EuroSun2026

BENCHMARKING MACHINE LEARNING MODELS FOR THERMAL LOAD FORECASTING: RELEVANCE OF BUILDING-COUNT NORMALIZATION IN GROWING DISTRICT HEATING NETWORKS

Theda Zoschke[1], Cosima Wörle[2], Lilli Frison[3], Felix Webhofer[1], Dhruv Patil[1], Gregor Rohbogner[4], Henrik Landersjö[5]

[1]*Fraunhofer Institute for Solar Energy Systems, Freiburg, Germany*
[2]*Fraunhofer Institute for Building Physics, Holzkirchen, Germany*
[3]*University of Freiburg, IMTEK, Freiburg, Germany*
[4]*HBG GmbH, Zell, Germany*
[5]*E.ON Energiinfrastruktur AB, Malmö, Sweden*

theda.zoschke@ise.fraunhofer.de, cosima.woerle@ibp.fraunhofer.de, lilli.frison@imtek.uni-freiburg.de, felix.webhofer@ise.fraunhofer.de, dhruv.patil@ise.fraunhofer.de, g.rohbogner@waldwaerme.com, henrik.landersjoe@eon.se

Summary

Short-term thermal load forecasting is an important part for advanced control strategies in district heating networks. As these networks grow continuously, connecting new buildings alters the aggregated heat load and complicates model training. This paper benchmarks five machine learning models (XGBoost, LightGBM, LSTM encoder-decoders with and without attention, and CNNs) on datasets from Germany and Sweden to investigate whether normalizing the heat load by the number of active buildings improves forecasting. Results show that normalization consistently reduces errors, particularly during the heating season. Tree-based models achieved the highest overall accuracy, while CNNs proved robust when building-count data was missing. LSTM architectures were more sensitive to data irregularities. Normalization is thus a simple, effective preprocessing step for growing networks.

Keywords: District heating, Thermal load forecasting, Machine learning, Benchmarking

Field of application for the findings: Load forecasting for advanced control and operational optimization.

Addressed audience: Utilities, Operators, Planners, Data Scientists.

1. Introduction

District heating networks are becoming increasingly complex to operate as they incorporate renewable and waste heat sources as well as sector coupling technologies such as CHP units and heat pumps. These dynamics make the use of advanced control strategies increasingly important. Most strategies like model predictive control, demand side management and fault detection require accurate short term thermal load forecasts. Many machine learning algorithms use historical consumption and ambient temperature data for this purpose (Geysen et al. (2018); Pressa et al. (2023); Runge and Saloux (2023); Frison et al. (2024)).

In district heating networks, the heat load can either be predicted individually by building, if detailed substation data is available or the aggregated heat load of the network can be predicted. Aggregation reduces computational effort and simplifies data handling, as fewer, less heterogeneous time series need to be trained and maintained. At the same time, aggregation inherently smooths individual consumption patterns, potentially helping to leveling out real data outliers and non-demand specific spikes which can increase the forecasting accuracy. Moreover, individual building-level data is not always available in practice, e.g. due to metering infrastructure or data privacy constraints. Even though the individual distribution of the heat demand in the network gets lost, there can be many applications where aggregation of heat load is helpful or necessary.

A particular challenge arises from the fact that district heating networks are rarely static: buildings are continuously connected or disconnected, changing the composition of consumers over time. Consequently, the aggregated heat load reflects not only changes in specific consumption behavior or weather conditions, but also the varying number of connected buildings. This complicates the training of forecasting models, as load increases in historical data may stem from newly connected buildings rather than genuine demand changes.

A straightforward approach to account for this effect is to normalize the aggregated heat load by the number of currently active buildings, yielding a load metric that is more independent of network growth and connection status. However, it remains unclear whether this normalization uniformly improves forecasting performance, as individual load patterns of newly connected buildings may differ and if its benefit applies across different machine learning models and datasets, or whether it depends on the operating conditions of the network. Furthermore, the structure of heat demand may differ between winter, when space heating dominates that is strongly driven by ambient temperature, and summer, when the demand is largely independent of weather conditions and instead reflects comparatively stable, user-driven hot water consumption patterns. It is an open question whether normalizing by building count offers seasonally consistent benefits or varies throughout the year.

2. Objective

This paper investigates whether normalizing the aggregated heat load of a district heating network by the number of active buildings improves short-term load forecasting accuracy, and whether this benefit is subject to seasonal variation. Several machine learning models, including decision tree-based methods and deep learning approaches, are benchmarked on multiple datasets from district heating networks that supply a mix of space heating and domestic hot water demand. For each model and dataset, forecasts are evaluated both with and without normalization by the number of active buildings. In addition, the evaluation is performed monthly in order to determine whether the benefit of normalization is more pronounced during the heating period or during the summer period, when hot water demand dominates. To reach a meaningful comparability the available monitoring data of two district heating networks (Weil am Rhein, Germany and Malmö, Sweden) were preprocessed with similar approaches and resampled to the same time step of one hour. Similar durations for training (depending on data availability) and identical durations for validation and testing were applied across both datasets. Furthermore, the individual features and input data to each forecasting model were aligned as closely as possible. The different methods are introduced and results presented in the following chapters.

3. Method

All compared models are trained and evaluated on the same datasets, which include historical thermal load measurements and ambient temperature. Obvious outliers are capped, and short gaps with NaNs are filled using an interpolation routine. Hyperparameters are optimized for most models to further improve the results and the computational performance required for training and validation is contrasted. In the following sections the models as well as the forecasting and computational performance metrics are introduced.

Given the strong seasonal asymmetry between heating-driven winter demand and domestic-hot-water-dominated summer demand, metrics are additionally reported separately for heating and summer regimes.

3.1 Forecasting models

In this study, several machine learning and deep learning models for thermal load forecasting in a district heating network are benchmarked. The models considered include Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), Long Short-Term Memory (LSTM) encoder-decoder networks with (AEDL) and without (EDL) attention mechanisms and Convolutional Neural Networks (CNNs).

Both XGBoost and LightGBM are gradient boosting decision tree algorithms applied to engineered tabular feature vectors. XGBoost (Chen and Guestrin (2016)) constructs ensembles of shallow trees sequentially to reduce residual errors, using regularisation to prevent overfitting. LightGBM (Ke et al. (2017)) follows the same principle but employs a leaf-wise growth strategy and histogram-based split finding for enhanced computational efficiency. Both models use a direct multi-horizon strategy, training a dedicated model for each forecast horizon

on lagged load and temperature values, rolling aggregates, and calendar indicators. Hyperparameters are tuned to balance complexity and accuracy. For LightGBM, hyperparameters (learning rate, number of leaves, minimum child samples, subsample and column-subsampling ratios, L1/L2 regularization) are tuned once per aggregation variant via a single Optuna pilot search at a representative horizon (h = 3) and reused across all 24 horizon-specific models. Early stopping determines the effective number of boosting rounds.

The LSTM encoder-decoder models are designed to handle sequential time series data, capturing temporal dependencies in the thermal load (EDL). In a further development an attention mechanism is additionally implemented in one variant (AEDL), allowing the decoder to weigh past time steps differently, potentially improving performance on longer sequences with varying dynamics, different time scales and exogenous covariates. For more information, see Frison et al. (2024). Hyperparameters are tuned in the case of AEDL.

Another approach are Convolutional neural networks (CNNs) that apply learned sliding filters to extract local patterns from sequential data. They use causal, dilated convolutions inspired by WaveNet (van den Oord et al. (2016)) to restrict filters to past time steps while exponentially expanding the receptive field. CNNs process raw 48-hour input sequences without manual feature engineering, jointly predicting all 24 horizons from a single model. Architecturally, input streams pass through 1D convolutional layers, which are specialized blocks that normalize data and apply non-linear transformations like ReLU. These features are then summarized and fed into a final prediction layer with dropout. Hyperparameters are tuned once per aggregation variant via Optuna, and models are trained using the Adam optimizer with early stopping.

To account for the growth of the network over time, the heat load is predicted in two variants: as an absolute value $Q(t)$ and as a load normalized by the number of currently active buildings $n(t)$,

$$q(t) = \frac{Q(t)}{n(t)}, \tag{1}$$

where $q(t)$ denotes the per-building load used as the target variable in the normalized case. For the tree-based models, all load-derived features (e.g., lagged and rolling values) are computed on the basis of $q(t)$ rather than $Q(t)$. For the deep learning models, the normalized load $q(t)$ is used directly as the target time series, without requiring an explicit feature engineering step. In both cases, predictions are rescaled to absolute values by multiplying with the number of active buildings at the respective forecast time step, $\hat{Q}(t) = \hat{q}(t) \cdot n(t)$, allowing a direct comparison between the normalized and non-normalized modeling approach. A further common characteristic across all models is that the ambient temperature is not only used during training and validation, but is also assumed to be known over the 24-hour forecasting horizon, corresponding to a perfect weather forecast. The lookback window for heat demand and ambient temperature is set to 48 hours for all models.

3.2 Forecasting performance

Four metrics are used to evaluate forecasting performance: the Mean Absolute Percentage Error (MAPE), the Weighted Absolute Percentage Error (WAPE), the Root Mean Squared Error (RMSE), and the Mean Percentage Error (MPE).

MAPE is reported as it remains the most widely used and interpretable metric in the literature, facilitating comparison with other studies (Nassif et al. (2021); Pressa et al. (2023); Bujalski et al. (2021)). It measures percentage error at the individual observation level. Since true heat-load values can become very small, a small constant ε was added to the denominator to avoid numerical instability.

$$\text{MAPE} = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i + \varepsilon} \right| \tag{2}$$

WAPE is considered the most suitable metric for comparing performance across different models and datasets, as it weights the absolute error by the overall level of demand, avoiding the disproportionate influence of periods with very low load that affects MAPE (Hyndman (2025)). It expresses the total absolute error relative to the total observed load and remains stable even when individual target values are small.

$$\text{WAPE} = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{\sum_{i=1}^{n} |y_i|} \tag{3}$$

RMSE is included since it corresponds to the loss function used during model training and validation, and therefore provides a consistency check between training objective and evaluation (Nassif et al. (2021); Bujalski et al. (2021); Monzani et al. (2024)). It penalizes larger errors more strongly than MAPE and is therefore useful for identifying models with occasional large deviations.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4}$$

Finally, MPE is reported to identify systematic bias in the forecasts, i.e. whether a model tends to over- or underestimate the thermal load.

$$\mathrm{MPE} = \frac{100}{n}\sum_{i=1}^{n}\frac{y_i - \hat{y}_i}{y_i + \varepsilon} \tag{5}$$

where y_i denotes the observed value, $\hat{y}_i$ the predicted value, $\bar{y}$ the mean of the observed values, n the number of observations, and $\varepsilon = 10^{-6}$ is a small constant used to stabilize percentage-based metrics.

In addition to the point estimate of each metric, a standard deviation is reported to characterize how strongly the forecast error fluctuates over the test period. It is computed by first evaluating the respective metric per time step (or per month) and then taking the sample standard deviation across these values, thus reflecting temporal variability rather than uncertainty across random seeds. For WAPE, no standard deviation is reported, since WAPE is defined as a ratio of aggregated sums rather than an average of per-observation errors. Therefore, it cannot be meaningfully decomposed per time step, and in fact reduces algebraically to MAPE when each time step contains a single observation. Only the pooled point estimate of WAPE is therefore reported.

3.3 Computational performance

To assess the computational cost of each model, the cost of a single iteration of the innermost training loop, denoted Φ_{window} (or Φ_{round} for the boosted-tree models), is derived first. The symbols used within this chapter are summarized in Table 1. For the neural models (EDL, AEDL, CNN), the total training cost follows directly as $T_{\text{train}} = O(E\,N\,\Phi_{\text{window}})$, since the per-window cost is incurred once per training sample and once per epoch. For the gradient-boosted models, the analogous relation is $T_{\text{train}} = O(F\,B_{\text{tree}}\,\Phi_{\text{round}})$, reflecting one boosting round per tree and one model per forecast horizon.

Tab. 1: Symbols in the runtime-complexity equations (6)–(8).

Symbol	Meaning	Scope
N	rolling-origin training samples	all models
m	number of input features	all models
F	forecast horizon length; for XGBoost also the number of per-horizon models	all models
H	LSTM hidden size	EDL, AEDL
P	history / look-back length (encoder steps)	EDL, AEDL, CNN
L	number of stacked LSTM layers (encoder + decoder)	EDL, AEDL
E	training epochs	EDL, AEDL, CNN
L_{CNN}	number of stacked (dilated) convolutional layers	CNN
K	kernel size of the 1D convolution	CNN
C_{in}	number of input channels per convolutional layer	CNN
C_{out}	number of output channels (filters) per convolutional layer	CNN
B_{tree}	number of boosting rounds	XGBoost, LightGBM
$d_{\max}$	maximum tree depth	XGBoost, LightGBM
B_{bin}	histogram bins per feature	XGBoost, LightGBM

The EDL model is implemented as an encoder–decoder LSTM (sequence-to-sequence architecture) without attention. Since each LSTM step incurs a cost of $O(H^2)$ (Sak et al. (2014)), the per-window cost is dominated purely by the recurrent encoder and decoder computations,

$$\Phi^{\text{EDL}}_{\text{window}} = O\big((P+F)\, L\, H^2\big), \tag{6}$$

yielding a total training cost that is linear in the number of samples and epochs, and quadratic in the hidden size.

The AEDL model extends the EDL architecture with a dual-stage attention mechanism, comprising both input and temporal attention. The per-window cost therefore gains two additive terms beyond the recurrent computation: the temporal attention score matrix, scaling as $O(PFH)$ (Vaswani et al. (2017)), and the input attention mechanism, scaling as $O\big(Pm(P+H)\big)$ and quadratic in the look-back length P,

$$\Phi^{\text{AEDL}}_{\text{window}} = O\big((P+F)\, L\, H^2 + PFH + P\, m\, (P+H)\big). \tag{7}$$

As with EDL, the resulting training cost is linear in samples and epochs and quadratic in the hidden size; the recurrent term continues to dominate overall cost unless P or m become large relative to LH^2.

The CNN model applies 1D causal, dilated convolutions following the WaveNet architecture (van den Oord et al. (2016)) to extract local and mid-range temporal patterns, without residual connections. Following the standard convolutional-layer complexity formulation (Goodfellow et al. (2016)), the per-window cost scales linearly with the number of layers L_{CNN}, kernel size K, channel widths C_{in}, C_{out}, and the look-back length P,

$$\Phi^{\text{CNN}}_{\text{window}} = O\big(L_{\text{CNN}}\, K\, C_{\text{in}}\, C_{\text{out}}\, P\big). \tag{8}$$

Unlike the LSTM-based models, this cost is not quadratic in a hidden-state size but is set instead by the convolutional filter parameters, and scales only linearly in P. Because L_{CNN} is fixed by architectural design rather than grown with P, no additional coupling term between depth and window length arises, making the architecture highly parallelizable along the time axis, in contrast to the inherently sequential LSTM encoder.

The decision tree-based models LightGBM and XGBoost both follow a direct multi-horizon approach, training one gradient-boosted ensemble per horizon $h = 1, \ldots, F$, with each ensemble grown using histogram-based split-finding (Chen and Guestrin (2016); Ke et al. (2017)). A single boosting round requires constructing per-feature histograms over all N rows, at cost $O(d_{\max} mN)$, followed by a split search over B_{bin} bins, at cost $O(d_{\max} mB_{\text{bin}})$,

$$\Phi^{\text{GBM}}_{\text{round}} = O\big(d_{\max}\, m\, (N + B_{\text{bin}})\big). \tag{9}$$

The resulting total training cost scales linearly with the number of samples, features, horizons, and boosting rounds; the histogram-construction term dominates whenever $N \gg B_{\text{bin}}$, which is typically the case in practice.

The gradient-boosted ensembles (XGBoost, LightGBM) are the cheapest to train: every factor enters linearly, histogram binning replaces a per-sample split search by a per-bin one, and, unlike the neural models, the data is traversed once per boosting round rather than once per epoch. Training a separate model per horizon partly offsets this advantage, but each of those models is individually inexpensive. Among the neural network architectures the CNN is the least expensive, because its cost is set by the comparatively small convolutional filter parameters instead of a quadratic hidden-size term, and because its convolutions are evaluated in parallel across the sequence. The two recurrent models (EDL, AEDL) are the most expensive: their cost grows quadratically with the hidden size and, more importantly in practice, their encoder and decoder steps must be evaluated sequentially, which prevents parallelization along the time axis and leaves the hardware underutilized. AEDL is in turn more expensive than EDL, as the two attention stages add further terms to the per-window cost, with the input-attention contribution growing quadratically with the look-back length.

4. Case Studies

The evaluation considers two case studies, where measurement data is available for very similar periods. The first is the district heating network in Weil am Rhein, Germany, comprising up to 125 consumers that are mostly residential, but also include a small number of schools, commercial, and industrial buildings. The second is the district heating network in Malmö, Sweden, comprising up to 118 residential consumers. The available

measurement data is preprocessed to filter gaps and outliers and is resampled to hourly resolution. Figures 1 and 2 show the aggregated heat demand and the number of active buildings for both networks over the considered period of January and April 2020 to December 2025. Both networks exhibit a significant growth in the number of connected buildings, increasing by about 60 buildings in Malmö and about 100 buildings in Weil am Rhein. Despite a mean ambient temperature only about 2 K lower in Malmö, the heat demand per building is in both Winter and Summer considerably higher there than in Weil am Rhein, suggesting that the connected buildings in the Malmö network tend to be larger multi-family homes than in Weil am Rhein. As a result, the summer base load in Weil am Rhein lies notably closer to zero (about 200 kW, compared to about 500 kW in Malmö), which may lead to stronger relative gradients and fluctuations in the data which could pose a greater challenge for the forecasting algorithms.

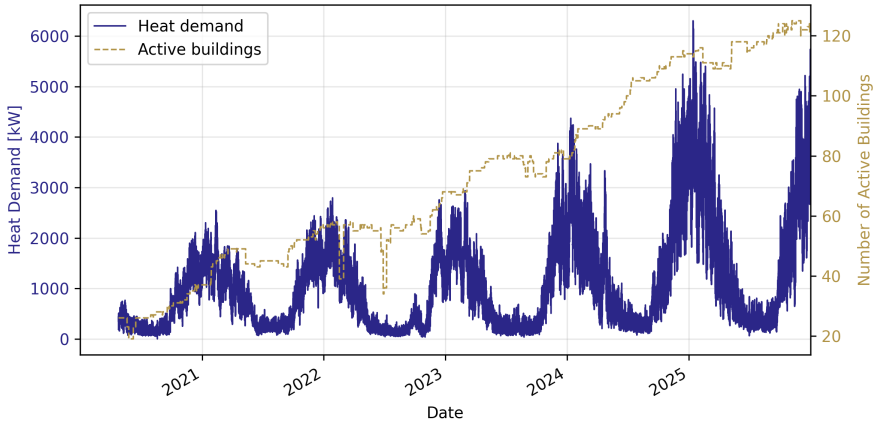

Fig. 1: Heat demand in kW and number of active buildings in district heating network of Weil am Rhein, Germany, from April 2020 to December 2025.

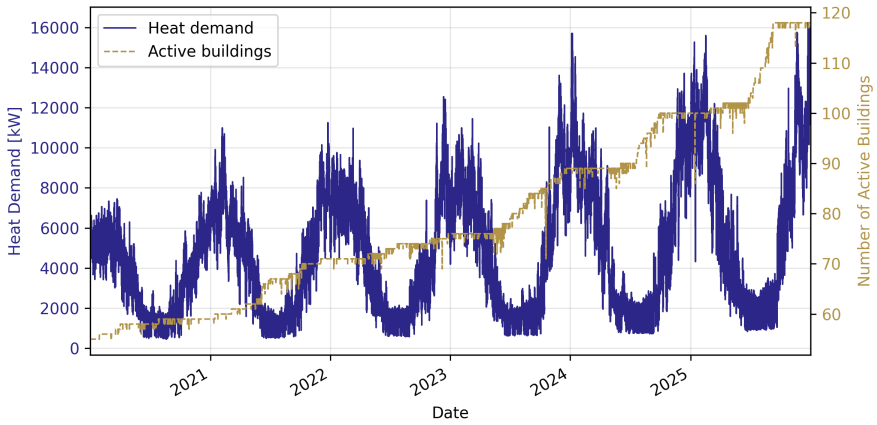

Fig. 2: Heat demand in kW and number of active buildings in district heating network of Malmö, Sweden, from January 2020 to December 2025.

For both networks the training was done with data up until the end of 2023, the validation was done with data from January 2024 to December 2024 (one full year) and testing was done from January 2025 to December 2025 (again a full year).

5. Results

For all introduced models, the evaluation has been carried out for both the Malmö and Weil am Rhein datasets, comparing forecasts based on the load normalized by active buildings and the unnormalized raw load.

5.1 Overall evaluation metrics

Table 2 and 3 summarize MAPE, WAPE, RMSE, and MPE for all evaluated models, both with and without normalization by the number of active buildings, for both case studies. Among the reported metrics, WAPE is considered the most suitable for an overall ranking of model performance, as it weights errors by the actual demand level and thereby better reflects the pronounced seasonal shift in absolute load and relative errors than MAPE.

For every model and both networks, normalizing the aggregated heat load by the number of active buildings reduces WAPE compared to the unnormalized case, confirming that accounting for the growing number of connected buildings removes a substantial part of the error that would otherwise be misattributed to genuine demand changes. LightGBM with normalization achieves the lowest WAPE for both Weil am Rhein (9.1) and Malmö (6.5) and thus represents the best-performing model overall in terms of this metric. XGBoost with normalization performs very similarly to LightGBM in the Malmö network (WAPE 6.8 vs. 6.5) but worse in Weil am Rhein (WAPE 10.9 vs. 9.1). At the same time, XGBoost consistently exhibits smaller standard deviations across MAPE and RMSE than LightGBM in both networks, indicating that, despite a slightly larger systematic bias (as indicated by MPE), XGBoost produces a more consistent, less fluctuating error, which can be advantageous for operational applications where predictability of the forecast error is important. This can be attributed to the fact that XGBoost's depth-limited, level-wise tree growth produces more balanced and smoother predictors, so its forecast error fluctuates less over the test period, whereas LightGBM's leaf-wise growth fits sharper local structure at the cost of a more uneven temporal error profile (Chen and Guestrin (2016); Ke et al. (2017)). At the same time, this may also, at least partly, be an artifact of the chosen tree capacity (result of HPO) rather than an intrinsic property, since a more strongly regularized LightGBM, or a deeper XGBoost, would likely narrow the gap.

Tab. 2: Overview of MAPE, WAPE, RMSE, and MPE for all models in Weil am Rhein, with and without normalization by the number of active buildings (minimum values bold).

Model	Norm.	MAPE	WAPE	RMSE	MPE
AEDL	Yes	17.1 ± 17.5	13.2	317.4 ± 218.2	−0.8 ± 24.3
AEDL	No	25.0 ± 12.5	24.7	588.6 ± 417.2	21.8 ± 16.8
EDL	Yes	18.5 ± 17.9	13.6	323.3 ± 210.3	−4.5 ± 24.7
EDL	No	18.4 ± 12.0	18.5	476.7 ± 354.4	9.0 ± 19.5
CNN	Yes	13.0 ± 11.2	10.4	254.9 ± 170.4	**0.2** ± 16.7
CNN	No	15.6 ± 12.1	14.6	377.1 ± 270.8	4.7 ± 18.9
LightGBM	Yes	**11.3** ± 10.6	**9.1**	**228.9** ± 161.1	−1.4 ± 15.3
LightGBM	No	17.1 ± 11.0	19.9	544.8 ± 425.8	12.4 ± 16.0
XGBoost	Yes	13.5 ± 6.3	10.9	256.5 ± 146.1	−4.6 ± 9.4
XGBoost	No	19.4 ± 7.4	23.2	598.3 ± 432.2	13.1 ± 11.5

Tab. 3: Overview of MAPE, WAPE, RMSE, and MPE for all models in Malmö, with and without normalization by the number of active buildings (minimum values bold).

Model	Norm.	MAPE	WAPE	RMSE	MPE
AEDL	Yes	11.7 ± 8.8	10.0	840.8 ± 522.2	5.1 ± 13.6
AEDL	No	12.7 ± 7.6	13.3	1157.8 ± 779.8	7.7 ± 12.6
EDL	Yes	12.0 ± 8.6	11.5	997.4 ± 649.2	6.2 ± 12.9
EDL	No	15.8 ± 8.5	15.7	1326.4 ± 845.1	12.1 ± 12.8
CNN	Yes	9.7 ± 8.3	7.3	631.1 ± 371.9	**-0.2** ± 12.4
CNN	No	12.9 ± 10.6	9.1	749.9 ± 417.4	3.3 ± 16.1
LightGBM	Yes	**8.4** ± 9.7	**6.5**	**580.8** ± 385.2	−3.6 ± 12.2
LightGBM	No	11.9 ± 6.9	12.2	1070.7 ± 720.4	9.4 ± 9.9
XGBoost	Yes	8.8 ± 6.0	6.8	614.3 ± 307.8	−3.7 ± 8.2
XGBoost	No	15.5 ± 4.6	15.3	1279.8 ± 709.2	11.6 ± 6.6

Among the deep learning models, the CNN clearly performs best in both networks and also achieves the MPE value closest to zero of all models (0.2 in Weil am Rhein, −0.2 in Malmö), indicating that over- and underestimation of the load largely cancel out. The encoder-decoder LSTM models (EDL and AEDL) show markedly higher errors and larger standard deviations than all other models, an effect that is considerably more pronounced in Weil am Rhein than in Malmö (e.g., AEDL MAPE 17.1 ± 17.5 in Weil am Rhein vs. 11.7 ± 8.8 in Malmö). This suggests that these models are noticeably more sensitive to less regular, "messier" input data and are more easily disturbed by load values close to zero. Although EDL and AEDL follow qualitatively similar patterns, their absolute results still differ considerably; since no dedicated hyperparameter optimization was performed for the EDL model, this may be one contributing factor. However, AEDL - for which HPO was performed - still shows comparatively high errors and variance, raising the question of whether recurrent architectures are inherently harder to tune reliably for this forecasting task.

Considering the case without normalization, the CNN achieves the lowest WAPE of all models in both networks (14.6 in Weil am Rhein, 9.1 in Malmö), outperforming even the tree-based models in this setting. Also in the case with normalization the performance is comparable to those of the decision tree models. This makes the CNN a suitable choice, especially when information on the number of active buildings is not available.

5.2 Seasonal performance

To assess whether the benefit of normalization by the number of active buildings depends on the season, MAPE is additionally evaluated separately for each month of the test year 2025. Figure 3 shows the mean MAPE together with its standard deviation across the forecasting horizon for each month, for all models with and without normalization. To further improve direct comparability between models, the MAPE values of all normalized models are additionally combined into a single graph for each network, shown in Figure 4 for Weil am Rhein and Figure 5 for Malmö.

For all models and both networks, forecasting accuracy is consistently higher during the winter months than during summer, which is expected since the absolute heat demand is lower in summer, so that even comparatively small absolute deviations translate into larger percentage errors. In Weil am Rhein, both forecasting accuracy and its standard deviation are notably poorer in the summer months than in Malmö, consistent with the more heterogeneous and irregular consumption behavior observed in this network. The benefit of normalization by the number of active buildings is most pronounced during the heating period, where it consistently reduces forecast error across all models and both networks. During summer, this benefit is less uniform. For the tree-based models in the Malmö network there are one to two months in which normalization results in a slightly worse forecast than the unnormalized case. Similarly, this occurs with the EDL in the Weil am Rhein network from July to September. This effect does not appear systematically in the respective other network and can therefore be regarded as an isolated occurrence rather than a general limitation of the normalization approach. In all other months, normalization proves beneficial, in particular during the heating period. The CNN stands out among the deep learning models by showing a very stable and largely monotonic seasonal trend, with better performance in winter and worse performance in summer, without pronounced outliers in either network.

5.3 Performance over the forecast horizon

Figure 6 and Figure 7 show the MAPE as a function of the forecast horizon, from the first predicted time step ($h = 0\,\mathrm{h}$) up to the full 24-hour horizon, for all models.

Across all models, the forecast error increases with the length of the forecast horizon, as expected. The decision tree-based models, LightGBM and XGBoost, consistently achieve lower errors than the AEDL/EDL and CNN architectures across the full horizon, in line with their advantage in the overall evaluation metrics. This advantage is particularly pronounced for the very first forecasted time steps, where the accuracy of both tree-based models is roughly 10-20% higher than their average accuracy over the full 24-hour horizon, suggesting that they are especially effective at capturing the most recent consumption trend. This advantage is particularly relevant in applications such as model predictive control, since the first time step of a forecast has a disproportionately higher influence on subsequent control actions.

For this class of district-heating forecasting problems, tree-based models such as LightGBM and XGBoost thus appear particularly well suited, achieving the best overall accuracy while showing comparatively low sensitivity to heterogeneous and imperfect operational data. This advantage is most apparent for the Weil am

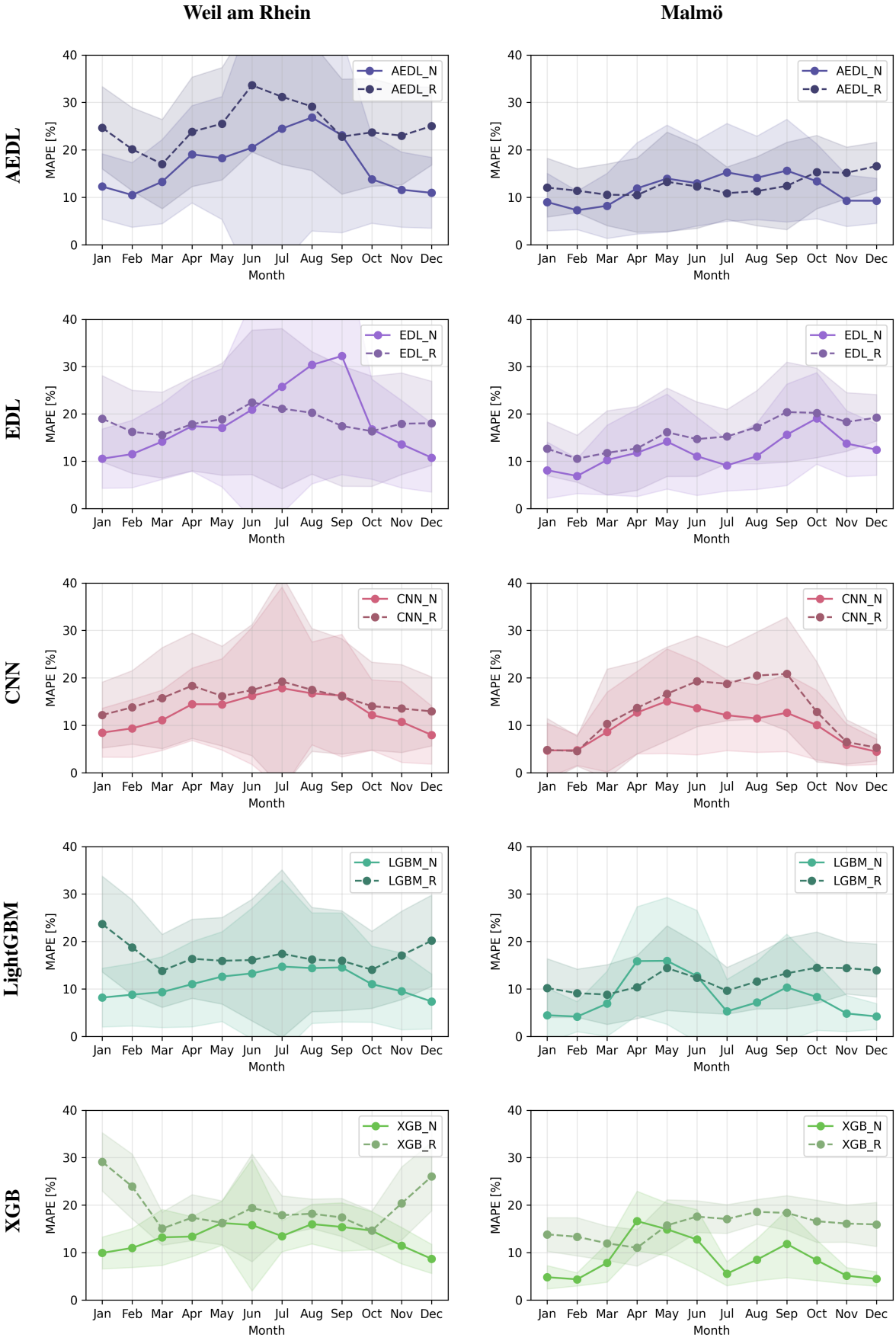

Fig. 3: Comparison of seasonal MAPE development for various forecasting models with normalization(_N) and raw load data without normalization (_R) in the networks Weil am Rhein and Malmö.

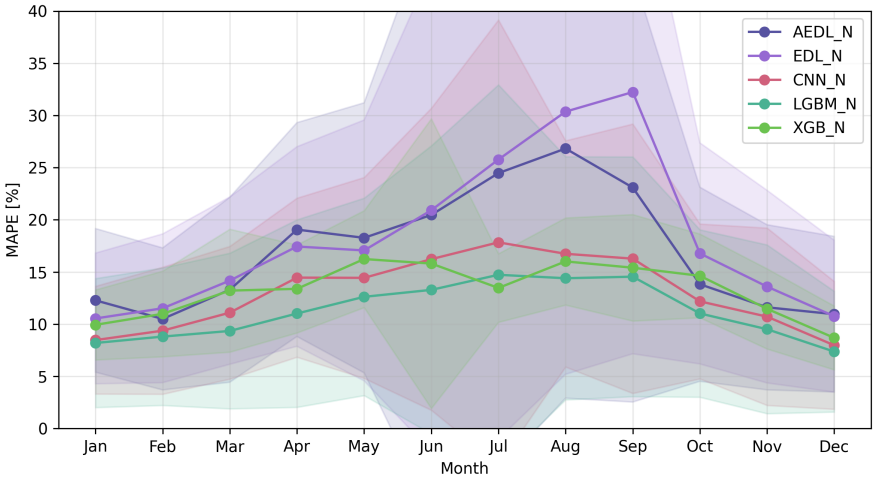

Fig. 4: Mean MAPE per month over the test year 2025, with standard deviation, for all models with normalization in district heating Network of Weil am Rhein, Germany.

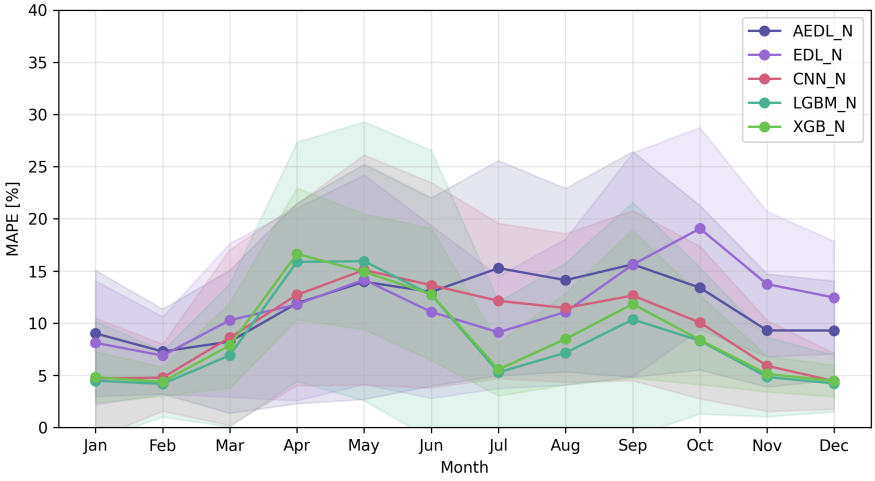

Fig. 5: Mean MAPE per month over the test year 2025, with standard deviation, for all models with normalization in district heating Network of Malmö, Sweden.

Rhein network, where the data contain stronger heterogeneity, irregular consumption patterns and values closer to zero during summer months. Among the deep learning models, the CNN performs best, suggesting that much of the predictive information is contained in local and recurring temporal patterns that convolutional filters can extract efficiently. The encoder-decoder LSTM models (EDL and AEDL) perform worse and less consistently, indicating that their additional recurrent and attention-based complexity does not provide a clear benefit for this relatively structured 48 h input / 24 h forecast problem and may instead make them more sensitive to irregularities in the input data and to error propagation over the forecast horizon.

6. Conclusions

Overall, the results demonstrate that normalizing the aggregated heat load by the number of active buildings is a simple yet effective preprocessing step, particularly during the heating period, where it consistently improves forecasting accuracy across all models and both case studies. During summer, its benefit tends to depend on the specific network characteristics: in networks with more volatile, control-driven hot water demand, such as Weil am Rhein, normalization can occasionally slightly deteriorate performance for individual models, whereas in networks with larger, more stable per-building demand, such as Malmö, it tends to remain beneficial throughout the year. Among the benchmarked models, decision tree-based methods generally achieve the best

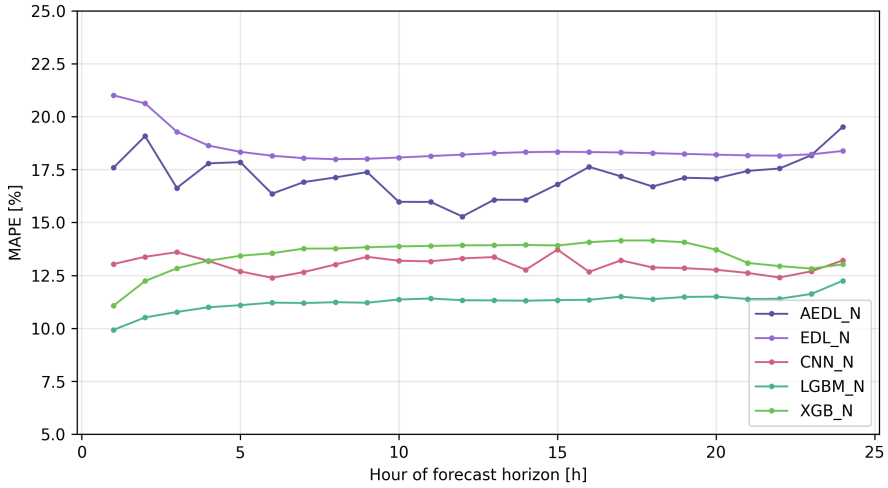

Fig. 6: MAPE as a function of the forecast horizon (0 to 24 h) for all evaluated models in network of Weil am Rhein.

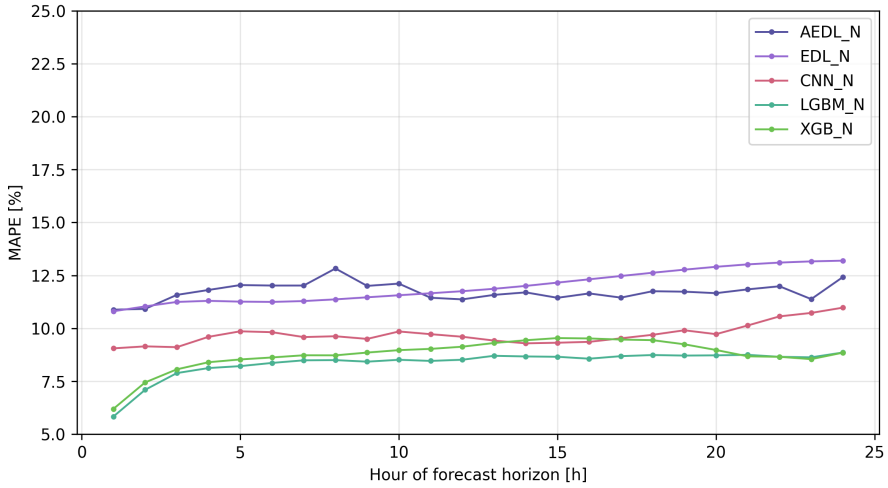

Fig. 7: MAPE as a function of the forecast horizon (0 to 24 h) for all evaluated models in network of Malmö.

overall performance, with LightGBM combined with normalization typically preferable when forecast accuracy is the primary criterion, and XGBoost combined with normalization representing a favorable alternative for operational applications, where its comparatively lower standard deviation may outweigh a slightly higher average error. It should be noted, however, that this lower standard deviation is likely at least partly attributable to the hyperparameter optimization rather than an intrinsic property of the algorithm, so that the observed gap in prediction stability between XGBoost and LightGBM should not be over-generalized. Both tree-based models show a notable advantage in the early forecast horizon, which is especially relevant for control-oriented applications. Furthermore, they lead to the least computational cost among all models. Another good alternative are CNNs that tends to be the most suitable choice among the evaluated models and are particularly strong, if information on the number of active buildings is unavailable while still being fairly computational inexpensive. Although based on only two case studies with differing network characteristics, the consistent benefits observed across both networks are a sign for the broader applicability of building-count normalization, which future studies should further validate.

Acknowledgments

This work was funded by the BMWE (German Federal Ministry of Economic Affairs and Energy) under grant 03EN3054A (“WOpS - Heat Flow Optimization for Sector Coupling”) and 03EN3104A (“CoolDown – Optimization of District Heating in Existing Buildings”).

References

Maciej Bujalski, Paweł Madejski, and Krzysztof Fuzowski. Heat demand forecasting in district heating network using xgboost algorithm. *E3S Web Conf.*, 323:4, 2021. doi: 10.1051/e3sconf/2021323000004.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

Lilli Frison, Simon Gölzhäuser, Moritz Bitterling, and Wolfgang Kramer. Evaluating different artificial neural network forecasting approaches for optimizing district heating network operation. *Energy*, 307:132745, 2024. doi: 10.1016/j.energy.2024.132745.

Davy Geysen, Oscar de Somer, Christian Johansson, Jens Brage, and Dirk Vanhoudt. Operational thermal load forecasting in district heating networks using machine learning and expert advice. *Energy and Buildings*, 162:144–153, 2018. doi: 10.1016/j.enbuild.2017.12.042.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

Rob J. Hyndman. Wape: Weighted absolute percentage error, 2025. URL `https://robjhyndman.com/hyndsight/wape.html`.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Neural Information Processing Systems*, pages 3146–3154, 2017. URL `https://mlanthology.org/neurips/2017/ke2017neurips-lightgbm/`.

C. Monzani, G. Cerino Abdin, G. Montanari, and A. Poggio. Modelling hourly thermal energy demand: A machine learning approach of residential district heating substations in turin. *RE and PQJ*, pages 39–44, 2024. doi: 10.52152/4003.

Ali Bou Nassif, Bassel Soudan, Mohammad Azzeh, Imtinan Attilli, and Omar AlMulla. Artificial intelligence and statistical techniques in short-term load forecasting: A review. *International Review on Modelling and Simulations*, 2021.

Christian Pressa, Stefan Leiprecht, Fabian Behrens, Verena Jetzinger, Hendric Popma, and Matthias Finkenrath. Benchmarking of state-of-the-art machine learning methods for highly accurate thermal load forecasting in district heating networks. In *Proceedings of the 36th International Conference on Efficiency, Cost, Optimization, Simulation, and Environmental Impact of Energy Systems (ECOS 2023)*, 2023.

Jason Runge and Etienne Saloux. A comparison of prediction and forecasting artificial intelligence models to estimate the future energy demand in a district heating system. *Energy*, 269:126661, 2023. doi: 10.1016/j.energy.2023.126661.

Haşim Sak, Andrew Senior, and Françoise Beaufays. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *Proc. Interspeech*, pages 338–342, 2014.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio, 2016. URL `https://arxiv.org/abs/1609.03499`.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.