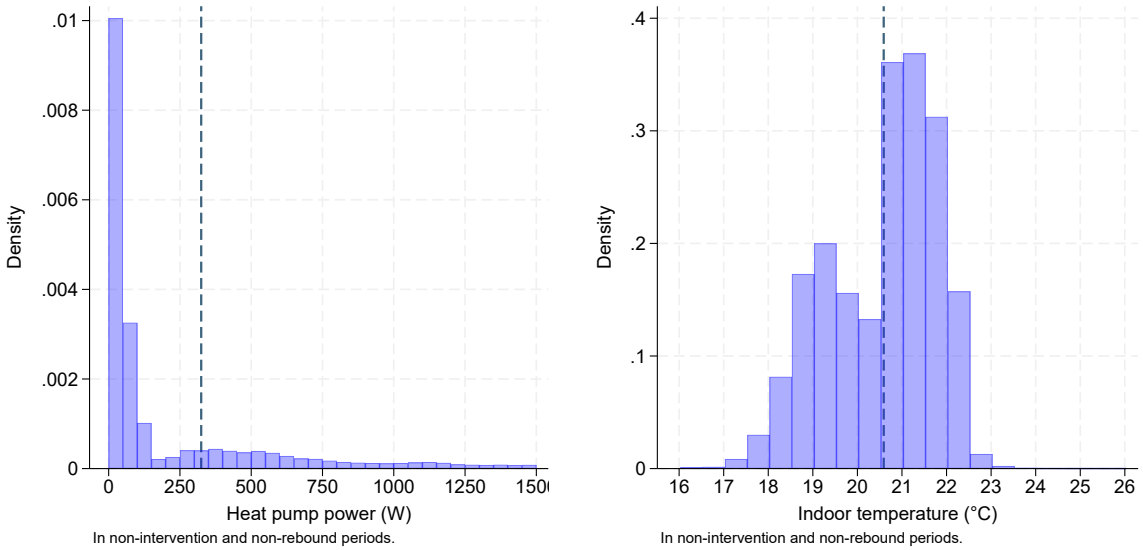
Yilmaz, S., Chanez, C., Cuony, P., & Patel, M. K. (2022). Analysing utility-based direct load control programmes for heat pumps and electric vehicles considering customer segmentation. *Energy Policy*, *164*, 112900. https://doi.org/10.1016/j.enpol.2022.112900

Yilmaz, S., Cuony, P., & Chanez, C. (2021). Prioritize your heat pump or electric vehicle? Analysing design preferences for Direct Load Control programmes in Swiss households. *Energy Research & Social Science*, *82*, 102319. https://doi.org/10.1016/j.erss.2021.102319

Zhang, F., De Dear, R., & Candido, C. (2016). Thermal comfort during temperature cycles induced by direct load control strategies of peak electricity demand management. *Building and Environment*, *103*, 9–20. https://doi.org/10.1016/j.buildenv.2016.03.020

A Additional figures and tables

A.1 Histograms of indoor temperature and heat pump power in non-intervention periods



(a) (b)

Figure 10: Heat pump variables during non-intervention periods. Left: power (in watts), truncated at 1500 W. Right: indoor temperature (in °C), between 16 and 26 °C to remove outliers. The vertical dotted lines show the sample averages.

A.2 Average heat pump daily profile

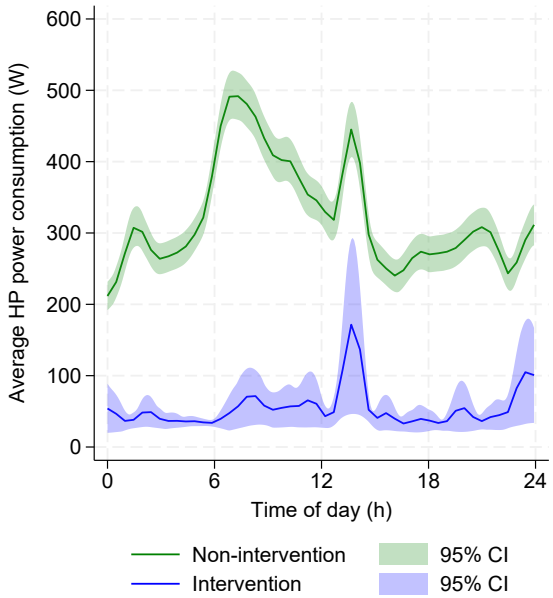


Figure 11: Daily profile of average heat pump power during and outside intervention periods, averaged across all heat pumps and heating seasons. The profiles are smoothed using local polynomial regression of degree 0 for the mean and confidence intervals. Standard errors reflect the variability of the mean in 5-min-of-day bins, assuming independence among observations.

A.3 Comparison of average heat pump daily profiles across different outdoor temperature ranges

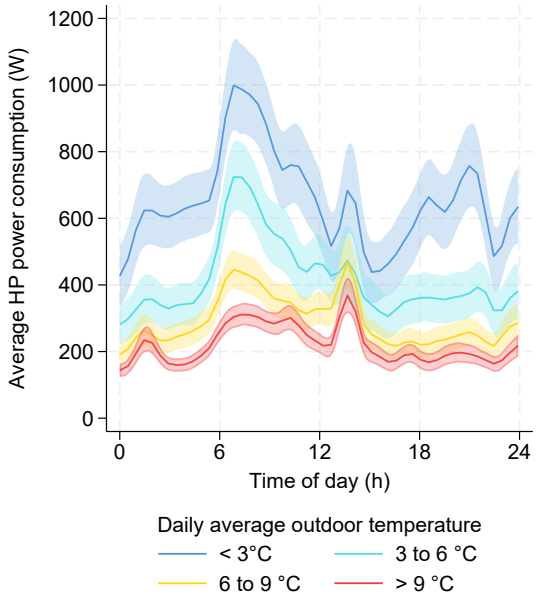


Figure 12: Daily profile of average heat pump power during non-intervention periods, averaged across all heat pumps and heating seasons, and categorized by different daily average outdoor temperatures. The profiles are smoothed using local polynomial regression of degree 0 for the mean and confidence intervals. Standard errors reflect the variability of the mean in 5-min-of-day bins, assuming independence among observations.

A.4 Rebound consumption of heat pumps resuming normal operation at the fleet level

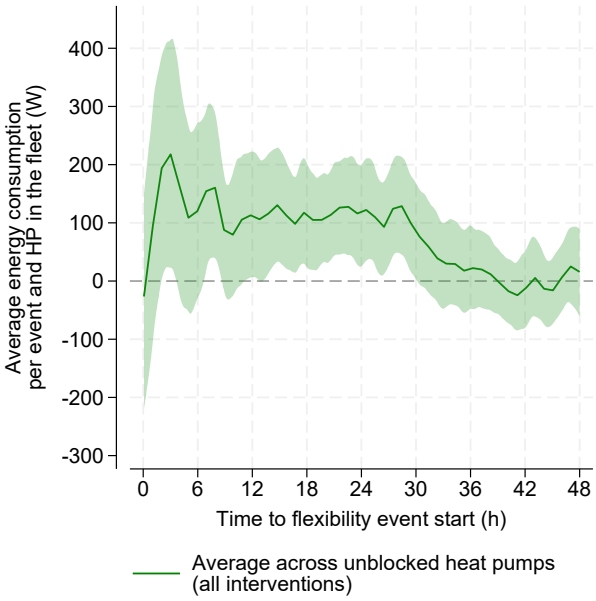
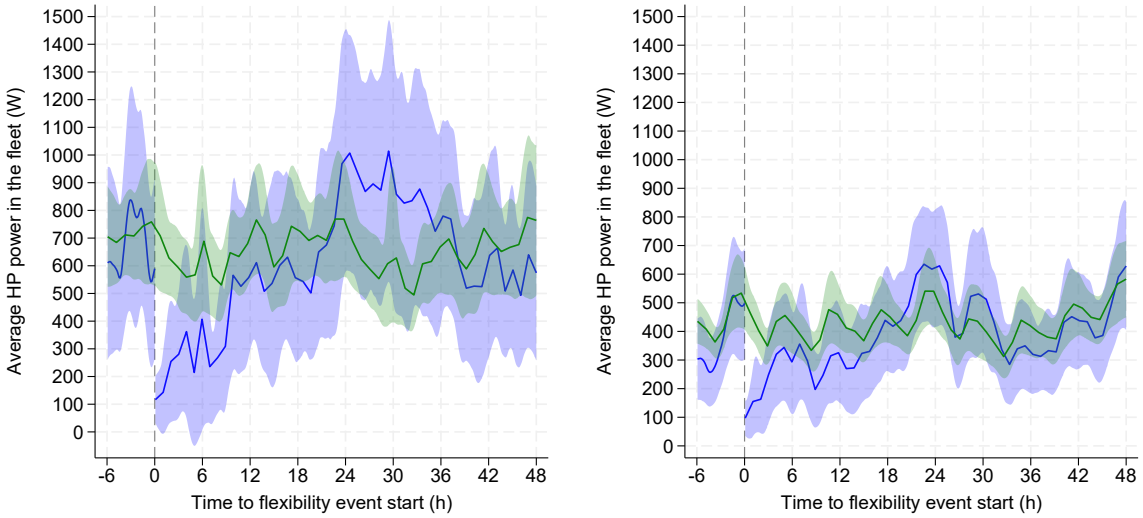


Figure 13: Average heat pump power consumption per unit in the fleet relative to the time of intervention start for unblocked heat pumps that have completed their intervention and undergo rebound consumption. Standard errors reflect the variability of the means in 5-minute-to-intervention-start bins, assuming independence among observations (eq. (5)). The means and confidence intervals are smoothed using a local polynomial of degree 0.

A.5 Fleet-level power consumption profiles during flexibility events: heterogeneity across average outdoor temperature



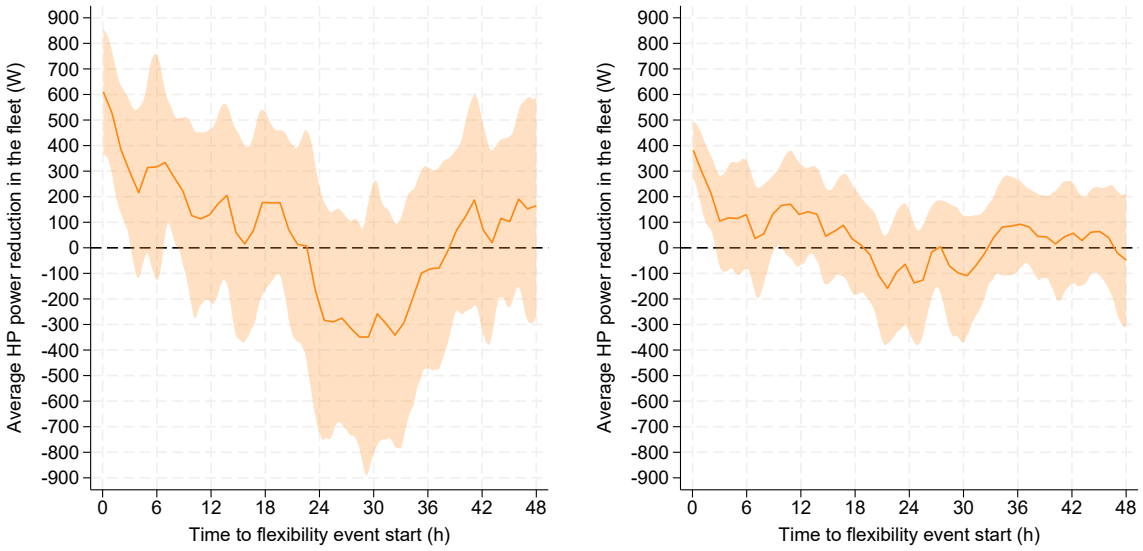
(a) Below 3 °C

(b) Between 3 and 6°C

(c) Between 6 and 9 °C

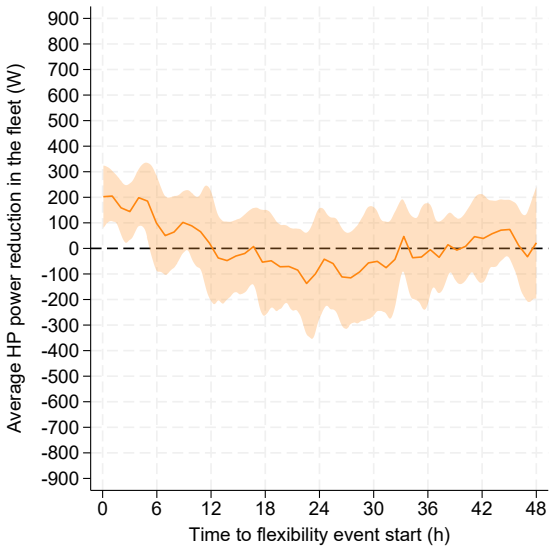
(d) Over 9 °C

Figure 14: Average heat pump power consumption per unit in the fleet relative to the time of flexibility event start, categorized by four outdoor temperature ranges. The temperature categories are based on the average outdoor temperature within the first 18 hours after the event starts: below 3 °C, 3 to 6 °C, 6 to 9 °C, and above 9 °C. The control curve is computed using the average daily heat pump consumption profile as the counterfactual, by outdoor temperature categories, and aligned to the time of event start (eq. (4)). Standard errors reflect the variability of the means in 5-minute-to-intervention-start bins, assuming independence among observations (eq. (5)). The means and confidence intervals are smoothed using a local polynomial of degree 0, with the optimal bandwidth of the intervention curve calculated over the entire plotted period.

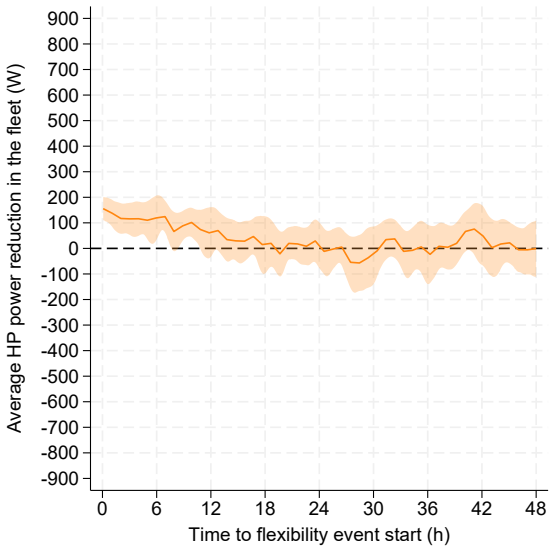


(a) Below 3 °C

(b) Between 3 and 6 °C



(c) Between 6 and 9 °C



(d) Over 9 °C

Figure 15: Net average heat pump power reduction in the fleet relative to the time of flexibility event start, categorized by four outdoor temperature ranges. The temperature categories are based on the average outdoor temperature within the first 18 hours after the intervention starts: below 3 °C, 3 to 6 °C, 6 to 9 °C, and above 9 °C. The net reduction is calculated as the difference between the observed power consumption and the counterfactual, where the counterfactual is based on the average daily heat pump consumption profile for each temperature category. Standard errors reflect the variability of the means in 5-minute-to-intervention-start bins, assuming independence among observations (eq. (5)). The means and confidence intervals are smoothed using a local polynomial of degree.

A.6 Average monthly temperatures during the experimental period

Table 4: Average monthly temperatures (in °C)

	Oct	Nov	Dec	Jan	Feb	Mar	Apr	Average
Heating season 1 [a]	N/A	7.9	4.8	6.6	7.3	8.7	10.6	7.5
Heating season 2 [a]	11.8	8.4	7.5	4.7	9.4	10.3	N/A	8.3
Historic average [b]	11.3	7.2	4.3	3.7	4.2	7.1	10.4	5.4

[a] Source: own data.
[b] Average monthly temperature for the period 1991-2020. Source: Royal Meteorological Institute, https://www.meteo.be/fr/climat/climat-de-la-belgique/normales-climatiques-a-uccle/temperature/temperature-moyenne, 26 September 2024.

A.7 Sample composition

Table 5: Households participation in first and second heating seasons

ID	Decoupled dummy	HS1 participation	HS2 participation	Total number of interventions
1	0	●	●	52
2	0	●	●	28
3	0	●	●	56
4	1	●	○	19
5	0	●	◐	34
6	0	●	●	55
7	1	○	●	10
8	0	◑	●	12
9	1	●	●	21
Total	3	8	8	287

○ indicates that the household did not actively participate in the heating season (HS). ◑ (◐) indicates that the household actively participated in the first (second) part of the HS. The black area size does not represent the number of interventions.

B Notification e-mail in heating season 2

During heating season 2, half of the 32 scheduled interventions were notified a day-ahead to all participating households. This notification consisted in the following e-mail (translated from Dutch), sent by the research team:

Object: Notification FlexSys test

Content:

Dear FlexSys participant

Thank you once again for participating in the tests we are conducting in the FlexSys project with your heat pump. We would like to remind you that half of the the tests conducted this heating season will be preceded by a notification, one day in advance. With this message, we are sending you such a notification: a FlexSys test will be conducted tomorrow.

The operation of your heat pump will be temporarily blocked, while, of course, the internal safety mechanisms of the device remain unaffected.

The blockage will stop automatically when the temperature of the house or the buffer tank for sanitary hot water reaches a predetermined lower limit, or will stop immediately when you use the override button via the COFY-box platform.

We refer to the e-mail from [*experiment coordinator at the cooperative*] dated 26/10/2023 for more information about the tests.

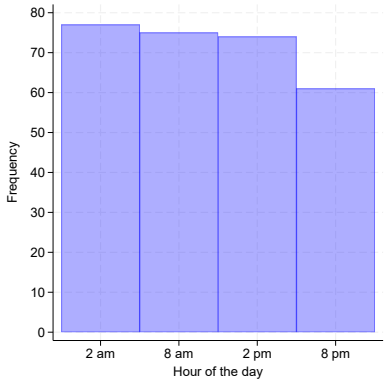
Do not reply to this automatically sent e-mail. If you have any questions or comments, you can e-mail them to:

[*research team's coordinator name and contact details*]

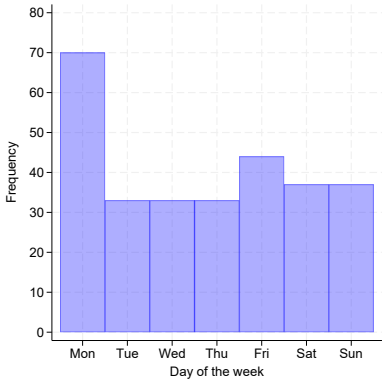
By participating in this test, you are contributing to our research on the potential of smart demand response, for which we thank you sincerely.

The FlexSys team

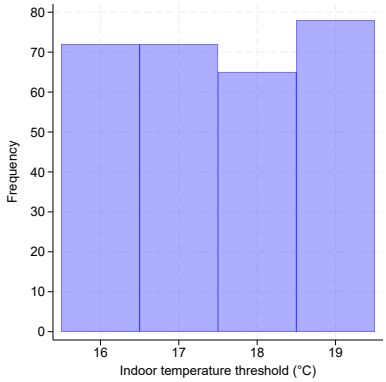
C Distribution of the interventions per start time, indoor temperature threshold value and day of the week



(a) Distribution of intervention's starting hour of the day, TOD



(b) Distribution of intervention's starting day of the week, day



(c) Distribution of intervention's indoor temperature threshold, $T_{in}^{\text{thres.}}$

Figure 16: Distribution of interventions by their characteristics in the intervention schedule

The software SPSS[25] was used to construct an orthogonal design to ensure no correlation between the main effects of the following intervention characteristics: the hour of the day (four levels) and the day of the week (seven levels) at which the intervention is initiated, as well as the indoor temperature threshold used for the automatic trigger back to normal operation (four levels). Additionally, all interventions in HS1 were notified a day in advance, and the schedule in HS2 included a notification status (two levels, with half of the HS2 interventions being notified a day-ahead). While 28 is the least common multiple of the levels of each characteristic, it results in a large imbalance in the combinations, as the 16 combinations of time of day and temperature threshold cannot be evenly distributed among 28 interventions. Instead, SPSS generated a 32-size array, repeating the first value of the day of the week characteristic several times to minimize the imbalance while keeping correlations in the main effects statistically insignificant. As a result, twice as many interventions were scheduled to start on Mondays compared to any other day.

In the dataset of interventions successfully achieved in practice, the Pearson correlation coefficients between the main effects remain statistically insignificant, consistent with the theoretical schedule design. Across the entire intervention sample, the correlation coefficients are: $r(T_{in}^{\text{thres.}}, TOD) = 0.006$, $r(T_{in}^{\text{thres.}}, day) = 0.018$, $r(TOD, day) = 0.051$, $r(T_{in}^{\text{thres.}}, D_{\text{notif}}) = -0.095$, $r(day, D_{\text{notif}}) = 0.023$, $r(TOD, D_{\text{notif}}) = -0.053$. All six correlation coefficients are insignificant at the 5% level, meaning that the 287 studied interventions are random across the starting hour of the day, day of the week, and indoor temperature threshold and notification status.

[25] IBM Corp. Released 2023. IBM SPSS Statistics for Windows, Version 29.0.2.0. Armonk, NY: IBM Corp.

D Likert-scale questions in the pre-experiment survey

D.1 Participants' comprehension of flexibility-related concepts

The awareness of pre-survey respondents towards concepts related to residential electricity flexibility were probed similarly to (Li et al., 2017). Specifically, we used a similar 1 - 4 scale (1: "Never heard of it", 2: "I have heard of it, but I do not understand the concept", 3: "I know a bit about the concept", 4: "I know a lot about the concept") and asked about the following items (randomly ordered and in Dutch in the survey):

- "Energy transition",
- "Smart home",
- "Electricity flexibility",
- "Demand-response programs".

D.2 Participants' attitudes towards the environment

To probe respondents' attitudes towards the environment, the following four relevant items were extracted from a previous related study on Belgian energy cooperative members (Bauwens & Devine-Wright, 2018) and probed via a 1 - 5 scale (1: "Strongly disagree", 2: "Disagree", 3: "Neither agree, nor disagree", 4: "Agree", 5: "Strongly agree") (randomly ordered and in Dutch in the survey) on the following items:

- "I want to feel that I am personally contributing to the protection of the environment.",
- "I am concerned about climate change.",
- "I am the type of person who cares about the environment.",
- "I see myself as an environmentally conscious consumer.".

D.3 Participants' propensity to engage with electricity-saving habits

The frequency at which respondents engage with typical electricity-conserving gestures were probed, following what was done in (Herabadi et al., 2021) from which we extracted the following nine items (probed on a 1 - 5 scale: 1: "Never", 2: "Rarely", 3: "Sometimes"; 4: "Often", 5: "Always" ; randomly ordered and presented in Dutch in the survey):

- "I make sure the lights are off before I leave a room.",
- "I use natural light as a light source.",
- "I use energy-saving lamps (e.g., LED lamps).",
- "I unplug the power plug when not in use.",
- "I turn off PCs/laptops when they are not in use (turned off, not in sleep mode).",
- "I choose electronic devices (not lighting) that use the least energy even if they are a bit more expensive to purchase.",
- "I make sure that the refrigerator door is not open too long.",
- "I set a moderate temperature for my heating system.",
- "I reduce the use of warm water for bathing (e.g. use cold water in warm/hot weather)."

E Manual overrules: comments left at overrule

E.1 Heating season 1

Table 6: Reasons reported at manual overrules (heating season 1). Translated from Dutch.

ID	At overrule:				Reason mentions:	
	TOD	Duration (h)	T_{in} (°C)	Reason reported	Heating	Water
3	15h00	1.0	21.0	Our son (14 months) came home sick from daycare, so today and tomorrow we would prefer that the temperature doesn't drop.	1	0
	08h03	0.0	20.5	Sick son at home, so we would prefer that no test takes place at the moment. (*)	0	0
	14h02	0.0	20.6	We currently have a child recovering at home after a week in the hospital, so we would prefer no temperature drops these days. (*)	1	0
4	15h40	7.6	18.2	Too cold to start the evening.	1	0
	15h20	7.3	18.5	The temperature in the house is too cold, so the heat pump can be turned back on.	1	0
	07h25	11.4	18.7	Colder than 19 degrees is too uncomfortable when we're all at home.	1	0
	07h00	4.9	19.6	Too chilly. The whole family is home, and there is someone studying.	1	0
	14h15	6.2	20.0	We are expecting visitors.	0	0
	18h40	4.6	20.0	Getting too chilly.	1	0
	08h15	12.2	19.4	Too chilly.	1	0
	07h35	17.6	18.1	Too cold.	1	0
	05h30	3.5	18.7	Too cold.	1	0
	06h15	4.2	19.3	Too cold.	1	0
	05h30	3.4	19.3	Too chilly, and someone is home all day today.	1	0
	07h00	16.9	19.3	Heating is too low.	1	0
	06h40	22.7	18.9	Too chilly.	1	0
	02h00	0.0	20.9	Too chilly to come home to. (*)	1	0
	15h40	37.6	19.3	Too cold.	1	0
5	07h40	41.6	19.5	The water temperature could not be boosted.	0	1
	13h40	17.6	19.8	The heating is too low.	1	0
	04h20	14.3	19.5	The water temperature got very low this morning.	0	1
	05h20	3.3	20.2	Low temperature.	1	0
	06h45	16.7	20.3	Temperature is too cold now.	1	0
9	09h55	13.9	18.7	Too cold - weekend.	1	0
	18h50	28.8	17.9	Too cold.	1	0
	08h40	18.6	17.7	Cold morning. Someone works from home in the afternoon.	1	0

Manual overrules performed before the start of an intervention are marked with an asterisk (*).

E.2 Heating season 2

Table 7: Reasons reported at manual overrules (heating season 2). Translated from Dutch.

ID	At overrule:				Reason mentions:	
	TOD	Duration (h)	T_{in} (°C)	Reason reported	Heating	Water
3	16h55	2.9	20.1	Too cold for a sick child.	1	0
	07h35	11.6	19.8	Cold in the house.	1	0
	16h35	8.1	20.1	Too cold in the house.	1	0
	14h10	12.1	21.6	Child at home with stomach flu.	0	0
	02h00	0.0	21.1	Sick child. (*)	1	0
	10h00	20.0	19.1	Day off at home. With freezing temperatures outside and only 19 degrees inside, it's starting to feel too cold.	0	0
	02h00	30.0	20.0	Flu in the house.	0	0
	12h00	4.0	20.8	Newborn baby in the house.	0	0
	08h30	6.5	20.4	Newborn baby in the house, so it's not ideal to let the temperature drop at home right now.	1	0
5	17h10	9.1	20.1	Too low temperature.	1	0
	14h00	0.0	21.5	Water temperature was very low this morning. (*)	0	1
	05h45	9.7	20.1	Too low temperature.	1	0
	08h00	0.0	20.9	Too cold. (*)	1	0
	08h35	18.5	19.8	Too low temperature.	1	0
	06h50	4.8	20.5	Temperature gets too low.	1	0
6	05h15	15.2	19.5	The heat pump is showing an error [*description of the error message*]. It's also cold in the house... I'm not sure if this error is related to the test on October 3rd?	1	0
	07h15	17.2	19.0	We were away for a few days and had lowered the temperature, but now it's cold in the house and we want to warm up again.	0	0
	09h25	13.4	20.6	Someone is sick at home and feeling too cold...	1	0
7	20h40	12.6	18.1	Tomorrow at home with children.	0	0
	08h10	24.2	16.5	Temperature already dropped to 16.4 °C. Now switched back on, so it will be warmer again by tonight. [Because of the] inertia, [it] will otherwise take too long and we won't have decent temperatures again until tomorrow.	1	0
	14h00	0.0	19.7	Daughter is ill. (*)	0	0
9	17h35	39.5	19.0	Living room too cold (18.8 °C but feels lower).	1	0
	09h35	43.5	19.1	Two days at a conference, [*partner*] home alone, unable to adjust the COFY-box, so turn it off in advance.	0	0
	11h05	33.0	18.3	Game night with friends.	0	0
	14h40	48.6	20.0	The heating has been off for 6 days. Living room temperature is fine, but we want to warm up the house a bit.	1	0
	09h05	1.1	18.3	Living room too cold (working from home).	1	0
	23h50	135.8	19.5	Weekend comfort in the living room.	0	0
	14h00	0.0	21.5	Weekend comfort in the living room. (*)	0	0

Manual overrules performed before the start of an intervention are marked with an asterisk (*).

F Regression analysis of the rebound energy consumption in the post-intervention period

In this appendix, we analyze the factors influencing rebound energy consumption in the post-intervention period ($E^{16h}_{\text{rebound},i}$, in kWh), defined as the additional electricity required within 16 hours after an intervention i for the HP of household h to return to user setpoints. We specify the regression model as:

$$\begin{aligned} E^{16h}_{\text{rebound},ih} =& \beta_1 \cdot \overline{T}^{\leq 16h}_{\text{out},ih} + \beta_2 \cdot \Delta T^{f}_{\text{setpoint},ih} + \beta_3 \cdot T^{f}_{\text{in},ih} \\ &+ \beta_4 \cdot \text{TOD}^{f}_{\text{AM},ih} + \beta_5 \cdot \text{TOD}^{f}_{\text{evening},ih} + \beta_6 \cdot \text{TOD}^{f}_{\text{night},ih} \\ &+ \mathbb{1}(\text{FE} = 0) \cdot \beta_0 + \mathbb{1}(\text{FE} = 1) \cdot \alpha_h + \varepsilon_{ih} \end{aligned} \tag{11}$$

Where $E^{16h}_{\text{rebound},ih}$ denotes the additional energy consumption (in kWh) observed within 16 hours after the intervention ends, compared to the counterfactual derived from the HP-specific daily consumption profile during non-intervention periods (eq. (3)). $T^{f}_{\text{in},ih}$ is the indoor temperature at the end of the intervention, while $\Delta T^{f}_{\text{setpoint},ih} = T_{\text{setpoint,ih}} - T^{f}_{\text{in},ih}$ measures its deviation from the thermostat setpoint specified by the household. $\overline{T}^{\leq 16h}_{\text{out},ih}$ represents the average outdoor temperature within the 16-hour rebound period [26]. The three TOD dummies indicate whether the intervention ended in the morning (6 a.m. - 12 p.m.), evening (6 p.m. - 12 a.m.) or night (12 a.m. - 6 a.m.), with the afternoon as the baseline category. Finally, α_h represents household FE, capturing within-household variation in the parameters by controlling for household characteristics that remain invariant across interventions.[27]

We estimate the model specified in eq. (11) via a linear regression and account for the small number of clusters (nine heat pumps) by using wild cluster bootstrap (100,000 repetitions) at the HP-level to compute cluster-robust standard errors. We report the p-values derived from the empirical distribution of the bootstrapped estimates. Table 8 presents the results for four models, from a parsimonious specification to the full model with household FE in eq. (11). All estimates show the expected signs and remain robust across the four specifications, with the exception of the intercept in Model (1). Rebound energy consumption is found to decrease as outdoor temperatures increase (as heat loss during the rebound period is then reduced) and to increase when the indoor temperature at the end of an intervention is below the setpoint.

Interestingly, comparing Models (1) and (2) (without household FE), the estimate for T^{f}_{in} in Model (1) is insignificant, while replacing it with $\Delta T^{f}_{\text{setpoint},ih}$ in Model (2) results in a significant parameter. This suggests that the rebound is driven by the actual difference between the indoor temperature at the end of an intervention and the user setpoint, rather than the actual value of the indoor temperature. Hence, $\Delta T^{f}_{\text{setpoint},ih}$ better captures the true data generation process.

Including household FE results in higher adjusted R^2 in Models (3) and (4). In Model (3) (achieving highest adjusted R^2), $\Delta T^{f}_{\text{setpoint},ih}$ has the highest marginal effect on $E^{16h}_{\text{rebound},ih}$. A one-degree increase in the difference between the user's setpoint temperature and the indoor temperature at the end of an intervention increases rebound by 0.88 kWh on average within a household (significant at the 5% level), as HPs compensate for larger temperature deviations. Additionally, a one-degree increase in the average outdoor temperature within the rebound window reduces rebound by 0.72 kWh on average within a household (significant at the 1% level). Once these parameters are controlled for, the TOD dummies included in Model (4) are insignificant, indicating no evidence that the period of the day when the intervention ends affects rebound consumption.

[26] The average outdoor temperature marginally improved adjusted R^2 compared to other parametrizations, such as the minimum temperature.

[27] Additional specifications were tested, including models with explicit parameters for DHW temperature at the end of the intervention. However, the corresponding estimates were found to be insignificant for both decoupled and non-decoupled HPs. This aligns with expectations: for decoupled units, the rebound is unrelated to DHW reheating, as the hot water buffer is already allowed by the system to be reheated during the intervention if needed. For non-decoupled units, as most interventions stopped due to the DHW automatic threshold, there is only limited variability in T^{f}_{DHW} around 40 °C at the end of an intervention, resulting in an insignificant parameter.

Table 8: Linear regression results for energy consumption rebound 16 h after intervention stop (in kWh)

	(1)	(2)	(3)	(4)
$\overline{T}_{\text{out}}^{\leq 16\text{h}}$	-0.768**	-0.746**	-0.721**	-0.740**
	(0.000)	(0.004)	(0.000)	(0.000)
T_{in}^{f}	-0.557			
	(0.318)			
$\Delta T_{setpoint}^{f}$		1.062**	0.881*	0.823*
		(0.007)	(0.022)	(0.030)
$\text{TOD}_{\text{AM}}^{f}$				0.768
				(0.311)
$\text{TOD}_{\text{evening}}^{f}$				-0.117
				(0.699)
$\text{TOD}_{\text{night}}^{f}$				0.462
				(0.457)
Constant	19.601+	8.057**		
	(0.092)	(0.000)		
Household-FE	No	No	Yes	Yes
Adj. R-Square	0.427	0.480	0.499	0.498
N observations	261	261	261	261

Linear regression estimates. Models (3) and (4) include household-fixed effects. The reference category for TOD^{f} is afternoon (12 a.m. - 6 p.m.). P-values (in parentheses) are derived from wild cluster bootstrapped standard errors (100,000 repetitions) clustered at the household level (nine clusters for all models). $^{+}p < 0.1$,$^{*}p < 0.05$, $^{**}p < 0.01$.