buildings

MDPI

Article

Heating Buildings Only When Needed: A Data-Driven Based Framework for Comfort-Aware HVAC Scheduling

Jevgenijs Telicko [1,*], Mihails Danilovs [2] and Andris Krumins [2]

[1] Institute of Numerical Modelling, University of Latvia, Jelgavas Street 3, LV-1004 Riga, Latvia
[2] Lafivents SIA, Kārļa Ulmaņa gatve 1B, LV-1004 Riga, Latvia; mihails@intellify.lv (M.D.); andris@lafivents.lv (A.K.)
* Correspondence: jevgenijs.telicko@lu.lv

Abstract

Optimal- heating-start strategies can reduce unnecessary heating, ventilation, and air conditioning (HVAC) operation while maintaining acceptable indoor temperatures at occupancy. However, forecasting models are often evaluated using aggregate prediction errors that may not reflect operational performance. This study compares four data-driven approaches: gated recurrent unit (GRU) and long short-term memory (LSTM), recurrent networks with different memory mechanisms; temporal convolutional network–GRU (TCN-GRU), combining temporal convolutions and recurrence; and extreme gradient boosting (XGBoost), using gradient-boosted decision trees. Using measurements from an occupied office building, these models were compared with four conventional optimal-start strategies using 24 h of historical inputs and a 6 h prediction horizon. Performance was assessed through temperature prediction errors, recovery-time errors, temperature-deficit exposure, and unnecessary heating runtime. Uncertainty was evaluated using paired day-level bootstrap resampling of 22 recovery events. XGBoost achieved the lowest temperature mean absolute error (MAE) (0.182 °C), while LSTM achieved the lowest recovery-time MAE (25 min) and unnecessary heating runtime (5.7 min/day). TCN-GRU achieved the lowest temperature-deficit exposure (4.9 K·min/day), despite a higher aggregate forecasting error. With considerable uncertainty in some comparisons, no model was consistently superior across objectives. These findings show that optimal-start model selection should consider operational trade-offs, not forecasting accuracy alone.

Keywords: building energy efficiency; data-driven modelling; HVAC scheduling; model predictive control; optimal start; thermal comfort

Check for updates

Academic Editor: Yimin Chen, Yunyang Ye

Received: 2 October 2026
Revised: 23 October 2026
Accepted: 26 October 2026
Published: 28 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Buildings account for a substantial share of global energy consumption, while indoor environmental conditions directly affect occupant comfort and well-being [1,2]. Improving heating, ventilation, and air-conditioning (HVAC) control is therefore an important means of reducing building energy demand without necessarily requiring major changes to the building envelope or equipment [3,4]. In office buildings, where occupancy generally follows a predictable schedule, reducing HVAC operation during unoccupied periods while restoring acceptable indoor temperatures before occupants arrive is a particularly relevant control objective [5,6].

A central challenge is determining when heating should be activated following a night-time temperature setback. The recovery time varies with the initial indoor temperature,

outdoor conditions, solar and internal gains, and the thermal response of the building. A fixed heating-start schedule must accommodate unfavourable conditions and may consequently activate heating earlier than necessary on other days. Conversely, starting too late risks failing to reach the target temperature when required. Optimal-start strategies seek to reduce this mismatch by adjusting the activation time to prevailing conditions [7–9].

Existing optimal-start approaches range from conventional adaptive rules to physics-based and data-driven models. Classical methods estimate recovery duration using a small number of variables and update their parameters from previous operation [8,9]. Physics-based and grey-box models represent building thermal dynamics explicitly, but their development and calibration can require substantial engineering effort [5,10,11]. Data-driven approaches provide an alternative by learning the relationship between operating conditions and thermal response from building management system measurements. Previous work has investigated neural networks, regression methods, and learning-based start-time controllers, demonstrating the potential of adaptive preconditioning and scheduling [12–15]. However, their relative advantages depend on the building, operating conditions, and evaluation objective.

An important unresolved issue concerns the relationship between forecasting accuracy and control-oriented performance. Indoor-temperature prediction models are frequently assessed using aggregate errors such as mean absolute error (MAE) or mean squared error (MSE). Such metrics characterize the quality of predicted temperature trajectories, but can be dominated by periods when indoor temperature changes little. Optimal heating-start decisions, in contrast, depend primarily on predicting the duration of thermal recovery and the time at which an acceptable temperature threshold is reached. Consequently, a model with low overall temperature error may not produce the most favourable heating-start estimate. Related distinctions between prediction quality and control performance have also been discussed in building model predictive control (MPC) research [16–18].

Previous studies have evaluated the energy and comfort-related benefits of adaptive HVAC scheduling [13–15,19]. Nevertheless, direct comparisons of different forecasting architectures and conventional optimal-start strategies using identical measured heating-recovery events and common operational indicators remain limited. Such comparisons are valuable because an optimistic estimate of recovery duration may be associated with a temperature deficit, whereas a conservative estimate may imply unnecessary early heating. These consequences cannot be fully characterized by aggregate forecasting MAE alone. Furthermore, the limited number of suitable recovery events in operational datasets makes uncertainty in model rankings an important consideration.

The aim of this study is to determine whether aggregate indoor-temperature forecasting accuracy is a reliable criterion for selecting predictive models for optimal heating-start scheduling. Using historical measurements from an occupied office building, four data-driven forecasting architectures, namely gated recurrent unit (GRU), long short-term memory (LSTM), temporal convolutional network–GRU (TCN-GRU), and extreme gradient boosting (XGBoost), are compared with four conventional optimal-start reference strategies. The assessment distinguishes among pointwise temperature prediction accuracy, heating-recovery duration estimation, estimated temperature-deficit exposure, and potentially unnecessary heating runtime. Model comparisons are supplemented by paired day-level uncertainty analysis.

The contribution of this work is a common, control-oriented evaluation of forecasting and conventional optimal-start strategies under the same observed recovery conditions. Rather than claiming energy savings from an implemented controller, the study investigates how forecasting-model rankings change when recovery-time and retrospective operational indicators are considered alongside conventional prediction errors. This provides evi-

dence for model selection according to application-specific comfort–runtime trade-offs and identifies the limitations of relying on forecasting accuracy alone.

2. Methods

2.1. Problem Formulation

We consider the short-horizon prediction of indoor air temperature in a multi-zone commercial building operated with night setback. Let $y_t \in \mathbb{R}$ denote the air temperature of the zone of interest at discrete time t, sampled at a uniform interval $\Delta = 10\,\text{min}$. Let $\mathbf{x}_t \in \mathbb{R}^{d_x}$ denote the vector of measured process and weather variables at time t, and let $\mathbf{u}_t \in \mathbb{R}^{d_u}$ denote the vector of zone temperature setpoints (the controllable inputs) at time t. Given a history of length $H = 144$ steps (24 h) and a planned setpoint schedule over a prediction horizon of $P = 36$ steps (6 h), the task is to learn a mapping

$$f_\theta : \underbrace{\mathbb{R}^{H\times d_x}}_{\text{history}} \times \underbrace{\mathbb{R}^{P\times d_u}}_{\text{future control}} \longrightarrow \underbrace{\mathbb{R}^{P}}_{\text{predicted trajectory}}, \qquad \hat{\mathbf{y}}_{t+1:t+P} = f_\theta(\mathbf{X}_{t-H+1:t}, \mathbf{U}_{t+1:t+P}), \tag{1}$$

where $\mathbf{X}_{t-H+1:t}$ stacks the H most recent feature vectors and $\mathbf{U}_{t+1:t+P}$ stacks the P future setpoint vectors. Conditioning the prediction on the future control sequence is what makes f_θ usable as the internal model of a model predictive controller: the same history can be evaluated under alternative setpoint schedules. The intended downstream application is *optimal start* — determining how early heating must be enabled so that a zone reaches its comfort setpoint by the beginning of occupancy. Accordingly, the quantity of practical interest is not the pointwise temperature error but the error in the predicted *time to comfort*, formalised in Section 2.10.

2.2. Model Choice Among Black-Box Architectures

A wide range of architectures has been proposed for time-series forecasting, from classical recurrent networks such as long short-term memory (LSTM) and Gated Recurrent Unit (GRU) models, through convolutional sequence models such as the temporal convolutional network (TCN), to Transformer-based architectures such as Informer [20–23]. Larger and more expressive architectures can capture more complex temporal dependencies, but their higher capacity and parameter count require correspondingly larger training sets in order to generalise rather than overfit [24]. This trade-off is particularly restrictive in building applications. As discussed above, the thermal dynamics of a building are not stationary: the weather differs from year to year, the thermal properties of the envelope change as the building ages, occupant preferences evolve, and the HVAC plant is re-commissioned or repaired. Older records therefore do not necessarily describe the current dynamics, so simply accumulating more history does not guarantee better accuracy and may even degrade it. In practice the usable training set is limited to a small number of heating seasons, which places the present problem firmly in the low-data regime and argues against attention-based architectures with state-of-the-art capacity. For these reasons, three time-series architectures of moderate capacity were selected for comparison, together with one non-sequential reference:

- **GRU.** A gated recurrent network with three gates per cell. GRU models have previously shown strong performance in comparable building energy forecasting tasks while requiring substantially fewer trainable parameters than attention-based architectures [21,23,25]. It is used here as the primary recurrent architecture.
- **LSTM.** The same topology with LSTM cells, which carry four gates and a separate cell state. The LSTM is the most popular recurrent architecture in the building-control literature [20]; including it with an identical topology isolates the effect of the gating mechanism.

- **TCN-GRU.** A hybrid architecture combining parallel temporal convolutional and recurrent branches. Both branches process the same augmented historical input, allowing the TCN to extract local and multi-scale temporal patterns while the GRU captures sequential dependencies. Their outputs are combined through a learned convolutional fusion layer. This architecture has shown promising potential in building-dynamics modelling [25].
- **XGBoost.** A gradient-boosted regression tree ensemble, included as a non-sequential baseline. It receives exactly the same information as the three sequential models, but with the time axis flattened. Any difference between XGBoost and the sequential models is therefore attributable to the treatment of the temporal structure and not to the choice of input features.

These four data-driven models are compared against the classical optimal-start formulations of Section 2.9.

2.3. Data Acquisition and Preprocessing

The experiments were conducted using an office building located in Riga, Latvia. The building has a known occupancy schedule for its office workers. It features a glass facade and is situated alongside a river (see Figure 1). Due to its construction and location, the indoor climate of the building may be influenced by both wind and solar radiation. The main characteristics of the building are summarized in Table 1.



Figure 1. Office building.

Table 1. Main technical parameters of the building HVAC system.

Parameter	Value
Building floor area	12,278 m^2
Total ventilation airflow	24,555 m^3/h
Installed cooling capacity	475 kW
Ventilation system	Central Air Handling Units (AHUs)
Cooling system	Water chillers + chilled ceiling system
Heating system	Two natural-gas boilers (800 kW heating capacity), radiators
Building control	Centralized Building Management System (BMS)

A building management system (BMS) was installed during the construction of the building. The BMS controls all centralized HVAC equipment and the climate conditions in individual rooms. Each room is equipped with a BMS controller and room panel, allowing users to adjust the temperature setpoint. The controller operates the room heating radiator during the heating season and the fan coil unit during the cooling season. The BMS enforces minimum and maximum temperature setpoints and uses a time schedule to activate comfort mode. This schedule is static and was previously modified manually by the facility manager. All BMS data points are available through native Building Automation and Control Network (BACnet), enabling the integration of a cloud-based solution for time-schedule adjustment. A dedicated BACnet controller was installed and programmed in the CODESYS programming language. A custom library was developed to read data from the BMS and transmit it to the cloud database. Raw records were logged at 1 min resolution and augmented with on-site meteorological observations. Records were aggregated to $\Delta = 10$ min by arithmetic mean within each bin, yielding 144 samples per day. Residual gaps, arising from transient communication losses, were filled by linear interpolation. This choice is consistent across all models and splits. The BMS exposes, for each zone z, an occupancy flag $o_t^z \in \{0, 1\}$, a daytime setpoint $s_t^{z,\mathrm{day}}$, and a night minimum $s_t^{z,\mathrm{night}}$, but not the setpoint actually in force. The effective setpoint was reconstructed as

$$u_t^z = \begin{cases} s_t^{z,\mathrm{day}}, & \bar{o}_t^z \geq 0.5, \\ s_t^{z,\mathrm{night}}, & \text{otherwise}, \end{cases} \tag{2}$$

where $\bar{o}_t^z \in [0, 1]$ is the mean of the binary occupancy flag over the 10 min aggregation bin. The threshold in Equation (2) is required because averaging a binary signal produces fractional values in bins that straddle a mode change; a strict equality test would leave such bins undefined and, after interpolation, would introduce spurious setpoint ramps that do not exist in the plant. Two cosine features encoding daily and weekly phases were appended:

$$c_t^{(\tau)} = \cos\left(\frac{2\pi\,\mathrm{ts}(t)}{\tau}\right), \qquad \tau \in \{86\,400\,\mathrm{s},\ 7 \cdot 86\,400\,\mathrm{s}\} \tag{3}$$

with $\mathrm{ts}(t)$ the POSIX timestamp of sample t. All variables were mapped to $[0, 1]$ by min–max scaling against *fixed physical ranges* $[\ell_j, h_j]$ rather than empirical statistics:

$$\tilde{v}_{j,t} = \frac{v_{j,t} - \ell_j}{h_j - \ell_j}. \tag{4}$$

The ranges (Table 2) were chosen from instrument specifications and engineering limits. Fixed ranges are preferred over statistics estimated from the training split because they are invariant to the split definition and can be applied unchanged to a deployed controller that encounters conditions absent from the training period. The denormalised prediction follows by inverting Equation (4).

The resulting input feature vector has $d_x = 23$ components: four plant-level measurements, ten zone-level measurements (temperature and valve position for five zones), two meteorological variables, two cyclic time features, and the five reconstructed setpoints. The control vector comprises the $d_u = 5$ zone setpoints, and the label is the single air temperature of the modelled zone ($d_y = 1$).

Table 2. Normalisation ranges $[\ell_j, h_j]$ applied to each variable group.

Variable Group	Unit	Range $[\boldsymbol{\ell}, \boldsymbol{h}]$
Outdoor air temperature (BMS)	°C	$[-25, 40]$
District heating supply temperature	°C	$[0, 90]$
Supply air temperature	°C	$[15, 35]$
Supply air flow rate	%	$[0, 100]$
Zone air temperature (×5)	°C	$[15, 35]$
Heating valve position (×5)	–	$[0, 1]$
Wind speed at 10 m	$\mathrm{m\,s^{-1}}$	$[0, 70]$
Direct normal irradiance	$\mathrm{W\,m^{-2}}$	$[0, 900]$
Effective setpoint (×5)	°C	$[15, 27]$
Cyclic time features (×2)	–	$[-1, 1]$

2.4. Window Construction and Data Splits

Training examples were generated by a sliding window with a stride of one over each split. For every admissible index t, the triple $(\mathbf{X}_{t-H+1:t}, \mathbf{U}_{t+1:t+P}, \mathbf{y}_{t+1:t+P})$ forms one example, so that consecutive windows overlap in $H + P - 1 = 179$ of their 180 samples. The available measurements were divided into three non-overlapping subsets according to calendar dates. The test dataset covered 2 January to 3 February 2025, while the training dataset covered 4 February to 2 April 2025. The validation dataset comprised observations from 13 January to 27 February 2026. This division was selected to ensure that model training, hyperparameter selection, and final performance evaluation were performed on distinct sets of observations.

The decision to use the subsequent year's measurements for validation rather than final testing was motivated by the limited number of suitable heating-recovery events during that period. Changes in occupant requirements and the associated operating schedules reduced the number of transitions satisfying the experimental selection criteria. Consequently, the earlier heating-season period was retained as the test dataset to provide a larger sample of relevant transition events.

The separation was applied before generating sliding-window samples to prevent the input or prediction windows from crossing dataset boundaries. A total of 22 eligible heating-recovery events were retained for the final operational evaluation. The resulting conclusions are conditional on the selected operating periods and should not be interpreted as evidence of generalization across independent buildings or multiple heating seasons.

2.5. Recurrent Model (GRU)

The recurrent architecture addresses a structural mismatch between its two inputs: the history has $H = 144$ steps whereas the control schedule has $P = 36$. Rather than padding or truncating, the control sequence is first encoded by a gated recurrent unit whose hidden width equals the history length,

$$\mathbf{g} = \mathrm{GRU}_H(\mathbf{U}_{t+1:t+P}) \in \mathbb{R}^H, \tag{5}$$

taking only the terminal hidden state. The resulting vector is reshaped to $\mathbb{R}^{H\times 1}$ and concatenated to the history along the feature axis,

$$\mathbf{Z} = [\, \mathbf{X}_{t-H+1:t} \,\|\, \mathrm{reshape}(\mathbf{g}) \,] \in \mathbb{R}^{H\times(d_x+1)}. \tag{6}$$

A second gated recurrent unit of width P collapses the augmented sequence to the prediction directly,

$$\hat{\mathbf{y}}_{t+1:t+P} = \mathrm{reshape}(\mathrm{GRU}_P(\mathbf{Z})) \in \mathbb{R}^{P\times 1}. \tag{7}$$

The control information is thus injected as a single additional channel whose value at history index k is the k-th component of the encoded schedule. We note that this channel carries no temporal alignment with the remaining d_x channels; it acts as a distributed, learned representation of the schedule rather than as a time-indexed signal. Training minimises the Huber loss over the horizon,

$$\mathcal{L}_\delta(e) = \begin{cases} \frac{1}{2}e^2, & |e| \le \delta, \\ \delta\left(|e| - \frac{1}{2}\delta\right), & |e| > \delta, \end{cases} \qquad e = \hat{y} - y, \tag{8}$$

which is quadratic near zero and linear in the tails, limiting the influence of the occasional large excursion produced by sensor artefacts. Optimisation uses AdamW with decoupled weight decay and gradient-norm clipping at 1.0. Batches are drawn in chronological order (`shuffle=False`) so that the temporal correlation structure within a batch mirrors deployment.

2.6. Long Short-Term Memory Variant (LSTM)

To separate the contribution of the gating mechanism from that of the surrounding architecture, a second recurrent model was constructed that is identical to Section 2.5 in every respect—input representation, control-encoding scheme, layer widths, output reshaping, loss, optimiser and training protocol—except that both recurrent units are replaced by long short-term memory cells.

2.7. Hybrid Convolutional–Recurrent Model (TCN-GRU)

The recurrent models of Sections 2.5 and 2.6 compress the entire H-step history into a single terminal state, and process it strictly sequentially. A temporal convolutional network (TCN) instead applies dilated causal convolutions in parallel across the sequence, giving every output position an explicit, fixed receptive field and a gradient path whose length is logarithmic rather than linear in H. Because the two inductive biases are complementary—recurrence favouring accumulated state, convolution favouring localised multi-scale patterns such as the shape of a warm-up ramp—we evaluate a hybrid that computes both and learns to combine them. The control-encoding stage is unchanged from Equations (5) and (6): a GRU of width H encodes the future setpoint schedule, and the encoding is appended to the history as one additional channel, producing $\mathbf{Z} \in \mathbb{R}^{H\times(d_x+1)}$ as in [25]. Both branches read this same tensor. The TCN is a stack of L residual blocks with exponentially increasing dilation b^ℓ, $\ell = 0, \dots, L-1$. Each block applies two dilated causal convolutions of kernel size κ with layer normalisation and rectified linear unit (ReLU) activation, adds a residual connection (through a 1×1 projection where channel counts differ), and contributes to a summed skip pathway. Causal padding ensures that the activation at position t depends only on positions $\le t$, so no future information leaks backwards within the history window. The depth is set automatically to the smallest L whose receptive field covers the full history,

$$L = \left\lceil \log_b\left(\frac{(H-1)(b-1)}{2(\kappa-1)} + 1 \right) \right\rceil, \qquad R = 1 + 2\,(\kappa-1) \sum_{\ell=0}^{L-1} b^\ell \;\ge\; H, \tag{9}$$

where R is the receptive field of the TCN stack. For $b = \kappa = 2$ and $H = 144$, the minimum required depth is $L_{\min} = 7$, corresponding to a receptive field of $R = 255$.

However, the implemented architecture uses $L = 8$ residual blocks, with dilations $1, 2, 4, 8, 16, 32, 64$, and 128, resulting in $R = 511$. Thus, the receptive field exceeds the historical input length of $H = 144$ samples. Only the activation at the final position is retained, giving $\mathbf{a} \in \mathbb{R}^F$ with $F = 64$ filters, which a dense layer maps to the horizon,

$$\mathbf{t} = \mathbf{W}_t\mathbf{a} + \mathbf{b}_t \in \mathbb{R}^P. \tag{10}$$

In parallel with the TCN branch, the augmented sequence $\mathbf{Z}$ is processed by a GRU layer with P hidden units, producing a vector in $\mathbb{R}^P$. The outputs of both branches are reshaped and concatenated into a tensor of shape $P \times 2 \times 1$.

A convolutional fusion layer with one filter and kernel size $(1, 2)$ combines the two representations at each forecast step. The resulting tensor is reshaped to obtain the final P-step indoor temperature forecast.

2.8. Gradient-Boosted Tree Model (XGBoost)

As a non-recurrent reference we train an independent gradient-boosted regressor for each step of the horizon. The sequential inputs are flattened and concatenated into a single design vector,

$$\boldsymbol{\phi}_t = [\, \mathrm{vec}(\mathbf{X}_{t-H+1:t}) \,\|\, \mathrm{vec}(\mathbf{U}_{t+1:t+P}) \,] \in \mathbb{R}^{Hd_x+Pd_u} = \mathbb{R}^{3492}, \tag{11}$$

and $P = 36$ separate models $\{g_{\theta_k}\}_{k=1}^{P}$ are fitted, with g_{θ_k} predicting the temperature at horizon step k:

$$\hat{y}_{t+k} = g_{\theta_k}(\boldsymbol{\phi}_t), \qquad k = 1, \ldots, P. \tag{12}$$

Each model is trained with the pseudo-Huber objective $\mathcal{L}_\delta^{\mathrm{ph}}(e) = \delta^2\big(\sqrt{1 + (e/\delta)^2} - 1\big)$, a twice-differentiable surrogate for Equation (8) required by the second-order boosting algorithm, with $\delta = 0.5$ held fixed to match the recurrent baseline. The per-step formulation imposes no smoothness across the horizon: each $\hat{y}_{t+k}$ is produced independently, and any temporal coherence in the predicted trajectory is a consequence of the shared input rather than of the model structure.

2.9. Classical Optimal-Start Baselines

Four conventional optimal-start strategies were evaluated as reference methods. Their initial parameters were calibrated on the training dataset. Unlike the data-driven forecasting models, these strategies directly estimate the required heating-recovery duration rather than predicting a temperature trajectory. Let ϑ denote the measured warm-up duration (Section 2.10), $\Delta T_z = u^\star - \varepsilon - y_{t_0}$ the zone temperature deficit at the heating-start instant t_0, where $u^\star$ is the daytime setpoint and ε the comfort tolerance, and $\Delta T_{\mathrm{oa}} = u^\star - T_{\mathrm{oa}}(t_0)$ the indoor–outdoor driving temperature difference.

Constant. $\hat{\vartheta} = \bar{\vartheta}_{\mathrm{train}}$, the mean training warm-up duration. This is the floor that any condition-dependent method must outperform to justify its complexity.

Guideline 36. Following the American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) Guideline 36 optimal-start sequence [26], $\hat{\vartheta} = \min(i \cdot \Delta T_z,\ \vartheta_{\max})$, with the coefficient i [min/K] estimated by ordinary least squares through the origin,

$$i = \frac{\sum_n \Delta T_{z,n}\, \vartheta_n}{\sum_n \Delta T_{z,n}^2}. \tag{13}$$

Guideline 36, adaptive. Identical in form, but the coefficient continues to adapt during evaluation, as a deployed self-learning controller would. After each observed warm-up the coefficient is updated by exponential smoothing,

$$i \leftarrow (1 - \alpha)\, i + \alpha \frac{\vartheta_n}{\Delta T_{z,n}}, \qquad \alpha = \tfrac{1}{3}, \tag{14}$$

applied strictly *after* the prediction for day n has been recorded, so that no future information enters any prediction. This baseline is therefore an online learner and is reported separately from the fixed-parameter methods.

Seem regression. A multivariate form incorporating the outdoor driving force through an interaction term,

$$\hat{\vartheta} = a_0 + a_1 \Delta T_z + a_2 \Delta T_z \Delta T_{oa}, \tag{15}$$

with coefficients obtained by least squares on the training transitions.

All variants are clipped to $[0, \vartheta_{max}]$ with $\vartheta_{max} = P\Delta = 360$ min, so that no baseline is credited with information beyond the horizon available to the learned models. Calibration days with non-positive deficit ($\Delta T_z \leq 0.1$ K), missing outcome, or warm-up exceeding 600 min are excluded from the fit as degenerate.

2.10. Evaluation on Night–Day Transitions

Aggregate temperature prediction errors over all sliding windows are strongly influenced by quasi-steady-state operation. Therefore, model performance was additionally evaluated using heating-recovery events associated with the transition from nighttime setback to daytime heating operation.

The BMS provides a binary mode flag that controls activation of the daytime heating regime. Although this signal is labelled as an occupancy flag in the BMS, its transition from 0 to 1 represents heating activation, not the actual arrival of occupants.

Let o_t denote the forward-filled binary heating activation signal. A candidate event was identified when $o_{t-1} = 0$ and $o_t = 1$. To exclude short-duration switching events, the signal was required to remain active for at least 60 min.

For each calendar day, the first valid transition occurring between 00:00 and 13:00 local time was selected as the heating activation instant t_0. The subsequent temperature trajectory was analyzed to determine the time required to reach the daytime comfort threshold.

Comfort is declared reached at the first crossing of the threshold $u^\star - \varepsilon$ with $\varepsilon = 0.2$ K. A uniform tolerance of $\varepsilon = 0.2$ K was applied to the daytime temperature setpoint when determining comfort attainment. Therefore, both historically measured and model-predicted temperature trajectories were evaluated against the same effective threshold $u^\star - \varepsilon$.

This tolerance shifts the temperature threshold slightly below the nominal setpoint and is applied identically to all models and reference measurements, ensuring a consistent comparison of recovery durations. Because the signal is sampled at $\Delta = 10$ min, the crossing is refined by linear interpolation between the bracketing samples: if k is the first index with $y_k \geq u^\star - \varepsilon$, then

$$\vartheta = \Delta\left(k - 1 + \frac{u^\star - \varepsilon - y_{k-1}}{y_k - y_{k-1}}\right), \tag{16}$$

clipped to $[k-1, k]\Delta$. The same operator is applied to the measured trajectory, yielding ϑ, and to each model's predicted trajectory, yielding $\hat{\vartheta}$. The per-event signed error is

$$e_n = \hat{\vartheta}_n - \vartheta_n. \tag{17}$$

A negative error ($e_n < 0$) indicates underestimation of the required recovery duration, whereas a positive error ($e_n > 0$) indicates overestimation. In optimal heating-start applications, these errors have asymmetric operational consequences. Underestimation may result in delayed comfort attainment, whereas overestimation may lead to premature heating activation and additional heating runtime.

Over N evaluated transitions we report the mean absolute error MAE $= \frac{1}{N}\sum_n |e_n|$. For forecasts without a threshold crossing within the prediction horizon, the recovery-time MAE was assigned a penalty of 360 min. For the operational indicators, the predicted recovery duration was set to the maximum prediction horizon of 360 min.

2.11. Operational Performance Metrics

In addition to recovery-time prediction MAE, two operational indicators were evaluated to characterize the consequences of underestimating or overestimating the required heating-recovery duration.

For each evaluated event, the historically measured indoor-temperature trajectory was used as the reference for determining the actual time required to reach the daytime comfort threshold. The predicted recovery duration obtained from each model was then compared with this measured recovery duration.

The signed recovery-time prediction error was defined in Equation (17) as $e_{m,n} = \hat{\vartheta}_{m,n} - \vartheta_n$, where m denotes the evaluated model and n the heating-recovery event.

Positive errors indicate an overestimation of recovery duration, which may lead to premature heating activation in an optimal-start application. The corresponding unnecessary heating runtime proxy was defined as

$$R_{m,n} = \max(0, \hat{\vartheta}_{m,n} - \vartheta_n). \tag{18}$$

Negative errors indicate that the model underestimates the required recovery duration. In this case, the model predicts comfort attainment before the historically observed temperature reaches the required threshold.

To quantify this discrepancy, the temperature-deficit exposure was calculated using the historically measured indoor temperature trajectory:

$$E_{m,n} = \begin{cases} \int_{t_{0,n}+\hat{\vartheta}_{m,n}}^{t_{0,n}+\vartheta_n} \max(0, u_n^\star - \varepsilon - y_n(t))\, dt, & \hat{\vartheta}_{m,n} < \vartheta_n, \\ 0, & \hat{\vartheta}_{m,n} \geq \vartheta_n. \end{cases} \tag{19}$$

Here, $t_{0,n}$ is the recorded heating activation time, $u_n^\star$ is the daytime temperature setpoint, and $y_n(t)$ is the historically measured indoor temperature during recovery event n.

The resulting exposure, expressed in K·min, accounts for both the magnitude and duration of the measured temperature deficit during the period between predicted and observed comfort attainment.

Both indicators were calculated separately for each model and heating-recovery event. Cumulative values were obtained by summing the event-level metrics, while mean values were calculated over the $N = 22$ evaluated events.

These indicators represent retrospective, measurement-based proxies for the operational consequences of recovery-time prediction errors. They do not constitute direct measurements of energy consumption or occupant thermal discomfort under alternative heating-control strategies.

2.12. Statistical Uncertainty and Paired Comparisons

Statistical uncertainty was assessed using a nonparametric bootstrap at the calendar-day level. The test dataset comprised $N = 22$ heating-recovery events, with all eight algorithms evaluated on the same days.

For each of 20,000 bootstrap iterations, 22 days were sampled with replacement. Identical sampling indices were applied to all models, preserving the paired structure of the observations. Activation-time MAE, mean daily temperature-deficit exposure, and mean unnecessary heating runtime were recalculated for each resample.

Two-sided 95% confidence intervals (CIs) were obtained from the 2.5th and 97.5th percentiles of the bootstrap distributions. For pairwise comparisons, differences between the mean metric values of two models were calculated within each bootstrap iteration.

Paired randomization tests based on sign-flipping the daily differences were additionally performed. To account for multiple comparisons, the resulting p-values were adjusted using Holm's method across all 28 unique model pairs, separately for each metric.

Given the limited number of events and potential temporal dependence between consecutive days, the statistical results were interpreted cautiously. The analysis quantifies uncertainty within the investigated operating conditions and does not establish generalizability to other buildings or heating seasons.

For the learned models, the mean absolute error of temperature—computed on the whole *validation* split—serves as the early stopping criterion and the hyperparameter selection criterion. Validation transition windows are identified automatically by four conditions: a setpoint step-up of at least 0.5 K relative to the end of the history; a setpoint stable to within 0.35 K after the first three horizon steps; an initial zone temperature at least 0.3 K below the target; and a measured rise of at least 0.3 K that attains the threshold within the horizon. Steady-state windows, for which the time to comfort is undefined, are thereby excluded.

3. Results and Experimental Evaluation

3.1. Experimental Protocol

All models predict the air temperature of one zone over a 6 h horizon from 24 h of history. Training and hyperparameter selection use the training and validation splits exclusively; the test split is evaluated once, for the selected configuration of each model family, and is not used at any point for selection.

All neural network models (GRU, LSTM, and TCN-GRU) were implemented using TensorFlow (version 2.15, Google LLC, Mountain View, CA, USA) within the Python programming environment (version 3.11, Python Software Foundation, Wilmington, DE, USA). The gradient-boosted tree models (XGBoost) were implemented using the XGBoost Python package (version 3.3.1)

3.2. Hyperparameter Search

For each neural network architecture a grid of $3 \times 2 \times 4 \times 3 = 72$ configurations was evaluated (Table 3). The architecture itself is fixed: the recurrent widths are determined by H and P, so the search covers the optimisation and regularisation regime rather than model capacity. Each candidate trains until early stopping triggers on the validation mean absolute error, with a patience of 150 epochs and a warm-up of 50 epochs before the criterion is activated. Weights are restored to the best-scoring epoch.

Table 3. Search space for the recurrent model. Fixed: Huber loss, AdamW, gradient clipping at 1.0, chronological batching.

Hyperparameter	Values
Learning rate	$\{1 \times 10^{-3}, 5 \times 10^{-4}, 1 \times 10^{-4}\}$
Weight decay	{0.01, 0.05}
Batch size	{50, 100, 200, 500}
Huber δ	{0.25, 0.50, 0.75}

For XGBoost a grid of $3 \times 3 \times 2 = 18$ configurations was evaluated over tree depth $\{3, 5, 7\}$, learning rate $\{0.01, 0.05, 0.1\}$ and ensemble size $\{200, 500\}$. Each configuration requires $P = 36$ independent fits, giving 648 fitted ensembles in total. The smaller grid reflects both the lower sensitivity of boosted trees to these settings and their substantially greater cost per configuration at 3492 input features.

3.3. Obtained Values

In Table 4, we can see the normalized mean squared error (MSE) and mean absolute error (MAE) of temperature forecasts produced by the different models on the test dataset. It should be noted that the evaluation in this table is performed across all windows in the test set; in other words, the errors are not evaluated specifically on samples in which the building changes its thermal state. First, it can be observed that, at first glance, all models exhibit relatively small errors. Even the model with the largest error, TCN-GRU, has a mean absolute error below 0.4 °C. This value is close to the resolution of the temperature sensor (0.25 °C). Based solely on this information, it might appear that all models provide nearly perfect predictions. However, these small error values are partly explained by the fact that the test dataset contains many segments in which the temperature remains relatively stable, making such situations easier for the models to predict.

Table 4. Test-set MAE and MSE of temperature forecasts produced by the evaluated models.

Model	MSE (°C^2)	MAE (°C)
GRU	0.095	0.230
LSTM	0.102	0.235
XGBoost	**0.066**	**0.182**
TCN-GRU	0.194	0.359

Table 4 shows that XGBoost achieved the lowest prediction errors, with an MSE of 0.066 °C^2 and an MAE of 0.182 °C. Compared with GRU, the second-best model according to both metrics, XGBoost reduced MSE by approximately 30.5% and MAE by approximately 20.9%.

Nevertheless, these results characterize prediction accuracy across the complete test dataset, including periods of nearly constant indoor temperature, and do not necessarily reflect performance during heating recovery.

However, if we look at Table 5, which presents the mean absolute error of the activation time, the maximum error leading to delayed activation and, consequently, failure to reach comfortable conditions when occupants arrive at the office (denoted as Max. late), and the maximum error associated with excessively early activation, which leads to inefficient operation (denoted as Max. early), the ranking of the leading models may change. Table 5 shows that, on average, the activation time predicted by the LSTM model is the closest to the actual activation time in this comparison. It can also be observed that the evaluated data-driven models have a lower mean absolute error of the building transition time from the night mode to the day mode compared with the classical baseline methods. To assess uncertainty in recovery-time prediction accuracy, three data-driven models were selected for paired comparison: XGBoost (lowest pointwise temperature MAE), LSTM (lowest recovery-time MAE), and TCN-GRU (highest pointwise temperature MAE).

Table 5. Recovery-time prediction errors across 22 heating-recovery events.

Model	MAE (min)	Max. Late (min)	Max. Early (min)	MAE 95% CI (min)
GRU	28	123	80	[15.2, 42.8]
LSTM	**25**	150	54	[11.4, 41.2]
TCN-GRU	41	**102**	101	[31.0, 51.8]
XGBoost	31	109	60	[22.1, 42.1]
G36	33	179	30	[18.7, 51.2]
G36 adaptive	35	162	81	[20.9, 52.6]
Seem	33	181	43	[21.7, 50.0]
Constant	42	237	**23**	[21.0, 67.5]

Paired day-level bootstrap analysis ($N = 22$, 20,000 resamples) yielded recovery-time MAE differences of −6.72 min (95% CI: [−13.56, 0.71]) for LSTM versus XGBoost, −16.27 min ([−27.30, −3.89]) for LSTM versus TCN-GRU, and −9.55 min ([−17.25, −1.49]) for XGBoost versus TCN-GRU.

The latter two intervals excluded zero, whereas the interval for LSTM versus XGBoost included zero. These results provide nominal evidence of lower recovery-time MAE for LSTM and XGBoost than for TCN-GRU, although the intervals were not adjusted for multiple comparisons and the number of evaluated events was limited.

Another observation is that, in this test, the smallest maximum duration of discomfort (i.e., a situation in which the daytime temperature should have been reached but was not achieved due to an inaccurate prediction) is obtained with the TCN-GRU model. This duration is approximately one-third shorter than that of LSTM. This suggests that this criterion does not directly correlate with the results presented in Table 4. Several conventional baseline methods exhibited smaller maximum positive recovery-time prediction errors than the data-driven models. Under the adopted sign convention, positive errors correspond to overestimation of the required recovery duration and may result in premature heating activation.

Consequently, smaller positive errors indicate less extreme overestimation in the evaluated events, rather than a larger safety margin. However, maximum errors alone are insufficient to characterize the overall conservativeness of a strategy, which also depends on the frequency and distribution of positive and negative prediction errors.

If we examine Figure 2, we can see how the prediction error of the transition time varies from day to day across the different models. Figure 2 shows that there are several outlier days in the test month (e.g., 13 January) on which all models exhibit their largest errors. This is most likely caused by the unusually low nighttime temperature on that day, which resulted in a longer time being required for the building to reach the target temperature. The information presented in Figure 2 indicates that maximum and minimum error values are poor indicators for selecting a model, as individual days, such as 13 January, can have a substantial influence on these values.

For reference, Table 6 shows the average, maximum, and minimum time required for the zone to transition from the night mode to the day mode across the validation and test sets.

Based on the information presented in Figure 3, it can be observed that the TCN-GRU model has the lowest total duration across all temperature ranges, which indicates that using this model would result in more comfortable conditions. This finding also does not correlate with the MSE and MAE metrics. It is also important to note that a small temperature deviation may cause discomfort that is barely noticeable to occupants, whereas a larger temperature deviation is likely to result in more noticeable discomfort. At the same time, the level of discomfort may also depend on its duration. Ultimately, the selection of the model providing the most comfortable conditions is likely to depend on the specific requirements and preferences of the users.

Table 6. Measured heating-recovery duration in the validation and test datasets (min).

Zone	Validation	Test
Average	97	122
Min	35	70
Max	210	330

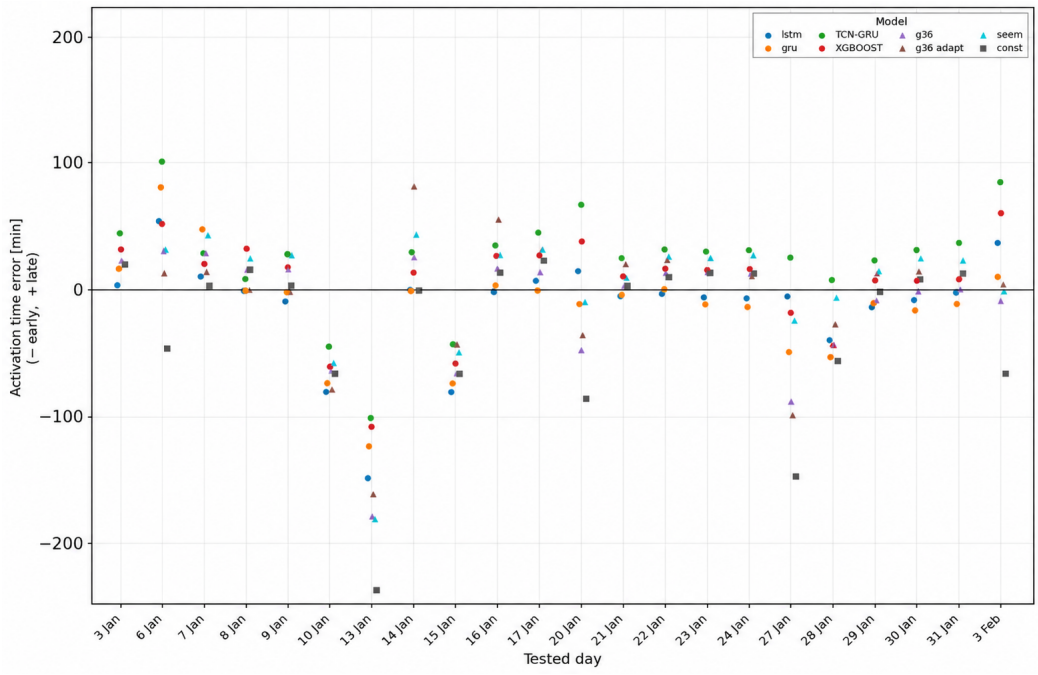


Figure 2. Models activation error across all tested days.

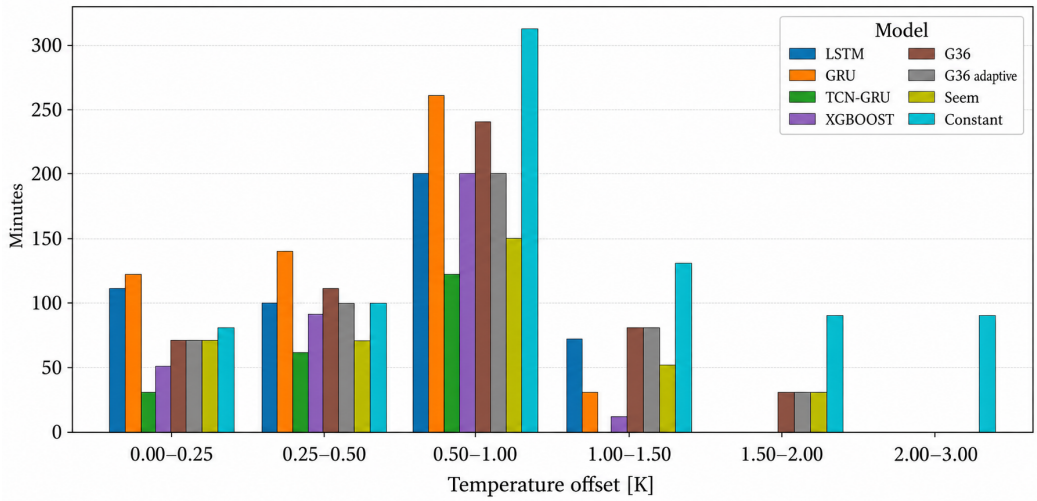


Figure 3. Temperature offset describing discomfort of all models across different temperature intervals in the test set.

Assessing the energy consumption associated with different predictive models is challenging because the relationship between heating duration and energy consumption is not necessarily linear. In particular, the thermal response of the building depends on whether the indoor temperature approaches a quasi-steady state during the nighttime setback period or whether heating resumes before such a state is reached. Consequently, differences in heating activation time cannot be directly translated into proportional differences in energy consumption.

For the investigated building, the average gas consumption during nighttime operation was approximately 20% lower than during daytime operation. Although this difference suggests potential energy savings from extending the nighttime setback period, it does not provide a sufficient basis for accurately estimating the energy savings associated with individual models. Nevertheless, assuming a simplified relationship between heating

duration and energy consumption, the timing of setpoint attainment can be used as an indirect indicator of potential energy inefficiency.

To assess the operational consequences of recovery-time prediction errors, two complementary indicators were considered. Unnecessary heating runtime was estimated from positive recovery-time prediction errors, representing the potential duration of premature heating activation.

Temperature-deficit exposure was calculated for negative recovery-time prediction errors by integrating the measured temperature shortfall relative to the effective comfort threshold over the interval between predicted and observed comfort attainment. Both metrics represent retrospective operational estimates rather than directly measured outcomes under alternative control strategies.

Both metrics are presented in Figure 4 to illustrate the trade-off between premature setpoint attainment and thermal discomfort across the evaluated models. While the results provide a comparative assessment of the models' operational performance, they should not be interpreted as a direct quantification of energy savings.

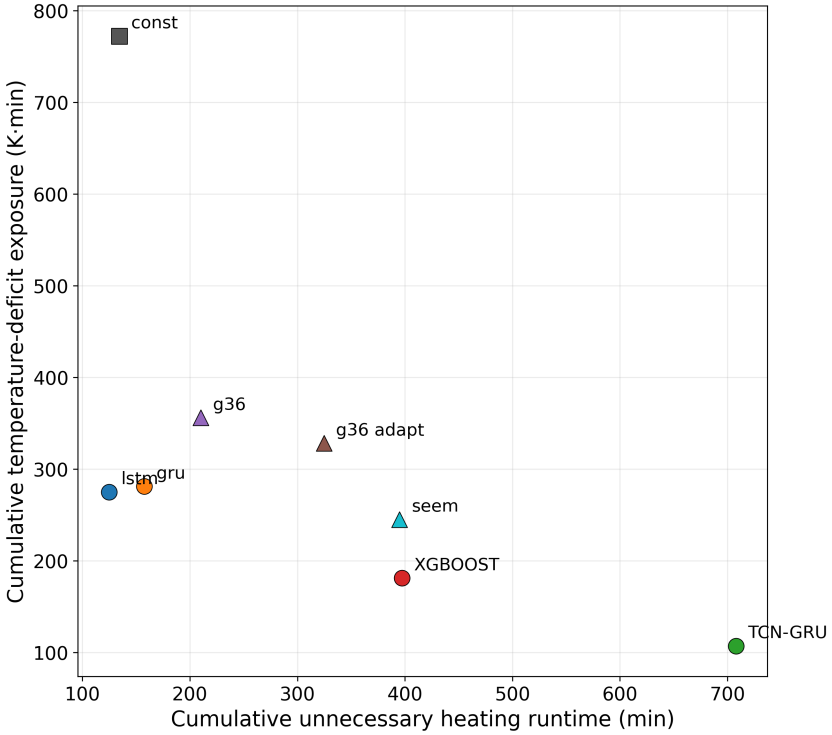


Figure 4. Trade-off: cumulative unnecessary heating runtime vs. cumulative temperature-deficit exposure.

The operational indicators were aggregated over the $N = 22$ evaluated heating-recovery events. Figure 4 presents cumulative values over the entire evaluation period, whereas Table 7 reports the corresponding mean values per evaluated event, together with their 95% bootstrap confidence intervals. Therefore, the cumulative values shown in Figure 4 are $N = 22$ times the unrounded means reported in Table 7.

Figure 4 further illustrates the difficulty of identifying a single best-performing model, as none of the evaluated models simultaneously minimizes both operational metrics. Instead, different models exhibit advantages depending on the optimization objective. From an operational efficiency perspective, LSTM achieves the lowest cumulative unnecessary heating runtime of 125 min over the test period. In contrast, TCN-GRU achieves the lowest cumulative temperature-deficit exposure of 107 K·min, indicating the smallest overall

thermal discomfort. These results highlight the trade-off between minimizing unnecessary heating and maintaining the required indoor temperature.

Table 7. Mean daily temperature-deficit exposure and unnecessary heating runtime with 95% bootstrap confidence intervals ($N = 22$ daily transitions).

	Exposure (K·min/Day)		Unnecessary Runtime (min/Day)	
Model	**Mean**	**95% CI**	**Mean**	**95% CI**
LSTM	12.5	[2.2, 26.4]	**5.7**	[1.1, 11.8]
GRU	12.8	[4.8, 22.8]	7.2	[0.8, 16.1]
TCN-GRU	**4.9**	[0.0, 11.9]	32.2	[22.4, 42.9]
XGBoost	8.3	[1.6, 17.0]	18.1	[11.6, 25.0]
G36	16.2	[3.9, 33.6]	9.6	[5.3, 14.0]
G36 adaptive	14.9	[2.6, 31.4]	14.8	[7.7, 23.7]
Seem	11.1	[1.0, 28.4]	18.0	[11.9, 23.8]
Constant	35.1	[9.4, 68.9]	6.1	[3.2, 9.3]

Table 7 presents the mean daily temperature-deficit exposure and unnecessary heating runtime for each model, together with their corresponding 95% confidence intervals (CIs), estimated using paired day-level bootstrap resampling of the 22 daily transitions. The results reveal substantial overlap between the confidence intervals of several models. For example, LSTM and GRU exhibit nearly complete overlap in both metrics, suggesting similar operational performance. However, overlapping confidence intervals alone do not establish statistical equivalence or the absence of a statistically significant difference.

More pronounced differences can be observed between certain models. For instance, TCN-GRU achieves a substantially lower mean daily temperature-deficit exposure of 4.9 K·min/day compared with 35.1 K·min/day for the constant activation-time baseline, representing an approximately 86% reduction. The corresponding 95% CIs range from 0.0 to 11.9 K·min/day for TCN-GRU and from 9.4 to 68.9 K·min/day for the constant baseline, showing only limited overlap. Nevertheless, the relatively wide intervals indicate considerable uncertainty in the estimated mean values, particularly for the constant baseline. Furthermore, the overlap of individual model confidence intervals cannot establish the statistical significance of their differences.

To examine the operational trade-off in greater detail, LSTM and TCN-GRU were selected for focused paired comparisons. These models achieved the lowest observed unnecessary heating runtime and temperature-deficit exposure, respectively.

Their comparisons with the remaining models are presented in Figure 5. The analysis preserves pairing between observations from the same evaluation days. The displayed confidence intervals are unadjusted percentile bootstrap intervals, while statistical significance is assessed separately using Holm-adjusted paired randomization tests.

For all comparisons, differences were calculated as the metric value of the reference model (LSTM or TCN-GRU) minus that of the comparison model. Negative values therefore favour the reference model, whereas positive values favour the comparison model. Gray indicates that the interval crosses zero, which in turn means there is no clear advantage for either of the compared models.

The paired comparison between LSTM and GRU supports the observation of similar performance. The mean difference in temperature-deficit exposure was only -0.3 K·min/day, with a 95% CI of $[-5.3,\ 4.9]$ K·min/day. LSTM also achieved a slightly lower unnecessary heating runtime than GRU, although the confidence interval for this difference included zero. Therefore, the available data do not provide sufficient evidence to establish a statistically significant performance difference between these models.

In contrast, the comparison between LSTM and TCN-GRU reveals a pronounced trade-off between thermal comfort and unnecessary heating runtime. TCN-GRU achieved a mean daily temperature-deficit exposure 7.6 K·min/day lower than LSTM, with an unadjusted paired 95% CI of [2.1, 14.6] K·min/day for the difference calculated as LSTM minus TCN-GRU. This difference remained statistically significant after Holm correction (adjusted $p = 0.004$), providing evidence that TCN-GRU achieves lower temperature-deficit exposure under the investigated conditions. However, this improvement was accompanied by considerably longer unnecessary heating runtime. Compared with TCN-GRU, LSTM reduced unnecessary runtime by approximately 26.5 min/day, with an unadjusted paired 95% CI of [−32.7, −20.2] min/day. Thus, TCN-GRU achieved lower estimated temperature-deficit exposure, whereas LSTM offered an advantage in limiting unnecessary heating runtime.

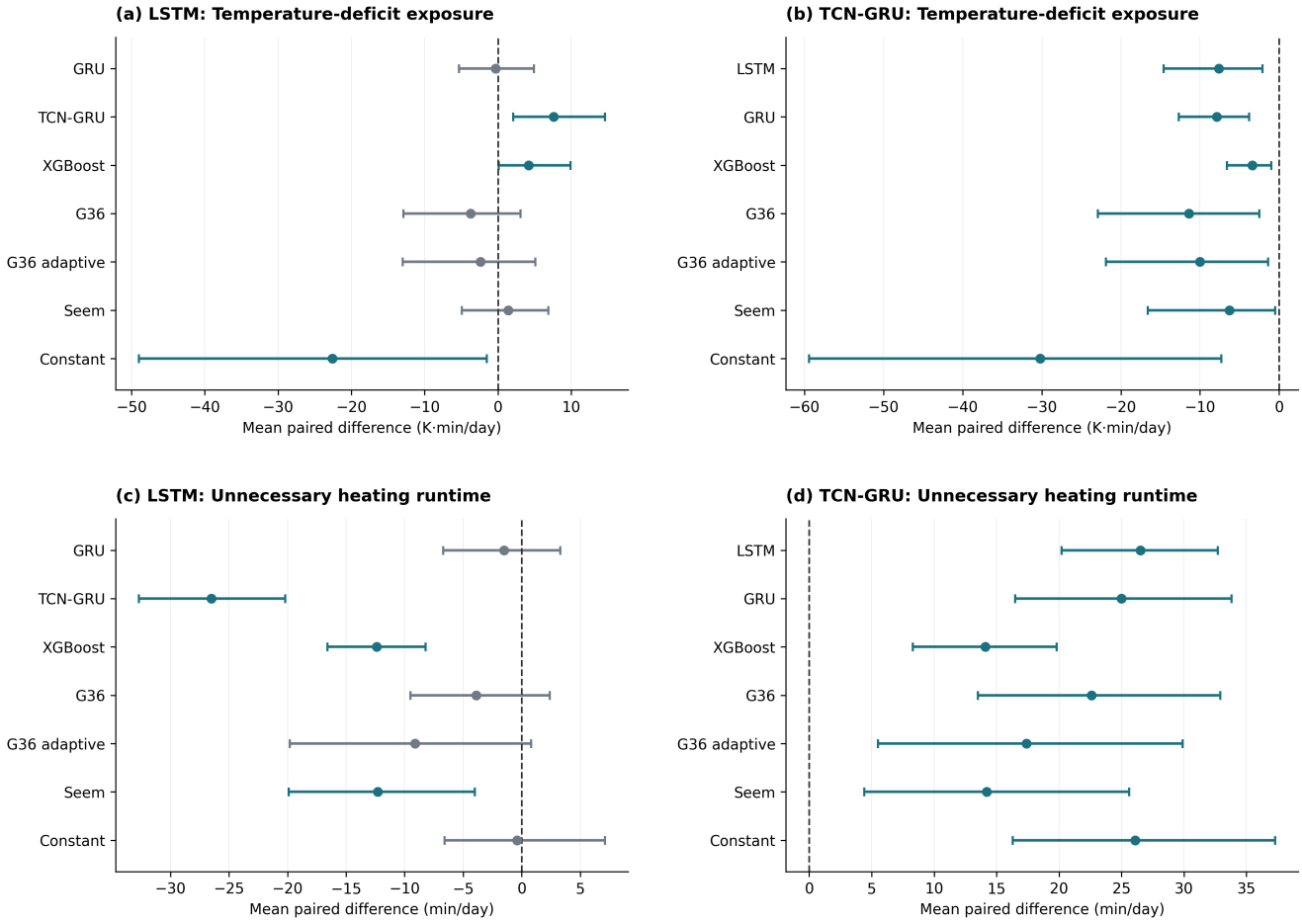


Figure 5. Paired day-level comparisons of operational outcomes of TCN-GRU and LSTM against other models.

A similar trade-off can be observed when comparing TCN-GRU with the constant activation-time baseline. TCN-GRU achieved a 30.2 K·min/day lower mean temperature-deficit exposure, with an unadjusted paired 95% CI of [−59.4, −7.3] K·min/day. However, this difference did not remain statistically significant after Holm correction across all 28 model pairs (adjusted $p = 0.192$). At the same time, TCN-GRU resulted in approximately 26.1 additional minutes of unnecessary heating runtime per day compared with the constant baseline, with an unadjusted 95% CI of [16.3, 37.3] min/day. These results suggest that the constant activation-time strategy limits premature heating but may expose occupants to greater thermal discomfort, whereas TCN-GRU prioritizes comfort at the expense of additional heating time.

Overall, the results demonstrate that predictive accuracy alone is insufficient to identify the most suitable model for optimal heating activation. Model selection depends on the relative importance assigned to thermal comfort and operational efficiency. TCN-GRU achieves the lowest observed temperature-deficit exposure, although this advantage is accompanied by substantially increased unnecessary heating runtime. Conversely, LSTM achieves the lowest unnecessary runtime while maintaining relatively moderate temperature-deficit exposure. Consequently, neither model can be considered universally superior across both objectives. These findings emphasize the importance of evaluating predictive models using operationally relevant metrics and explicitly accounting for the comfort–efficiency trade-off when designing building heating control strategies.

More complex architectures may require larger datasets for effective training, which could explain why, for example, TCN-GRU exhibited the highest MSE and MAE in the initial evaluation, which is not consistent with some other studies [25].

4. Discussion

4.1. Forecasting Accuracy Versus Operational Performance

The results indicate that conventional temperature prediction metrics are not sufficient to characterize the suitability of forecasting models for optimal heating-start applications. Although XGBoost achieved the lowest pointwise temperature prediction error, LSTM produced the lowest observed recovery-time MAE, while TCN-GRU achieved the lowest temperature-deficit exposure. This difference in model rankings demonstrates that prediction performance depends on the evaluation objective.

One possible explanation is the disproportionate contribution of quasi-steady-state observations to aggregate prediction-error metrics. During these periods, indoor temperature changes relatively slowly, and most forecasting models can reproduce the observed behaviour with comparatively low error. In contrast, optimal heating activation depends primarily on predicting the transient thermal response and the time required to reach the target temperature.

The results therefore support a distinction between models that reproduce typical temperature trajectories accurately and models that provide favourable control decisions. This distinction is particularly relevant to data-driven optimal-start control, where errors of similar magnitude may have different operational consequences depending on their direction and timing.

4.2. Comfort–Efficiency Trade-Off

The operational evaluation revealed that none of the investigated models performed best across all considered metrics. TCN-GRU achieved the lowest observed mean daily temperature-deficit exposure, while LSTM required substantially less unnecessary heating runtime.

This trade-off reflects the asymmetric consequences of recovery-time prediction errors. Overestimating the required recovery time may lead to premature heating, whereas underestimating it may result in insufficient indoor temperatures at the beginning of occupancy. Consequently, minimizing the mean absolute recovery-time error does not necessarily minimize either operational consequence independently.

The paired statistical analysis supports the importance of distinguishing between these objectives. In particular, the difference in mean daily temperature-deficit exposure between LSTM and TCN-GRU remained statistically significant after correction for multiple comparisons. However, the relatively wide confidence intervals for several other comparisons indicate that model rankings should not be generalized without considering statistical uncertainty.

From a practical perspective, model selection should therefore depend on the relative costs associated with delayed comfort attainment and premature heating. Buildings with strict requirements for indoor-temperature availability may favour conservative activation strategies, whereas buildings with greater flexibility may prioritize reductions in unnecessary heating operation.

4.3. Relationship to Previous Research

Previous studies have demonstrated the potential of data-driven optimal-start control to reduce HVAC energy consumption while maintaining thermal comfort. Esrafilian-Najafabadi and Haghighat [13] investigated neural network-based preconditioning, highlighting trade-offs between energy consumption, thermal comfort, and peak demand. Similarly, Park and Kim [14] reported 6.5–7.6% annual energy savings using supervised learning and iterative control, while Dai et al. [15] achieved 5.1–17.8% reductions in daily precooling energy consumption using reinforcement-learning-assisted control. These studies primarily evaluated the effectiveness of specific control strategies rather than comparing the operational consequences of different forecasting architectures under identical conditions.

The relationship between prediction accuracy and control performance was further investigated by Langtry et al. [17], who demonstrated that simpler forecasting approaches can achieve competitive accuracy in building MPC applications. However, their investigation focused on energy-system forecasting and MPC performance rather than heating-recovery prediction and optimal-start timing.

The present study complements these findings by comparing four forecasting architectures and four conventional optimal-start strategies using the same measured heating-recovery events. The results demonstrate that the model with the lowest temperature prediction error does not necessarily achieve the most favourable operational performance: XGBoost achieved the lowest pointwise temperature MAE, LSTM minimized observed unnecessary heating runtime, and TCN-GRU achieved the lowest estimated temperature-deficit exposure. Furthermore, paired bootstrap analysis quantified uncertainty in the model comparisons, highlighting the limitations of rankings based on a small number of events.

These findings emphasize the importance of control-oriented, uncertainty-aware evaluation when selecting forecasting models for optimal heating-start applications. Nevertheless, the retrospective nature of the analysis requires prospective validation before conclusions about actual energy savings or occupant comfort improvements can be established.

4.4. Limitations

Several limitations should be considered when interpreting the results.

First, the evaluation was based on measurements from a single office building and a limited number of heating-recovery events. Although the dataset captured different outdoor conditions and operating states, the statistical analysis involved only 22 eligible transitions. This limits the precision of estimated performance differences and the generalizability of the model rankings to other heating seasons, building types, or HVAC configurations.

Second, the experimental comparison was retrospective rather than a full closed-loop validation of all candidate controllers. The operational metrics therefore represent estimated consequences of alternative heating activation decisions rather than directly measured outcomes following their implementation in the building. Depending on the reconstruction procedure, changes in initial indoor temperature, weather conditions, or heating-system operation may affect the actual recovery trajectories.

Third, unnecessary heating runtime was used as a proxy for operational inefficiency. The relationship between additional heating duration and actual fuel or thermal-energy

consumption depends on heating-system efficiency, thermal losses, available heating capacity, and operating conditions. Consequently, the observed differences in runtime cannot be translated directly into equivalent energy savings.

Fourth, thermal discomfort was represented using temperature-deficit exposure relative to a predefined temperature setpoint. Although this indicator accounts for both the magnitude and duration of temperature deficits, it does not directly measure occupant thermal sensation or satisfaction. Actual thermal comfort depends on additional factors, including mean radiant temperature, air movement, humidity, clothing, and metabolic activity.

Fifth, the evaluated forecasting approaches depend on the quality and representativeness of the available building measurements. Changes in HVAC configuration, control parameters, occupancy schedules, equipment condition, or sensor calibration may modify the relationship between the inputs and the thermal response. Such changes may require retraining or recalibration before the models can be applied reliably.

Finally, the use of calendar-day bootstrap resampling assumes that daily transition events provide a sufficiently representative basis for estimating uncertainty. Temporal dependencies and shared weather patterns between consecutive days may reduce the effective sample size. The confidence intervals and hypothesis tests should therefore be considered approximate, particularly given the limited number of events.

Future research should extend the evaluation to multiple buildings, zones, and heating seasons, investigate the robustness of the model rankings under changing operating conditions, and validate the predicted operational outcomes through prospective closed-loop experiments. Integrating measured heating-energy consumption and additional thermal comfort indicators would also enable a more complete evaluation of the relationship between model selection, energy efficiency, and occupant comfort.

5. Conclusions

This study investigated the relationship between indoor-temperature forecasting accuracy and the operational performance of optimal heating-start strategies. Four data-driven forecasting architectures were evaluated: GRU and LSTM, representing recurrent neural networks with different memory mechanisms; TCN-GRU, combining convolutional feature extraction with recurrent processing; and XGBoost, using gradient-boosted decision trees instead of recurrent processing. These models were compared with four conventional optimal-start strategies using measured data from an occupied office building.

The results demonstrate that architectural differences do not translate into a consistent performance ranking across evaluation objectives. XGBoost, the tree-based model, achieved the lowest pointwise temperature MAE (0.182 °C), whereas LSTM, a recurrent model, achieved the lowest observed recovery-time MAE (25 min) and unnecessary heating runtime (5.7 min/day). In contrast, TCN-GRU, the hybrid convolutional–recurrent model, achieved the lowest mean daily temperature-deficit exposure (4.9 K·min/day), despite exhibiting the highest aggregate temperature prediction error among the four data-driven models. These findings indicate that no consistent operational advantage can be attributed to greater architectural complexity or lower conventional forecasting error.

The statistical analysis, based on 22 heating-recovery events, revealed considerable uncertainty in several model comparisons. Although some differences in operational performance remained statistically significant after correction for multiple comparisons, the relatively wide confidence intervals indicate that the observed model rankings should be interpreted cautiously. Consequently, no single model could be identified as universally superior across all performance objectives.

The main contribution of this study is the demonstration that forecasting accuracy alone is insufficient for selecting models for optimal heating-start applications. Model

evaluation should additionally consider the operational consequences of prediction errors, particularly the trade-off between delayed comfort attainment and unnecessary heating operation. The proposed comparative evaluation approach provides a basis for selecting forecasting models according to application-specific control objectives rather than solely minimizing prediction error.

Further research should investigate the generalizability of these findings across different buildings and heating seasons and validate the estimated operational performance through prospective closed-loop experiments.

Author Contributions: Conceptualization, J.T., A.K. and M.D.; methodology, J.T., A.K. and M.D.; software, J.T. and M.D.; validation, J.T. and M.D.; formal analysis, J.T., M.D. and A.K.; investigation, J.T. and M.D.; resources, J.T.; data curation, J.T., A.K. and M.D.; writing—original draft preparation, J.T., A.K. and M.D.; writing—review and editing, J.T., A.K. and M.D.; visualization, J.T., M.D. and A.K.; supervision, J.T. and A.K.; project administration, A.K.; funding acquisition, A.K. All authors have read and agreed to the published version of the manuscript.

Funding: The present research is being conducted based on an agreement with SIA "ETKC" (Centre of Competence for Energy and Transportation) within the framework of project Nr. 5.1.1.2.i.0/2/24/A/CFLA/002 co-funded by The Recovery and Resilience Facility.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Acknowledgments: During the preparation of this manuscript, the authors used GPT-5.5 for the purposes of assisting with grammar correction, language editing and improving the clarity of the text. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: Authors Andris Krumins and Mihails Danilovs were employed by the company Lafivents SIA. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Nomenclature

Symbol/Abbreviation	Description
Mathematical symbols	
H	Historical input length (144 samples; 24 h)
P	Prediction horizon (36 samples; 6 h)
N	Number of evaluated heating-recovery events (22)
t	Discrete time index
t_0	Recorded heating activation time
$\mathbf{x}_t$	Historical process and weather feature vector
$\mathbf{u}_t$	Zone temperature setpoint vector
y_t	Measured indoor air temperature (°C)
$\hat{y}_t$	Predicted indoor air temperature (°C)
d_x	Number of historical input features (23)
d_u	Number of future control features (5)
$\mathbf{X}$	Historical input matrix
$\mathbf{U}$	Future setpoint schedule matrix
$\mathbf{Z}$	Augmented historical input representation
f_θ	Parametrized temperature forecasting model
$u^\star$	Daytime temperature setpoint (°C)
ε	Comfort threshold tolerance (0.2 K)
ϑ	Measured heating-recovery duration (min)
$\hat{\vartheta}$	Predicted heating-recovery duration (min)
$e_{m,n}$	Signed recovery-time prediction error for model m and event n (min)

$R_{m,n}$	Estimated unnecessary heating runtime (min)
$E_{m,n}$	Estimated temperature-deficit exposure (K·min)
ΔT_z	Initial zone temperature deficit (K)
ΔT_{oa}	Indoor–outdoor driving temperature difference (K)
ϑ_{max}	Maximum allowed recovery duration (360 min)
α	Adaptive baseline updating coefficient
δ	Huber loss threshold parameter
L	Number of TCN residual blocks
R	TCN receptive field (samples)
κ	TCN convolution kernel size
b	TCN dilation base
F	Number of TCN filters
Abbreviations	
ASHRAE	American Society of Heating, Refrigerating and Air-Conditioning Engineers
AHU	Air Handling Unit
BMS	Building Management System
BACnet	Building Automation and Control Network
CI	Confidence Interval
G36	ASHRAE Guideline 36
GRU	Gated Recurrent Unit
HVAC	Heating, Ventilation, and Air Conditioning
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MPC	Model Predictive Control
MSE	Mean Squared Error
ReLU	Rectified Linear Unit
TCN	Temporal Convolutional Network
TCN-GRU	Hybrid Temporal Convolutional Network and Gated Recurrent Unit Model
XGBoost	Extreme Gradient Boosting

References

1. Cao, X.; Dai, X.; Liu, J. Building energy-consumption status worldwide and the state-of-the-art technologies for zero-energy buildings during the past decade. *Energy Build.* **2016**, *128*, 198–213. [CrossRef]
2. Klepeis, N.; Nelson, W.; Ott, W.; Robinson, J.; Tsang, A.; Switzer, P.; Behar, J.; Hern, S.; Engelmann, W. The National Human Activity Pattern Survey (NHAPS): A resource for assessing exposure to environmental pollutants. *J. Expo. Anal. Environ. Epidemiol.* **2001**, *11*, 231–252. [CrossRef] [PubMed]
3. Wang, S.; Ma, Z. Supervisory and Optimal Control of Building HVAC Systems: A Review. *HVAC&R Res.* **2008**, *14*, 3–32. [CrossRef]
4. Afram, A.; Janabi-Sharifi, F. Theory and applications of HVAC control systems—A review of model predictive control (MPC). *Build. Environ.* **2014**, *72*, 343–355. [CrossRef]
5. Dong, B.; Lam, K.P. A real-time model predictive control for building heating and cooling systems based on the occupancy behavior pattern detection and local weather forecasting. *Build. Simul.* **2014**, *7*, 89–106. [CrossRef]
6. Fernandez, N.; Katipamula, S.; Wang, W.; Xie, Y.; Zhao, M.; Corbin, C. *Impacts of Commercial Building Controls on Energy Savings and Peak Load Reduction*; Technical Report PNNL-25985; Pacific Northwest National Laboratory: Richland, WA, USA, 2017.
7. Bloomfield, D.P.; Fisk, D.J. The optimisation of intermittent heating. *Build. Environ.* **1977**, *12*, 43–55. [CrossRef]
8. Seem, J.E.; Armstrong, P.R.; Hancock, C.E. A new algorithm for computing the optimal start time of HVAC equipment (comparison of methods for forecasting the time to return from night setback). In Proceedings of the ASHRAE Transactions, Chicago, IL, USA, 28 January–1 February 1989; Volume 95.
9. Fraisse, G.; Virgone, J.; Yezou, R. A numerical comparison of different methods for optimizing heating-restart time in intermittently occupied buildings. *Appl. Energy* **1999**, *62*, 125–140. [CrossRef]
10. Shin, M.S.; Rhee, K.N.; Jung, G.J. Optimal heating start and stop control based on the inferred occupancy schedule in a household with radiant floor heating system. *Energy Build.* **2020**, *209*, 109737. [CrossRef]

11. Chen, Q.; Li, N. Thermal response time prediction-based control strategy for radiant floor heating system based on Gaussian process regression. *Energy Build.* **2022**, *263*, 112044. [CrossRef]
12. Yang, I.H.; Yeo, M.S.; Kim, K.W. Application of artificial neural network to predict the optimal start time for heating system in building. *Energy Convers. Manag.* **2003**, *44*, 2791–2809. [CrossRef]
13. Esrafilian-Najafabadi, M.; Haghighat, F. Occupancy-based HVAC control using deep learning algorithms for estimating online preconditioning time in residential buildings. *Energy Build.* **2021**, *252*, 111377. [CrossRef]
14. Park, M.; Kim, S.H. Supervised learning based iterative learning control platform for optimal HVAC start-stop in a real building context. *Case Stud. Therm. Eng.* **2024**, *61*, 105055. [CrossRef]
15. Dai, M.; Li, H.; Wang, S. A reinforcement learning-enabled iterative learning control strategy of air-conditioning systems for building energy saving by shortening the morning start period. *Appl. Energy* **2023**, *334*, 120678. [CrossRef]
16. Drgoňa, J.; Arroyo, J.; Cupeiro Figueroa, I.; Blum, D.; Arendt, K.; Kim, D.; Ollé, E.P.; Oravec, J.; Wetter, M.; Vrabie, D.L.; et al. All you need to know about model predictive control for buildings. *Annu. Rev. Control* **2020**, *50*, 190–232. [CrossRef]
17. Langtry, M.; Wichitwechkarn, V.; Ward, R.; Zhuang, C.; Kreitmair, M.J.; Makasis, N.; Xuereb Conti, Z.; Choudhary, R. Impact of data for forecasting on performance of model predictive control in buildings with smart energy storage. *Energy Build.* **2024**, *320*, 114605. [CrossRef]
18. Rätz, M.; Müller, D. Poster Abstract: The Accuracy Paradox in Data-Driven Model Predictive Control. In Proceedings of the ACM Sustainability Week 2026, Banff, AB, Canada, 22–25 June 2026; pp. 579–580. [CrossRef]
19. Salsbury, T.I.; Devaprasad, K.; Lutes, R.; Rogers, A.P. Smarter building start – A distributed solution. *Energy Build.* **2023**, *282*, 112776. [CrossRef]
20. Paré, M.C.; Dermardiros, V.; Lesage-Landry, A. Efficient Data-Driven MPC for Demand Response of Commercial Buildings. *arXiv* **2024**, arXiv:2401.15742. [CrossRef]
21. Dey, R.; Salem, F. Gate-Variants of Gated Recurrent Unit (GRU) Neural Networks. *arXiv* **2017**, arXiv:1701.05923. [CrossRef]
22. Bai, S.; Kolter, J.; Koltun, V. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. *arXiv* **2018**, arXiv:1803.01271. [CrossRef]
23. Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; Zhang, W. Informer: Beyond efficient transformer for long sequence time-series forecasting. In Proceedings of the AAAI conference on artificial intelligence, Virtual, 2–9 February 2021; Volume 35, pp. 11106–11115.
24. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
25. Telicko, J.; Krumins, A.; Nikitenko, A. Development and Evaluation of Neural Network Architectures for Model Predictive Control of Building Thermal Systems. *Buildings* **2025**, *15*, 2702. [CrossRef]
26. ASHRAE. *Guideline 36-2024: High-Performance Sequences of Operation for HVAC Systems*; Technical Report; ASHRAE: Peachtree Corners, GA, USA, 2024.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.