Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

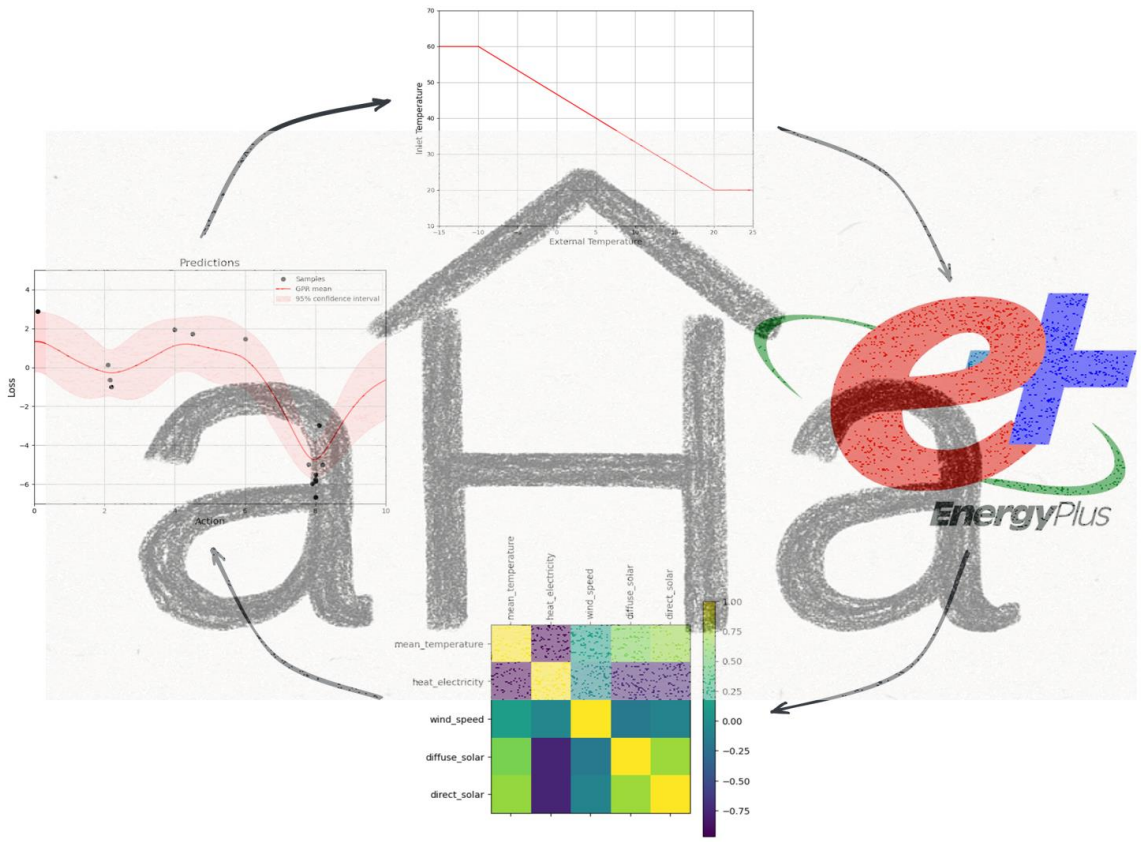
Federal Department of the Environment,
Transport, Energy and Communications DETEC

Swiss Federal Office of Energy
Energy Research and Cleantech

Final report from 16 October 2025

Self-learning adaptive heating curve adjustment for intuitive optimization

AHA



Source: own illustration

Date: 16.10.2025

Location: Dübendorf/Grabs

Subsidy Provider:
Swiss Federal Office of Energy SFOE
Energy Research and Cleantech
CH-3003 Berne
www.energy-research.ch

Subsidy Recipients:
Eidgenössische Material Prüfungsanstalt (EMPA)
Urban Energy System LAB (UES LAB)
Ueberlandstrasse 129, 8600 Dübendorf
www.empa.ch

Autors:
Michael Locher, Urban Energy System Lab @ EMPA, michael.locher@empa.ch

SFOE project coordinators:
Andreas Eckmanns, andreas.eckmanns@bfe.admin.ch
Martin Ménard, menard@lowtechlab.ch

SFOE contract number: SI/502672-01

The authors bear the entire responsibility for the content of this report and for the conclusions drawn therefrom.

Summary

This research focuses on optimizing heating systems in buildings, particularly to reduce energy consumption and CO2 emissions, which are significantly impacted by heating demand. In Switzerland, buildings are responsible for a large portion of the nation's energy use and emissions, with heating accounting for most of this. Traditional heating systems rely on static heating curves, set during installation, which often become outdated over time and lead to inefficiencies. This project introduces an adaptive approach that continuously adjusts the heating curve to optimize energy consumption while maintaining thermal comfort.

The main objective of this approach is to autonomously adjust the heating curve based on minimal data inputs such as room temperature, outdoor temperature and heating energy. By doing so, the system aims to minimize energy consumption and comfort deviations using an optimization algorithm. The method involves optimizing the heating curve parameters to minimize a score function that combines energy consumption and comfort deviations. Gaussian Process regression surrogate models are employed to adapt the system and find via Contextual Bayesian Optimization the optimal heating curve parameters over time.

Simulations of the heating system, based on detailed EnergyPlus models for buildings like Bülsweg - a multi-family residential building – and NEST, Sprint – a multi-office building – have been conducted to validate the system's effectiveness. These simulations use real data, including measurements from heat meters and sensors, to assess the heating energy usage and comfort levels. This method includes estimating the average outdoor temperature for the next 24 hours, using this estimate to predict the optimal inlet temperature, and continuously updating the model after each cycle to improve system performance over time.

Field testing of the algorithm is being conducted in collaboration with Lippuner AG in Bülsweg and NEST, Sprint. The real-world test assesses the impact of the algorithm on energy consumption, with comparisons made using linear regression, normalization by heating degree days, and machine learning predictions. Two digital twins are used for validation – one with the algorithm and one without – as control. This allows for a more robust comparison of the effects of the optimization algorithm.

The adaptive heating approach achieved energy savings of approximately 4 – 6% compared to the static baseline while maintaining nearly unchanged comfort levels (Comfort Score: 0.24 → 0.23). Among the four evaluation methods: Linear Regression, Heating Degree Day normalization, Machine Learning prediction, and Digital Twin analysis. The HDD-based estimate proved most reliable, balancing weather normalization and model robustness. Although the Digital Twin failed to reproduce these gains due to limited extrapolation beyond its training range (30 – 45 °C), real-world data confirmed measurable efficiency improvements. Further analysis suggests that extending the optimization range downward to 20 °C could unlock additional savings potential. The Bayesian Optimization model demonstrated stable convergence after 30 – 40 iterations, but robustness tests revealed asymmetric sensitivity: tolerant to strong positive outliers yet vulnerable to negative ones. This finding highlights the need for improved noise handling and regularization in future iterations.

Future work will focus on scaling and refining the adaptive heating framework to enable autonomous, data-driven optimization across multiple buildings. The next development phase aims to automate data processing and model retraining, enhance robustness against sensor noise and outliers, and expand the optimization range to fully capture the system’s energy-saving potential. Additional field tests, including applications in single-family houses with heat pumps, will assess the method’s transferability and reliability under diverse operating conditions. Furthermore, integrating additional contextual factors—such as wind, solar radiation, and occupancy dynamics, directly into the learning model will improve both predictive accuracy and comfort control. These developments will strengthen the adaptive heating approach as a scalable and robust alternative to static control strategies, advancing the transition toward intelligent, energy-efficient building operation.

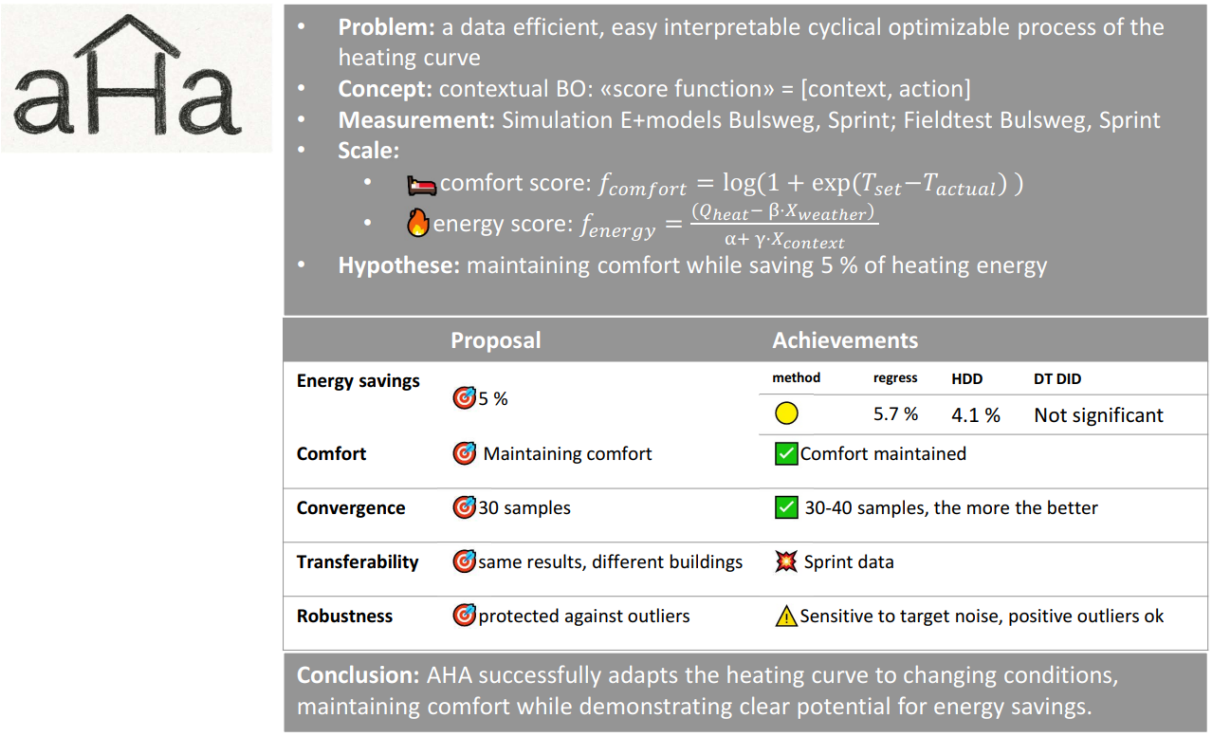


Figure 1: This figure provides a comprehensive overview of Project AHA, illustrating the core concept of optimizing heating curves using contextual Bayesian optimization. It visualizes the problem statement, the proposed approach, and the underlying hypothesis. The diagram also outlines the measurement strategy, the scale of the study, and both the proposed and achieved results. The optimization process is based on a Gaussian Process Regression model that is iteratively updated with contextual data (outdoor temperature) and system actions (inlet temperature) to minimize a score function. This score combines normalized values of energy consumption and comfort deviations relative to a chosen reference point, demonstrating how the adaptive method balances comfort maintenance with energy savings in practice.

Zusammenfassung

Dieses Forschungsprojekt konzentriert sich auf die Optimierung von Heizsystemen in Gebäuden, mit dem Ziel, den Energieverbrauch und die CO_2-Emissionen zu reduzieren, die wesentlich durch den Heizbedarf beeinflusst werden. In der Schweiz sind Gebäude für einen grossen Anteil des nationalen Energieverbrauchs und der Emissionen verantwortlich, wobei das Heizen den grössten Teil ausmacht. Herkömmliche Heizsysteme basieren auf statischen Heizkurven, die bei der Inbetriebnahme festgelegt werden. Diese werden im Laufe der Zeit oft obsolet und führen zu Ineffizienzen. Das vorliegende Projekt stellt einen adaptiven Ansatz vor, der die Heizkurve kontinuierlich anpasst, um den Energieverbrauch zu optimieren und gleichzeitig den thermischen Komfort sicherzustellen.

Das Hauptziel dieses Ansatzes besteht darin, die Heizkurve autonom auf Basis weniger Eingangsdaten – wie Raumtemperatur, Aussentemperatur und Heizenergie – anzupassen. Dadurch soll der Energieverbrauch und die Abweichung vom Komfortniveau mithilfe eines Optimierungsalgorithmus minimiert werden. Die Methode beruht auf der Optimierung der Heizkurvenparameter zur Minimierung einer Zielfunktion, die Energieverbrauch und Komfortabweichungen kombiniert. Hierfür werden surrogatbasierte Modelle mittels Gaussian Process Regression eingesetzt, welche über kontextuelle Bayes'sche Optimierung die optimalen Heizkurvenparameter im Zeitverlauf bestimmen.

Zur Validierung der Wirksamkeit des Systems wurden Simulationen des Heizsystems auf Basis detaillierter EnergyPlus-Modelle für Gebäude wie Bülsweg – ein Mehrfamilienhaus – und NEST Sprint – ein

Bürogebäude – durchgeführt. Diese Simulationen nutzen reale Daten, darunter Messungen von Wärmezählern und Sensoren, um den Heizenergieverbrauch und den Komfort zu bewerten. Die Methode umfasst zudem die Prognose der durchschnittlichen Aussentemperatur für die nächsten 24 Stunden, die Nutzung dieser Schätzung zur Vorhersage der optimalen Vorlauftemperatur sowie die kontinuierliche Aktualisierung des Modells nach jedem Zyklus, um die Systemleistung schrittweise zu verbessern.

Feldtests des Algorithmus werden in Zusammenarbeit mit Lippuner AG in Bülsweg und NEST Sprint durchgeführt. Diese Praxistests bewerten den Einfluss des Algorithmus auf den Energieverbrauch, wobei Vergleiche anhand linearer Regression, Normalisierung über Heizgradtage und maschinelles Lernen gezogen werden. Zwei digitale Zwillinge dienen der Validierung – einer mit und einer ohne Algorithmus – um die Effekte der Optimierung robust zu quantifizieren.

Der adaptive Heizansatz erzielte Energieeinsparungen von etwa 4 – 6 % gegenüber der statischen Referenz, bei nahezu unverändertem Komfortniveau (Comfort Score: 0.24 → 0.23). Von den vier Bewertungsmethoden – Lineare Regression, Heizgradtage-Normalisierung, Machine-Learning-Vorhersage und Digital Twin Analyse – erwies sich der HDD-basierte Ansatz als am zuverlässigsten, da er Wettereinflüsse berücksichtigt und eine robuste Modellgüte aufweist. Obwohl der Digitale Zwilling diese Effizienzgewinne aufgrund begrenzter Extrapolation (30 – 45 °C) nicht reproduzieren konnte, bestätigten reale Felddaten messbare Verbesserungen. Weitere Analysen deuten darauf hin, dass eine Erweiterung des Optimierungsbereichs bis 20 °C zusätzliches Einsparpotenzial freisetzen könnte. Das Bayes'sche Optimierungsmodell zeigte eine stabile Konvergenz nach 30 – 40 Iterationen, jedoch auch eine asymmetrische Sensitivität: robust gegenüber positiven, aber anfällig gegenüber negativen Ausreissern. Dies unterstreicht die Notwendigkeit einer verbesserten Rauschbehandlung und Regularisierung in zukünftigen Versionen.

Die nächste Entwicklungsphase konzentriert sich auf die Skalierung und Verfeinerung des adaptiven Heizrahmens, um eine autonome, datengetriebene Optimierung über mehrere Gebäude hinweg zu ermöglichen. Ziel ist die Automatisierung der Datenverarbeitung und des Modell-Updates, die Erhöhung der Robustheit gegenüber Sensorausfällen und Ausreissern sowie die Erweiterung des Optimierungsbereichs, um das volle Energieeinsparpotenzial des Systems zu erschliessen. Weitere Feldversuche – insbesondere in Einfamilienhäusern mit Wärmepumpen – werden die Übertragbarkeit und Zuverlässigkeit der Methode unter unterschiedlichen Betriebsbedingungen bewerten. Darüber hinaus soll die Integration zusätzlicher kontextueller Faktoren wie Wind, solare Einstrahlung und Belegungsdynamik direkt in das Lernmodell die Vorhersagegenauigkeit und den Komfort weiter verbessern. Diese Weiterentwicklungen stärken den adaptiven Heizansatz als skalierbare und robuste Alternative zu statischen Regelstrategien und fördern den Übergang zu einer intelligenten, energieeffizienten Gebäudebewirtschaftung.

Resumé

Cette recherche se concentre sur l'optimisation des systèmes de chauffage dans les bâtiments, dans le but de réduire la consommation d'énergie et les émissions de CO_2, fortement influencées par la demande de chauffage. En Suisse, les bâtiments représentent une part importante de la consommation énergétique nationale et des émissions, le chauffage en constituant la plus grande part. Les systèmes de chauffage traditionnels reposent sur des courbes de chauffe statiques, définies lors de l'installation, qui deviennent souvent obsolètes au fil du temps et entraînent des inefficacités. Le présent projet introduit une approche adaptative qui ajuste en continu la courbe de chauffe afin d'optimiser la consommation d'énergie tout en maintenant le confort thermique.

L'objectif principal de cette approche est d'ajuster de manière autonome la courbe de chauffe à partir d'un minimum de données d'entrée, telles que la température intérieure, la température extérieure et l'énergie thermique. L'algorithme vise à minimiser la consommation d'énergie et les écarts de confort à l'aide d'un processus d'optimisation. La méthode consiste à optimiser les paramètres de la courbe de chauffe afin de minimiser une fonction de coût combinant la consommation d'énergie et les écarts de

confort. Des modèles de substitution basés sur la régression par processus gaussiens (Gaussian Process Regression) sont utilisés, permettant, via l'optimisation bayésienne contextuelle, d'identifier au fil du temps les paramètres optimaux de la courbe de chauffe.

Pour valider l'efficacité du système, des simulations EnergyPlus détaillées ont été réalisées pour des bâtiments tels que Bülsweg – un immeuble résidentiel multifamilial – et NEST Sprint – un bâtiment de bureaux. Ces simulations utilisent des données réelles, incluant les mesures de compteurs de chaleur et de capteurs, afin d'évaluer la consommation de chauffage et les niveaux de confort. La méthode inclut également l'estimation de la température extérieure moyenne sur les 24 heures suivantes, l'utilisation de cette prévision pour déterminer la température d'entrée optimale, et la mise à jour continue du modèle après chaque cycle, améliorant ainsi progressivement les performances du système.

Des tests sur le terrain du nouvel algorithme sont menés en collaboration avec Lippuner AG sur les sites de Bülsweg et de NEST Sprint. Ces essais évaluent l'impact de l'algorithme sur la consommation énergétique, à l'aide de comparaisons fondées sur la régression linéaire, la normalisation par degrés-jours de chauffage et des prédictions issues de l'apprentissage automatique. Deux jumeaux numériques servent à la validation – l'un intégrant l'algorithme et l'autre non – afin de comparer de manière robuste les effets de l'optimisation.

L'approche de chauffage adaptative a permis d'obtenir des économies d'énergie de l'ordre de 4 à 6 % par rapport à la configuration statique, tout en maintenant un niveau de confort quasi identique (Comfort Score : 0.24 → 0.23). Parmi les quatre méthodes d'évaluation – régression linéaire, normalisation par degrés-jours de chauffage, prédiction par apprentissage automatique et analyse par jumeau numérique – l'estimation basée sur les degrés-jours s'est révélée la plus fiable, équilibrant la normalisation météorologique et la robustesse du modèle. Bien que le jumeau numérique n'ait pas reproduit ces gains, en raison de sa capacité limitée d'extrapolation (30 – 45 °C), les données réelles ont confirmé des améliorations mesurables en efficacité. Une analyse complémentaire indique qu'étendre la plage d'optimisation jusqu'à 20 °C pourrait libérer un potentiel d'économie supplémentaire. Le modèle d'optimisation bayésienne a montré une convergence stable après 30 à 40 itérations, mais également une sensibilité asymétrique : tolérant aux fortes valeurs aberrantes positives mais vulnérable aux valeurs négatives. Cela souligne la nécessité d'un meilleur traitement du bruit et de la régularisation dans les futures versions.

La prochaine phase de développement visera à étendre et perfectionner le cadre de chauffage adaptatif, afin de permettre une optimisation autonome et fondée sur les données à l'échelle de plusieurs bâtiments. Elle se concentrera sur l'automatisation du traitement des données et de la réactualisation des modèles, le renforcement de la robustesse face aux bruits de capteurs et aux valeurs aberrantes, ainsi que l'élargissement de la plage d'optimisation pour exploiter pleinement le potentiel d'économie d'énergie du système. Des essais supplémentaires – notamment dans des maisons individuelles équipées de pompes à chaleur – permettront d'évaluer la transférabilité et la fiabilité de la méthode dans des conditions d'exploitation variées. Enfin, l'intégration de facteurs contextuels supplémentaires, tels que le vent, le rayonnement solaire et la dynamique d'occupation, directement dans le modèle d'apprentissage, améliorera la précision prédictive et la gestion du confort. Ces développements renforceront l'approche de chauffage adaptatif en tant qu'alternative évolutive et robuste aux stratégies de contrôle statiques, contribuant ainsi à la transition vers une exploitation des bâtiments plus intelligente et plus économe en énergie.

Take-Home Message

The AHA framework offers a simple, interpretable, and transferable optimization concept. Its lightweight design and minimal data requirements make it adaptable to various heating systems, enabling scalable deployment across Switzerland's building stock. By improving control logic rather than hardware, the approach directly contributes to national goals of reducing building-related energy use and emissions.

Adaptive heating curves enable measurable energy savings without compromising comfort. Field tests demonstrated a potential reduction in heating energy consumption of approximately 4–6%, confirming the potential of data-driven control to lower CO_2 emissions in Swiss buildings while maintaining thermal comfort.

Automation and robust data infrastructure are key to scalability. Manual data handling, unreliable sensors, and delayed system feedback highlighted the need for a fully automated and validated data pipeline to ensure reliable optimization across multiple buildings and heating systems.

Contents

List of abbreviations

SFOE	Definition: Swiss Federal Office of Energy
TB	Definition: Testbed. As Testbeds we can make use of NEST Umar, NEST SPRINT and Bülsweg.
DT	Definition: Digital Twin. When we refer to Digital Twins, we are actually referring to EnergyPlus models, which we use as a framework to simulate energy efficiencies. We use 'EnergyPlus models' and 'Digital Twins' interchangeably.
BO	Definition: Bayesian Optimization. We use Bayesian Optimization for a smart selection process to efficiently identify optimal parameters for the heating system.
CI	Definition: Confidence Interval quantifies the uncertainty associated with an estimated effect or parameter. In this work, the CI represents the 95% range within which the true value of the estimated energy savings is expected to lie, based on the variability observed in the data. The CIs were obtained using a bootstrapping procedure, which repeatedly resamples the data to construct an empirical distribution of the Difference-in-Differences estimates.
DID	Definition: The Difference-in-Differences method estimates the causal effect of an intervention by comparing the change in outcomes over time between a treatment and a control group. In this work, DID quantifies the impact of the adaptive heating tuner relative to the static baseline by evaluating the differential change in heat energy consumption predicted by two digital twins (DT1: adaptive, DT2: static). The approach is implemented in a bootstrapped framework to incorporate sampling variability and model uncertainty from the testbed.
GPR	Definition: Gaussian Process Regression. We use Gaussian Process Regression to approximate the underlying score function of heating energy consumption and comfort score, enabling a seamless implementation of Bayesian Optimization
Model	Definition: Our surrogate model, a Gaussian Process Regression model used within the Bayesian Optimization framework, learns the combined score (comfort + energy) as a function of inlet temperature (action) and context (24 hour mean outdoor temperature)
HDD	Definition: Heating Degree Days method, see Appendix for further explanations.
Inlet	Definition: Inlet water temperature of the heating circuit.
Action	Definition: In general, actions are the controllable parameter of a setting. In our specific implementation of a 1-dimensional linear heating curve action refers to the adjustment of the inlet temperature.
Context	Definition: In general, contexts are the non-controllable parameters of a setting e.g. weather conditions. In our specific implementation context refers to the mean outdoor temperature.
Energy Score	Definition: A normalized metric that quantifies the energy efficiency of the heating system over a 24-hour period, corrected for external weather influences such as sunshine duration and wind speed. Calculation: The raw heating energy consumption is denoised by subtracting the estimated influence of sunshine and wind, based on a linear regression model trained on historical data. This isolates the influence of outdoor temperature. The result is then normalized by dividing it by the expected energy consumption predicted by the denoising model for the given weather context. Interpretation: • A score of **1** means the adaptive heating curve performs around the estimated performance of the static configuration

- A score **< 1** indicates improved efficiency with respect to static configuration
- A score **> 1** indicates reduced efficiency with respect to static configuration

Formula:

$$f_{energy} = \frac{(Q_{heat} - \beta \cdot X_{weather})}{\alpha + \gamma \cdot X_{context}}$$

Q_{heat} = measured heating energy consumption

$X_{weather}$ = weather variables: sunshine duration and wind speed

$X_{context}$ = contextual variables, mean outdoor temperature

Comfort score

Definition: A penalty score that reflects how far the actual indoor temperature falls below a defined comfort threshold (20°C). It penalizes only underheating, as exceeding the comfort temperature is not considered a penalty because a heating system cannot cool when temperatures exceed the target.

Calculation: The score is computed using a smooth penalty function (log-sum-exp) applied only when the actual temperature is below the comfort threshold. This ensures that small deviations are tolerated, while large deviations are penalized increasingly.

Interpretation:

- A low score indicates good thermal comfort: temperature close to or above setpoint.
- A high score indicates poor comfort: temperature significantly below the threshold.

Formula:

$$f_{comfort} = \log(1 + \exp(T_{setpoint} - T_{actual}))$$

$T_{setpoint}$ = defined comfort threshold

T_{actual} = measured indoor temperature

1 Introduction

1.1 Context and motivation

Buildings account for approximately 42% of Switzerland's final energy consumption and 26% of total CO2 emissions, with heating demand being the primary driver at 68%. Heating systems are typically optimized using static heating curves set manually by technicians during installation. These initial settings are often suboptimal, degrade over time, and are rarely adjusted for changing conditions, such as tenant turnover. This leads to higher energy consumption, comfort issues, and increased scores.

Our adaptive heating curve tuning offers a continuous, self-adjusting solution. It relies on minimal data inputs – room temperature, heating and cooling energy use, outdoor temperature, and solar radiation – and optimizes using only 30 days of data. Daily updates require the building's total energy consumption and comfort deviations. This approach is lightweight, adaptable, and can run in the cloud or locally on the heating system's computer, ensuring efficient, responsive heating optimization.

1.2 Project objectives

The goal is to maintain thermal comfort while reducing energy consumption. The method autonomously adjusts the heating curve without altering the system structure. In this optimization problem, we aim to find the best heating curve parameters s to minimize our score function:

$$\min_{s \in S} \; y(\boldsymbol{s}, z) = a \cdot f_{energy}(\boldsymbol{s}, z) + b \cdot f_{Comfort}(\boldsymbol{s}, z) + \epsilon \tag{1.0}$$

- **Action s**: the heating curve parameters that control the heating system, in our setting the inlet temperature.
- **Context z**: the daily mean outdoor temperature, which influences heating energy use

We define a score as the value returned by the score function, which combines heating energy score and comfort score:

- **Heating energy score:** the function models the energy required for heating based on the parameters s and outside temperature z.
- **Comfort score:** the function measures how much the indoor temperature deviates from the desired setpoint.

The objective is to optimize s to minimize energy consumption while maintaining comfort, accounting for uncertainties through a noise term ϵ.

By incorporating Gaussian Process regression with an exponential kernel, the approach effectively models the relationship between heating energy, comfort, and ambient temperature. The method converges to the optimal heating curve settings, improving energy efficiency and comfort over time.

1.2.1.Comfort Score

To estimate comfort score, we calculate penalties for indoor temperatures that fall below the specified set point temperature. The calculation is performed separately for daytime (5:00–22:00) and nighttime (0:00–5:00 and 22:00–24:00) periods, recognizing their distinct comfort requirements. Temperatures above the setpoint are not penalized, and the total comfort score is normalized by the number of measurements taken during the respective period. For a clear illustration, refer to Figure 3 (comfort score function). The function we use to estimate the degree of daily comfort scores is as follows:

$$f_{comfort} = \log\left(1 + exp\left(T_{setpoint_{\frac{night}{day}}} - T_{actual}\right)\right) - \log(2) \tag{1.1}$$

This formula was chosen for its smoothness, as it avoids sharp transitions that could complicate model estimation of the underlying score function. Its design effectively penalizes deviations below the comfort

target while maintaining interpretability regarding the severity of comfort score. Small deviations from the set point result in minor penalties, whereas larger deviations are penalized exponentially, highlighting significant comfort score more prominently.

1.2.1.1. Comfort Objective: Comparison with ASHRAE

Fanger's model is the scientific standard for estimating indoor thermal comfort, using the metrics PMV (Predicted Mean Vote) and PPD (Predicted Percentage of Dissatisfied) to quantify how comfortable people feel in each environment. We did come up with our own comfort function due to simple reasons:

- PMV and PPD are nonlinear functions and can lead to problems while optimizing our objective function
- Fanger evaluates both temperature exceedances and deficits, which is problematic in our context since a heating system cannot cool when temperatures exceed the target, and such cases should therefore not be penalized.

In contrast, our comfort violation formula provides a continuous and more efficient approach, focusing on the key variable temperature, making it better suited for optimization problems.

A direct comparison of PMV and PPD for some selected days during our heating period display that is with respect to correlation highly similar. The Base assumption for comparison is a standard case in winter: activity = "Typing", garments = ["Sweatpants", "Long sleeve shirt (thick)", "Thick trousers", "Calf length socks", "Slippers"]

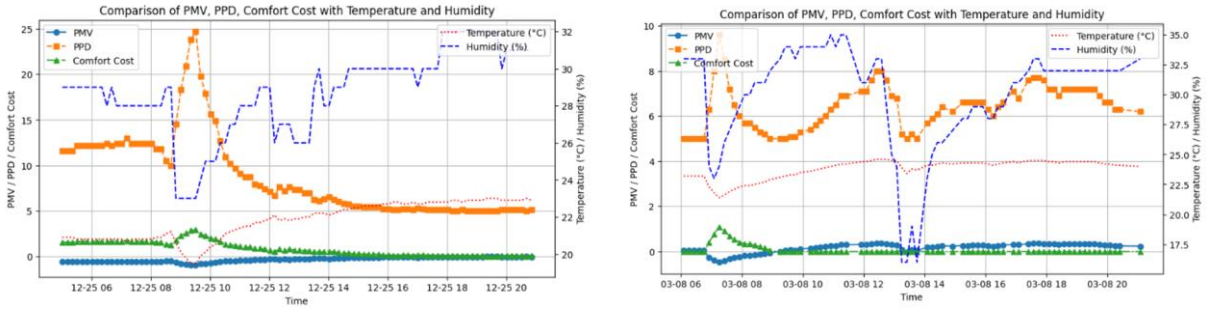


Figure 2: Displayed is a direct comparison of PMV, PPD, and our comfort score metric over the course of the selected days: 25.12.24 and 08.03.2025. Aside from the obvious differences in scale, the overall behaviour is notably similar.

For these two days, we calculated the Pearson correlation coefficients and obtained the following results:

- PMV and Comfort Score: Very strong negative correlation of -0.96
- PPD and Comfort Score: Very strong positive correlation of +0.98

These high correlations indicate that both PMV and PPD are strongly aligned with our comfort score metric. This suggests that the proposed comfort score approximation captures the essential information of traditional comfort indicators and is therefore an efficient candidate for estimating occupant comfort.

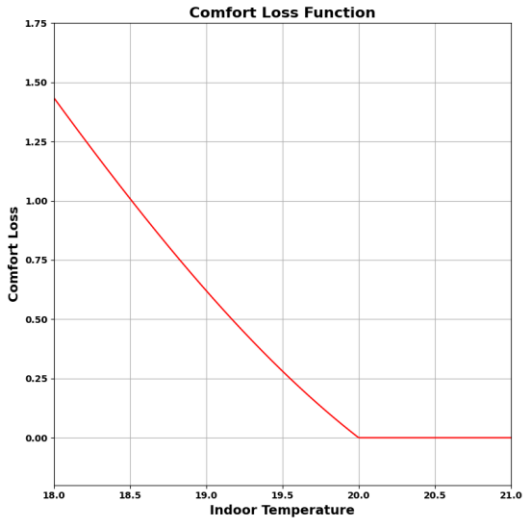


Figure 3: Shows comfort score estimated by our comfort loss function. We only consider comfort scores which are temperatures below a defined threshold, because the heating system can only warm but not cool.

2 Approach and Method

In our approach a simple two-point parametrization of the heating curve is proposed together with the adaptive heating curve strategy GPR and BO, see for more details appendix 5.1. The adaptive solutions are discussed regarding the tracking of the defined score based on the desired indoor temperature $T_{setpoint}$ and heating energy in simulations and field tests.

2.1 Parameterization of the heating curve

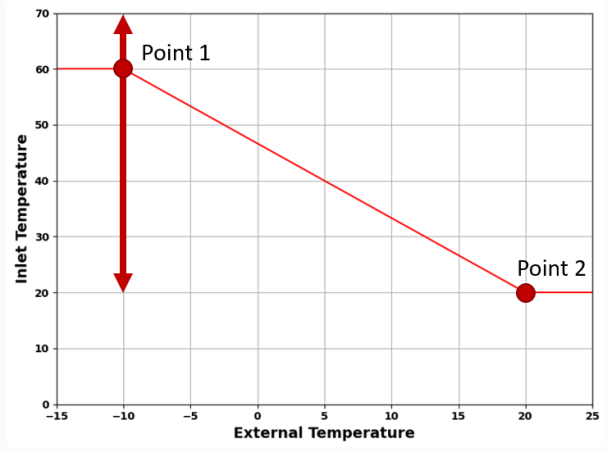


Figure 4: Shown is a 2 point linear heating curve defined by Point 1 and Point 2. We focus on a 1-dimensional linear heating curve, specifically on reference point 1 on $T1_{inlet}$, which represents the inlet temperature. Highlighted for reference is the potential safe range of parameter values considered during our Bayesian optimization process, ranging from 20 up to 70 degrees Celsius. This ensures that the optimization remains within safe operational limits while seeking the optimal solution. Right:

The heating curve defines the relationship between the outdoor temperature and the inlet temperature of the heating system, ensuring efficient, weather-controlled operation. For heat pump-based floor

heating, optimizing the heating curve helps maintain thermal comfort, reduce operational scores, and maximize the heat pump's efficiency. Typically, the heating curve is linear, as heating demand is approximately proportional to outdoor temperature [5].

We simplified our optimization to a one-dimensional linear heating curve, see figure 4, parameterized using two reference points based on the mean external temperature:

- Reference Point 1 = $(T1_{external}, T1_{inlet})$
- Reference Point 2 = $(T2_{external}, T2_{inlet})$

This approach leads to a strong simplification, ensuring easier interpretation, reduced data requirements, and straightforward adaptation for optimal system performance.

2.2 Adaptive optimization of the heating curve

Our adaptive heating strategy automatically adjusts the heating curve daily, without requiring detailed prior modelling of the building or heating system. To do this, we define a generic score function (1.0), which is initially estimated for each new building and then gradually optimized over time. This function calculates a score based on thermal comfort score and energy score, which depend on the selected action – the inlet temperature $T1_{inlet}$ – and the prevailing context – the mean outdoor temperature. Intuitively, certain combinations of inlet and outdoor temperatures will result in a higher (i.e., more scorely) score – for example, high inlet temperatures during warm outdoor conditions, or very low inlet temperatures when it's cold. In contrast, high inlet temperatures during cold weather are more reasonable and should lead to lower scores.

The exact behaviour and magnitude of this score are building-specific and must be learned through continuous sampling and evaluation of the relationship between inlet temperature, context, and resulting energy and comfort scores which defines the score of a building. To guide the sampling process, we define a “safe space” – a range of inlet temperatures that are considered safe for the given building and heating system. For example, for underfloor heating, an inlet range of 25 - 45°C is typically safe. We begin by sampling 10 predefined points from this safe space. Example for Bülsweg: [30–45°C].

After this initial exploration, the actual optimization begins. The data collected from the initial samples is used to fit a first version of the scoring function. This fitted model can then be used to determine the optimal inlet temperature for a given context – specifically, the estimated outdoor temperature for the next day – to minimize the score.

This process of sampling, updating the model, and selecting the best possible inlet temperature is repeated iteratively each day.

In detail, the procedure for adapting and optimizing the algorithm works as follows: We use Bayesian Optimization (BO) to guide this process. BO systematically predicts based on the latest measurement the next best $T1_{inlet}$ most likely to reduce the score function. Each measurement contributes a new data point to evaluate the score function, which is modelled using GPR. This approach enables efficient exploration of the available parameter space of the heating curve.

The adaption of the reference point 1 based on the mean outdoor temperature is performed according to the problem iteration cycle [2, 3]:

2.2.1.Problem Iteration Cycle

Pseudocode of our contextual Bayesian Optimization approach to optimize the heating curve, see also figure 5 for a summarized description:

1. Estimation of Context: mean outdoor temperature of previous day:

We take contextual information into account only above a certain threshold of samples, hard coded 10, and focus initially solely on a safe space for $T1_{inlet}$ itself. This space of safe points contains 10 entries ranging from [30–45°C] for Bülsweg.

2. **Estimation of next $T1_{inlet}$ based on context from 1 and existing surrogate model with Bayesian Optimization algorithm, upper confidence bound.**

We start to use upper confidence bound only after a list of safe points have been sampled first. BO using the upper confidence bound strategy uses the updated surrogate model to make it's prediction of the next optimal inlet temperature.

3. **Update heating curve parameters with next $T1_{inlet}$ and start simulation**

Energyplus engine creates an energyplusoutput file which is processed by Simulationdriver class to be used for the ML process

4. **The heating energy from the Energyplusmodel is processed by our Denoiser to denoise and normalize the heating energy. This result is then combined with the estimated comfort violation to estimate our score, which represents our new sample.**
5. **The new sample is used to update our GPR**
6. **The iteration cycles starts over, by estimating the new context via mean temperature of the previous day**

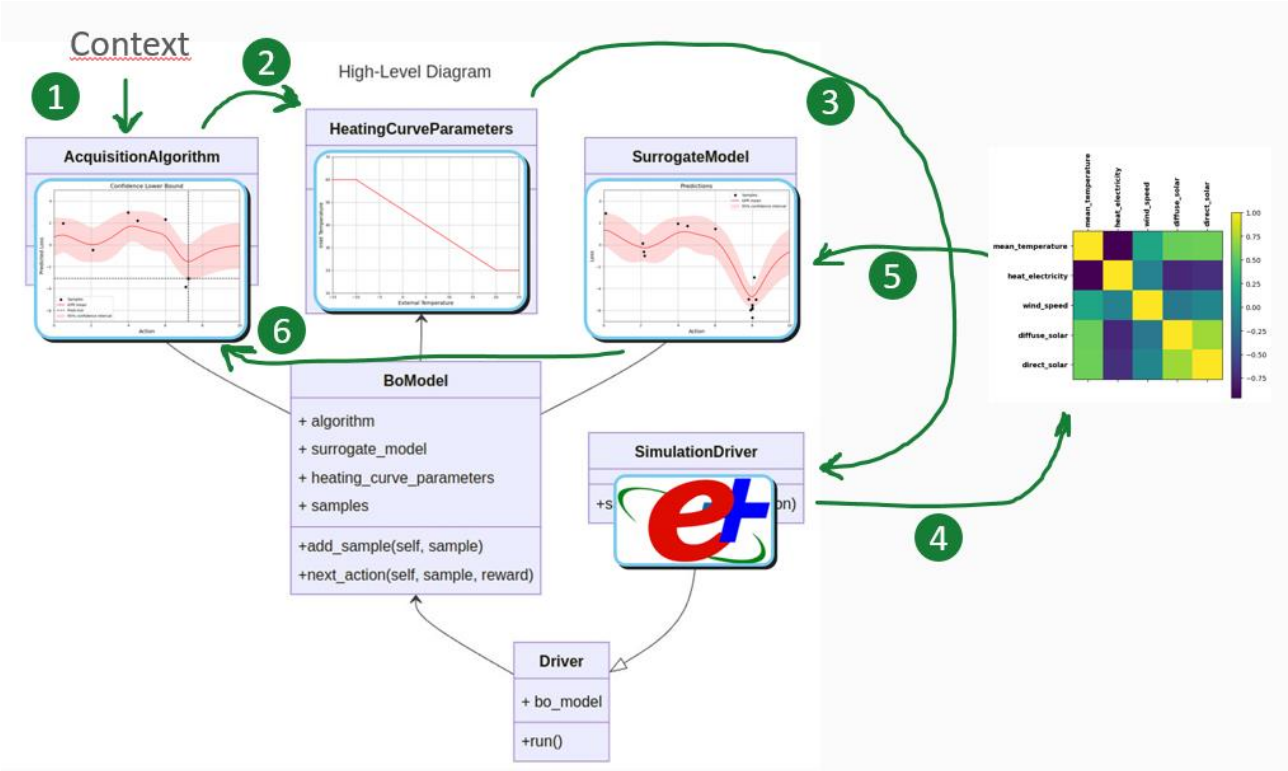


Figure 5: The process begins with the estimation of the context, using the mean temperature of the previous day. Next, the algorithm estimates the next action using the context and an existing surrogate model, applying the Upper Confidence Bound method to obtain the new heating parameters. With the heating curve parameters updated the EnergyPlus simulation is run to generate heating energy and comfort violation data. The heating energy data is denoised, normalized and combined with comfort score to estimate the score. This new sample is used to update the Gaussian Process regressor – which mimics our understanding of the score function, and the cycle repeats by estimating the new context.

3 Results and Discussion

To demonstrate the performance of the proposed adaptive strategy and validate our claim that a dynamic, adaptive heating tuner can achieve a 5% reduction in energy consumption while maintaining comfort, the following objectives must be shown:

- Energy Efficiency: Achieve a 5% reduction in energy consumption while maintaining comfort levels throughout the entire heating period.
- Convergence Behaviour: After 30 iterations of optimizing the heating curve per context, there should be no significant changes in regret, difference between an optimal.
- Transferability: The algorithm must be applicable across different building types and heating systems, delivering comparable results.
- Robustness: The system should effectively respond to extreme climate and user scenarios and handle missing data without performance degradation.

To achieve these goals, we will conduct simulations followed by real-world field tests. The simulations are based on detailed EnergyPlus models to replicate various building and system configurations. For further details on the simulation models, see chapter 7.2. Some of the objectives listed above will be analysed only within the simulations: convergence and robustness. To analyse convergence properly, we need an idea of an optimal curve, which can be found through grid search using contextual data. This will allow us to specifically test whether our adaptive strategy converges toward this curve. We will test robustness by applying sudden temperature changes, which are rare in real-world scenarios and difficult to predict.

3.1 Simulation

3.1.1. Energy Score

To assess the normalized energy consumption – the energy score, see problem iteration cycle 2.2.1. point 4 – we fixed the inlet temperature $T1_{inlet}$ to a specific value e.g., 30°C for one heating period and simulated the entire period for NEST, see figure 6. This process was repeated for inlet temperatures ranging from 30°C to 60°C. Each data point for a given inlet temperature corresponds to the total daily heating energy consumption. Additionally, each day is color-coded based on the average outdoor temperature: **blue** indicates colder conditions (around 0 °C), while **red** indicates warmer conditions (around 8 °C).

It is evident that the Energy score is lowest for the lowest inlet temperature and increases as the inlet temperature rises. Another observable trend is that the increase in scores levels off above 50°C, primarily because higher inlet temperatures reduce the time required to heat the space to the desired set point temperature.

What the Energy Score does not reveal, however, is that at inlet temperatures of 30°C or lower, comfort is often significantly compromised, see next chapter.

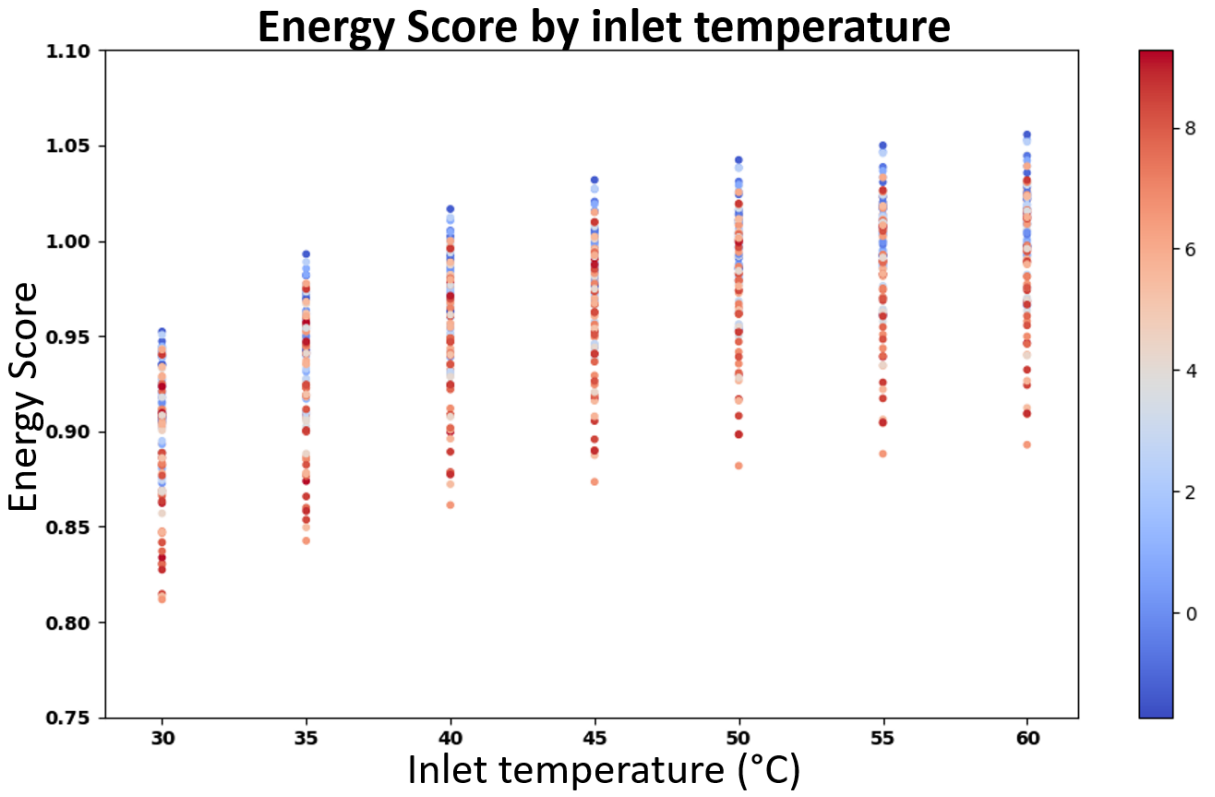


Figure 6: We compute the total consumed heating energy remove noise and normalize the heating energy consumption to a score independent of external conditions, as described in section 2.2.1, point 4. Normalization is based on the default heating curve, where the consumed energy is divided by the energy consumption of the default heating curve. Because the default heating curve is defined at $(T1_{external}, T1_{inlet}) = (-10, 36)$ and $(T2_{external}, T2_{inlet}) = (20, 20)$, heating curves with a lower T_{inlet} will have a reduced heating energy score.

3.1.2.Comfort Score

To assess the comfort score, see 1.2.1, we fixed the inlet temperature to a specific value (e.g., 30°C) for one heating period and simulated the entire period for NEST, see figure 7. This process was repeated for inlet temperatures ranging from 30°C to 60°C. Each data point for a given inlet temperature corresponds to the total daily consumption. Additionally, individual days are color-coded based on context – specifically, the average outdoor temperature: blue represents an average temperature around 0°C, and red represents an average temperature around 5°C.

We can clearly observe that at inlet temperatures above 40°C, there are virtually no comfort scores. Below 40°C, however, significant comfort scores occur, particularly on colder days, while on warmer days, comfort scores are largely absent. These comfort scores arise primarily for two reasons: first, the target temperature at cold outdoor temperature cannot longer be reached with lower inlet temperatures, and second, it takes longer to heat the building to the desired temperature.

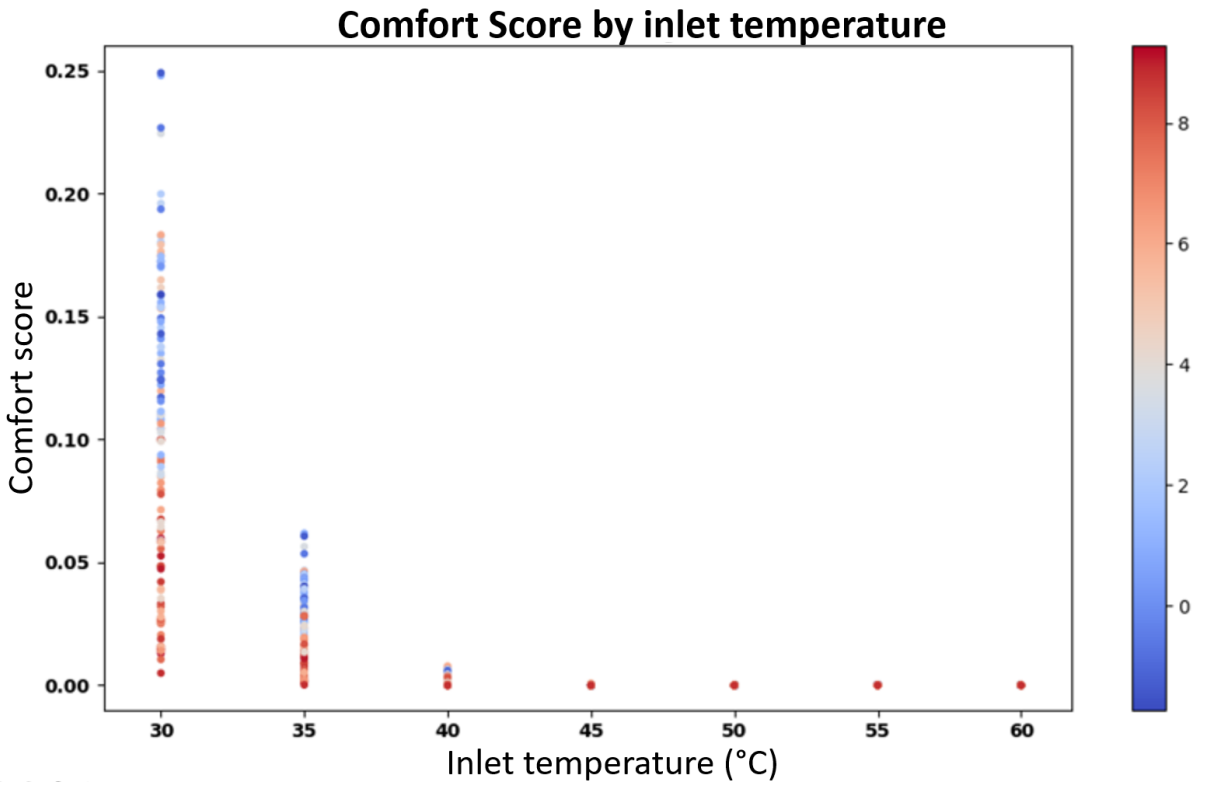


Figure 7: Comfort score for varying inlet temperatures (30°C–60°C). Daily energy consumption is plotted with color coding based on the average outdoor temperature: blue for 0°C and red for 5°C. Comfort score are minimal above 40°C but increase below, particularly on colder days, due to slower heating and difficulty reaching the target temperature.

3.1.3.Combined Score and Score Function

We assumed a static heating curve and varied the inlet temperature between 25°C and 60°C, calculating the resulting scores as the sum of normalized heating energy and daily comfort score, combined score. This analysis was performed for all days within the selected heating period, with each individual point representing a single day, see figure 8 on the left.

Using GPR, we estimated the score function for each context, resulting in a series of curves. These curves are color-coded based on their context: dark blue represents an average outdoor temperature of -2°C, while dark red corresponds to 10°C. The score functions enable us to identify the optimal inlet temperature settings for the heating curves. For instance, the minimum of the dark blue curve is around 40°C, see figure 8 on the right.

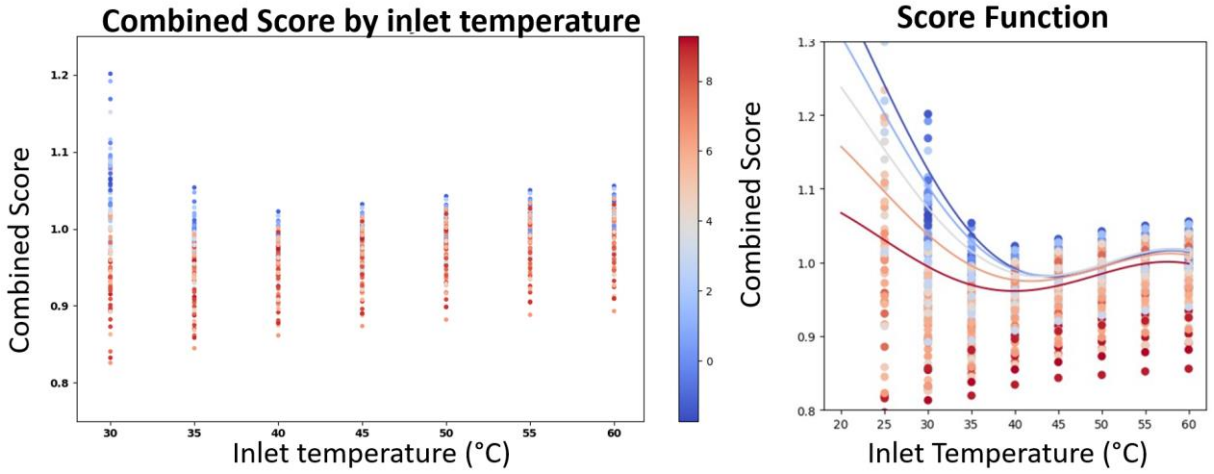


Figure 8: On the left, we show the combined score based on our score function from Chapter 1.2, which is a weighted linear combination. The weights, set as trade-off parameters, significantly impact the potential energy savings, see Appendix chapter 5.4. Here, a weight of 1 is applied, prioritizing low comfort score, especially at lower temperatures, which limits the energy-saving potential. On the right, the trained model made predictions based on context and inlet temperatures. The obtained curves clearly illustrate a distinct minimum, indicating the optimal inlet temperature.

3.1.4.Results Bülsweg and NEST

The adaptive results were calculated using the proposed strategy – the Problem Iteration Cycle 2.2.1. Key findings are illustrated in figures 9, 10 and summarized in the following subsection:

The results of the adaptive heating strategy for NEST and Bülsweg are similar regarding the development of the score function. Thus, selected results for NEST are presented, comparing the adaptive heating curve with two static heating curves with reference points: ($T1_{external}$, $T1_{inlet}$) = $(-10, 60)$, and $(-10, 40)$. We conducted three simulations, one for each heating curve strategy: adaptive, high and medium static heating curves. The daily scores were calculated and plotted for each heating curve to enable a direct comparison. The adaptive heating curve shows that score reductions can be measured as early as the third day, and the adaptive strategy outperforms both static curves. In this example, the adaptive heating strategy achieves energy savings of approximately 1.5% for the chosen heating period of January – March 2021 compared to the lower curve and ~5.9% compared to the higher curve, quickly converging to the optimal solution for each mean outdoor temperature.

Figure 10 presents the results for the best possible choice of $\mathrm{T1_{inlet}}$ based on the context and outdoor temperature. The context is color-coded: dark blue corresponds to an average outdoor temperature of -2°C, and dark red to 10°C, compare chapter 3.1.3. We marked the minimum point for each curve. For example, the estimated score function for NEST - Umar and Bülsweg indicates that the minimum for average daily temperatures of -2°C occurs at an inlet temperature of 40°C.

A direct comparison with the initial heating curve values for Bülsweg: ($T1_{external}$,$T1_{inlet}$) = $(-10, 36)$ and ($T2_{external}$,$T2_{inlet}$) = $(20, 20)$ reveals that the heating curve is only optimal for an average outdoor temperature of 0°C. In contexts deviating from this temperature, score reductions can be achieved by adapting $T1_{inlet}$, with particularly significant savings in warmer contexts. For NEST - Umar, a similar pattern is observed.

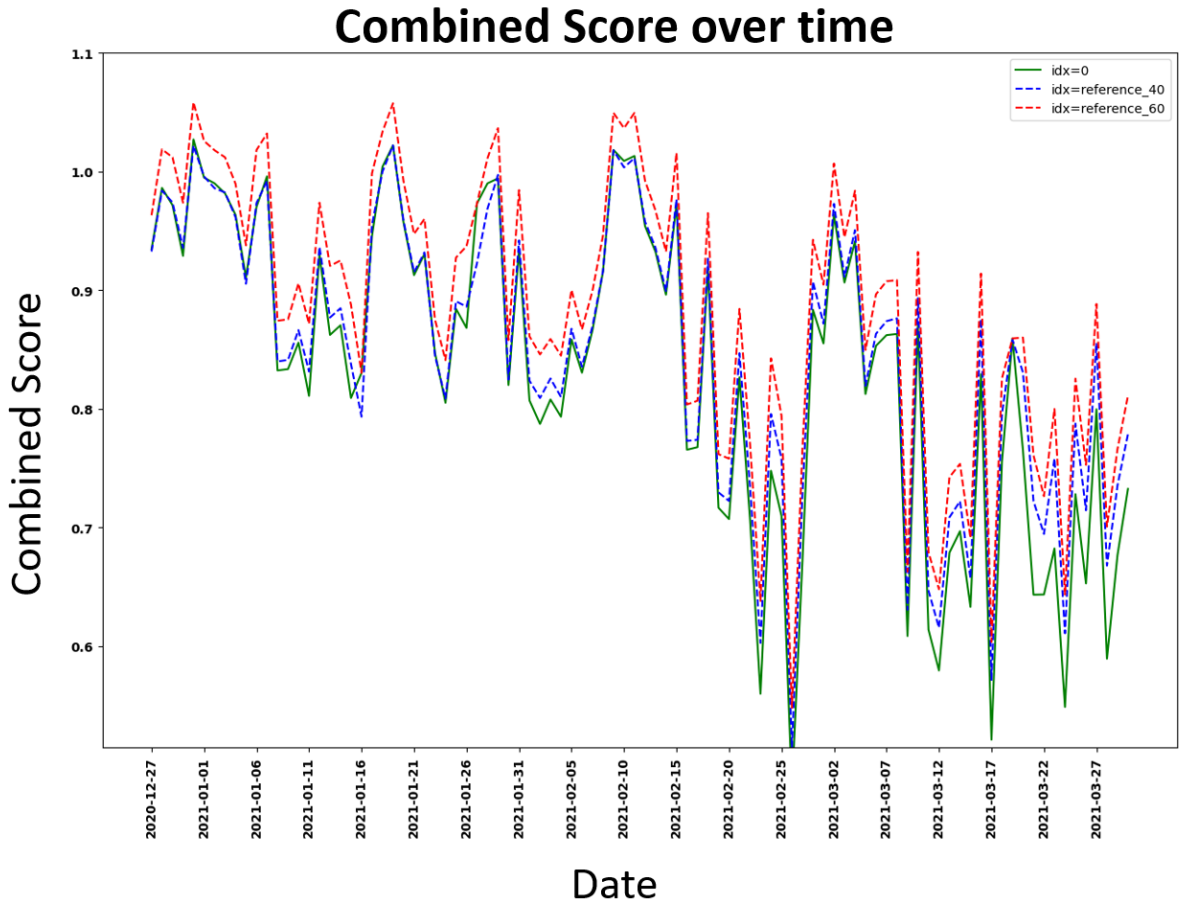


Figure 9: The results for a 3 month simulation of NEST Umar with our adapted heating curve, idx=0, and two reference heating curves fixed at an inlet temperature at 40 and 60 degrees.

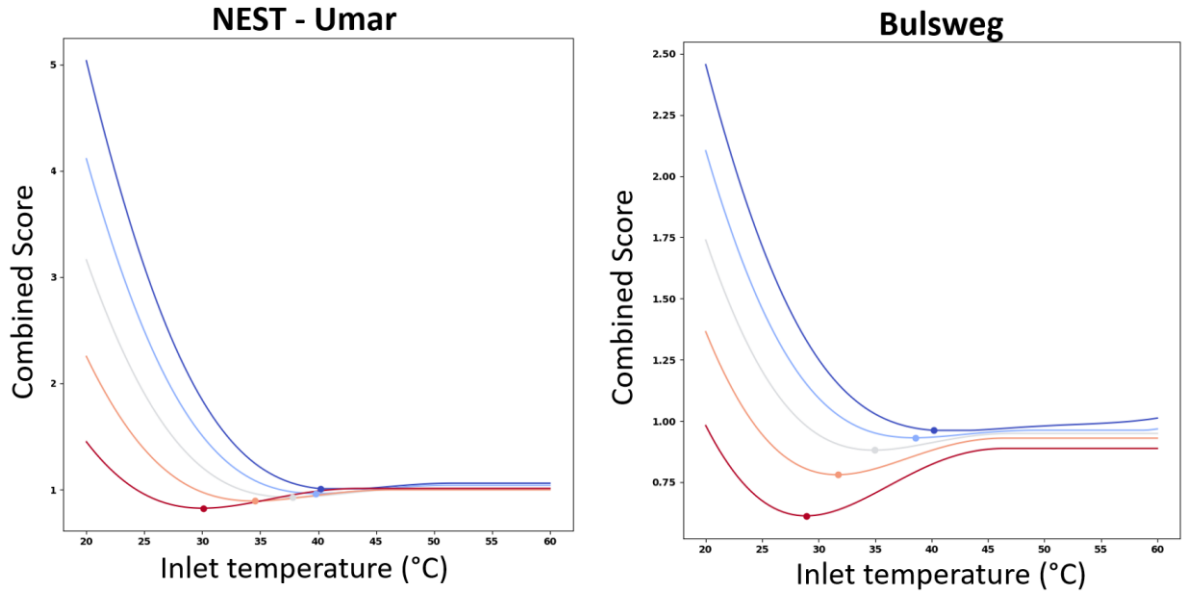


Figure 10: Shows score functions obtained by GPR. A direct comparison of the two DT for NEST resp. Bülsweg. Highlighted are 5 different graphs corresponding to the context: -2, 0, 2.5, 6.25, 10 degrees Celsius from dark blue to dark red, based on the surrogate estimation of the score per action.

3.2 Field Test / Benchmark

To assess the situation real building environment with a defined initial setting of the heating curve, we are collaborating with Lippuner AG, who have provided us with a multi-family residential building, Bülsweg, as the test site, see Appendix 8. This building consists of three 2.5-room apartments and six 3.5-room apartments. Bülsweg is connected to the district heating network of the Buchs waste incineration plant, and all units are heated through underfloor heating. Heat consumption per apartment can be regularly monitored via digital heat meters.

During the project, each apartment will also be equipped with a multifunction sensor to measure VOC concentration, indoor temperature, and relative humidity. Although a photovoltaic system is installed on the building's roof, we do not have access to its data. Instead, we rely on data from local weather stations – specifically, the Vaduz station – to determine solar irradiation. The outdoor sensor of the heating system provides information about the ambient temperature.

To assess the energy-saving potential of the adaptive heating curve derived from Bayesian Optimization, we carry out a field test under real building conditions. Figure 11 illustrates the dynamic adjustment of the heating curve of Bülsweg in response to outdoor temperature variations. The subsequent field test aims to quantify whether this adaptive strategy reduces energy consumption compared to the conventional static heating curve.

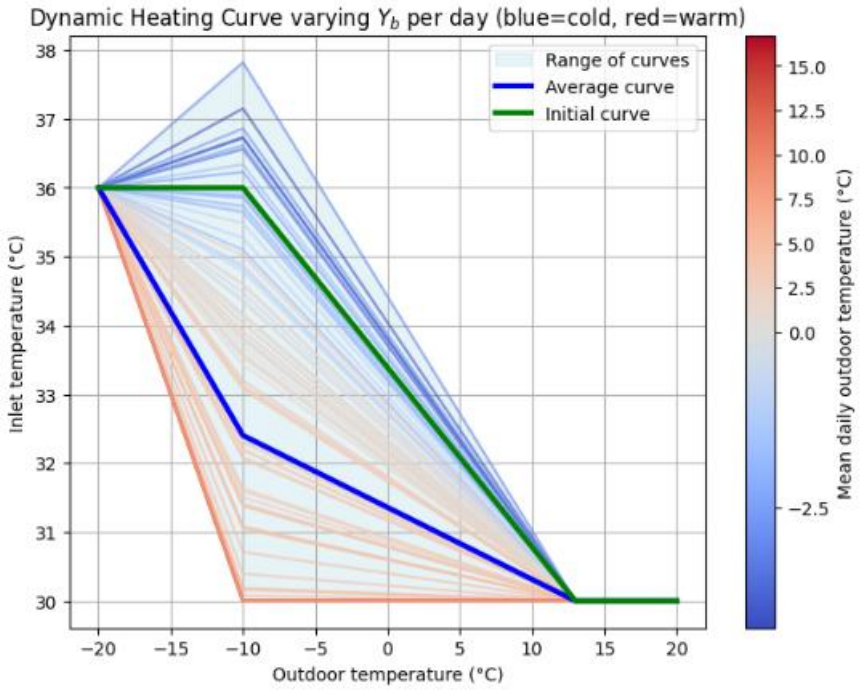


Figure 11: Dynamic heating curves generated by Bayesian Optimization. The green line shows the static baseline, the blue line the average BO-adjusted curve, and individual daily curves are colored by mean daily outdoor temperature (blue = cold, red = warm), as indicated by the colorbar. The light blue shaded area shows the full range of curves. On average, the BO curve lowers the inlet temperature to ~32.4 °C at –10 °C outdoor temperature (~31 °C at 0 °C). During November – December 2024 and some days in January 2025, ~63% of curves remain below 32 °C, while occasional upward adjustments reach up to 37.8 °C to maintain comfort. The figure illus-

trates that for warmer outside temperatures, BO tends to lower the inlet temperature, while for colder temperatures, it occasionally increases it. Overall, this adaptive adjustment suggests potential energy savings by reducing heating demand while still ensuring comfort.

3.2.1. Benchmark Analysis

Our study employs a "single-group pre-post design," meaning we conduct measurements before and after an intervention within the same group to capture changes and evaluate the intervention's effectiveness. To compare scores before and after the intervention, we aim to apply various approaches to validate performance:

- Regression analysis
- Heating degree day method
- Machine learning
- Digital twin

We assume that user behaviour will not be significantly affected by the intervention and that it is primarily influenced by the context, including outdoor temperature and sunshine hours.

Therefore, the intervention will be planned to maximize days with similar temperature and sunshine profiles. Validation of the digital twins will occur before the intervention, with the accuracy of the validation error improving with longer validation periods. A critical metric for the intervention is convergence, where the goal is to complete the training phase within 30 days, which represents an upper limit for our heating curve model. However, we anticipate a much shorter convergence period (see discussion in Section 3.1). The days required for convergence will be excluded from comparable days. As a result, a continuous intervention is preferred.

Based on the symmetrical climate chart for Vaduz (see figure 12), it makes sense to conduct the intervention in January 2025.

Our algorithm optimizes the objective function 1.0, which combines normalized heating energy and comfort score into a single, dimensionless score. To better interpret the results, we present the individual effects on energy use and comfort.

Our measurements at Bülsweg cover the heating season from November 15, 2024, to March 31, 2025. The adaptive heating curve was activated on January 23, 2025, and remained active for two months. During this period, outdoor temperatures ranged from -2 °C to 15 °C, defining the effective temperature regime for our analysis. To allow the GPR model to approximate the underlying score function over the action-context space, we required at least 10 data points. The first 10 samples for the adaptive configuration were therefore excluded from our analysis. This 10-day initialization phase included an initial 5-day safe exploration of predefined inlet temperatures [30 to 45 °C], followed by 5 days of additional sampling to support model training. Further details on the data and recording frequencies are provided in Appendix 8.

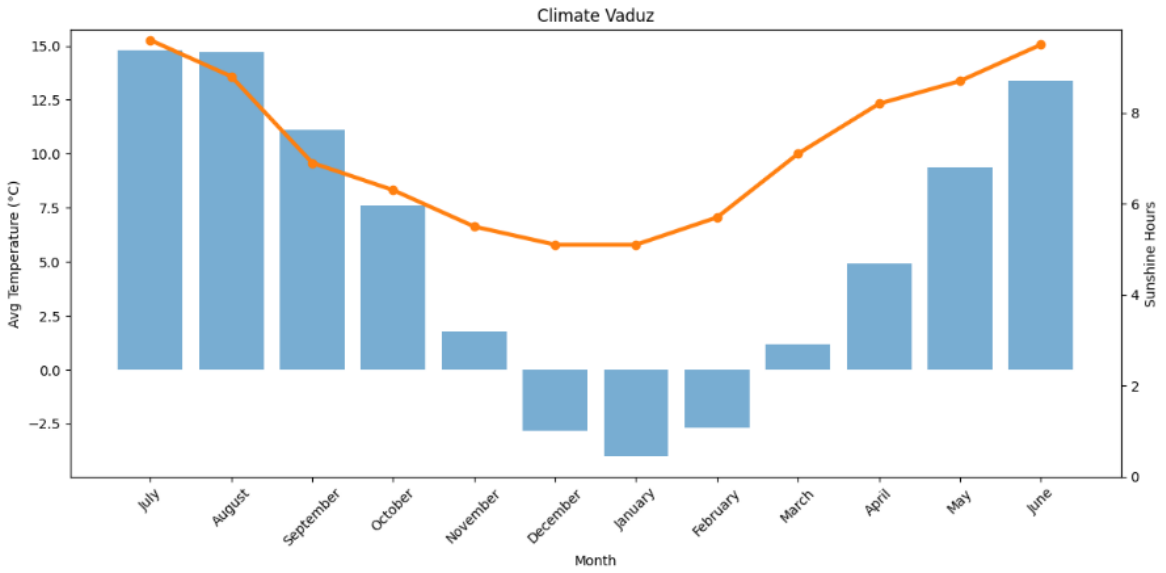


Figure 12: The image shows Vaduz's climate diagram for the climate reference period 1981−2010. It clearly illustrates the symmetry from January onward between the first and second halves of winter concerning major external influencing factors: hours of sunlight and temperature.

3.2.2. Regression Analysis

We will begin by comparing the data before and after the intervention using linear regression. Two linear curves will represent the average heating scores against the average daily temperature, pre- and post-intervention.

Before intervention:

$$y_{static} = m_{static} * x + b_{static} \tag{3.0}$$

After intervention:

$$y_{adaptive} = m_{adaptive} * x + b_{adaptive} \tag{3.1}$$

Where: x = average daily outdoor temperature and y = objective score (heating energy plus comfort score) for the observed day.

The difference between these two lines tells us how much energy we save by using the adaptive algorithm at a specific outdoor temperature x

$$\Delta_{Effect} = \Delta_{slope} * x + \Delta_b \tag{3.2}$$

$$\Delta_{slope} = m_{static} - m_{adaptive}$$

$$\Delta_b = b_{static} - b_{adaptive}$$

This provides an efficient estimation of the expected effect of the algorithm over a heating period [4].

We begin by focusing on the heating energy component: From figure 13 we can see that the static heating configuration is less efficient overall but reacts more aggressively to warmer outdoor temperatures leading to larger score drops in warm conditions. The adaptive strategy is more stable: saving more in cold conditions, possibly less in warm ones. Comparing the two regression we get:

$$y_{static} = -8.1366 * x + 201.54$$

$$y_{adaptive} = -4.1578 * x + 167.14$$

$$\Delta_{slope} = -3.9789$$

$$\Delta_b = 34.40$$

From these results we can calculate the weighted energy savings by context distribution:

$$\Delta Energy_{weighted} = \sum_{x \in context} p(x) \cdot (\Delta_{slope} \cdot x + \Delta_b)$$

$$p(x) = \frac{n_x}{N} = \text{relative weight (in \%)}$$

$$n_x = \text{number of observations per context x}$$

$$N = \text{total number of observations}$$

and obtain an estimated energy saving of 146 kWh over the temperature range during the active time of the adaptive heating system and an average weighted energy saving per degree Celsius of 11 kWh. The total estimated heating energy consumption over the temperature range from −2°C to 15°C for the static configuration amounts to approximately 2527 kWh. Based on our regression-based estimate of 146 kWh in savings, this corresponds to an energy reduction potential of about 5.7% (= $\frac{146\ \text{kWh}}{2527\ kWh}$) over the evaluated range.

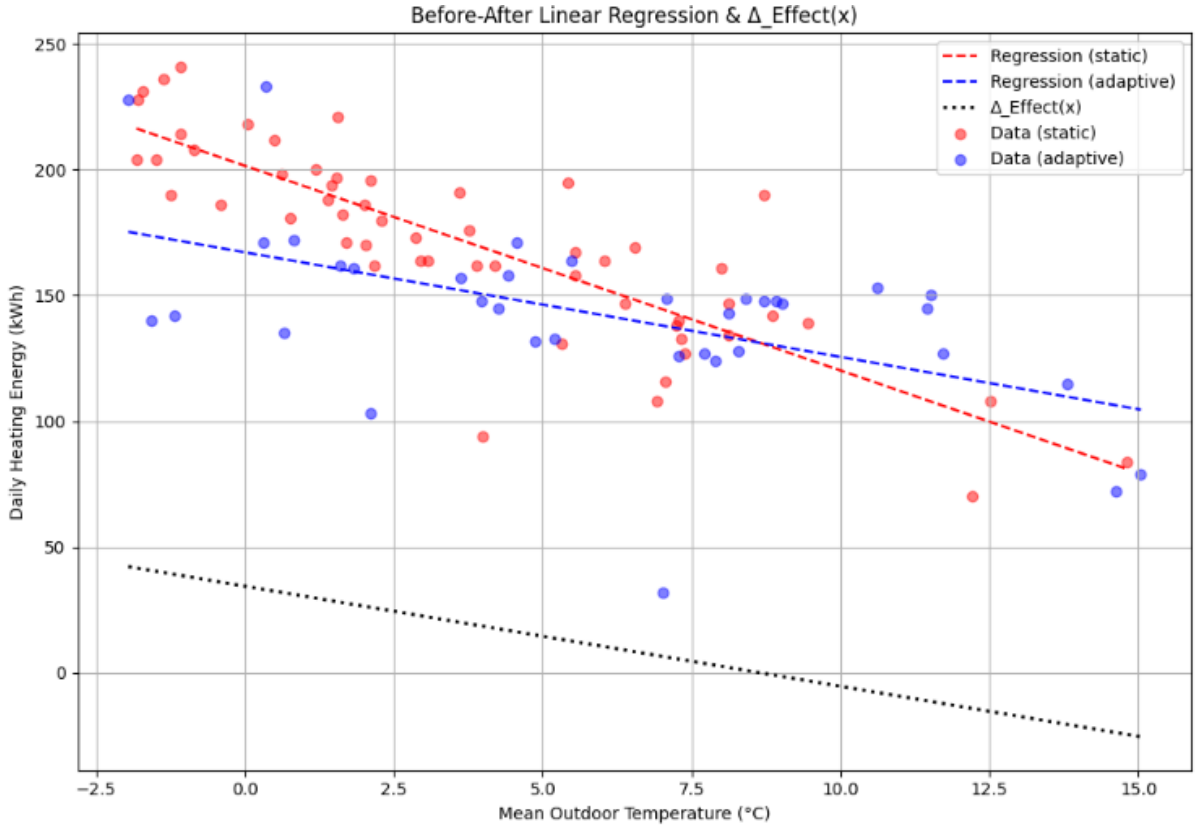


Figure 13: The scatterplot displays the analyzed data points for Bülsweg, including the fitted regression lines for the static (red), adaptive (blue), and net effect (black) configurations. To ensure a fair comparison and minimize extrapolation errors, we only included data points from the pre- and post-intervention periods that fall within the temperature range where the adaptive heating curve was active. Additionally, the first 10 samples from the adap-

tive configuration were excluded to account for the initial exploration phase of the optimization. The net effect regression suggests an estimated total energy saving of approximately 146 kWh across the explored temperature range, corresponding to about 11 kWh per degree Celsius.

3.2.3. Heating degree day method:

To estimate the net energy-saving effect of the new adaptive heating control implementation, we used a HDD normalization approach. This method corrects for differences in outdoor temperature by quantifying heating demand using the HDD metric:

$$HDD = \sum_{k=0}^{n} \max(0, \mathrm{T_{set}} - \mathrm{T_{outside}})$$

Based on this, we calculated the relative change in specific energy consumption per HDD before and after the intervention, see for more details Appendix 5.2:

$$\Delta_{Effect} = \left[1 - \frac{\frac{\mathrm{Heatingenergy_{post}}}{HDD_{post}}}{\frac{\mathrm{Heatingenergy_{pre}}}{HDD_{pre}}}\right] \cdot 100$$

$$\mathrm{Heatingenergy} = \sum_{k=0}^{n} Heatingenergy_k$$

This approach yielded an estimated energy-saving potential of approximately 4.1±1.1%.

3.2.4.Machine Learning:

We trained machine learning models on pre-intervention data to predict the heat energy consumption in kWh. Since detailed occupant information (e.g., presence, window usage, or shading behaviour) is not directly available, we infer these effects from the historical data. We assume that such behavioural patterns remain consistent over time and are unaffected by the intervention.

We used weather features as inputs to three models: a Generalized Additive Model, our BO and XGBoost. Each model was trained using data from September 2023 to March 2024 and evaluated on a hold-out test set. The XGBoost model, optimized with Optuna for hyperparameter tuning, achieved the best performance with an average RMSE of 15.35 and R^2 of 0.89 based on 18 optimization runs with R^2 values ranging from 0.884 to 0.903 and RMSE values between 14.46 and 15.83.

To quantify the algorithm's impact, we compared predicted consumption with actual post-intervention measurements. The difference estimates the net energy saving:

$$\Delta_{Effect} = Heatingenergy_k - \mathrm{Prediction_k}$$

To further assess the energy-saving potential, we performed a regression analysis on both the model predictions and the actual measurements as functions of outdoor temperature. Linear regression lines were fitted to the predicted baseline and the measured consumption data. We then integrated the area under each regression line over the temperature range from −2°C to 15°C, which represents the active operating range of the adaptive heating system. The difference between these areas provides an estimate of the total energy savings. The XGBoost model estimates an average consumption of 2418 kWh. Based on our regression analysis, the intervention led to an estimated average energy saving of 40 kWh, corresponding to an average reduction potential of approximately 1.6% (= $\frac{40\ \mathrm{kWh}}{2418\ kWh}$) over the evaluated range.

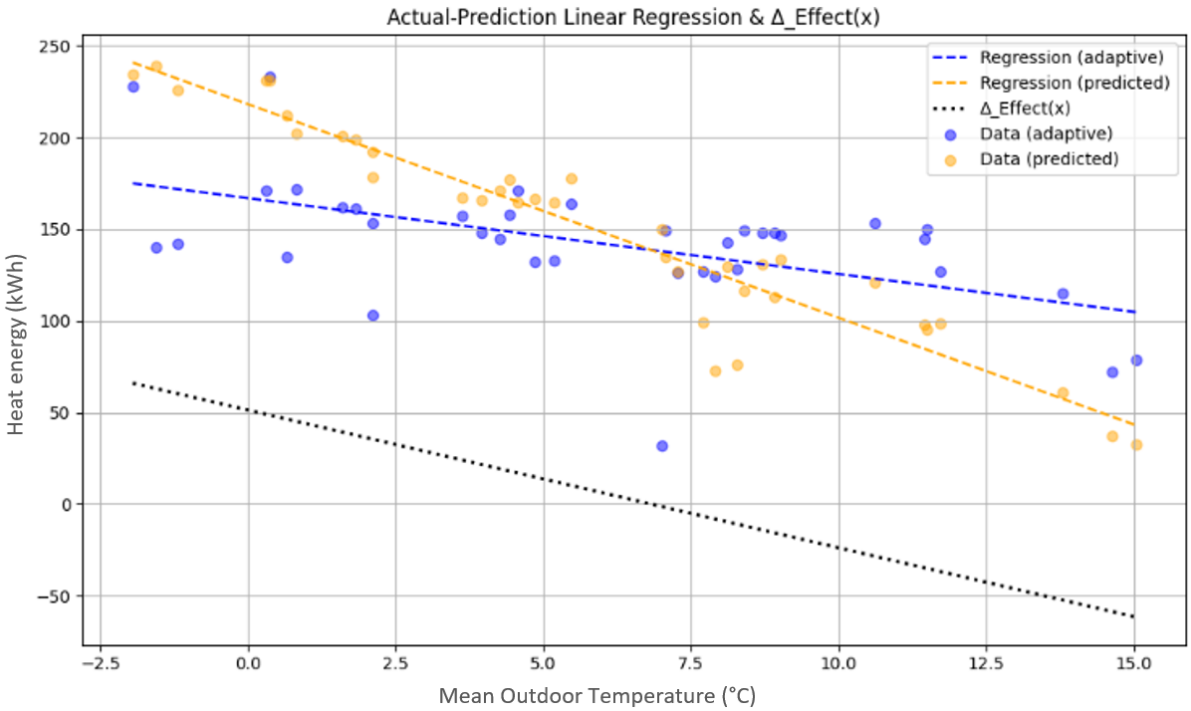


Figure 14: The scatterplot displays the actual data points for the adaptive configuration and the predicted data points for a simulated control group for Bülsweg, including the fitted regression lines: adaptive (blue), simulated (orange) and net effect (black). The first 10 samples from the adaptive configuration were excluded to account for the initial exploration phase of the optimization. The net effect regression suggests an estimated total energy saving of approximately 40 kWh across the explored temperature range.

3.2.5.Digital Twin

Initially, we aimed to use an EnergyPlus-based physical model of the building. However, discrepancies in the heating system's control behavior prevented it from accurately reproducing the real dynamics. The main issue was the reduced heat output near setpoint temperatures. Therefore, we adopted a statistical digital twin approach based on cascaded machine learning models, calibrated and validated on measured data, see figure 15.

A digital twin is considered valid if it

- responds correctly to key inputs: weather, control actions,
- predicts building behavior with sufficient accuracy, and
- remains transparent, reproducible, and documented.

The twin consists of six interconnected models representing successive physical processes, from inlet temperature to indoor comfort. Each model uses weather, control, and hydraulic features to predict its target variable. Validation on unseen data shows consistent predictive performance (R^2 = 0.70 – 0.95 across models) and on the hold-out test set (10.01. - 23.01.2025), the digital twin achieved RMSE = 2.43 kWh, MAE = 1.81 kWh, and R^2 = 0.82, outperforming the GAM baseline (R^2 = 0.69). To assess physical plausibility, we further applied sanity checks using extreme input conditions. Under the assumption of calculated zero inlet temperature, we expect a shutdown of heat transfer throughout the cascade:

$$\text{inlet temperature} = 0 \rightarrow \text{return temperatur} \sim 0 \rightarrow \text{heat energy} \sim 0$$

Within the test window (01.02. – 22.02.2025), the digital twin responded consistently with this expectation, predicting inlet temperatures around 7 °C, return temperatures around 20 °C, and a reduced total heat energy of approximately 1900 kWh. In contrast, baseline regressors collapsed toward their training mean, producing inlet ≈ 35 °C, return ≈ 30 °C, and total energy ≈ 2500 kWh.

A discussion of the full architecture, feature selection process, and detailed performance metrics are provided in Appendix G: Machine Learning-Based Digital Twin and for further details regarding the sanity checks, see H.1.3 Sanity Checks: Perturbation Experiment.

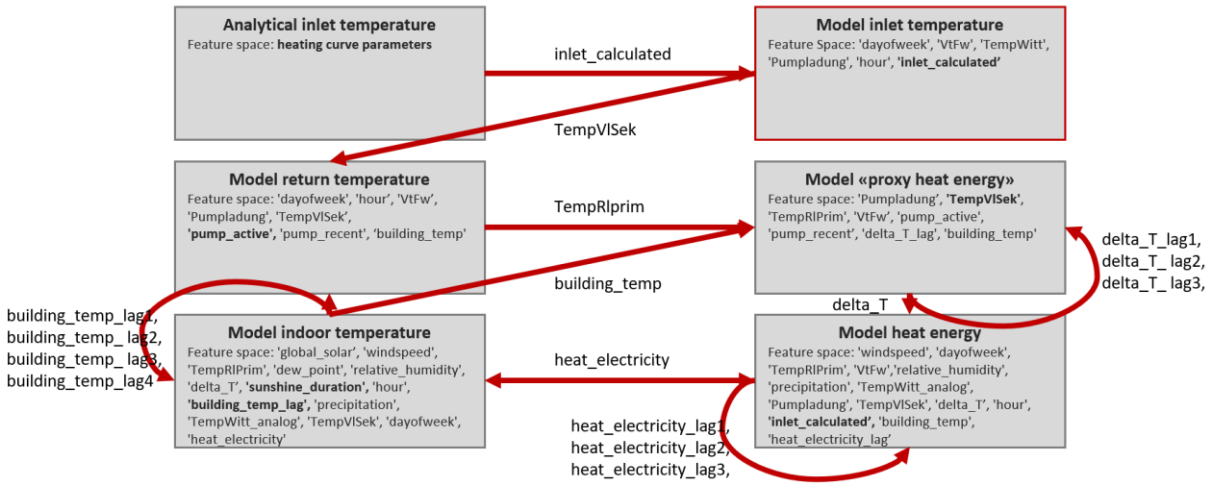


Figure 15: Architecture of cascaded models: The figure illustrates the hierarchical architecture of the data-driven digital twin, composed of six interconnected models predicting hourly temperature and energy dynamics. Each block represents one model, with its primary feature space listed below the block; the most influential input variables are highlighted in bold. The sequence follows the causal flow of the heating process: (1) the Analytical inlet temperature module computes the theoretical supply temperature (inlet_calculated) based on the heating curve parameters; (2) the Model inlet temperature refines this estimate to predict the actual measured supply temperature (TempVlSek), accounting for control noise and pump activity; (3) the Model return temperature predicts the return water temperature (TempRlPrim), reflecting the thermal energy extracted by the building; (4) the Model proxy heat energy estimates an intermediate proxy variable (ΔT = TempVlSek – TempRlPrim) as a measure of effective heat transfer; (5) the Model heat energy predicts the total thermal energy consumption (heat_electricity); and (6) the Model indoor temperature predicts the average indoor temperature (building_temp), capturing the resulting comfort response. Directed links between the models indicate the flow of predicted variables that serve as inputs for subsequent models, ensuring physical consistency and causal interpretability. For instance, TempVlSek predicted in step (2) feeds into step (3), and TempRlPrim into step (4), while building_temp also influences the proxy energy model to reflect reduced heat transfer at higher indoor temperatures. Lagged variables are used in the proxy, energy, and indoor temperature models to capture temporal dependencies.

Using two digital twins: DT1 (adaptive) and DT2 (static), we estimate the intervention's effect via a bootstrapped Difference-in-Differences approach that accounts for sampling and model uncertainty with respect to our testbed. For a detailed description incl. discussion of accounting for model uncertainty / bias of the digital twin, see Appendix H: Pseudocode of the bootstrapped Difference-in-Differences algorithm.

- **DT1:** Synchronized with the real building and **adapted** with the algorithm after intervention
- **DT2:** Operates without intervention, serving as a control / **static** baseline
- **TB:** measurements of real building Bülsweg
- **pre** (before intervention) and **after** (after intervention).

Accounting for sampling variability: We build a sampling distribution for the difference-in-differences between the two digital twins while injecting model uncertainty and bias estimates. Specifically, DT1 and DT2 are used to predict the hourly heat energy consumption of Bülsweg for the pre (15.11.24 – 22.01.25) and after (02.02. – 17.03.25) periods. These hourly predictions are then aggregated into daily energy usage values, using 5 pm–to–5 pm intervals. The resulting arrays of daily energy values are resampled via a bootstrap procedure to generate a distribution of DID estimates.

$\overline{DT1}_{pre}$ = the mean of daily DT1 predictions in the bootstrap generated array (bootstrap sample)

$\overline{DT2}_{pre}$ = the mean of daily DT2 predictions in the bootstrap generated array (bootstrap sample)

$\overline{DT1}_{after}, \overline{DT2}_{after}$ similarly for after-period

One bootstrap difference-in-difference draw, denoted by (b), is:

$$DID^{(b)} = \left(\overline{DT2}_{after}{}^{(b)} - \overline{DT1}_{after}{}^{(b)}\right) - \left(\overline{DT2}_{pre}{}^{(b)} - \overline{DT1}_{pre}{}^{(b)}\right)$$

And is generated 5000 times. The mean effect is:

$$\overline{DID} = \frac{1}{B}\sum_{b=1}^{B} DID^{(b)}$$

and reported together with the percentile-based confidence interval studied for two cases, Case 1 and Case 2 discussed below.

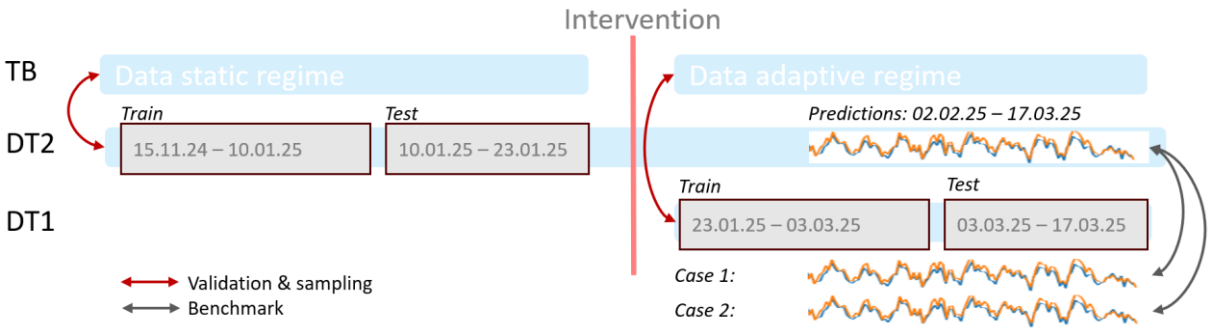


Figure 16: Overview of the DT1–DT2 comparison setup. DT1 and DT2 were trained on distinct operating regimes to represent the adaptive and static heating behaviors of the testbed Bülsweg. DT2 serves as the control digital twin, trained on data collected between November 2024 and January 2025, while DT1 was trained on data from January to March 2025 to reflect the adaptive configuration. A separate hold-out test set was used for validation. Two comparison cases are analyzed: Case 1: DT1 synchronized with the real testbed operation. Case 2: DT1 operating with fully converged Bayesian Optimization parameters and reduced lower bounds (down to 20 °C). The evaluation horizon for both digital twins covers the period 01.02.2025–15.03.2025. The red arrows indicate relationships between a digital twin and the testbed, representing validation (via R^2 and error metrics) and sampling (via mean offset and standard deviation of predictions). The dark grey arrows indicate benchmarking relationships between the two digital twins (DT1 ↔ DT2), comparing their heat energy predictions.

The analysis shows:

- **Case 1** (direct comparison with experiment curve): mean DID = –11.3 kWh/day, 95% CI = (–39.7, +17.5)
- **Case 2** (extrapolated curve lower bound, more flexible dynamic curve): mean DID = –3.7 kWh/day, 95% CI = (–30.9, +24.1)

Interpretation: Statistically, we cannot conclude that our strategy leads to significant energy savings, as the confidence interval overlaps with zero.

In Case 1, we used DT2 to predict the potential daily heat energy consumption for the period 02.02. – 17.03.25, i.e., after the calibration phase of our algorithm and in synchronization with the applied inlet temperatures for the real building (see Section 3.2.6.1 Results Bülsweg: Convergence & Robustness). Considering the known limitations discussed in Section 3.2.6 Discussion Energy Net Savings and illustrated in Figure 17, the model indicates an average loss of 11 kWh/day, with possible outcomes ranging from a loss of 40 kWh/day (worst case) to a gain of 18 kWh/day (best case).

In Case 2, we compared DT2 and DT1, where DT1 was reparametrized to match the static heating curve configuration (see table in Section 3.2.6 Discussion Energy Net Savings). The parameters $(\mathbf{X_3}; \mathbf{Y_3})$ and $(\mathbf{X_4}; \mathbf{Y_4})$ were adjusted such that the curve could reach temperatures as low as 20 °C, depending on outdoor conditions. Furthermore, the converged heating tuner predictions were used to estimate the

optimal $\mathbf{Y_2}$. Under this setup, the results show near parity between the static and adaptive configurations, indicating that the new parameterization effectively mitigates most of the previous energy penalty.

Appendix I provides the full pseudocode for evaluating heat energy savings using DT1 and DT2, including the implemented uncertainty treatment.

3.2.6.Discussion Energy Net Savings

Method	Energy Net Savings
Linear Regression	By comparing the average energy consumption before and after the intervention and fitting a linear regression to sampled data points, we estimated a **net saving of 5.7%**.
HDD	Using the HDD normalization technique, which adjusts for weather variability, we calculated a **net saving of 4.1%.** This method is robust for such comparisons and likely provides the most conservative and reliable estimate.
Machine Learning	We trained an ML model to predict energy consumption based on weather and room conditions, using it as a virtual control group. The model achieved an average **net saving of 1.6%.** This approach to estimate
Digital Twin DID	**No statistically significant effect** was observed (confidence intervals: –39.7 to +17.5 and **–30.9 to +24.1**). Based on the manual alignment with the static heating curve in **Case 2**, we had expected comparable performance under mild temperature conditions and reduced energy use during colder periods. However, with the current model, this hypothesis cannot be confirmed.

While linear regression, HDD and Machine Learning point to a positive impact of the algorithm, the HDD-based estimate is likely the most reliable due to its ability to normalize for weather influences without being overly sensitive to model assumptions. The ML model supports this finding, validating the presence of a small but measurable effect.

The most plausible explanation for the discrepancy between the Digital Twin results and the other evaluation methods lies in the training characteristics and extrapolation limitations of the machine learning models used to construct the digital twin. These models are optimized to reproduce observed system behavior within the range of their training data, meaning their boundaries and response functions are tuned to minimize error for known patterns. As a result, they perform well within the data domain but struggle to extrapolate to unseen parameter configurations, such as heating curve settings that were not represented during training (see Figure 18).

In our case, DT1 was trained predominantly on data corresponding to a heating curve parameterization that did not include values below 20 °C, which reflects the system's historical operating range. Consequently, DT1 lacks exposure to configurations representing milder outdoor conditions and lower inlet temperatures, making it difficult to accurately predict potential efficiency gains in those regimes.

Additionally, the Bayesian Optimizer was tuned within a restricted temperature range (30 – 45 °C), further limiting exploration of potentially more energy-efficient configurations below 30 °C. Consequently, while real-world measurements confirm that the adaptive strategy yields modest energy savings (~5 %), the current statistical digital twin framework cannot yet reproduce this outcome consistently due to model asymmetry and extrapolation bias (DT1, BO). This asymmetry, manifested as unequal sensitivity to lower versus higher inlet temperatures and corresponding changes in heating energy, is further discussed in Section H.1.3, Sanity Checks: Perturbation Experiment.

We believe that the actual saving potential is positive and probably higher than the estimated savings estimated for the HDD and Linear Regression method. This is mainly because our current heating curve configuration imposes a lower limit of 30°C for the inlet temperature, even during periods of mild outdoor conditions. In contrast, the previous static heating curve configuration allowed the inlet temperature to drop to as low as 20°C, resulting in lower energy consumption.

We observe inefficiencies in the high temperature range. These inefficiencies are believed to be caused due to the configuration of our heating curve in comparison with the old configuration. Our configuration didn't allow to go below 30 degrees, while the old configuration allowed to go down till 20 degrees, see table below. The table below highlights this difference in configuration.

Point	Old Setting (Temperature °C)	Our Configuration (Temperature °C)
$(\mathbf{X_1}; \mathbf{Y_1})$	(-10; 36)	(-20; 36)
$(\mathbf{X_2}; \mathbf{Y_2})$	(0; 32)	(-10; **variable**)
$(\mathbf{X_3}; \mathbf{Y_3})$	(10; 28)	(10; 30)
$(\mathbf{X_4}; \mathbf{Y_4})$	(20; 20)	(20; 30)

A rough estimation of incurred energy losses can be made by comparing the theoretical inlet temperatures from the adaptive and static heating curves under mild conditions, as shown in Figure 12. We looked at a warm day with an average outdoor temperature of around 13 °C. The area under the inlet temperature curves, which is proportional to heating energy, suggests a loss of approximately 16% with respect to the adaptive configuration, see figure 17.

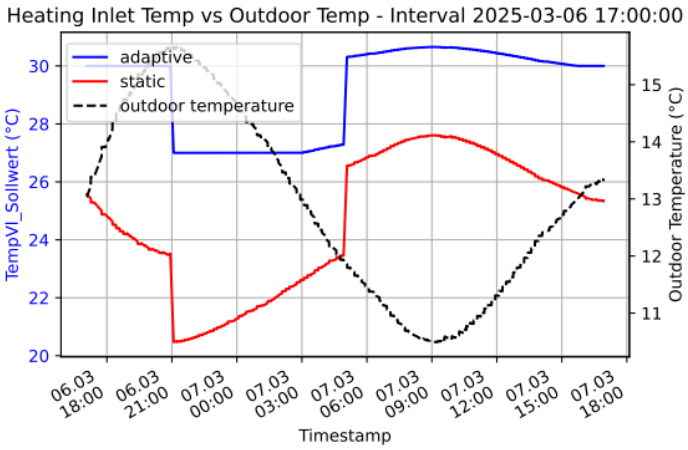


Figure 17: We plotted the theoretical inlet temperatures based on outdoor temperature and heating curve parameterizations. While the adaptive curve (TempVl_Sollwert) can be directly retrieved from the heating system, the inlet temperatures for the static configuration were computed for comparison, red curve. The difference in the area under both curves, proportional to heating energy, amounts to approximately 16% (= $\frac{area_{adaptive}}{area_{static}}$).

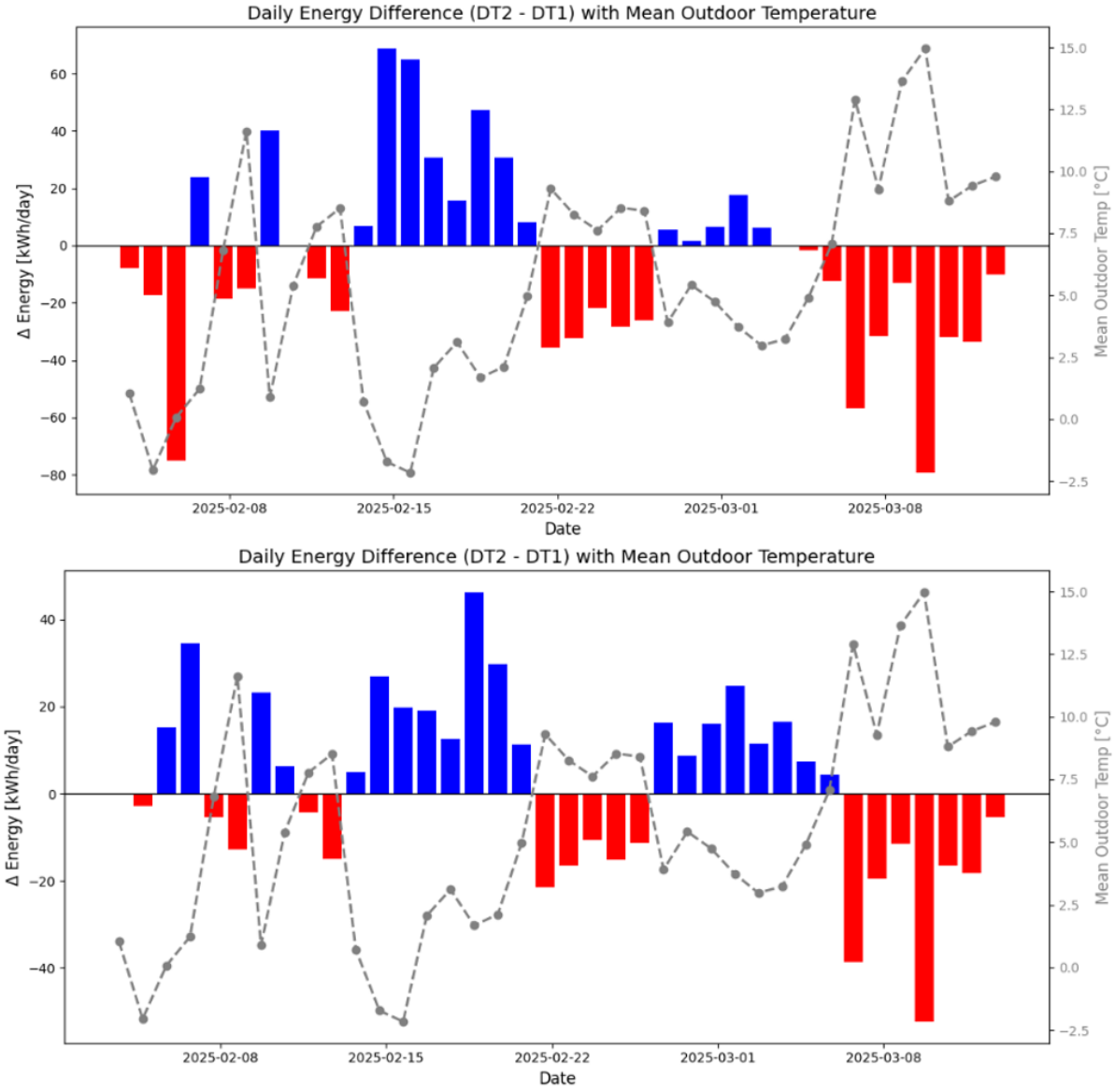


Figure 18: Daily differences in predicted heat energy consumption between the static (DT2) and adaptive (DT1) digital twins. Positive (blue) bars indicate energy savings of the adaptive strategy, while negative (red) bars represent increased energy use relative to DT1. The gray dashed line shows the mean outdoor temperature for the corresponding days. A consistent pattern emerges across both scenarios: at mild outdoor temperatures, the adaptive control tends to consume more energy than the static configuration, while at lower temperatures, it achieves higher efficiency and energy savings. The upper panel corresponds to **Case 1**, where DT1 is synchronized with the real testbed, and the lower panel to **Case 2**, where the lower bound of the heating curve was reduced to 20 °C and the predictions were derived from a fully converged Bayesian Optimization configuration.

Comfort: static comfort score 0.24 vs. adaptive comfort score of 0.23

We now turn our focus to thermal comfort. To assess it, we estimated the daily comfort score for both the static and adaptive heating configurations using equation (1.1). This calculation was based on indoor temperature measurements from the living spaces in each apartment. Due to irregular and sometimes missing data, we included only those apartments with at least one valid reading during both day and night periods. For each qualifying sensor reading, we calculated the deviation from the comfort thresholds: 23 °C during the day and 20 °C at night. These deviations were averaged per apartment and then aggregated across all apartments to estimate the overall building-level comfort.

The resulting comfort score values were binned for visualization, see figure 19, and we fitted a Gaussian distribution to the histograms for both heating configurations. Based on the average comfort score, we find that comfort levels remained nearly unchanged after switching from the static to the adaptive heating curve, demonstrating that thermal comfort for residents is maintained.

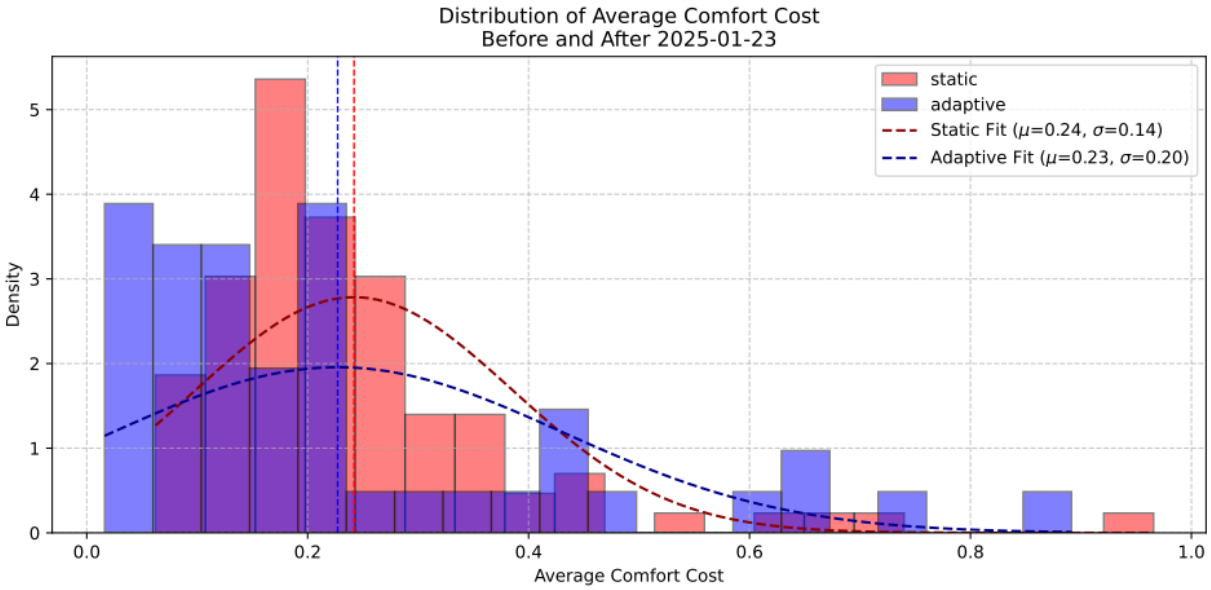


Figure 19: The histograms show the distribution of daily comfort score per apartment for both heating configurations. A Gaussian distribution is fitted to each histogram to highlight the mean and spread of comfort scores. The comparison shows that the average comfort score remains nearly unchanged between the static and adaptive configurations, indicating that occupant thermal comfort is maintained following the implementation of the adaptive heating curve.

3.2.6.1. Results Bülsweg: Convergence & Robustness

We perform an empirical convergence analysis in the setting where the true objective function is unknown. Specifically, we use a local convergence test that evaluates the stability of the model's predictions within discretized context–action squares. The analysis is based on GPR models trained on incrementally larger datasets.

Experimental Setup

- We define a grid consisting of 54 context – action squares, determined by:
    - Context values: [-10, -7.5, -5, -2.5, 0, 2.5, 5, 7.5, 10, 12.5]
    - Action values: [31, 34, 37, 40, 43, 46]
- For each square, we train a new model with progressively increasing data sizes from 10 till 60 by additional 5 datapoints: [10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60].
- At each stage, we predict the mean and standard deviation at the center point of each square and observe how predictions evolve with more data.

Convergence after 40 datapoints

At low data densities (e.g., 10 data points for 56 squares → ~0.18 points/square), the model's prediction surface is relatively flat and only sensitive in areas with lower actions. As more data is added, particularly in underrepresented regions (e.g., context 0, action 46), significant shifts in predicted means occur.

We visualize this progression by inspecting prediction grids over time. The shift in the prediction values becomes apparent when comparing early stages (with almost constant combined score values) to later stages (where the combined scores exhibits structure and local extrema).

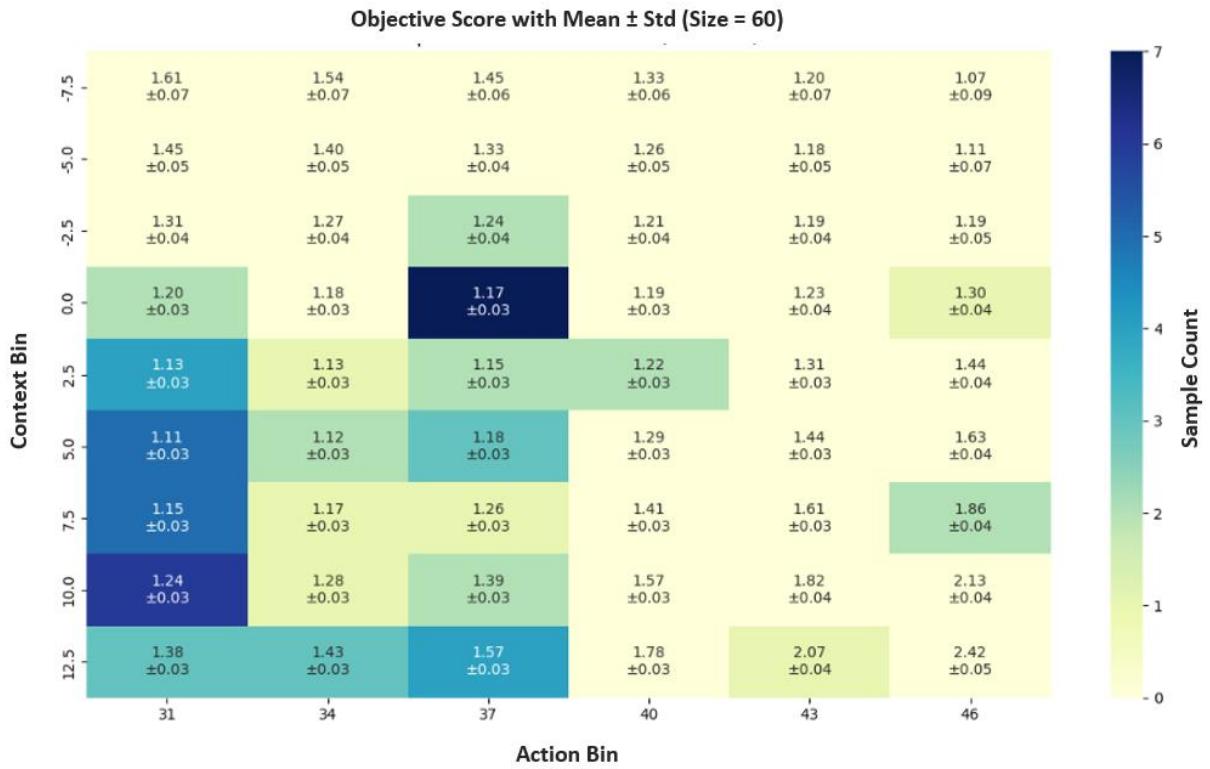
We observe that after approximately 30 – 40 datapoints for our setting, the predictions stabilizes for the already explored areas and there is no significant improvement, suggesting that the model has converged to a robust heating policy.

To understand the model's exploration behavior over time, we visualize the frequency of predictions across a discretized 6×9 context – action grid. The bar on the right represents how often each square in the grid was selected by the model, the values inside each square are the predictions for the mean combined score and it's standard deviation at three stages: early (10 datapoints), intermediate (40 datapoints), and late (60 datapoints). For example, the square at Action Bin 37 and Context Bin 0.0 contains 3 values, with a predicted mean combined score of 1.19 and a standard deviation of ±0.04. Even when standard deviation estimates are low in early phases, the mean prediction may still vary significantly, suggesting that small uncertainty does not guarantee convergence in low-data regimes.

Early Stage (10 datapoints): uniform values across squares.

Middle Stage (40 datapoints): emerging variation, e.g., square at (46, 0) shows a notable change.

Final Stage (60 datapoints): smoother, more structured values indicating a more refined model.



The increasing density and spread of samples across the grid demonstrate progressive exploration followed by exploitation. Notably, high-utility regions (e.g., column 2 across several rows) are consistently reinforced, while seemingly low-performing regions remain unvisited. This transition is controlled by adjusting the relative weights of exploration and exploitation of our BO implementation. As more data points are added, the relative importance of exploitation increases, while exploration decreases, speeding up

convergence. This dynamic weighting is a distinctive feature of our implementation, differing from standard BO.

To quantify convergence, we compute the average prediction change across all squares of this grid between consecutive training steps. The results can be looked up in figure 20 and the visualization reveals:

- High instability in early iterations (especially step 1 and 2, from 10 to 15 and 15 to 20 data points).
- A stark reduction in mean prediction variation after 30 – 40 datapoints, indicating convergence behavior.
- Even when standard deviation estimates are low in early phases, the mean prediction may still vary significantly, suggesting that small uncertainty does not guarantee convergence in low-data regimes.

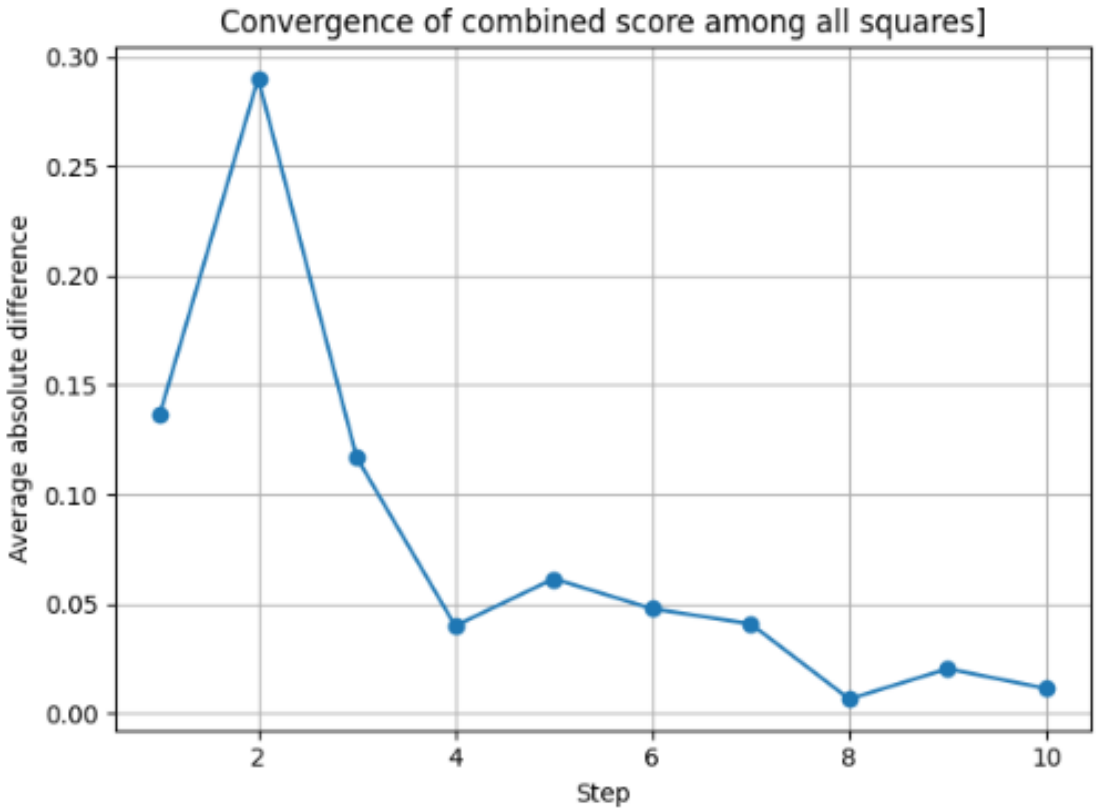


Figure 20: Plotted are the average differences between the predictions made for different steps: Step 1 stands for the average difference between 10 to 15 datapoints, step 2 stands for the average difference between 15 to 20 datapoints, step 3 stands for the difference between 20 to 25 and so forth. The average prediction change reveal: high instability in early iterations (especially from 10 to 15, and 15 to 20 data points). A stark reduction in mean prediction variation after 30 – 40 datapoints, indicating convergence behavior.

Robustness ok for positive outliers bad for negative outliers

To evaluate the robustness of our model when optimizing the heating curve, we manipulated existing samples and added them to the data used to train our model. Specifically, we randomly selected 8 distinct indices from our original dataset, namely: [3, 15, 24, 31, 37, 41, 50, 57] and multiplied their objective score by a distortion factor. The specified samples were distorted by a factor of 3 and replaced the old value. A factor of 3 corresponds approximately to a standard deviation of 3 with respect to our assumed normally distributed data, which classifies them as strong positive outliers. After each replacement, the model was retrained, and its prediction was evaluated across a fixed set of contexts ([-10, -7.5, -5, -2.5, 0, 2.5, 5, 7.5, 10] °C) and actions (inlet temperatures ranging from 28 °C to 45 °C in 0.1 °C

steps), see figure 21. We did the same for a factor of 0.1 to obtain negative outliers and mixed between factor of 3 and 0.1.

The tables below summarize the model's predicted optimal inlet temperature per context as the number of manipulated samples increases (Case 0: no distortion, Case 6: 6 distortions added):

Factor 3:

Context (°C)	Case 0	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6
-10	45	45	45	43.11	43.11	45	45
-5.0	43.11	45	39.33	39.33	39.33	39.33	41.22
-2.5	37.44	41.22	35.56	35.56	35.56	37.44	39.33
0.0	33.67	33.67	33.67	33.67	33.67	35.56	37.44
2.5	29.89	28	31.78	31.78	31.78	33.67	35.56
5.0	28	28	29.89	29.89	31.78	31.78	31.78
10	28	28	28	28	29.89	29.89	28

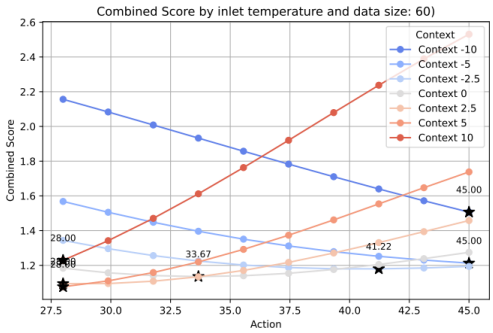
Factor 0.1 and Mixed:

Context (°C)	Case 0	Case 1, 0.1	Case 4, 0.1	Case 6, 0.1	Case 6, Mix
-10	45	45	45	45	45
-5.0	43.11	45	45	28	45
-2.5	37.44	35.56	45	28	45
0.0	33.67	31.78	45	28	45
2.5	29.89	31.78	45	28	45
5.0	28	31.78	45	28	45
10	28	31.78	29.89	45	28

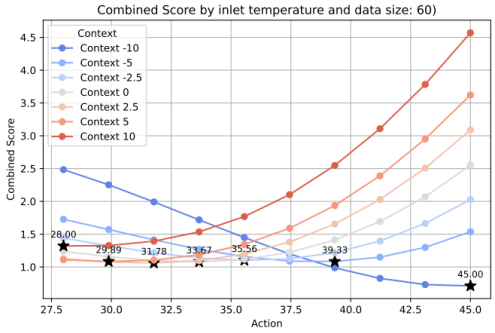
The results reveal a critical asymmetry in the system's robustness:

- Factor 3: seem to add noise, but they do not remove the signal, the model could still reliably converge toward a reasonable solution.
- Factor 0.1: seem to destroy the signal, if we reduce the rewards to near zero for some samples, the model predictions flatten, the model believes that almost no actions yield to any optimal solution. After enough negative distortions have been added the model starts to make bad decisions. Concrete example for a bad decision is observed in Case 6 (factor 0.1) where at 10°C, the system wrongly predicts the maximum action of 45°C when it should recommend much lower inlet temperatures based on the undistorted case of 28°C.
- Mixed: seem that the negative effects dominated. Although half of the distortions were positive, the system still suffered from bad decisions, reinforcing the finding that underestimations are more harmful than overestimations.

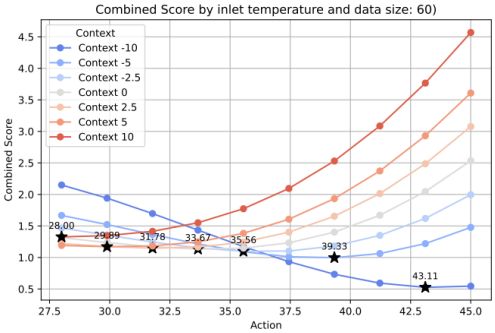
Case 1:



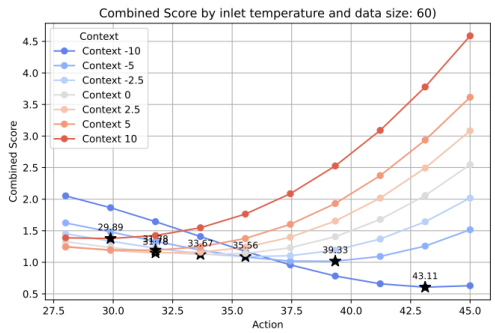
Case 2:



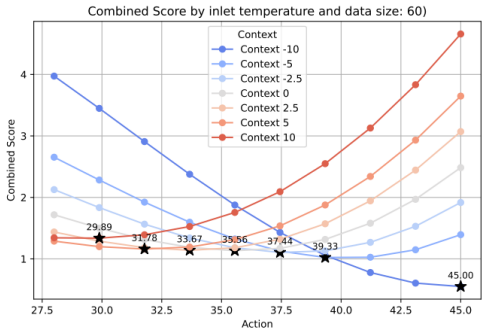
Case 3:



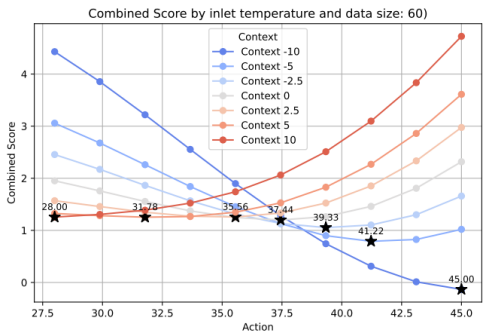
Case 4:



Case 5:



Case 6:



Case 0:

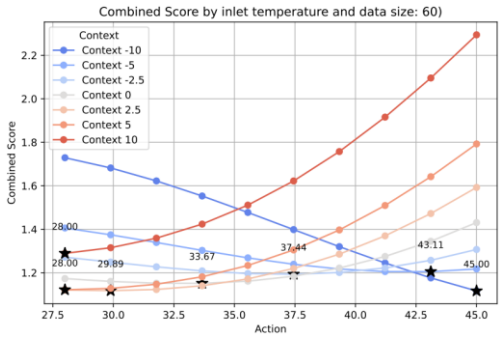


Figure 21: Model predictions for fixed contexts across the full range of inlet temperatures (actions) are shown for varying numbers of distorted data points, from case 0 to case 6. Notably, the curve progression in case 1 stands out from the others, which exhibit more consistent and similar behavior.

4 Summary and Conclusions

4.1 Energy Performance and Savings Potential

To quantify the energy-saving potential of the adaptive heating curve compared to the static baseline, we applied four complementary evaluation methods: Linear Regression, Heating Degree Day (HDD) normalization, Machine Learning prediction, and a Digital Twin difference-in-differences (DID) approach.

- **Linear Regression** suggested a net energy saving of 5.7%, comparing average energy consumption before and after the intervention.
- **HDD normalization**, which corrects for weather variability, yielded a slightly lower and robust estimate of 4.1%.
- **Machine Learning–based virtual control** indicated an average saving of 1.6%, supporting the existence of a measurable but smaller positive effect.
- **Digital Twin DID**, by contrast, did not reveal a statistically significant result, Case 1[1]: 95% CI: –39.7 to +17.5 and Case 2[2]: –30.9 to +24.1.

The HDD-based estimate is likely the most reliable, as it normalizes for weather while remaining robust to model assumptions. By contrast, the Digital Twin shows discrepancies, likely due to its limited ability to extrapolate beyond the training data and the restricted BO optimization range (30 – 45 °C). While it accurately captures system dynamics within observed conditions, it struggles to predict performance for untested heating curve configuration

Despite these modeling limitations, the real-world field data confirm a modest yet meaningful energy-saving potential of approximately 4 – 6%. Importantly, this result is likely conservative: the adaptive configuration enforced a minimum inlet temperature of 30 °C, whereas the previous static curve allowed temperatures to drop as low as 20 °C. This design constraint likely reduced the achievable energy savings.

Furthermore, analyses of high-temperature inefficiencies (see Figure 17) suggest that the current configuration results in up to 16% additional energy use under mild conditions, underscoring the importance of extending the optimization range downward. Future experiments should therefore include a broader inlet temperature range (20 – 45 °C) to unlock further energy-saving potential.

4.2 Thermal Comfort

Thermal comfort was evaluated through a daily comfort score based on indoor temperature deviations from 23 °C (day) and 20 °C (night). Results show that comfort levels remained nearly unchanged after implementing the adaptive heating curve:

- Static configuration: **0.24**
- Adaptive configuration: **0.23**

This near-equivalence demonstrates that the adaptive control strategy maintains the occupant comfort.

[1] Case 1: The static baseline model (DT2) – representing the original Bülsweg building without the adaptive heating curve tuner — was used to predict daily heat demand (02.02–17.03.25) after calibration, aligned with the real building's inlet temperatures.

[2] Case 2: The adapted digital twin (DT1 – synchronized with the real building and tuned after intervention) was compared with the static baseline (DT2). DT1 was reparametrized to match the static heating curve, adjusting (X_3; Y_3) and (X_4; Y_4) to allow supply temperatures down to 20 °C and using the converged tuner to estimate the optimal Y_2.

4.3 Convergence and Robustness of the Learning Model

The Bayesian Optimization model demonstrated stable convergence after approximately 30 – 40 datapoints, beyond which additional data yielded no substantial improvement in predicted outcomes. This indicates a robust learning process within the explored context–action space.

However, robustness tests revealed an asymmetry in sensitivity:

- The model tolerated strong positive outliers (factor 3 distortion) without losing predictive accuracy.
- Conversely, negative outliers (factor 0.1 distortion) severely degraded performance, flattening predictions and leading to incorrect recommendations (e.g., excessive inlet temperatures at mild outdoor conditions).

This behavior suggests that underestimating rewards (negative bias) is more harmful than overestimation, emphasizing the need for regularization and noise handling strategies in future iterations of the BO framework.

4.5 Overall Conclusion

Overall, the adaptive heating curve guided by Bayesian Optimization demonstrates promising energy savings of approximately 4 – 6%, while maintaining thermal comfort. The results confirm that real-world implementation is feasible and potentially beneficial. With demonstrated convergence, manageable robustness limitations, and strong industrial interest, the approach shows considerable potential for scaling toward practical deployment in smart heating systems.

5 Outlook

Our industrial partner Co4 / Lippuner AG expressed strong interest in continuing collaboration and further development of the adaptive control strategy. In their words:

“We have great interest in continuing this work. Our goal is to provide customers with a tool that not only monitors heating systems but actively optimizes their operation. Energy optimization is key to improving system performance, longevity, and ultimately reducing energy use and costs. Savings of around 5% are already significant. For future iterations, integrating predictive weather forecasting would be an exciting addition to further enhance performance.”

This industry perspective underlines the practical relevance and commercial potential of the adaptive heating control system.

The next development phase of the AHA project focuses on scalability, automation, analysis of the maximum potential of our approach, and methodological improvements through additional contextual information. The following key steps are planned:

- **Automation:** Currently, several steps in data acquisition, preprocessing, optimization, validation, and parameter updates still require manual execution. For large-scale deployment across multiple buildings, these processes will be automated through continuous data integration and retraining pipelines.
- **Maximum potential analysis:** The current implementation prioritizes interpretability and data efficiency. Using simulation studies, we will explore the full parameter space of the heating curve to estimate its theoretical optimization potential and compare it with our current lightweight implementation.
- **Additional field tests:** Beyond the Sprint test (Winter 2025/26), further validation in collaboration with Lippuner is planned to test transferability and robustness. These follow-up activities will be carried out outside the scope of the BFE-funded project and will be financed through internal resources.

- **Methodological improvements**: Future work will focus on enhancing the adaptiveness and data-driven tuning of key hyperparameters. This includes incorporating additional contextual factors, such as wind and solar gains, directly into the Gaussian Process Regression model; evaluating whether the fixed 24-hour update cycle is optimal or if adaptive cycles based on system dynamics improve performance; refining the exploration–exploitation balance, including step sizes in the acquisition function, to increase convergence speed and stability; and systematically optimizing Gaussian Process kernel functions and the relative weight factors of the normalized energy score and comfort score rather than defining them statically (currently both are 1), allowing the model to adapt to the observed data and building-specific characteristics.
- **Robustness:** Following our robustness analysis, we will implement filtering and robust optimization techniques (e.g., reward clipping, outlier detection) to ensure stable and reliable performance under imperfect data conditions.

6 National and international cooperation

Lippuner AG brings decades of expertise in energy and building technologies, with a strong commitment to advancing these fields through innovation. They provide the necessary infrastructure and practical experience to integrate and monitor our adaptive Heating Tuner within their systems. Additionally, they facilitate testing our algorithm on the Bülsweg building, which they operate, while sharing essential data and proactively addressing sensor failures.

7 Data management plan and open access/data/model strategy

In alignment with our commitment to open access and the requirements for project co-funding by the Swiss Federal Office of Energy (SFOE), we have developed a comprehensive data management plan. Our approach includes making parts of the algorithm's code, datasets, and digital twin models publicly accessible through the Renku platform [6].

By early next year, we plan to release the developed Python interface for direct interaction with EnergyPlus models. This interface, designed for seamless integration and ease of use, will be made available independently of ongoing collaboration with Lippuner AG, enabling broad usability and accessibility. The current development version can be accessed here: Files · development · Alessandro Tell / AHA · GitLab (renkulab.io) [6].

Furthermore, insights, results, and developed programs from the project will be published on the ARAMIS platform to ensure transparency and knowledge sharing within the community. This effort underlines our dedication to fostering open research and reproducibility while supporting innovation in energy efficiency and digital twin technologies.

8 References

ID	Reference description
1	Krause, Andreas and Ong, Cheng Soon (2011). Contextual Gaussian process bandit optimization, https://dl.acm.org/doi/10.5555/2986459.2986732.
2	Sui, Yanan and Gotovos, Alkis and Burdick, Joel and Krause, Andreas (2015). Safe Exploration for Optimization with Gaussian Processes, https://proceedings.mlr.press/v37/sui15.html.
3	Marcello, Fiducioso and Sebastian, Curi and Benedikt, Schumacher and Markus, Gwerder and Andreas, Krause (2019). Safe Contextual Bayesian Optimization for Sustainable Room Temperature PID Control Tuning, https://arxiv.org/abs/1906.12086
4	Sandmeier, E., Lobsiger-Kägi, E., Marek, R., Tomić, U., & Kälin, S. (2020). *2000-Watt-Gesellschaft leben: Reduktion des End-Energieverbrauchs durch Verhaltensänderungen – Nutzerinterventionen im Hüttengraben-Areal*. Schlussbericht vom 23. Dezember 2020. Eidgenössisches Departement für Umwelt, Verkehr, Energie und Kommunikation UVEK, Bundesamt für Energie BFE, Sektion Energieforschung und Cleantech. Available: https://www.aramis.admin.ch/Default?DocumentID=67511&Load=true, Chapter 4 Ergebnisse und Diskussion
5	PrimoŽ PotoČnik & Edvard Govekar (2019) Adaptive optimization of heating curves in buildings heated by a weather-compensated heat pump, Science and Technology for the Built Environment, 25:10, 1380-1393, DOI: 10.1080/23744731.2019.1616984. Available: https://doi.org/10.1080/23744731.2019.1616984
6	Data management Platform from SDSC: https://renkulab.io/

Appendix

- Appendix A: Bayesian Optimization and Gaussian Process Regression for Control Tuning
- Appendix B: Heating Degree Day
- Appendix C: Simulation
- Appendix D: Optimal Parameters for score function
- Appendix E: Data management plan and open access/data/model strategy
- Appendix F: Overview Bülsweg
- Appendix G: Data driven digital twin
- Appendix H: Pseudocode of the bootstrapped Difference-in-Differences algorithm

Appendix A: Bayesian Optimization and Gaussian Process Regression for Control Tuning

We defined the factors of our score function in Chapter 1.2. This function allows us to determine the scores associated with the optimization problem. However, we lack direct knowledge of how this function generally behaves or how it specifically applies to a given building. To understand and describe this function in order to optimize it later, we need to perform measurements and calculate the associated scores. Each measurement corresponds to a single data point. Once we have collected enough measurements, we can use these data points to create a surrogate model, see chapter 3.1.3.

We model our score function using GPR, which is particularly suited for this purpose due to its ability to approximate the true score function of a building with minimal prior knowledge. In the case of Contextual Bayesian Optimization, using GPR results in a three-dimensional surface, as illustrated in figure 22 and detailed in [1].

The goal of the next step is to use BO to identify the optimal action – in our case, the inlet temperature – that minimizes scores. BO proceeds as follows: since the context z is an external, uncontrollable variable, we fix the context at each step. In our example, we use the previous day's context value, assuming that the temperatures on two consecutive days do not differ significantly. By doing this, the optimization problem is reduced to a one-dimensional curve. This one-dimensional curve now describes the relationship between scores and the inlet temperature. This curve, along with its confidence intervals, is determined by the surrogate model determined by GPR.

Based on the newly acquired measurement, BO identifies the next best candidate for the inlet water temperature. This temperature is then set as the new value for the heating curve, and the next measurement is performed. Initially, each measurement will result in significant changes to the three-dimensional surface and, consequently, to the inlet water temperature. However, as more measurements are taken, these changes will decrease, eventually leading to convergence. This means we can approximate the effective underlying score function of an individual building with high accuracy.

In summary:

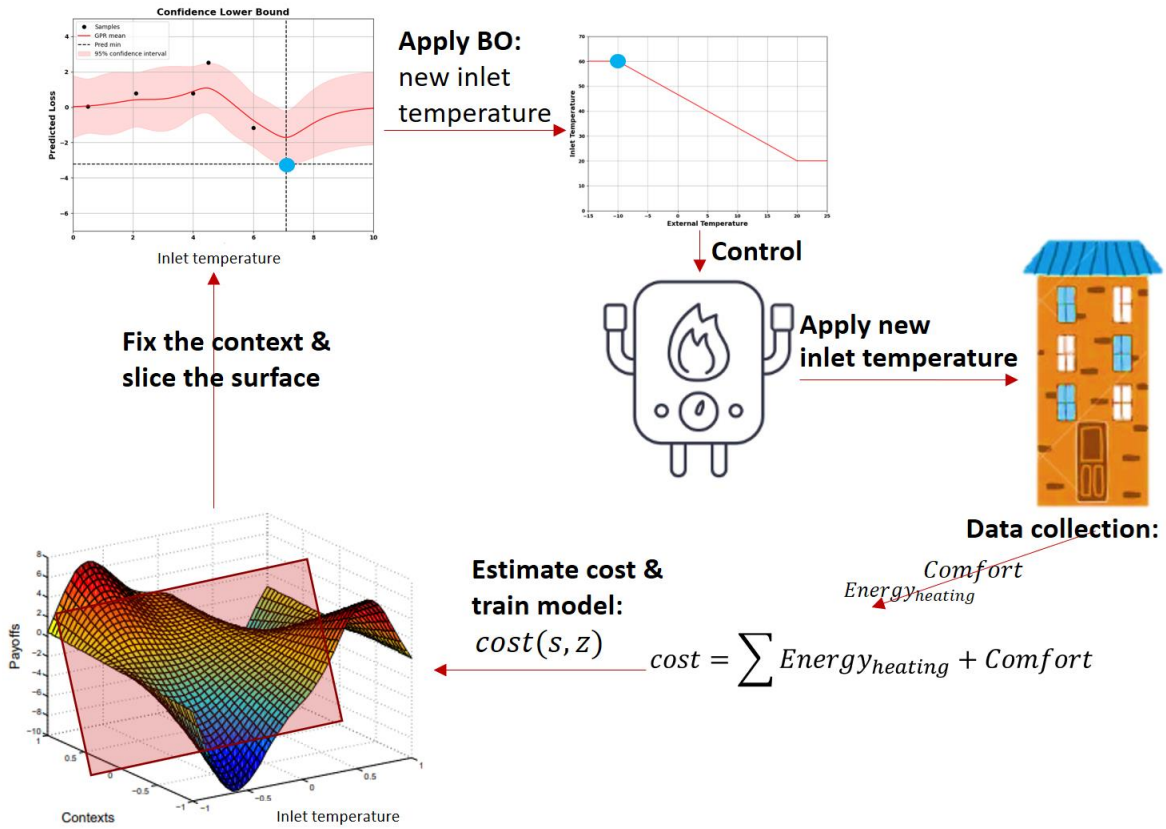


Figure 22: Depicted is the workflow for the contextual optimization of AHA. The experiment is initiated by setting the inlet temperature of the heating system, which leads to the Data Collection phase to gather observations needed to determine the score function. The scores are estimated, and the GPR is trained on the gathered information to obtain the 3D surface structure representing the score of the examined building. This provides the estimated scores (mean and variance) for each context and inlet temperature. Next, we fix the context based on past values and slice the surface, resulting in a 1-dimensional curve that depends only on the inlet temperature. We then apply BO to propose the next inlet temperature that minimizes the scores under the fixed context.

Appendix B: Heating Degree Day

The HDD method is a straightforward approach that normalizes heating energy based on heating degree days, providing a single value – the normalized heating energy consumption – for comparison. This allows for direct evaluation of energy use before and after an intervention.

Heating degree days reflect the climate-related heating energy demand by calculating, for each heating day (when the daily mean temperature is ≤ 12°C), how much the measured outdoor air temperature deviates from the desired indoor temperature of 20°C. The HDD 12/20 metric represents the sum of temperature differences from 20°C across all heating days within a specific period.

The HDD method enables a simple adjustment for outdoor air temperature by dividing the total heating energy consumption by the number of heating degree days. The resulting value indicates the average heating energy consumption per degree day, which can be compared across different periods (e.g., before and after an intervention). Using this approach, HDD values are determined for each cluster.

The main advantage of this method lies in its simplicity, as it distils heating energy performance into a single, easily comparable metric.

$$HDD = \sum_{k=0}^{n} \max(0, \mathrm{T_{set}} - \mathrm{T_{outside}})$$

$$normalized\ Heatingenergy\ = \frac{\sum_{k=0}^{n} Heatingenergy_k}{HDD}$$

Appendix C: Simulation

We have a complete setup to simulate EnergyPlus models using various heating curve models. For this, we have EnergyPlus models for three buildings: Bülsweg, NEST NEST, and NEST Sprint. We have also developed an interface and a data processing pipeline for our machine learning models. For these buildings, we can generate an unlimited number of weather data sets and create numerous simulation scenarios. Additionally, real data is available for these buildings. For Bülsweg, the heat output can be estimated through digitally readable heat meters. Each residential unit is also equipped with a multifunction sensor to measure VOC concentration, room temperature, and relative humidity throughout the project duration. Furthermore, we have access to weather data via a weather station in Vaduz. The outdoor sensor of the heating system is used to determine the ambient temperature. An overview of the AHA program can be found in figure 23. The table below explains the key classes and the source code: Files · development · Alessandro Tell / AHA · GitLab (renkulab.io)

Class	Description
SimulationWrapper	Interface to run Energyplus models containing specification of building / heating coefficients and variables used for our project which are extracted from the energyplus output.
SimulationDriver	Reading and processing the EnergyPlus output data so that it can be used for our machine learning models. Additionally, it includes the Controller class, which integrates all components (models, algorithms, and data storage) to perform effective sampling.
Denoiser	Initial measurements based on static heating curves indicate that the variance in heating requirements for each outdoor temperature can be quite large. This is primarily due to the impact of solar radiation. The effect of solar radiation is estimated and subtracted from the heating energy. The remaining heating energy is then standardized using a reference curve.
BoModel	The BO Heating model keeps track of all the samples and combines the various elements composing the model (the algorithm, the surrogate model, the heating curve)
SurrogateModel	The surrogate model is used to predict the score of a given action in a given context. We use the Gaussian Process Regression.
HeatingCurve	The HeatingCurve class models the parameter space of the heating curve, it contains the domain and the initial safe set.
UCBAlgo	We use the upper confidence bound to estimate the most promising next action using Gaussian Process Regression as surrogate model.

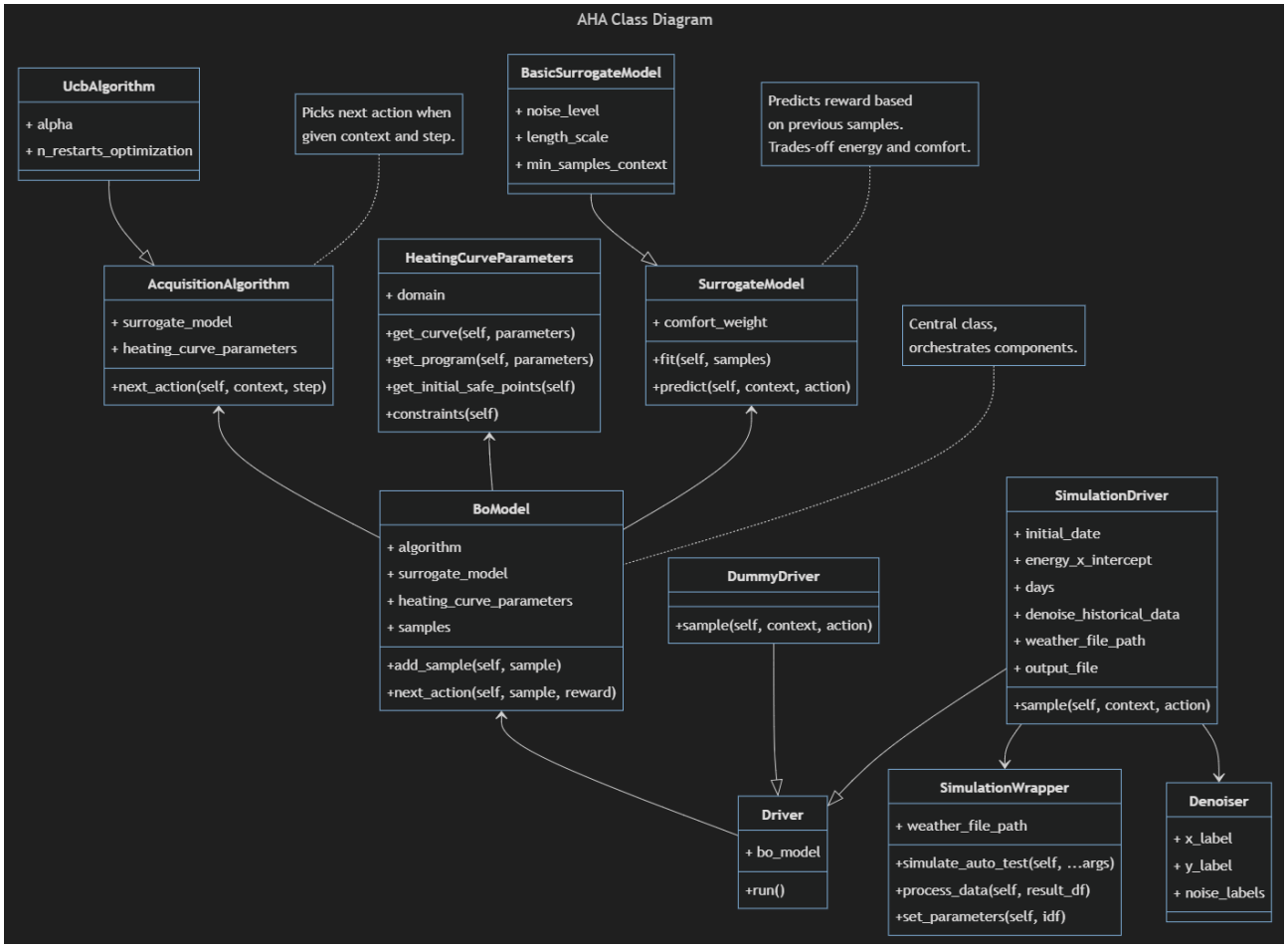


Figure 23: AHA Class Diagram displaying the different classes, attributes, methods and dependencies.

Appendix D: Optimal Parameters for score function

The definition of the score function in Chapter 1.2 allows us to experiment with the weighting between comfort and heating energy:

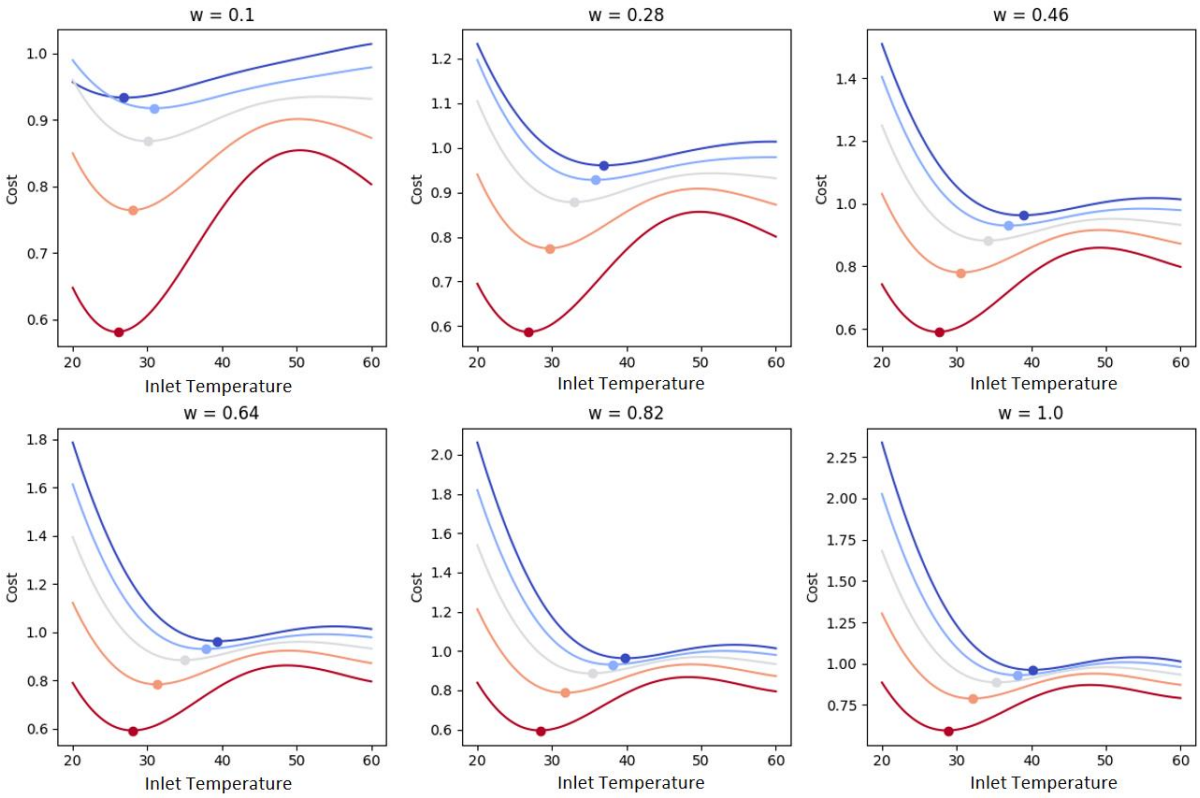


Figure 24: The simulations for NEST clearly demonstrate how adjusting the weighting between heating energy and comfort allows for additional energy savings. Of particular interest is the range between a weighting of 1, which emphasizes minimizing comfort score, and a slack at a weighting of approximately 0.8.

Appendix E: Data management plan and open access/data/model strategy

In alignment with our commitment to open access and the requirements for project co-funding by the Swiss Federal Office of Energy (SFOE), we have developed a comprehensive data management plan. Our approach includes making parts of the algorithm's code, datasets, and digital twin models publicly accessible through the Renku platform [6].

By early next year, we plan to release the developed Python interface for direct interaction with EnergyPlus models. This interface, designed for seamless integration and ease of use, will be made available independently of ongoing collaboration with Lippuner AG, enabling broad usability and accessibility. The current development version can be accessed here: Files · development · Alessandro Tell / AHA · GitLab (renkulab.io) [6].

Furthermore, insights, results, and developed programs from the project will be published on the ARAMIS platform to ensure transparency and knowledge sharing within the community. This effort underlines our dedication to fostering open research and reproducibility while supporting innovation in energy efficiency and digital twin technologies.

Appendix F: Overview Bülsweg

The Bülsweg building consists of three 2.5-room apartments and six 3.5-room apartments. Bülsweg is connected to the district heating network of the Buchs waste incineration plant, and all units are heated through underfloor heating. Heat consumption per apartment can be regularly monitored via digital heat meters. During the project, each apartment will also be equipped with a multifunction sensor to measure

VOC concentration, indoor temperature, and relative humidity. Although a photovoltaic system is installed on the building's roof, we do not have access to its data. Instead, we rely on data from local weather stations – specifically, the Vaduz station – to determine solar irradiation. The outdoor sensor of the heating system provides information about the ambient temperature.

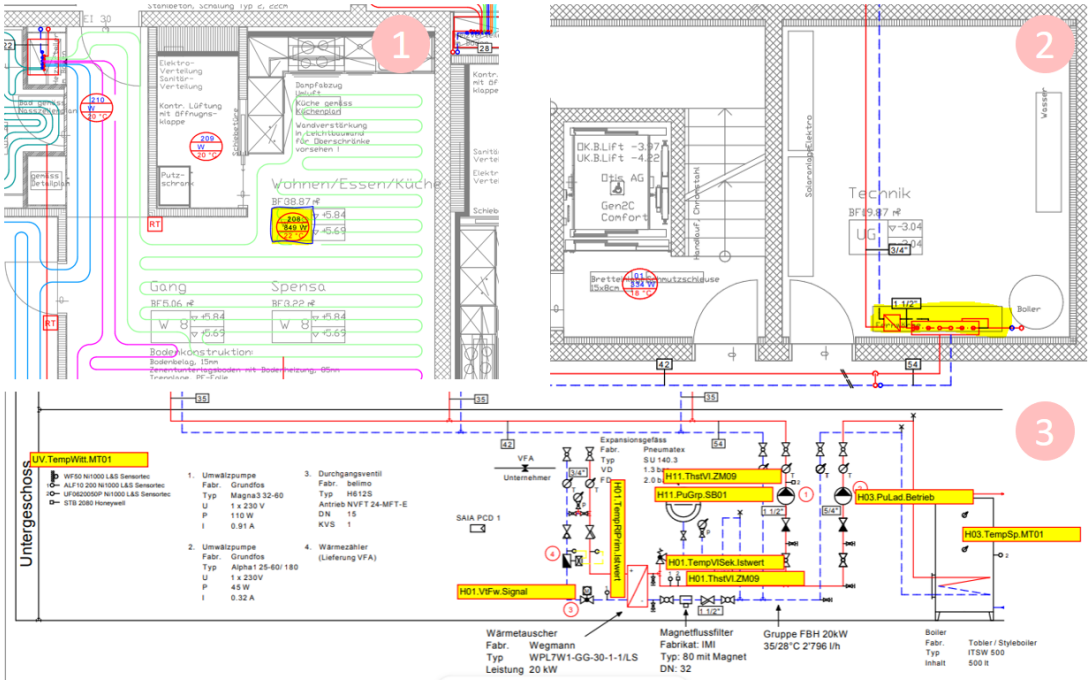


Figure 25: Subfigure 1 shows the floor plan with underfloor heating. Each room has an individual heating loop, which is configured to maintain the intended setpoint temperatures when the room thermostat is set to its default setting. For example, the reference room "Living/Dining/Kitchen" is heated to 22°C when the thermostat is set to level 3. Residents can fine-tune the flow through the room thermostat, allowing them to overheat or underheat the room according to personal preferences, thus adjusting the effective heating output. The heating power is supplied by the district heating system from Grabs and is transferred to a heat distributor in the basement, which then distributes the heat throughout the entire house. All key parameters for determining comfort and heating performance can be accessed for Bülsweg.

Key	Description	Dim	Freq	Installation
H01.TempVlSek.MT01	Secondary supply temperature	°C	15 min	Heating room
H01.TempRlPrim.MT01	Primary return temperature	°C	15 min	Heating room
H01.VtFw.VA01	Forward flow rate	m³/h	15 min	Heating room
H03.TempSp.MT01	Boiler temperature	°C	15 min	Heating room
UV.TempWitt.MT01	Outdoor temperature	°C	15 min	Weather sensor
H11.Anlage.Sollwert	System setpoint value	°C	15 min	Heating room
H11.Anlage.Errechnet	Calculated system output	°C	15 min	Heating room
H11.Anlage.Kurve_Yb_Empa	Heating curve setting Yb Empa	°C	15 min	Heating room
H03.PuLad.SB01	Loading pump command	On/Off	15 min	Heating room
H11.PuGrp.SB01	Pump group switching command	On/Off	15 min	Heating room
H11.PuGrp.Schaltung	Pump group switching status	OnOff	15 min	Heating room
H11.PuGrp.Stunden	Pump group operating hours	h	15 min	Heating room
UV.TempWitt.VA01	Outdoor temperature (analog value)	°C	15 min	Weather sensor

Key	Description	Dim	Freq	Installation
global_solar	Global solar radiation	W/m²	hourly	MeteoSwiss station Vaduz
diffuse_solar	Diffuse solar radiation	W/m²	hourly	MeteoSwiss station Vaduz
windspeed	Wind speed average	m/s	hourly	MeteoSwiss station Vaduz
precipitation	Precipitation sum	mm	hourly	MeteoSwiss station Vaduz
wind_direction_mean_deg	Mean wind direction (hourly)	°	hourly	MeteoSwiss station Vaduz
wind_direction_10min_deg	Wind direction (10-minute value)	°	10 min	MeteoSwiss station Vaduz
relative_humidity	Relative humidity	%	hourly	MeteoSwiss station Vaduz
sunshine_duration	Sunshine duration	min	hourly	MeteoSwiss station Vaduz
mean_temperature	Air temperature at 2m	°C	hourly	MeteoSwiss station Vaduz
dew_point	Dew point temperature at 2m	°C	hourly	MeteoSwiss station Vaduz
date	Timestamp	-	hourly	MeteoSwiss station Vaduz

Apartment	Key	Description	Dim	Freq	Installation
WHG1 EG Ost	temperature	Indoor temperature	°C	irregular	Entry living space, 1.5 m
	humidity	Indoor humidity	%	irregular	Entry living space, 1.5 m
	co2	CO_2 concentration	ppm	irregular	Entry living space, 1.5 m
	heat_electricity	Heating electricity consumption	kWh	irregular	Heat counter
WHG2 EG Mitte	temperature	Indoor temperature	°C	irregular	Entry living space, 1.5 m
	humidity	Indoor humidity	%	irregular	Entry living space, 1.5 m
	co2	CO_2 concentration	ppm	irregular	Entry living space, 1.5 m
	heat_electricity	Heating electricity consumption	kWh	irregular	Heat counter
WHG3 EG West	temperature	Indoor temperature	°C	irregular	Entry living space, 1.5 m
	humidity	Indoor humidity	%	irregular	Entry living space, 1.5 m
	co2	CO_2 concentration	ppm	irregular	Entry living space, 1.5 m
	heat_electricity	Heating electricity consumption	kWh	irregular	Heat counter
WHG4 1OG Ost	temperature	Indoor temperature	°C	irregular	Entry living space, 1.5 m
	humidity	Indoor humidity	%	irregular	Entry living space, 1.5 m

	co2	CO_2 concentration	ppm	irregular	Entry living space, 1.5 m
	heat_electricity	Heating electricity consumption	kWh	irregular	Heat counter
WHG5 1OG Mitte	temperature	Indoor temperature	°C	irregular	Entry living space, 1.5 m
	humidity	Indoor humidity	%	irregular	Entry living space, 1.5 m
	co2	CO_2 concentration	ppm	irregular	Entry living space, 1.5 m
	heat_electricity	Heating electricity consumption	kWh	irregular	Heat counter
WHG6 1OG West	temperature	Indoor temperature	°C	irregular	Entry living space, 1.5 m
	humidity	Indoor humidity	%	irregular	Entry living space, 1.5 m
	co2	CO_2 concentration	ppm	irregular	Entry living space, 1.5 m
	heat_electricity	Heating electricity consumption	kWh	irregular	Heat counter
WHG7 2OG Ost	temperature	Indoor temperature	°C	irregular	Entry living space, 1.5 m
	humidity	Indoor humidity	%	irregular	Entry living space, 1.5 m
	co2	CO_2 concentration	ppm	irregular	Entry living space, 1.5 m
	heat_electricity	Heating electricity consumption	kWh	irregular	Heat counter
WHG8 2OG Mitte	temperature	Indoor temperature	°C	irregular	Entry living space, 1.5 m
	humidity	Indoor humidity	%	irregular	Entry living space, 1.5 m
	co2	CO_2 concentration	ppm	irregular	Entry living space, 1.5 m
	heat_electricity	Heating electricity consumption	kWh	irregular	Heat counter
WHG9 2OG West	temperature	Indoor temperature	°C	irregular	Entry living space, 1.5 m
	humidity	Indoor humidity	%	irregular	Entry living space, 1.5 m
	co2	CO_2 concentration	ppm	irregular	Entry living space, 1.5 m
	heat_electricity	Heating electricity consumption	kWh	irregular	Heat counter

Appendix G: Machine Learning-Based Digital Twin

G.1 Architecture

The digital twin developed in this work is based on a **cascaded architecture** of machine learning models. The objective of this architecture is to predict the thermal and energetic dynamics of the building and its heating system in a physically consistent, data-driven manner. The single models are interlinked such that the output of one stage forms an input to the subsequent stage. This chaining of predictions reflects the causal dependencies between different physical quantities (supply temperature, return temperature, thermal balance, heat consumption, and indoor comfort).

G.1.1. Cascaded Model Structure

The cascaded approach was selected to reflect causal dependencies in the heating system:

- The analytical inlet temperature forms the starting point, reflecting control setpoints.
- Predicted inlet temperature propagates to return temperature, ΔT, and subsequently to heating energy.
- Finally, heating energy and system state determine indoor temperature and thus comfort.

This **hierarchical design** ensures that predictions remain physically interpretable and robust against unrealistic extrapolation. For example, setting the inlet signal to zero leads to downstream predictions converging at significantly lower heat energy predictions

The digital twin consists of six interconnected models, see Figure 15:

Model	Target & Concept	Features	Performance (MAE / RMSE / R^2)
1. Analytical Inlet Temperature	Target: inlet_calculated Concept: Analytical heating curve calculation, baseline control signal	Heating curve parameters	–
2. Supply Temperature (TempVlSek)	Target: Measured supply temp TempVlSek Concept: Residual approach: • Base: Linear regression on inlet_calculated_old • Residuals: XGBoost for nonlinear effects	Base: inlet_calculated_old Residuals: time encodings, weather, hydraulics, control inputs	Static: MAE 0.851 / RMSE 1.438 / R^2 0.918 Dynamic: MAE 3.120 / RMSE 4.794 / R^2 0.364
3. Return Temperature (TempRlPrim)	Target: TempRlPrim Concept: XGBoost conditioned on predicted TempVlSek and hydraulic/control variables	Time encodings, hydraulics (VtFw, Pumpladung), TempVlSek, building temp	Static: MAE 0.684 / RMSE 1.053 / R^2 0.953 Dynamic: MAE 0.915 / RMSE 1.598 / R^2 0.924
4. Temperature Difference (ΔT)	Target: TempVlSek – TempRlPrim Concept: XGBoost proxy for heat transfer to building	Pump activity, TempVlSek, TempRlPrim, lagged ΔT, building temp	Static: MAE 0.311 / RMSE 0.523 / R^2 0.862 Dynamic: MAE 0.626 / RMSE 0.968 / R^2 0.599
5. Heating Energy (heat_electricity)	Target: Thermal energy consumption heat_electricity Concept: XGBoost approximating heat counter	Weather/climate drivers, hydraulics, time encodings, predicted states (TempVlSek, TempRlPrim, ΔT), autoregressive lags	Static: MAE 1.516 / RMSE 2.033 / R^2 0.876 Dynamic: MAE 2.158 / RMSE 3.048 / R^2 0.703
6. Indoor Building Temperature	Target: building_temp (avg indoor temp) Concept: XGBoost linking heating system + weather to indoor comfort	Predicted supply/return temps, energy use, weather drivers	Static: MAE 0.149 / RMSE 0.211 / R^2 0.868 Dynamic: MAE 0.135 / RMSE 0.190 / R^2 0.821

G.1.2. Feature Selection Process

The features for each model were selected using a systematic **feature group search procedure**. Features were first grouped into physically meaningful categories:

- Weather drivers: global_solar, diffuse_solar, sunshine_duration, windspeed, wind_direction_mean_deg, relative_humidity, TempWitt_analog, precipitation, dew_point
- Building state and thermal inertia: building_temp
- Heating curve setpoints and control actions: inlet_calculated_old, action
- System measurements (supply/return): TempVlSek, TempRlPrim, delta_T, heat_electricity
- Pumps and hydraulics: Pumpladung, VtFw, pump_active, pump_recent
- Calendar encodings (cyclical time variables): hour, dayofweek, hour_sin, hour_cos, dow_sin, dow_cos
- Autoregressive lags: building_temp_lag1, building_temp_lag2, building_temp_lag3

All combinations of feature groups were evaluated with an **autoregressive XGBoost regressor** on a train/test split of the static regime. The best-performing feature set for each target was selected to ensure that the models balance predictive performance with interpretability.

G.1.3. Sanity Checks

To ensure the physical consistency and robustness of the cascaded model, several sanity checks were conducted. These tests assess whether the model behaves plausibly under controlled perturbations of key input variables.

Sensitivity Analysis with Respect to Inlet Temperature

As a first check, we analyzed how the model responds to variations in the inlet temperature (inlet_calculated), which directly affects the overall energy balance. From a physical standpoint, when setting the inlet temperature to zero, we would expect the predicted supply temperature (TempVlSek) to also approach zero.

The results are as follows:

- **Baseline model:** predicts approximately **35 °C** for TempVlSek, reflecting the average inlet temperature observed during training.
- **Digital twin:** predicts around **7 °C**, which aligns more closely with the expected physical behavior under such extreme conditions and leads to a substantial reduction in predicted heating energy.

Perturbation Experiment

We conducted a controlled sensitivity test by adjusting the inlet temperature ±5 °C around the baseline, using a representative test window (01.02.2025 – 22.02.2025). Three scenarios were simulated with the cascaded model:

Scenario	Description	Total Predicted Energy (kWh)	Δ vs. Baseline (kWh)
Baseline	Original inlet temperature	2571.26	—
Inlet – 5 °C	Reduced inlet temperature	1918.15	–653.11
Inlet + 5 °C	Increased inlet temperature	3579.34	+1008.08

Based on the fundamental physics of heat transfer, one might expect a roughly symmetrical response in energy use when adjusting the inlet temperature, since heating demand is proportional to the temperature difference between the supply and the environment ($Q = m \cdot \Delta T$). In other words, decreasing or increasing the inlet temperature by the same amount should, in theory, result in comparable reductions or increases in energy consumption. However, the observed response is asymmetric: while lowering the

inlet temperature reduces energy demand, increasing it leads to a disproportionately larger rise in energy consumption.This asymmetry indicates that the model does not yet fully capture the linear physical relationship between inlet temperature and thermal energy demand.

Nevertheless, this sanity check confirms that the model reacts qualitatively correctly to changes in inlet conditions, higher inlet temperatures increase, and lower inlet temperatures decrease, the predicted heating energy, providing confidence in the general directionality of the model's response.

G.1.4. Performance Comparison

On the hold-out test set (dates: 2025-01-10 and 2025-01-23), the digital twin architecture outperforms a baseline, e.g. GAM model:

GAM baseline: RMSE = 3.22, MAE = 2.31, R^2 = 0.69

Digital twin: RMSE = 2.43, MAE = 1.81, R^2 = 0.82

Beyond these accuracy improvements, the digital twin also supports extrapolation to counterfactual scenarios, see Sanity Checks above.

This example highlights how the digital twin improves accuracy and yields more realistic extrapolations, enabling meaningful exploration of scenarios outside the training distribution.

Appendix H: Pseudocode of the bootstrapped Difference-in-Differences algorithm

This section provides the pseudocode of the bootstrapped Difference-in-Differences algorithm used to quantify the effect of the adaptive heating tuner compared to the static baseline while accounting for model bias and uncertainty. The procedure combines the outputs of two digital twins, DT1 (adaptive) and DT2 (static), to estimate the distribution of possible energy savings.

Step 1: Aggregate predictions

- Convert hourly predictions of both digital twins: DT1 = adaptive and DT2 = static, into daily totals: kWh per day
- Split into pre-period (before intervention) and after-period (after intervention).

Notation: $\overline{DT1}_{pre}, \overline{DT2}_{pre}, \overline{DT1}_{after}, \overline{DT2}_{after}$

Step 2: Estimate bias and variance

- For each DT and period, compute residuals against the real testbed TB:

$$r_{m,period} = TB_{period} - DT_{m,period} \text{ , } m \in (1,2)$$

- From these residuals, estimate:

$$\text{Mean bias } b_{m,period} = mean(r))$$

$$\text{Standard deviation } s_{m,period} = std(r)$$

- Interpretation: bias = systematic offset, std = model uncertainty
- The estimated bias and standard deviation are used to adjust the model noise, replacing the normal assumption $\varepsilon \sim N(0,1)$ with a more realistic noise model $N\left(b_{m,period}, s^2_{m,period}\right)$

Step 3: Bootstrap resampling procedure

1. For each bootstrap draw $b = 1, \ldots, B$: Resample days: sample daily values with replacement from each digital twin's daily series in pre and after

2. Inject model noise: For each resampled day, add Gaussian perturbation: $\varepsilon \sim N\left(b_{m,period}, s^2_{m,period}\right)$
3. Compute bootstrapped mean differences:

$$\Delta^{(b)}_{pre} = \overline{DT2}^{(b)}_{pre} - \overline{DT1}^{(b)}_{pre}$$

$$\Delta^{(b)}_{after} = \overline{DT2}^{(b)}_{after} - \overline{DT1}^{(b)}_{after}$$

4. Compute DID draw:

$$DID^{(b)} = \Delta^{(b)}_{after} - \Delta^{(b)}_{pre}$$

Step 4: Collect bootstrap distribution

- Repeat steps 1 – 3 for B iterations, in our case B = 5000
- Obtain the empirical distribution DID: $\{DID^{(1)}, \ldots, DID^{(B)}\}$

Step 5: Summarize effect

- Report the mean effect:

$$\overline{DID} = \frac{1}{B}\sum_{b=1}^{B} DID^{(b)}$$

- Report uncertainty as percentile-based confidence interval:

$$CI_{95\%} = [quantile_{0.025}, quantile_{0.975}]$$