Mixed-Integer Nonlinear Model Predictive Control for Building Energy Systems Using the Combinatorial Integral Approximation (CIA) Decomposition Algorithm

Tobias Spratte[a,b], Steffen Eser[a], Nico Fuchs[a], Dirk Müller[a]

[a] *RWTH Aachen University, E.ON Energy Research Center,*
Institute for Energy Efficient Buildings and Indoor Climate, Germany
[b] *tobias.spratte@eonerc.rwth-aachen.de, CA*

Abstract:

Model predictive control (MPC) offers significant potential for reducing energy consumption in buildings through optimization-based control approaches. Precise descriptions of building energy systems often result in mixed-integer nonlinear optimization problems (MINLP). Conventional MINLP solvers are computationally intensive and often unsuitable for real-time applications. Since approximate solutions are typically sufficient in the building sector, this work investigates the Combinatorial Integral Approximation (CIA) decomposition algorithm, which partitions the MINLP into more easily solved subproblems consisting of two nonlinear problems (NLP) and one mixed-integer linear problem (MILP). The control strategy is applied to two energy systems of different complexity. Both systems consist of a heat pump, a heating rod, a water storage tank, and a cold water storage tank. The smaller system has two capacities with heating and cooling demands on the consumer side. The larger system includes an air handling unit, concrete core activation, and a thermal zone. The model size increases from 12 to 27 state variables and from 13 to 19 control variables. Nonlinear characteristics and switching behavior are modeled for selected components. Time-dependent disturbances include weather data and internal heat gains. Applying the MPC to the larger energy system results in high control quality and stability by intelligently controlling the system components at an early stage. MILP constraints reduce switching operations of selected components. With a 12-hour prediction horizon, the average MPC computation time is approximately 10 seconds. Larger prediction horizons improve control quality until the maximum computation time of 5 minutes is increasingly reached. For the smaller system, the maximum computation time is rarely reached, with average computation times between 3.5 and 42 seconds depending on prediction horizon and step size. Increasing the prediction horizon from 12 to 48 hours reduces thermal discomfort by more than 90 % while simultaneously reducing energy consumption by up to 28 %.

Keywords:

Heat Pump System; Concrete Core Activation; HVAC; Discrete Decisions; Slack Variables

1. Introduction

Over the past decades, substantial global efforts have been undertaken to reduce greenhouse gas emissions. In the European Union, current policy targets include at least a 55 % reduction in emissions by 2030 compared to 1990 and climate neutrality by 2050 [1]. Two key drivers are the transition to renewable energy systems and the reduction of energy demand, of which approximately 40 % in the EU is attributed to the building sector [2]. Advanced building control can decrease energy consumption while mitigating the variability of renewable generation, and model predictive control (MPC) has emerged as a particularly promising approach due to its optimization-based formulation [3].

The control performance of an MPC depends on the underlying model of real physical processes. In building energy systems, the underlying processes often exhibit nonlinear behaviour. Close-to-reality modeling of these processes, e.g., energy transfer via temperatures and mass flows, nonlinear component characteristics or data-driven models such as neural networks, leads to nonlinear optimization problems (NLPs) [4]. Furthermore, many components exhibit discrete on/off or mode-switching behaviour, including non-modulating or partially modulating heat pumps and reversible heat pump operation. A detailed representation of these phenomena leads to mixed-integer nonlinear optimization problems (MINLPs). State-of-the-art MINLP algorithms are typically computationally demanding and thus often unsuitable for real-time MPC of building energy systems [5].

In practice, the real processes are therefore frequently simplified. One common strategy is to linearise the dynamics and maintain integer variables, which yields a mixed-integer linear problem (MILP) that can be solved efficiently with dedicated solvers. Alternatively, the discrete behaviour may be omitted from the controller formulation, or integer constraints are

relaxed to continuous variables taking values in the unit interval. The latter results in a nonlinear problem, which can again be handled with well-established numerical solvers [6, 7]. While these approaches improve tractability, they either neglect important switching behaviour or rely on continuous relaxations that may produce control signals which are difficult to implement in practice.

In this study, we investigate a different, less commonly applied but promising approach, in which the detailed process description is retained and computational tractability is achieved through problem decomposition. For many building MPC applications, an approximate but feasible solution of the original MINLP is sufficient. The Combinatorial Integral Approximation (CIA) decomposition algorithm provides such a framework by splitting the MINLP into more tractable subproblems that are solved sequentially [6]. Successful applications of CIA-based mixed-integer nonlinear MPC have been demonstrated for solar-driven climate systems with thermal storage and traffic signal optimisation [7–9].

We apply the CIA-based MPC strategy to two white-box models of a building energy system of different size. The systems share the same physical topology but differ in the level of spatial aggregation on the demand side. Chapter 2 gives an overview of the two models. Chapter 3 formulates the mixed-integer nonlinear MPC and discusses the role of slack variables for solver robustness. Chapter 4 summarises the CIA decomposition and its implementation. The simulation setup and performance indices are described in Chapter 5. Chapter 6 presents results for a winter period and a parameter study on prediction horizon, step size and model size. Chapter 7 concludes the paper and outlines future work.

2. Building Energy Systems and Models

This chapter introduces the two building energy system models considered in this study. First, the complete system including all components relevant for the thermal zone is described in Section 2.1. Afterwards, the reduced system is presented in Section 2.2. Section 2.3 summarises the white-box dynamic models with emphasis on the nonlinear component characteristics and discrete operating modes that give rise to the mixed-integer nonlinear MPC formulation.

2.1. Complete Building Energy System

The complete building energy system consists of a reversible heat pump (HP), an electric heater (HS), a hot water storage (HWS), a cold water storage (CWS), an air handling unit (AHU), a concrete core activation (CCA) and a single thermal zone (TZ). The main hydraulic and air-side components and connections are shown in Figure 1.

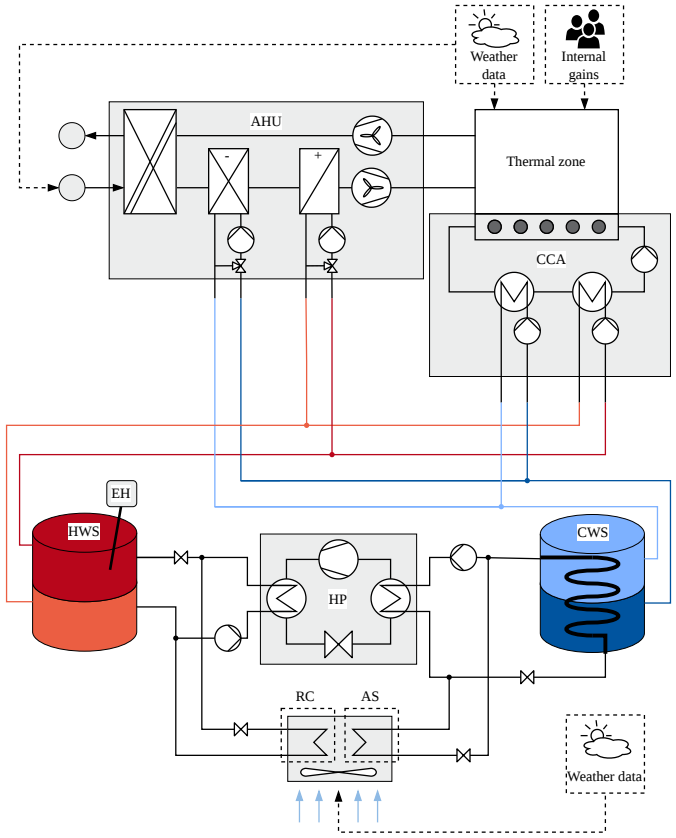

Figure 1. *Complete building energy system with heat pump (HP), air-source (AS), recooler (RC), hot and cold water storages (HWS, CWS), electric heater (EH), air handling unit (AHU), concrete core activation (CCA) and thermal zone.*

On the source side, the evaporator of the reversible heat pump is connected either to the CWS via a brine loop or to an air-source heat exchanger, depending on the operating mode. On the sink side, the condenser is connected to the HWS, which can be charged by the heat pump and, if needed, by the electric heater to reach higher temperatures than those achievable by the heat pump alone. The CWS and HWS supply the AHU cooling and heating coils via water–water heat exchangers. The CCA loop is supplied via two water–water heat exchangers connected to the CWS and HWS and exchanges heat with the embedded concrete core of the building.

The thermal zone model represents the air volume as well as lumped wall and window capacities. It is exposed to outdoor air temperature, solar radiation and internal gains from occupants, lighting and equipment. The AHU provides controlled ventilation and additional sensible heating or cooling of the supply air. Four operating modes are defined and shown in Figure 2: heating (H), cooling (C), combined heating and cooling (CHC), and standby (S). In heating mode, the heat pump extracts heat from the air source or the CWS and charges the HWS. In cooling mode, it extracts heat from the CWS and rejects it to the air source. In combined mode, both storages are charged simultaneously, whereas in standby the heat pump and circulation pumps are off and only passive thermal interactions and internal gains affect the states.

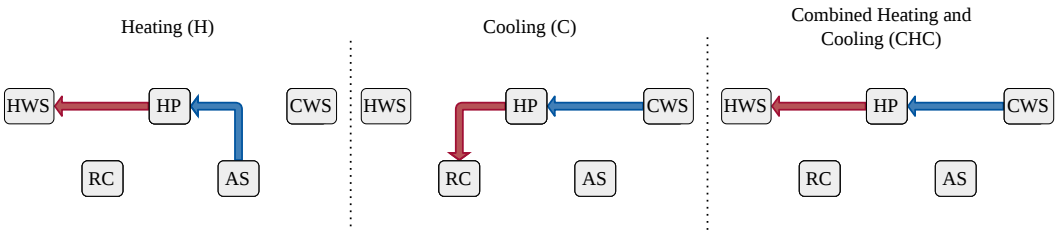

Figure 2. *Operating modes: heating (H), cooling (C) and combined heating and cooling (CHC).*

2.2. Reduced Building Energy System

In the reduced building energy system, the AHU, CCA and thermal zone are replaced by two lumped, constant heating and cooling loads that exchange heat with the storages via water–water heat exchangers. Hydraulically, the topology of the reduced system is identical to the generator-side part of the complete system.

The reduced system thus preserves the essential nonlinearities and mode-dependent behaviour of the heat pump and both storages while substantially decreasing the number of states and inputs. It is therefore well suited to assess the scalability of the CIA-based MPC with respect to prediction horizon and step size without the additional computational burden of the complete building model.

2.3. White-Box Dynamic Models

Both building energy systems are represented by dynamic white-box models derived from mass and energy balance. All heat exchangers are assumed to operate in counter-flow configuration and are modeled with lumped thermal capacities on each side. Water and brine are treated as incompressible fluids with constant specific heat capacities over the considered temperature ranges. The two storages are represented by stratified water volumes with additional mixing terms to obtain differentiable dynamics suitable for gradient-based optimisation [7, 10]. The same model structure is used for both system variants. Only the number of connected components differs.

Heat Pump and Efficiency Factor

The heat pump is modeled by an energy balance between evaporator and condenser heat rates and compressor electrical power according to Equations 1–3.

$$\dot{Q}_{\mathrm{Eva}} - \dot{Q}_{\mathrm{Con}} + P_{\mathrm{el,HP}} = 0 \tag{1}$$

$$\mathrm{COP} = \frac{\dot{Q}_{\mathrm{Con}}}{P_{\mathrm{el,HP}}} \tag{2}$$

$$\mathrm{EER} = \frac{\dot{Q}_{\mathrm{Eva}}}{P_{\mathrm{el,HP}}} \tag{3}$$

Here, $\dot{Q}_{\mathrm{Eva}}$ and $\dot{Q}_{\mathrm{Con}}$ denote the heat rates at evaporator and condenser, respectively, and $P_{\mathrm{el,HP}}$ is the compressor electrical power. In heating mode, the usable output is $\dot{Q}_{\mathrm{Con}}$ and the performance is quantified by the coefficient of performance (COP), whereas in cooling mode the relevant output is $\dot{Q}_{\mathrm{Eva}}$ and the energy efficiency ratio (EER) is used.

The ideal COP is obtained from the Carnot efficiency with condenser and evaporator temperatures T_{out} and T_{in} according to Equation 4 [11].

$$\mathrm{COP}_{\mathrm{ideal}} = \frac{T_{\mathrm{out}}}{T_{\mathrm{out}} - T_{\mathrm{in}}} \tag{4}$$

Real heat pumps are subject to electrical, mechanical and internal thermodynamic losses. To approximate the real performance, the ideal COP is multiplied by an efficiency factor ν according to Equation 5.

$$\mathrm{COP}_{\mathrm{real}} = \nu \cdot \mathrm{COP}_{\mathrm{ideal}} \tag{5}$$

Field measurements and manufacturer data show that the efficiency factor of modern heat pumps depends on the temperature lift $\Delta T = T_{\mathrm{out}} - T_{\mathrm{in}}$ [11, 12]. In this study, $\nu(\Delta T)$ is represented by a quadratic function according to Equation 6, which yields a maximum efficiency at $\Delta T = 40\,\mathrm{K}$ and decreases towards higher and lower temperature lifts (see Figure 3). This choice provides a conservative approximation of heat pumps in the building sector over the relevant operating range [11].

$$\nu(\Delta T) = -0.0001\,(\Delta T - 40)^2 + 0.4 \tag{6}$$

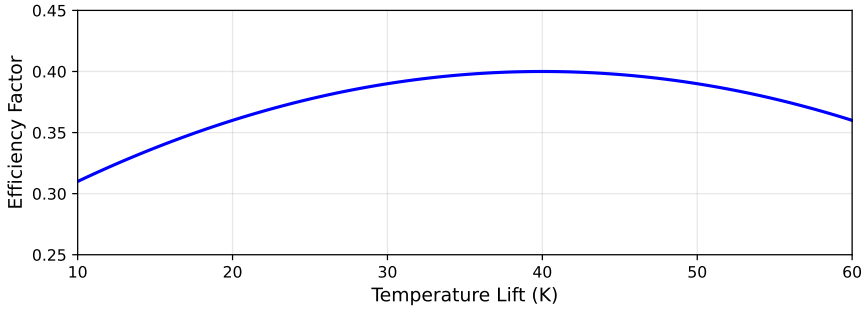

Figure 3. *Efficiency factor of the heat pump depending on the temperature lift based on [11].*

By combining Equations 1–6, both evaporator and condenser heat rates can be expressed as nonlinear functions of the compressor power and the temperature lift. This explicit dependence is crucial for the MPC, since it couples the electrical power consumption and the temperatures in the two storages.

Stratified Storages with Differentiable Mixing

The HWS and CWS are modeled as vertical cylindrical tanks with two water layers of equal volume. Each layer is represented by a lumped thermal capacity with temperature $T_{\mathrm{HWS,top}}, T_{\mathrm{HWS,bot}}$ and $T_{\mathrm{CWS,top}}, T_{\mathrm{CWS,bot}}$, respectively. In- and outlet mass flows from connected components (heat pump, AHU, CCA and constant loads) are assigned to the upper or lower layer according to the hydraulic design. Mass and energy balances for each layer yield first-order differential equations for the layer temperatures. Heat losses to the ambient are modeled by linear heat transfer coefficients.

A key modeling aspect is the treatment of vertical mixing between the two layers. A physically intuitive formulation uses a signed inter-layer mass flow $\dot{m}_{\mathrm{HWS,tb}}$ (from top to bottom layer) and its absolute value in the energy balance, which leads to a nondifferentiable term $|\dot{m}_{\mathrm{HWS,tb}}|$ [10]. To obtain differentiable dynamics, we follow the approach according to Sawant et al. and introduce a smoothed inter-layer flow $\dot{m}_{\mathrm{HWS,s}}$, with a small parameter $\varepsilon_{\mathrm{HWS}} > 0$ according to Equation 7 [7, 10].

$$\dot{m}_{\mathrm{HWS,s}} = \sqrt{\dot{m}^2_{\mathrm{HWS,tb}} + \varepsilon_{\mathrm{HWS}}} \tag{7}$$

For very small $\varepsilon_{\mathrm{HWS}}$, the difference between $\dot{m}_{\mathrm{HWS,s}}$ and $|\dot{m}_{\mathrm{HWS,tb}}|$ becomes negligible, while the square-root expression remains differentiable for all admissible flows. The same structure is applied to the CWS with its own smoothing parameter $\varepsilon_{\mathrm{CWS}}$.

In the complete system, additional water–water heat exchangers connect the storages to the AHU and CCA loops. Their lumped water-side volumes are modeled by first-order energy balances driven by the respective mass flows and storage temperatures. In the reduced system, two analogous heat exchangers connect the storages to constant heating and cooling loads.

Fans and Pumps

The electrical power of fans and pumps in the AHU and the CCA is modeled as a cubic function of the transported mass flow according to similarity laws for turbomachinery (see Equation 8) [13].

$$P_{\mathrm{el}} = P_{\mathrm{el,B}} \left(\frac{\dot{m}}{\dot{m}_{\mathrm{B}}} \right)^3 \tag{8}$$

Here, $P_{\mathrm{el,B}}$ and $\dot{m}_{\mathrm{B}}$ denote a reference operating point (design point) of the respective fan or pump. Equation 8 captures the strong nonlinear increase of electrical power with mass flow. It is applied to the AHU supply and exhaust fan mass flow $\dot{m}_{\mathrm{AHU}}$ and the CCA pump flow $\dot{m}_{\mathrm{CCA}}$, with component-specific base points chosen from design data.

This cubic relation is particularly relevant for the MPC, since it penalises high air and water flow rates and thereby promotes energy-efficient operation.

Thermal Zone and AHU Heat Exchangers

The thermal zone model is used only in the complete system. It consists of lumped capacities for zone air, exterior walls, interior walls and windows. The convective internal gains from occupants, lighting and equipment are computed from activity-based and schedule-based models and split into convective and radiative parts. The air energy balance is shown in Equation 9

$$\begin{aligned} C_{\mathrm{TZ,za}} \frac{\mathrm{d}T_{\mathrm{TZ,za}}}{\mathrm{d}t} = &\sum_j u_{za,j}(T_j - T_{\mathrm{TZ,za}}) + u_{\mathrm{wt,za}}(T_{\mathrm{CCA,wt}} - T_{\mathrm{TZ,za}}) \\ &+ \dot{m}_{\mathrm{AHU}} c_{p,l}(T_{\mathrm{AHU,sup}} - T_{\mathrm{TZ,za}}) + \dot{Q}_{\mathrm{IG,conv}}, \end{aligned} \tag{9}$$

where $C_{\mathrm{TZ,za}}$ is the thermal capacity of the zone air, $u_{za,j}$ denote heat transfer coefficients between zone air and surfaces j (walls and windows), $u_{\mathrm{wt,za}}$ couples the CCA wall temperature $T_{\mathrm{CCA,wt}}$ to the zone, $\dot{m}_{\mathrm{AHU}}$ and $T_{\mathrm{AHU,sup}}$ are the mass flow and the supply temperature of the AHU respectively, and $\dot{Q}_{\mathrm{IG,conv}}$ are the convective internal gains. Additional first-order balances describe wall and window temperatures and their interaction with ambient air and solar radiation.

Within the AHU, a rotary heat exchanger provides heat recovery between exhaust and supply air streams, and two air–water coils provide additional sensible cooling and heating. Each coil is modeled by two lumped capacities, one on the air side and one on the water side, with linear heat transfer between the volumes. The water-side inlet temperatures and mass flows originate from the storages and their respective control inputs, so that the AHU dynamics couple back to the reduced system.

Concrete Core Activation

The CCA loop consists of a circulating water mass flow that is tempered by two water–water heat exchangers connected to the CWS and HWS and exchanges heat with a lumped building concrete mass. The CCA is modeled by three thermal capacities for the fluid, the pipe wall and the surrounding concrete. First-order energy balances with linear heat transfer coefficients describe the interaction between these capacities and the zone air. The CCA pump power is computed via Equation 8, such that both thermal and electrical effects of activating the concrete core are captured.

Overall, the resulting dynamic models combine nonlinear heat pump characteristics, stratified storage dynamics with differentiable mixing, cubic fan and pump power laws and discrete operating modes. These features are directly reflected in the mixed-integer nonlinear MPC formulation in Chapter 3.

3. Mixed-Integer Nonlinear MPC Formulation

This section introduces the MPC formulation used for both system variants. We first present the formulation of objective function and constraints in Section 3.1. Afterwards, we discuss the role of slack variables and soft constraints in Section 3.2.

3.1. Objective Function and Constraints

The MPC problem is formulated on a discrete time grid with step index $k = 0, \ldots, N-1$ according to Equations 10–16.

$$\min_{x,u,s} J = \sum_{k=0}^{N-1} \left(\sum_{s \in S} w_s s_k^2 + \sum_{u \in U} w_u u_k^2 + q_{\mathrm{el}} P_{\mathrm{el,tot},k} \right), \tag{10}$$

subject to

$$x_{k+1} = f(x_k, u_k, d_k), \tag{11}$$
$$x_0 = x_{\mathrm{init}}, \tag{12}$$
$$x_{\min,k} \le x_k \le x_{\max,k}, \tag{13}$$
$$u_{\min} \le u_k \le u_{\max}, \tag{14}$$
$$Y_k \in \{0,1\}^{n_Y}, \tag{15}$$
$$0 \ge h(x_k, u_k, s_k, d_k). \tag{16}$$

The stage cost penalises violations of soft constraints, control effort and electrical power consumption. The total electrical power is given by Equation 17,with the latter two terms present only in the complete system.

$$P_{\mathrm{el,tot},k} = P_{\mathrm{el,HP},k} + P_{\mathrm{el,HS},k} + P_{\mathrm{el,V,HP},k} + P_{\mathrm{el,V,RLT},k} + P_{\mathrm{el,BKT},k} \tag{17}$$

The state vector x_k contains all temperatures and other dynamic states, the continuous control vector u_k comprises mass flows and electrical powers, and the binary control vector Y_k encodes the operating modes H, C, CHC, S. The disturbance vector d_k includes outdoor temperature, solar radiation and internal gains.

The four operating modes are constrained by a soft SOS1 condition according to Equation 18, ensuring that exactly one mode is active in each time step when the slack $s_{\mathrm{modi},k}$ is small.

$$H_k + K_k + CHC_k + S_k + s_{\mathrm{modi},k} = 1, \tag{18}$$

Additional constraints in $h(\cdot)$ enforce comfort bands for the zone temperature, bounds on heat pump temperature lift, and phase-change limits for water and refrigerant temperatures In the complete system, the state vector comprises 27 differential states and 19 control variables, including four binary mode variables, whereas the reduced system contains 12 states and 13 control variables, including four binary mode variables.

3.2. Slack Variables and Soft Constraints

Slack variables are introduced at several levels of the formulation to improve feasibility and numerical robustness without sacrificing physical interpretability:

- *Comfort constraints:* Violations of the zone air temperature band during occupied and unoccupied periods are represented by slacks and penalised in the cost function.
- *Component limits:* Temperatures close to freezing or boiling thresholds, as well as compressor power and temperature lift limits, are softened by slacks. This prevents the NLP subproblems from becoming infeasible when the model, disturbances or initial conditions are slightly inconsistent.
- *Mode constraint:* The SOS1 condition (see Equation 18) includes a slack $s_{\text{modi},k}$ to allow small violations in the relaxed NLP to ensure feasibility. A strong weight is applied to $s_{\text{modi},k}$ to ensure that the sum of modes is close to one which is important for the CIA problem.

From an optimisation viewpoint, the slacks widen the feasible set and help interior-point methods to find a solution with reasonable step sizes. By penalising the slacks quadratically with appropriate weights, the optimiser is steered towards solutions that satisfy the hard physical limits whenever possible. This design significantly reduces the number of failed iterations and "restoration failed" messages observed for large horizons and fine discretisations, while the remaining violations of the original hard constraints are small and physically acceptable.

4. CIA Decomposition and Implementation

The MPC formulation of Chapter 3 leads to a mixed-integer nonlinear problem with nonlinear dynamics and binary mode variables. Solving such MINLPs directly at each sampling instant is typically not feasible for the considered model sizes and horizons. Instead, the Combinatorial Integral Approximation (CIA) decomposition algorithm is used to obtain approximate solutions with reduced computational effort [6, 14].

At each sampling step, CIA proceeds in three stages:

1. **Relaxed NLP (NLP_{rel}):** All binary variables Y_k are relaxed to the interval $[0, 1]$ and treated as continuous controls. The original cost and constraints, including the slack-augmented SOS1 condition (see Equation 18), are maintained. Solving NLP_{rel} yields a continuous "mode trajectory" Y_k^{rel} that already approximates a piecewise constant mode sequence.
2. **Integer problem (MILP):** Based on Y_k^{rel}, a combinatorial problem is formulated that searches for integer mode trajectories Y_k^{int} minimising the integral deviation from Y_k^{rel} subject to the strict SOS1 constraint and optional switching constraints (maximum number of switches and minimum dwell times). The problem is a MILP and is solved using the open-source tool `Pycombina` with a branch-and-bound algorithm [8, 14].
3. **Fixed-binary NLP (NLP_{bin}):** The integer trajectories Y_k^{int} are then treated as fixed parameters, and the original nonlinear MPC problem is solved again for the continuous variables only. The solution of NLP_{rel} serves as initial guess. The first control action of the resulting sequence is applied to the system, and the horizon is shifted forward in receding-horizon manner.

To further enhance robustness for large products of horizon length and inverse step size, small continuous correction variables $\Delta Y_{H,K,CHC,k} \in [-10^{-4}, 10^{-4}]$ are introduced in the dynamic equations according to Equations 19–21, while Equation 18 uses the pure binary variables.

$$H_k = Y_{H,k} + \Delta Y_{H,k} \tag{19}$$

$$K_k = Y_{K,k} + \Delta Y_{K,k} \tag{20}$$

$$CHC_k = Y_{CHC,k} + \Delta Y_{CHC,k} \tag{21}$$

These corrections allow the NLPs to slightly adjust the effective mode signal to avoid numerical difficulties, while strong penalties in the objective keep the deviations negligible for the closed-loop behaviour [7].

The CIA-based MPC is implemented in Python using CasADi for automatic differentiation and code generation [15]. The nonlinear subproblems NLP_{rel} and NLP_{bin} are solved with IPOPT and the MUMPS linear solver [16]. The MILP is solved with `Pycombina`, which provides efficient C++ implementations of the branch-and-bound routines wrapped via `pybind11` [17]. The controller is integrated into the agentlib framework for agent-based MPC simulation and C-code export [18].

5. Simulation Setup

This section summarises the disturbance data, comfort specifications and scenarios used to evaluate the CIA-based MPC. The same disturbance data and comfort criteria are applied to both system variants to enable a fair comparison in the parameter study.

5.1. Weather Data and Internal Gains

Outdoor air temperature and solar radiation are taken from the German Test Reference Year (TRY) 2015 for the city of Aachen [19]. Hourly temperature values are linearly interpolated to the MPC sampling time. Solar gains for the thermal zone are computed from the TRY radiation data and basic geometric assumptions about window areas and orientations. Internal gains from occupants, lighting and equipment follow a weekly schedule with occupied hours from 07:00 to 18:00 on weekdays. During these hours, the gains are constant in time. Outside the occupied period they are set to zero.

5.2. Comfort Specifications and Performance Indices

During occupied hours, the zone air temperature is softly constrained to the interval $[21°C, 23°C]$. During unoccupied hours, a wider band of $[17°C, 27°C]$ is applied. Violations are represented by slack variables and penalised in the MPC cost function.

Thermal discomfort is quantified as the time integral of the absolute deviation of the zone air temperature from the corresponding comfort band in Kh. Electrical energy consumption is obtained by integrating the total electrical power $P_{\mathrm{el,tot}}$. Additional indices are the average coefficient of performance of the heat pump, the number of switching events and the average continuous run time of the compressor and selected fans.

5.3. Simulation Periods and MPC Configuration

The CIA-based MPC is evaluated in a two-week winter period representing the maximum heating demand (calendar weeks 51 and 52) and in a four-day transition period representing both, heating and cooling demands (calender weeks 39 and 40). Prior to the evaluation period an 8-day warm-up simulation is performed to reduce the influence of initial conditions.

Unless stated otherwise, the prediction horizon is $N = 24$ steps corresponding to 12 h and the control interval is $\Delta t = 30$ min. These settings are also used as reference for the parameter study. The maximum allowed computation time for the three CIA subproblems per sampling instant is set to 5 min. The allocation of this budget to $\mathrm{NLP_{rel}}$, CIA MILP and $\mathrm{NLP_{bin}}$ follows the values used in [14] with 55 %, 5 % and 40 %, respectively.

6. Results

This section presents closed-loop results for the winter period and results of a parameter study. Section 6.1 focuses on the complete building energy system with a fixed horizon and step size, whereas Section 6.2 analyses the impact of prediction horizon, step size and model size on comfort, energy use and computation time.

6.1. Winter Period: Complete Building Energy System

Figure 4 shows selected simulation results for the two-week winter period for the complete building energy system. The figure contains the zone air temperature including comfort band, the outdoor air temperature, the CCA and AHU inlet temperature, the HWS temperatures, the temperatures of condenser and evaporator of the heat pump, the electrical power of the heat pump and electric heater, the efficiency of the heat pump, the operating heating and cooling modes, the AHU and CCA mass flows, as well as solar and internal gains.

At the beginning of the first working day, the building mass has cooled down over the preceding weekend. The heat pump therefore operates at maximum power and is supported by the electric heater to raise the upper layer temperature of the hot water storage beyond the level achievable by the heat pump alone. This enables the supply air temperature constraint in the air handling unit to be satisfied even under low ambient temperatures. Thermal discomfort on the first working day is higher than on the remaining days, but remains limited due to the early activation of the generator-side components.

On subsequent weekdays, the CIA-based MPC exploits the thermal inertia of the building and storages. The heat pump is activated sufficiently early before occupancy to preheat the zone, while the electric heater is not needed. During occupied hours, the zone air temperature stays largely within the comfort band, and deviations remain small even under rapidly changing outdoor conditions. The supply and exhaust fan mass flows are mostly kept at their minimum values to reduce fan power.

The operating mode sequence shows that the heating mode dominates during winter, with occasional switches to the standby mode when thermal loads are low and the zone temperature can be maintained by internal gains and residual storage energy. The cooling and the combined heating and cooling modes are not required in this period because no cooling loads occur. Overall, the results indicate that the CIA-based MPC is able to coordinate heat pump, storage and air-side components such that comfort requirements are met with moderate electrical energy consumption and without excessive mode chattering.

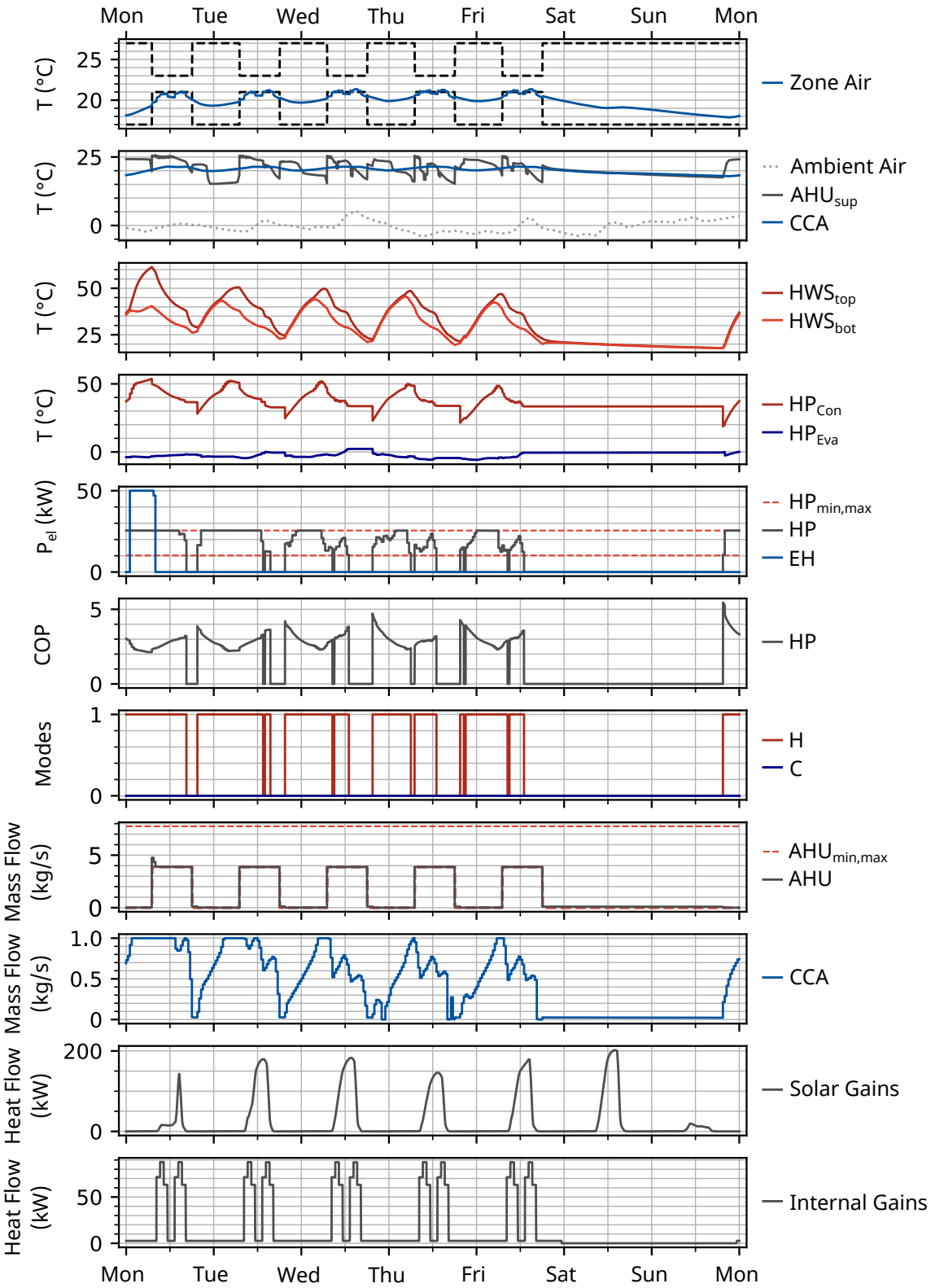

Figure 4. *Closed-loop results for the two-week winter period.*

6.2. Parameter Study: Horizon, Step Size and Model Size

The parameter study investigates the influence of prediction horizon, control interval and model size on control performance and computational effort. Both the reduced and the complete building energy system are simulated over the four-day period described in Section 5.3 with prediction horizons between 12 h and 72 h and control intervals between 15 min and 60 min. The simulated combinations are listed in Table 1.

Table 1. *Combinations of prediction horizon and step size used (1) and not used (0) in the parameter study.*

Step size / min	**Prediction horizon / h**					
	12	18	24	36	48	72
15	1	1	1	0	0	0
30	1	1	1	1	1	0
60	0	0	1	1	1	1

Figure 5 summarizes the results in terms of thermal discomfort, electrical energy consumption, average total computation time of the three CIA subproblems and the fraction of optimization steps hitting the maximum computation time. The left column shows the reduced system, the right column the complete system.

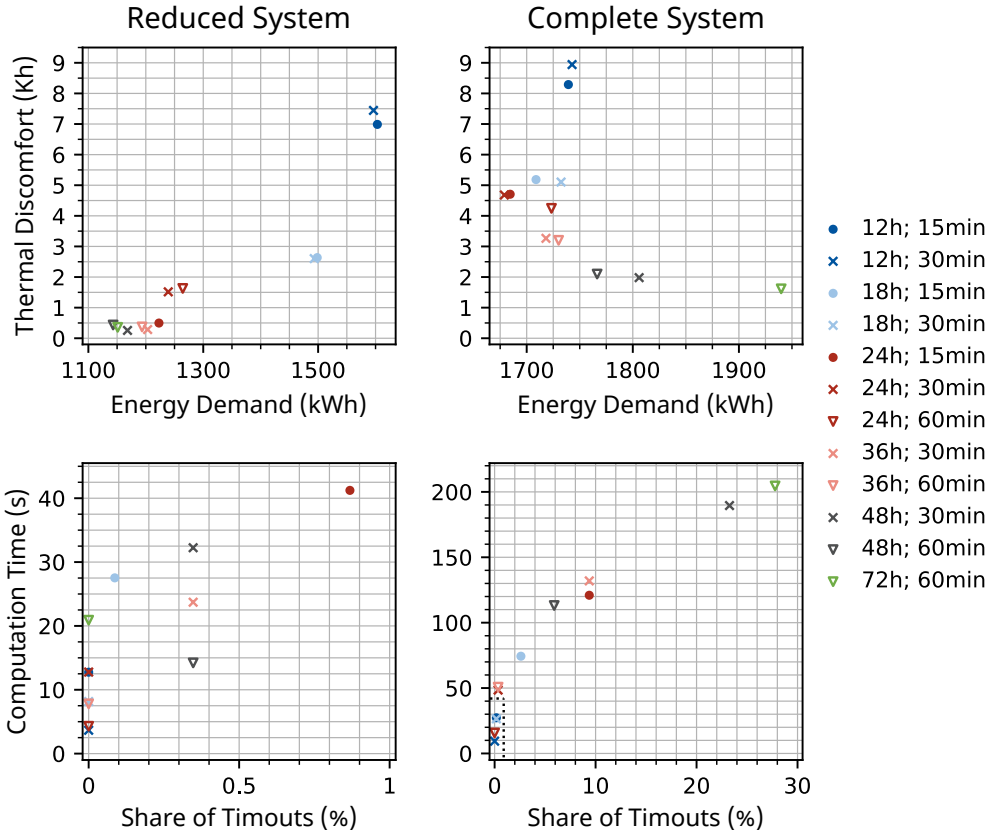

Figure 5. *Parameter study for reduced system (left) and complete system (right): thermal discomfort and electrical energy (top), average total computation time and average share of timeouts (bottom) for different prediction horizons and step sizes.*

For the reduced system, increasing the prediction horizon from 12 h to 36 h yields a pronounced reduction in both discomfort and energy consumption. Beyond 36 h, a saturation is observed: additional horizon length hardly improves the performance, particularly for horizons of 48 h and 72 h. The choice of control interval between 15 min and 60 min has no dominant effect on comfort or energy in this case, but shorter intervals increase the number of optimisation steps and thus the total computational effort.

For the complete system, increasing the horizon from 12 h to 24 h improves thermal comfort noticeably at the price of a moderate increase in electrical energy consumption. Beyond 24 h, the trend reverses: while discomfort can be further reduced, energy consumption and the frequency of timeouts rise significantly. The bottom-right subplot of Figure 5 shows

that for long horizons and small step sizes, a considerable share of optimisation steps reaches the maximum computation time, implying that the controller must occasionally operate with suboptimal or partially converged solutions. This effect explains the increase in energy use despite longer horizons, as the controller cannot always find an optimal solution in the allowed time interval.

The impact of model size on computation time is highlighted in Fig. 6, which plots the average total computation time of the CIA algorithm versus prediction horizon for selected step sizes. For identical controller settings, the computation time for the complete system is roughly 2–10 times larger than for the reduced system, depending on horizon and step size. The increase is mainly due to the larger number of states, controls and constraints in both NLP subproblems and the CIA MILP.

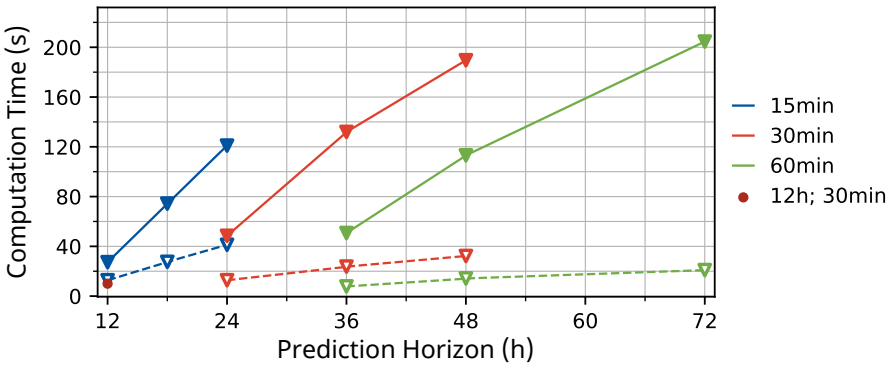

Figure 6. *Average total computation time of the CIA algorithm versus prediction horizon for selected step sizes, comparing reduced system (dashed) and complete system (solid).*

From a practical perspective, the results suggest that for the complete building energy system, prediction horizons between 12 h and 24 h with control intervals of 15–30 min provide a good compromise between comfort, energy use and computational cost. For the reduced system, longer horizons of up to 48 h can be utilised effectively, as the subproblems remain comparatively small and timeouts are rare. In both cases, the introduction of slack variables and CIA decomposition ensures that feasible solutions are obtained for all controller calls, even in the more challenging configurations.

7. Conclusion

This paper has presented a mixed-integer nonlinear MPC strategy for heat-pump-based building energy systems using the CIA decomposition algorithm. Two white-box models of different size were considered: a complete building energy system including an air handling unit, a concrete core activation and a thermal zone, as well as a reduced building energy system representing only the generator side of the complete building energy system. Both models feature nonlinear component characteristics and discrete operating modes.

The CIA-based MPC retains the detailed nonlinear dynamics and binary mode decisions of the underlying MINLP while decomposing the optimisation into two NLPs and one MILP. Slack variables at key constraints and small continuous corrections for the mode variables increased numerical robustness, particularly for large prediction horizons. Closed-loop simulations for a two-week winter period demonstrated that the controller can achieve high thermal comfort with moderate electrical energy consumption and without excessive mode chattering.

The parameter study quantified the trade-off between prediction horizon, control interval, model size, control performance and computational effort. For the complete system, horizons between 12 h and 24 h provided a favourable compromise, whereas the reduced system benefited from horizons up to 48 h. Model size was identified as a key driver of computation time, with the complete system requiring up to an order of magnitude more time than the reduced system.

Future work should focus on experimental validation of the CIA-based MPC on a real building demonstrator, comparison with alternative approximate MINLP-MPC formulations and conventional MPC schemes, and the investigation of distributed or hierarchical optimisation approaches to further improve scalability to large multi-zone building systems.

Acknowledgments

We gratefully acknowledge the financial support by the Federal Ministry for Economic Affairs and Energy (BMWE), promotional reference 03EN1102A

Nomenclature

Abbreviations

AHU	Air handling unit
AS	Air source
BES	Building energy system
CCA	Concrete core activation
CIA	Combinatorial Integral Approximation
COP	Coefficient of performance
CWS	Cold water storage
EER	Energy efficiency ratio
HP	Heat pump
HWS	Hot water storage
HVAC	Heating, ventilation and air conditioning
MILP	Mixed-integer linear programming
MINLP	Mixed-integer nonlinear programming
MPC	Model predictive control
NLP	Nonlinear programming
RC	Recooler
SOS1	Special ordered set of type 1
TRY	Test reference year
TZ	Thermal zone

Parameters

N	Prediction horizon (number of steps)
Δt	Control interval / sampling time
w_s, w_u	Weights of slacks and controls in stage cost
q_{el}	Weight / price factor for electrical power
$\nu(\Delta T)$	Heat pump efficiency factor
$\varepsilon_{\mathrm{HWS}}, \varepsilon_{\mathrm{CWS}}$	Smoothing parameters for HWS/CWS mixing flows
$P_{\mathrm{el,B}}, \dot{m}_{\mathrm{B}}$	Reference power and mass flow of fan or pump
$C_{\mathrm{TZ,za}}$	Thermal capacity of zone air

Variables

States, controls and disturbances

x_k, u_k, Y_k	State, continuous and binary control vectors at step k
d_k, s_k	Disturbance and slack vectors at step k
J	MPC cost over prediction horizon

Heat pump and storages

$\dot{Q}_{\mathrm{Eva}}, \dot{Q}_{\mathrm{Con}}$	Evaporator and condenser heat rates of the heat pump
$P_{\mathrm{el,HP}}, P_{\mathrm{el,HS}}$, $P_{\mathrm{el,tot},k}$	Electrical power of heat pump, electric heater and total power in step k
$T_{\mathrm{HWS,top}}, T_{\mathrm{HWS,bot}}$, $T_{\mathrm{CWS,top}}, T_{\mathrm{CWS,bot}}$	Layer temperatures of hot and cold water storages
$\dot{m}_{\mathrm{HWS,tb}}, \dot{m}_{\mathrm{HWS,s}}$, $\dot{m}_{\mathrm{CWS,tb}}, \dot{m}_{\mathrm{CWS,s}}$	Vertical and smoothed inter-layer mass flows in HWS and CWS

Thermal zone and AHU

$T_{\mathrm{TZ,za}}, T_j$	Zone air temperature and surface temperatures (walls, windows)
$T_{\mathrm{AHU,sup}}, T_{\mathrm{CCA,wt}}$	AHU supply air and CCA wall temperature
$\dot{m}_{\mathrm{AHU}}, \dot{m}_{\mathrm{CCA}}$	AHU air and CCA water mass flow rates
$\dot{Q}_{\mathrm{IG,conv}}$	Convective internal gains

Mode and CIA variables

H_k, K_k, CHC_k, S_k	Continuous mode variables (heating, cooling, combined, standby)
$Y_{H,k}, Y_{K,k}, Y_{CHC,k}$	Binary mode variables in CIA MILP
$\Delta Y_{H,k}, \Delta Y_{K,k}$, $\Delta Y_{CHC,k}$	Continuous mode correction variables
$s_{\mathrm{modi},k}$	Slack variable for SOS1 mode constraint
$Y_k^{\mathrm{rel}}, Y_k^{\mathrm{int}}$	Relaxed and integer mode trajectories in CIA algorithm

References

[1] European Commission. *'Fit for 55': delivering the EU's 2030 Climate Target on the way to climate neutrality*. Tech. rep. 550 final. Brüssel: European Commission, July 2021. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52021DC0550 (visited on 03/30/2026).

[2] European Commission. In focus: Energy efficiency in buildings. In: (Feb. 2020). URL: https://commission.europa.eu/news-and-media/news/focus-energy-efficiency-buildings-2020-02-17_en (visited on 03/30/2026).

[3] J. Drgoňa, J. Arroyo, I. Cupeiro Figueroa, D. Blum, K. Arendt, D. Kim, E. P. Ollé, J. Oravec, M. Wetter, D. L. Vrabie, and L. Helsen. All you need to know about model predictive control for buildings. en. In: *Annual Reviews in Control* 50 (2020), pp. 190–232. ISSN: 13675788. DOI: 10.1016/j.arcontrol.2020.09.001. URL: https://linkinghub.elsevier.com/retrieve/pii/S1367578820300584 (visited on 03/30/2026).

[4] O. Nelles, S. Ernst, and R. Isermann. Neuronale Netze zur Identifikation nichtlinearer, dynamischer Systeme: Ein Überblick. In: *at - Automatisierungstechnik* (June 1997). URL: https://www.degruyter.com/document/doi/10.1524/auto.1997.45.6.251/html (visited on 03/30/2026).

[5] R. Quirynen and S. Di Cairano. Sequential Quadratic Programming Algorithm for Real-Time Mixed-Integer Nonlinear MPC. In: *2021 60th IEEE Conference on Decision and Control (CDC)* (Jan. 2022). URL: https://ieeexplore.ieee.org/abstract/document/9683714 (visited on 03/30/2026).

[6] S. Sager, M. Jung, and C. Kirches. Combinatorial integral approximation. en. In: *Mathematical Methods of Operations Research* 73.3 (June 2011), pp. 363–380. ISSN: 1432-2994, 1432-5217. DOI: 10.1007/s00186-011-0355-4. URL: http://link.springer.com/10.1007/s00186-011-0355-4 (visited on 03/30/2026).

[7] A. Bürger, D. Bull, P. Sawant, M. Bohlayer, A. Klotz, D. Beschütz, A. Altmann-Dieses, M. Braun, and M. Diehl. Experimental operation of a solar-driven climate system with thermal energy storages using mixed-integer nonlinear model predictive control. en. In: *Optimal Control Applications and Methods* (May 2021). ISSN: 1099-1514. DOI: https://doi.org/10.1002/oca.2728. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/oca.2728 (visited on 03/30/2026).

[8] A. Bürger, C. Zeile, A. Altmann-Dieses, and M. Diehl. *Design, implementation and simulation of an MPC algorithm for switched nonlinear systems under combinatorial constraints | Elsevier Enhanced Reader*. en. May 2019. DOI: 10.1016/j.jprocont.2019.05.016. URL: https://www.sciencedirect.com/science/article/pii/S0959152419303592 (visited on 03/30/2026).

[9] S. Göttlich, A. Potschka, and U. Ziegler. Partial Outer Convexification for Traffic Light Optimization in Road Networks. en. In: *SIAM Journal on Scientific Computing* 39.1 (Jan. 2017), B53–B75. ISSN: 1064-8275, 1095-7197. DOI: 10.1137/15M1048197. URL: http://epubs.siam.org/doi/10.1137/15M1048197 (visited on 03/30/2026).

[10] P. Sawant, A. Bürger, M. D. Doan, C. Felsmann, and J. Pfafferott. Development and experimental evaluation of grey-box models of a microscale polygeneration system for application in optimal control. en. In: *Energy and Buildings* 215 (May 2020), p. 109725. ISSN: 03787788. DOI: 10.1016/j.enbuild.2019.109725. URL: https://linkinghub.elsevier.com/retrieve/pii/S0378778819331342 (visited on 03/30/2026).

[11] M. Köfinger, R. Geyer, and D. Basciotti. *Task 5.2 Methoden zur Auswahl, Auslegung und Integration von Wärmepumpen*. de. Sept. 2017. URL: https://energieforschung.at/wp-content/uploads/sites/11/2020/12/heat-portfolio-D5.2.pdf (visited on 10/28/2022).

[12] M. Zogg. *Wärmepumpen*. July 2009. URL: https://www.zogg-engineering.ch/Publi/WP_ETH_Zogg.pdf.

[13] D. Liepsch, F. Bajic, and C. Steger. 7. Pumpen und Ventilatoren. In: *Energie-, Gebäude-, Versorgungstechnik*. DE GRUYTER, Dec. 2014, pp. 179–206. ISBN: 9783486727692. DOI: 10.1524/9783486769678.179. URL: https://www.degruyter.com/document/doi/10.1524/9783486769678.179/html (visited on 03/30/2026).

[14] A. Bürger, C. Zeile, M. Hahn, A. Altmann-Dieses, S. Sager, and M. Diehl. pycombina: An Open-Source Tool for Solving Combinatorial Approximation Problems Arising in Mixed-Integer Optimal Control. en. In: *IFAC-PapersOnLine* 53.2 (2020), pp. 6502–6508. ISSN: 24058963. DOI: 10.1016/j.ifacol.2020.12.1799. URL: https://linkinghub.elsevier.com/retrieve/pii/S2405896320324083 (visited on 03/30/2026).

[15] J. A. E. Andersson, J. Gillis, G. Horn, J. B. Rawlings, and M. Diehl. CasADi: a software framework for nonlinear optimization and optimal control. en. In: *Mathematical Programming Computation* 11.1 (Mar. 2019), pp. 1–36. ISSN: 1867-2949, 1867-2957. DOI: 10.1007/s12532-018-0139-4. URL: http://link.springer.com/10.1007/s12532-018-0139-4 (visited on 03/30/2026).

[16] A. Wächter and L. T. Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. en. In: *Mathematical Programming* 106.1 (Mar. 2006), pp. 25–57. ISSN: 0025-5610, 1436-4646. DOI: 10.1007/s10107-004-0559-y. URL: http://link.springer.com/10.1007/s10107-004-0559-y (visited on 03/30/2026).

[17] J. Wenzel and H. Schreiner. *pybind11 — Seamless operability between C++11 and Python*. 2022. URL: https://github.com/pybind/pybind11 (visited on 03/30/2026).

[18] S. Eser, T. Storek, F. Wüllhorst, S. Dähling, J. Gall, P. Stoffel, and D. Müller. A modular Python framework for rapid development of advanced control algorithms for energy systems. In: *Applied Energy* 385 (2025), p. 125496. ISSN: 0306-2619. DOI: https://doi.org/10.1016/j.apenergy.2025.125496. URL: https://www.sciencedirect.com/science/article/pii/S0306261925002260 (visited on 03/30/2026).

[19] Deutscher Wetterdienst (DWD). *Wetter und Klima - Deutscher Wetterdienst - Leistungen - Testreferenzjahre (TRY)*. URL: https://www.dwd.de/DE/leistungen/testreferenzjahre/testreferenzjahre.html (visited on 03/30/2026).