

Real-World Short-Term District-Heating Load Forecasting: A Machine-Learning Benchmark Against an Operational Utility Model

Maximilian Beichter¹, Marian Turowski¹, Andreas Döring¹,
Alexander Kreusel¹, Nils Friederich¹, Simon Waczowicz¹, Jorge
Ángel González Ordiano^{1,3}, Andreas Iser², Jürgen Hitzel², René
Weschenfelder², Manuel Rink², Veit Hagenmeyer¹, Ralf Mikut¹

¹Institute for Automation and Applied Informatics (IAI), Karlsruhe Institute of
Technology (KIT)

Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany

E-Mail: {firstname.lastname}@kit.edu

²Stadtwerke Karlsruhe GmbH

Daxlander Straße 72, 76185 Karlsruhe, Germany

E-Mail: {firstname.lastname}@stadtwerke-karlsruhe.de

³Instituto de Investigación Aplicada y Tecnología, Universidad Iberoamericana
Ciudad de México

Prolongación Paseo de Reforma 880, Mexico City 01219, Mexico

E-Mail: jorge.gonzalez@ibero.mx

Abstract

In urban district-heating networks, operators plan unit dispatch, storage operation, and fuel procurement on the basis of short-term heat-demand forecasts. At the Stadtwerke Karlsruhe GmbH (Germany), this forecast has been generated since 2017 by a compact gray-box model that adjusts the load profile of a reference-day using a 72 h temperature forecast and a weekday-dependent calendar correction. We benchmark this operational reference model against machine-learning models (namely ordinary least-squares, Ridge, and ElasticNet

regression, a multi-layer perceptron, XGBoost, LightGBM, and the tabular foundation models TabPFN and TabFM). On the 2025/26 test season, the best-performing model, TabPFN, reduces the mean absolute error from 14.46 MW of the operational reference model to 10.69 MW, a reduction of 26 %. TabPFN obtains these results without a task-specific weight update, using the training seasons only as in-context examples. In comparison, the regularized linear model ElasticNet reaches 10.78 MW and thus achieves 97.6 % of the improvement achieved by TabPFN, while remaining interpretable: its coefficients reveal a plausible learned calendar correction and temperature forecast dependency. For similar applications, the reported results can serve as a guide to trade implementation effort against accuracy and explainability.

Keywords: district-heating load forecast, short-term heat load forecasting, data-driven forecasting, time-series forecasting, machine learning, tabular foundation models, day-ahead forecasting

1 Introduction

District-heating utilities need short-term forecasts of the aggregated network heat load, that is, the total hourly heat supplied to the district-heating network. These forecasts serve to schedule the following day's heat production [4] and to plan cogeneration and boiler operations [2,23] up to several days in advance [23]. Beyond the operation of the production units, they support the planning of thermal energy storage units [7, 19], ordering fuel such as natural gas [19,23], and maintenance [19]. Short-term heat load forecasts thus underlie a wide range of operational decisions, from the next day's heat production to storage operation, fuel purchases, and maintenance.

The Stadtwerke Karlsruhe GmbH (Germany), the municipal utility operating the district-heating network of Karlsruhe, uses a forecast of the total hourly heat supplied to the district-heating network for operational planning, so this forecast must be accurate and interpretable at once. To support this planning task, the Stadtwerke Karlsruhe GmbH (Germany) and the Karlsruhe Institute of Technology (KIT) cooperate on developing forecasting models. During this cooperation, they have jointly developed several forecasting models while the

requirements placed on these models have grown. The first of these models originated in 2015 as a parameter-linear model that forecasts the total daily heat supplied to the district-heating network. An evaluation of nonlinear and machine-learning methods confirmed the use of this parameter-linear model, since the evaluated methods did not yield a substantial improvement [24]. An hourly model was deployed in 2017 as an extension of the 2015 model and has remained in operation since then; we refer to it as the *operational reference model*.

By intention, the structure of the operational reference model was kept compact and interpretable: it follows the two main drivers of the heat demand, the outdoor temperature and the social behavior of the consumers, which differs between working days and weekends [6], so that each component has a clear physical or operational interpretation. Moreover, the model requires no dedicated training infrastructure and can be implemented and executed in a spreadsheet with negligible computational effort. During operation, the responsible operators use the forecast resulting from the operational reference model as decision support and review it before applying it in operational planning. Based on operational experience, the operators may adjust the forecast, in particular on public holidays and during school vacations. These adjustments are made case by case, are not part of the model, and are not logged.

In a broad sense, the design of the operational reference model follows the gray-box principle described by Nielsen and Madsen [20, 21]. In their terminology, gray-box modeling combines physical knowledge with data-based statistical modeling. Physical knowledge determines the main model structure, while statistical modeling estimates the model coefficients. In contrast, they characterize black-box modeling as an almost purely data-driven approach. However, the earlier evaluation [24] concerned the daily model. Since then, eight heating seasons of hourly data have accumulated, allowing a broader range of models to be evaluated on a common, longer data basis, and tabular foundation models such as TabPFN [10] and TabFM [15] have emerged that have not yet been evaluated for this task. It therefore remains unclear to what extent retaining the compact gray-box structure impacts the forecast accuracy.

This work therefore benchmarks the gray-box operational reference model against general-purpose machine-learning black-box models (from linear regression to tabular foundation models) using the same information available to the operational reference model at inference. Holding the input information constant allows differences in forecasting performance to be attributed to the model class rather than to differences in data availability at inference or feature sets.

Therefore, this work makes four contributions. First, it benchmarks the operational reference model as a gray-box model that has been in continuous operational use since 2017. Second, it establishes a controlled comparison by providing all benchmark models, commonly viewed as black-box modeling approaches, with exactly the same information as the operational reference model. Third, it evaluates tabular foundation models under these operational input constraints and compares them with conventional machine-learning models that are trained specifically for the task. Fourth, it interprets the best-performing linear model through its coefficients and input-block contributions, including its learned weekday-dependent calendar correction.

The remainder of this paper is structured as follows. Section 2 reviews related work on heat load forecasting. Section 3 describes the operational data, and Section 4 defines the forecast task, the evaluated models, and the evaluation. Section 4 presents the benchmark results, and Section 6 interprets the learned linear model through its coefficients and input-block contributions. Section 7 discusses the operational implications, the trade-off between accuracy and complexity, and the limitations of this work, and Section 5 concludes. Appendix 8 details the benchmark models.

2 Related Work

This section positions the paper within three areas of related work: short-term forecasting of aggregated loads in city-scale district-heating networks and substantial urban subnetworks, gray-box modeling approaches for district-heating load forecasting, and the application of tabular foundation models to district-heating forecasting.

The studies considered in this section focus on forecasting aggregated heat loads at the network or substantial-subnetwork level. They employ data-driven forecasting approaches based on historical operational data, weather information, calendar variables, or combinations thereof. Most studies forecast hourly heat loads over horizons ranging from one hour to several days, with day-ahead forecasting being particularly common.

Forecasting aggregated loads in city-scale district-heating networks. The studies collected in Table 1 investigate aggregated heat load forecasting in municipal district-heating networks and substantial urban subnetworks, including Aarhus [4], Cottbus [25], Flensburg and Munich [26], Helsinki [12], Ljubljana [23], Stockholm [6], Ulm [7, 19], Warsaw [18], and an unnamed European city [2]. The studies in the first part of Table 1 employ black-box modeling approaches in which the forecasting relationship is inferred from observed input and output data without the model structure itself being derived from explicit physical or operational knowledge. The applied methods range from statistical and regression-based approaches, such as linear regression, Generalized Additive Models (GAM), Generalized Additive Mixed Models (GAMM), Autoregression with Exogenous Inputs (ARX), and Seasonal Autoregressive Integrated Moving Average with Exogenous Regressors (SARIMAX), to machine-learning approaches, including neural networks, support vector regression (SVR), random forests, boosting, fuzzy models, ensembles, and foundation models (e.g., TabPFN-TS, Chronos-2). These black-box studies are relevant here because they forecast the same quantity with the same model classes that serve as benchmark models in this work. Unlike the present work, however, most of them use a longer load history or measured future weather as inputs and compare their models with generic baselines rather than with a forecasting model in long-term operational use. A study is assigned to the gray-box group only if physical reasoning determines the model structure itself, in the sense of Nielsen and Madsen [20, 21], while measured data are used to identify structural details or estimate model parameters.

Gray-box models in district-heating. Based on the literature reviewed, gray-box models appear to be relatively uncommon in short-term forecasting of the

aggregated load in district-heating networks. They are more commonly used to represent the thermal dynamics of individual buildings or to support the operational optimization of district-heating systems, including the operation of heat generators, thermal storage systems, and network components. One of the few identified forecasting approaches is that of Nielsen and Madsen, who developed a gray-box model for hourly forecasts of aggregated district-heating load with forecast horizons of up to 72 hours [20, 21]. Their approach combines physical knowledge with data-based statistical modeling: physical considerations determine the main model structure, while structural details and model coefficients are identified from measured data. Relations derived from heat transfer through walls and windows, solar gains, and ventilation or infiltration determine how the climate variables enter the model. Unlike detailed physical models, their specific approach uses physically motivated input–output relationships and does not require a detailed representation of individual buildings, physical states, or actual values of physical constants [20, 21].

Following a broader definition of gray-box modeling, the approach proposed by Dotzauer [6] can also be regarded as a gray-box model. The forecasting model combines a knowledge-based decomposition of heat demand into a temperature-dependent and a social component with data-based parameter estimation, using historical measurements to estimate the piecewise-linear temperature function and recurring calendar-based hourly load patterns [6].

The operational reference model by the Stadtwerke Karlsruhe GmbH (Germany) can be regarded as a gray-box model in a broad sense because domain knowledge determines its forecasting structure, whereas historical operational data are used to parameterize the individual model components.

Table 1: Studies forecasting the aggregated load of a district-heating network or of a substantial urban subnetwork, ordered alphabetically. All of them are data-driven; the gray-box studies are listed separately at the bottom. Reported error levels are deliberately not tabulated – network size, climate and load level differ too much across these studies for the numbers to be comparable. All models use an hourly resolution; Spoek et al. [26] additionally consider a 15-minute resolution. The horizon column reports the longest horizon considered in each study.

Reference	Model(s)	Forecast object	Features	Horizon
<i>Black-box forecasting of the aggregated network load</i>				
Bujalski et al. [2] ^a	GAM, GAMM (cubic splines, AR error)	Entire municipal network / central CHP plant output	Calendar variables (hour, day of week, day of year), lagged residuals (AR error); no weather inputs.	24 h
Dahl et al. [4]	OLS, MLP, SVR	Entire city-scale network (Aarhus)	Historical heat load, calendar and holiday variables, weather forecasts (air temperature, wind speed, solar irradiance), and measured weather.	38 h
Finkenrath et al. [7]	AdaBoost, RF, FNN, LSTM, Deep DHC ensemble	Substantial urban subnetwork (Ulm; 40 km, 1,100 households, 75 GWh/a)	Historical heat load (lags, rolling means), calendar variables, and weather forecasts (air temperature, dew point, wind, pressure, sunshine, visibility).	72 h
Hua et al. [12]	MLR, ANN	Multiple levels: entire city-scale network (Helsinki) and 5 individual buildings	Calendar variables (season, day, hour), past measured weather (temperature, solar radiation, wind speed), and measured future temperature.	1 h
Kurek et al. [18]	Linear/Ridge regression, ARX, fuzzy model, FFNN, AR-ANN	Multiple levels: entire city-scale network (Warsaw), regional areas, and substations	Historical heat load, calendar variables, total-network load forecast (hierarchical), measured and forecast weather (temperature, wind, humidity, solar radiation, cloudiness).	72 h

continued on next page

Table 1 – continued from previous page

Reference	Model(s)	Forecast object	Features	Horizon
Leiprecht et al. [19]	SARIMAX, AdaBoost, LSTM	Substantial urban subnetwork (Ulm; 40 km, 1,100 households, 75 GWh/a)	Historical heat load (lags, rolling means), calendar variables, and weather forecasts (air temperature, dew point, wind, pressure, sunshine, visibility).	72 h
Potočnik et al. [23]	Linear/robust regression, NN, ELM, RVFL, SVM, GPR	Entire city-scale network (Ljubljana; transmitted heat at plant border)	Historical heat load, calendar variables, past measured weather, weather forecasts (temperature, solar irradiance), and measured future weather.	48 h
Sakkas & Abang [25]	RNN, FFNN, 1D polynomial look-up table	Entire municipal network (Cottbus; 386 km piping, 20 substations)	Historical heat load, calendar variables, past measured weather (humidity, wind, sunshine, rain), and measured future temperature.	24 h
Spoek et al. [26]	TabPFN-TS, Chronos-2, AutoGluon suite, MRRC	Substantial urban subnetwork / district quarter (Munich) and city-scale network (Flensburg)	Historical heat load (rolling context), calendar variables, measured future temperature (primary), and weather forecasts.	24 h

Gray-box forecasting of the aggregated network load

Dotzauer [6]	Piecewise-linear regression + residual profile	Two municipal/regional networks (Stockholm; System 1: ~700 MW, System 2: ~300 MW)	Historical heat load, calendar variables (hour, day of week), and measured future temperature.	168 h
--------------	--	---	--	-------

continued on next page

Table 1 – continued from previous page

Reference	Model(s)	Forecast object	Features	Horizon
Nielsen & Madsen [20, 21]	Physical–statistical transfer functions, adaptive recursive least squares	Entire regional transmission network (VEKS Greater Copenhagen, > 8,000 TJ/a)	Calendar variables (day types), ambient temperature and solar radiation (low-pass transfer functions), wind speed, AR error lags	72 h

^aBujalski et al. [2] consider only the summer off-season from June to August; the models use calendar variables and lagged model residuals without weather inputs.

Abbreviations: AdaBoost, adaptive boosting; ANN, artificial neural network; AR-ANN, autoregressive ANN; ARX, autoregression with exogenous inputs; ELM, extreme learning machine; FFNN, feed-forward neural network (in Kurek et al. an ANN without autoregressive input); FNN, feed-forward neural network; GAM(M), generalized additive (mixed) model; GPR, Gaussian process regression; LSTM, long short-term memory network; MLP, multi-layer perceptron; MLR, multiple linear regression; MRRC, multi-resolution residual-correction forecaster; NN, neural network; OLS, ordinary least squares; RF, random forest; RNN, recurrent neural network; RVFL, random vector functional link network; SARIMAX, seasonal autoregressive integrated moving average with exogenous regressors; SVM, support vector machine; SVR, support vector regression.

Foundation models for district-scale and city-scale district-heating forecasting. Spoek et al. [26] test TabPFN-TS zero-shot against Chronos-2 and a suite of trained baselines for probabilistic heat load forecasting on two German networks: an expanding mixed-use district in Munich as the main proprietary case, and the city-scale Flensburg network with open data as a transfer check. Both zero-shot models outperform every trained baseline in the day-ahead setting, with Chronos-2 achieving the lowest aggregate error and TabPFN-TS providing better-calibrated prediction intervals than Chronos-2. Their baselines are, however, generic models fitted per network rather than a forecasting model in operational use, and the reported accuracies assume perfect future weather, with a separate sensitivity analysis quantifying the degradation under archived numerical weather forecasts.

Contributions.

- **A benchmark against a long-running operational gray-box model.** Previous studies have compared newly developed forecasting models with commercial forecasts used in district-heating operation [4,25]. In contrast, the present work uses a documented gray-box model, the operational

reference model, whose forecasting structure and inputs are physically and operationally interpretable. The model has remained in operation without reparameterization since 2017. Its performance is assessed over eight heating seasons, and its unchanged operational configuration serves as the reference in the held-out 2025/26 benchmark.

- **A controlled comparison under a common information set.** All benchmark models receive the same operational input information: the measured heat load and plant-level outdoor temperature of a reference-day, the operational 72 h outdoor-temperature forecast, and calendar information. Unlike time-series forecasting approaches that use a longer historical load context, this setup holds the available information constant across model classes.
- **Tabular foundation models under operational input constraints.** Complementing recent work on time-series foundation models for district-heating forecasting [26], this work evaluates TabPFN and TabFM as tabular regression models using only the compact input vector of the operational forecasting system. Their pretrained weights remain frozen, and no task-specific weight updates are performed.
- **Interpretation of the best-performing linear model.** The best-performing linear model is examined through its coefficients and input-block contributions. This analysis shows how the model combines the reference-day profile, temperature forecast, and calendar information over the forecast horizon.

3 Data

All results in this work are derived from the operational data of the Karlsruhe district-heating network. This section first introduces the network and its load–temperature relationship, then describes the sources of the data, and finally explains how the raw record is turned into forecast samples and split chronologically into training, validation, and test seasons.

The network. The Karlsruhe network, operated by the Stadtwerke Karlsruhe GmbH (Germany), is among the larger district-heating systems in Germany, with roughly 240 km of piping in 2025 [27] supplying more than 55,000 residential and commercial units [28]. For 2025, the Stadtwerke Karlsruhe GmbH (Germany) reports 885.8 GWh fed into the network and 778.8 GWh delivered to customers, corresponding to distribution losses of 12.1 % [27,28]. In 2025, the MiRO refinery supplied the largest share of heat to the Karlsruhe district-heating network at 57.9 %, followed by the EnBW Rheinhafen steam power plant at 26.6 % and the Schwarz Produktion paper mill in Maxau at 7.0 %. The Stadtwerke Karlsruhe GmbH (Germany) generated the remaining 8.5 % in its own facilities [27].

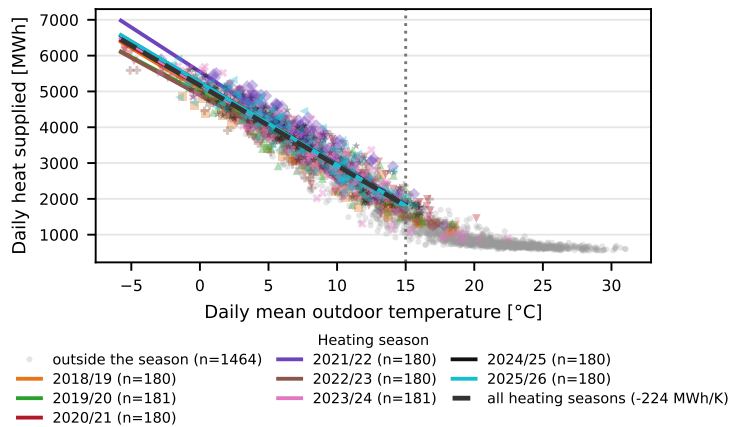


Figure 1: Daily heat supplied against daily mean outdoor temperature, with one point per complete day over the entire record. Both quantities are calculated from exactly 24 hourly values; thus, the daylight-saving days with 23 or 25 values are left out, as is the 2017/18 heating season. The dotted line marks 15 °C, above which the load approaches an approximately temperature-independent base level. Colored lines are per-season least-squares fits below that threshold, the dashed line the fit pooled over all seasons.

Sources. The first source is the measurement record. It is the hourly operational log of the network from November 2017 to April 2026, recorded by the Stadtwerke Karlsruhe GmbH (Germany): the total heat supplied to the network and the outdoor temperature measured at the plant site. The heat supplied to the

network is governed by the outdoor temperature. Figure 1 shows the daily load against the daily mean temperature for the full year, including both the summer months and the heating season. Below approximately 15 °C, the relation is almost linear; above it, the daily heat supplied levels off at a temperature-independent base level associated with domestic hot-water supply and network losses. The per-season fits are similar, so the load–temperature relationship underlying the operational reference model remains stable across the record. The second data source is the 72 h air-temperature forecast, a product provided to the Stadtwerke Karlsruhe GmbH (Germany) by an external provider.

From raw record to samples. Forecast samples are derived from these data. A calendar day is considered complete if it contains exactly 24 hourly values. Days containing 23 or 25 hourly values occur at the daylight-saving-time (DST) transitions and are excluded. A five-day window (reference-day, forecast-issue day, and the three-day horizon) becomes a sample if its days are consecutive and complete, if its forecast day falls in the heating season (i.e., October–March), if the temperature forecast exists, and if all 73 target values are present. Since the first season 2017/18 is incomplete, it is not considered. This procedure yields 1369 samples across the eight heating seasons from 2018/19 to 2025/26. Figure 2 resolves this count by day and gives the reason for every exclusion: the recurring gaps at the end of March and at the turn of October and November are forecast days whose five-day window contains a daylight-saving changeover. The remaining days are excluded for missing temperature forecasts.

Split. The split is chronological and cuts along season boundaries. The last season 2025/26 is considered the test season (164 forecast days), the second-to-last season (2024/25) the validation season (168 forecast days), and the six seasons from 2018/19 to 2023/24 the training seasons (1037 forecast days).

4 Experimental Setting

This section defines the forecast task, the operational reference and benchmark models, and the metrics used for their evaluation.

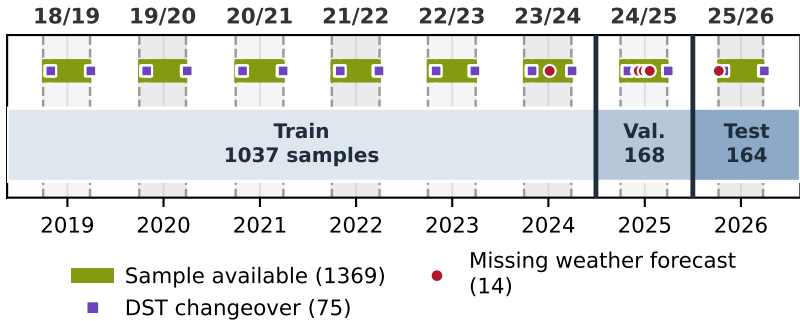


Figure 2: Every day of the eight heating seasons the benchmark models use, classified by whether it yields a sample and otherwise by the reason for its exclusion, with the chronological split on the same axis. It shows the days that yield no sample and those excluded from the training data for other reasons.

4.1 The forecast task

Each forecast day yields a five-day window. The 129-dimensional feature vector contains 25 hourly values of the reference-day heat load, 25 corresponding values of the outdoor temperature measured at the plant site, 72 hourly values of the air-temperature forecast, and a seven-dimensional one-hot encoding of the weekday of the first forecast day. Figure 3 illustrates this composition. The target is the 73 hourly values of the horizon.¹ We consider the task described above as a direct multi-output regression from a fixed feature vector onto the 73 hourly values of the horizon, rather than as an auto-regressive continuation of the load series. The reason is comparability: the operational reference model uses the reference-day two days before the forecast day, an air-temperature forecast, and calendar information, but no additional load history. For this reason, every benchmark model also receives the same inputs.

¹ For compatibility with other systems, the horizon spans 73 h, from 00:00 of the forecast day to 00:00 of forecast day +3, sampled at the 73 grid points that include both midnight endpoints (and thus also the last hour of the issue day); the temperature forecast therefore covers hours 1...72. Every value carries the timestamp at which its interval *ends*, so grid point 0 is the hour from 23:00 to 00:00 and thus the closing hour of the forecast-issue day. Grid point 0 has no temperature forecast of its own; the operational reference model reuses the value of hour 1 for that one input, while the benchmark models emit it as one of their 73 outputs.

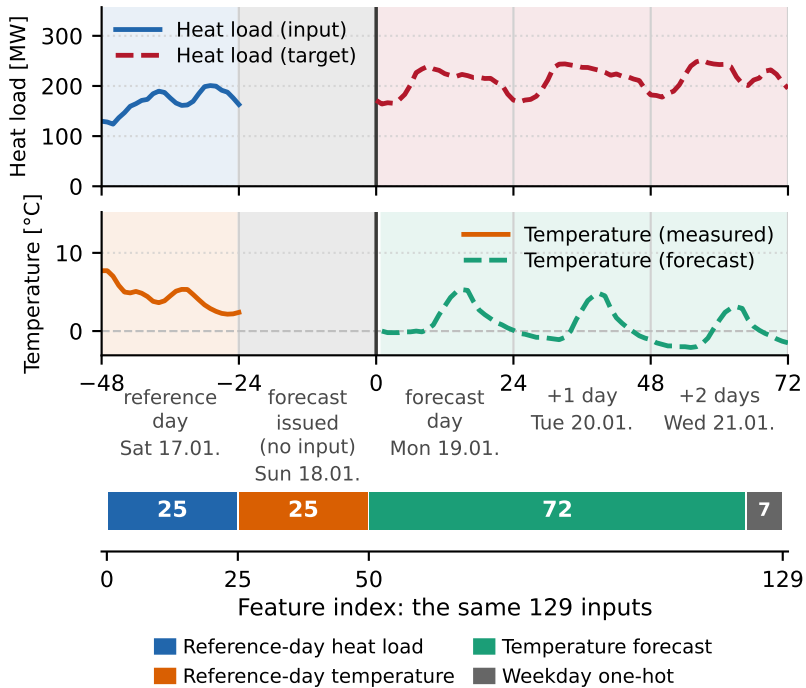


Figure 3: The forecast task. *Top and center:* one window from the test season, forecast day Monday 19 January 2026, with hour 0 at 00:00 of that day, so the reference-day spans hours $-48 \dots -24$; the gray band marks the issue day, which is not an input. *Bottom:* the same quantities as the 129-dimensional input vector. The feature index is the axis along which the coefficient matrix of Figure 5 (a) is read. In this case, the binary feature for the weekday is Monday.

This is also the reason why time-series foundation models are absent from the set of benchmark models. Chronos-2 [1], TimesFM 3.0 [5], and TabPFN-TS [11] forecast from a long history of the target series. Supplying that history would provide them with an input that the operational reference model does not have. It would further require the issue day to be masked and a decision on whether the history carries measured, forecast, or corrected temperature.

4.2 Operational reference and benchmark models

Table 3 in the Appendix lists all benchmark models, the operational reference model, and the naive baselines with their implementation and preprocessing.

The *operational reference model* always denotes the gray-box model in operation, evaluated in its deployed configuration; for this work, we re-implemented it from the operational tool with its current parameters. The operational reference model starts from the measured load profile of the reference-day and adjusts it hour by hour in two steps. First, a temperature correction shifts each hourly value in proportion to the difference between the forecast air temperature and the outdoor temperature measured at the same hour on the reference-day. Above 15 °C, the load is treated as independent of temperature. Second, a calendar correction maps the weekday of the reference-day onto the weekday of the forecast day using fixed hourly weekday corrections. The colder it is, the more weight this correction receives. Three scalar parameters are involved: a temperature factor, a constant correction of the temperature forecast, and a weight for the calendar term. These parameters and the calendar correction are estimated from operational data before deployment in 2017. Importantly, the operational reference model is not refitted using the data considered in this work.

The *naive baselines* repeat the load of the reference-day or the heat load of the same day one week earlier and thus deviate from the equal input data approach for broader comparability. In contrast to the operational reference model, all benchmark models are black-box models in the sense of Nielsen and Madsen [20, 21]: they learn the forecasting relationship from data, without a structure derived from physical or operational knowledge. The benchmark models are three linear

regressions (ordinary least squares, *OLS*; *Ridge*; and *ElasticNet*), a simple neural network (a multi-layer perceptron, *MLP*), and two tree-based models (extreme gradient boosting, *XGBoost*; and light gradient-boosting machine, *LightGBM*). The three linear models differ in their regularization: OLS uses none, Ridge uses an L_2 penalty on the coefficients, and ElasticNet uses a combination of L_1 and L_2 that drives some coefficients to exactly zero. In addition, we evaluate two tabular foundation models: the tabular prior-data fitted network *TabPFN* and the tabular foundation model *TabFM*. Although their weights stay frozen, they hold all 1037 training days in their context at inference time. They do not receive more information about the district-heating network, since all fitted models draw on the same training days. However, these models use the information differently: instead of condensing it into fitted parameters, they combine the full training set with a prior learned during pretraining. Supplying the training days as context can therefore be regarded as their form of training. Where Table 2 marks a benchmark model (*tuned*), its hyperparameters are selected based on its MAE on the validation season.

4.3 Evaluation

One forecast day is one sample and carries the $H = 73$ hourly values of its horizon. The metrics Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE) are first computed separately for each forecast day over its 73 target values. The values reported in Table 2 are then obtained by averaging the daily metric values across all forecast days of the evaluated season. For a day with forecast $\hat{y}$ and measurement y , they are defined as follows:

$$\begin{aligned} \text{MAE} &= \frac{1}{H} \sum_{t=0}^{H-1} |\hat{y}_t - y_t|, \\ \text{RMSE} &= \sqrt{\frac{1}{H} \sum_{t=0}^{H-1} (\hat{y}_t - y_t)^2}, \\ \text{MAPE} &= \frac{100}{H} \sum_{t=0}^{H-1} \left| \frac{\hat{y}_t - y_t}{y_t} \right|. \end{aligned} \tag{1}$$

Hyperparameters are selected by MAE on the validation season (Table 4 in the Appendix), and the scaler is fitted on the training seasons alone. No benchmark model is fitted or selected on the test season. All results on the test season reported in Table 2 are based on its 164 forecast days.

5 Results

We first examine the operational reference model over the eight available seasons and then compare the benchmark models on the test season.

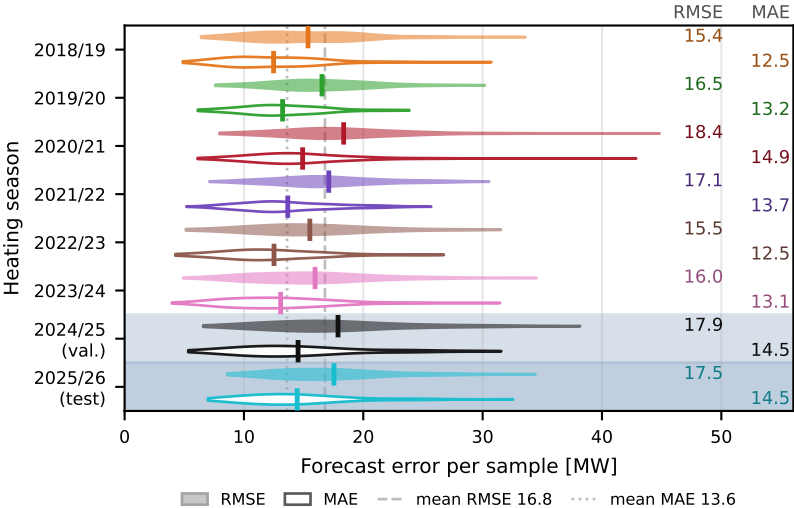


Figure 4: Per-day error of the operational reference model across the eight available seasons. Color identifies the season; the filled violin shows the RMSE and the outlined one the MAE; the vertical line within each violin indicates the mean. For the 2025/26 test season, these means correspond to the values of the operational reference model reported in Table 2. The vertical lines are the unweighted season means, against which a drift would show.

The operational reference model has held its level for eight seasons. The operational reference model was established in 2017 and has not been refitted since. Because its parameters were determined before the evaluation

period, every season can be seen as a valid test season for this model. Figure 4 shows the errors across the different seasons. Its seasonal mean MAE ranges from 12.5 MW to 14.9 MW, without an apparent temporal trend over the years. Differences between seasons are reflected in both the mean and the median of the daily errors.

Table 2: Forecast error on the held-out 2025/26 test season comprising 164 forecast days. MAE and RMSE are reported in MW, and MAPE in percent. Each value is the mean of the corresponding daily metric across the test season. The best value in each column is shown in bold. The final two blocks report the operational reference model and the naive baselines, respectively.

Model	MAE ↓	RMSE ↓	MAPE ↓
ElasticNet (tuned)	10.78	13.27	7.73
Ridge (tuned)	10.87	13.36	7.80
LightGBM (tuned)	11.16	13.89	7.75
XGBoost (tuned)	11.30	13.94	7.88
MLP (tuned)	11.74	14.37	8.17
Linear regression (OLS)	11.90	14.60	8.59
TabPFN (v3)	10.69	13.26	7.43
TabFM (v1.0)	10.97	13.55	7.61
Operational reference model	14.46	17.54	10.67
Naive baseline: reference-day	24.05	28.49	16.52
Naive baseline: previous week	29.90	34.47	20.67

Benchmark against the other models. Across the 164 days of the test season, every benchmark model outperforms the operational reference model (see Table 2), reducing the MAE from 14.46 MW to as low as 10.69 MW, corresponding to an improvement of up to 26 %. TabPFN achieves the lowest MAE, followed closely by the ElasticNet model (10.78 MW). Ridge and TabFM also achieve similar error levels, with MAEs of 10.87 MW and 10.97 MW, respectively. LightGBM (11.16 MW) and XGBoost (11.30 MW) follow; in MAE and RMSE, both gradient-boosting models are worse than the regularized linear models. The MLP, at 11.74 MW, ranks second to last among the benchmark models, ahead only of ordinary least squares (11.90 MW). The operational reference model, in turn, performs clearly better than both naive baselines.

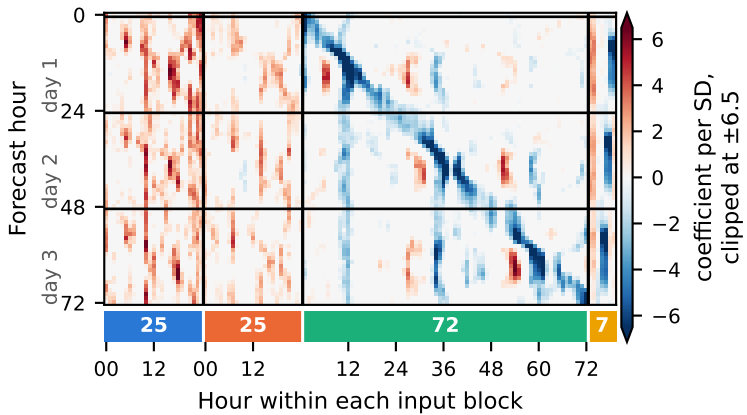
A linear model reaches almost the whole gain. ElasticNet exceeds the MAE of TabPFN by only 0.09 MW. Relative to the operational reference model, ElasticNet retains 97.6 % of the MAE reduction achieved by TabPFN, corresponding to 3.68 MW of the 3.77 MW reduction.

6 Interpretation of the ElasticNet Model

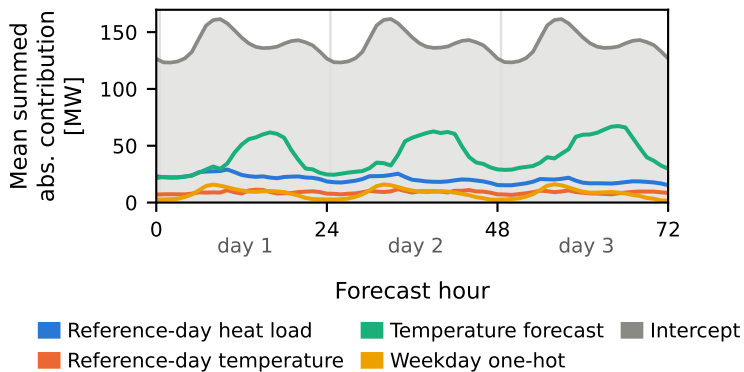
The three linear models differ mainly in their type of regularization. We examine ElasticNet in more detail because it achieves the lowest test-season MAE among the linear models and provides a sparse coefficient matrix. Its L_1 penalty sets coefficients exactly to zero, which makes it the sparsest of the three linear models with 58 % of the 9417 coefficients set to zero. We look at ElasticNet from two angles (Figure 5): the coefficient matrix and the mean absolute contribution of each input block over the horizon.

The coefficient matrix (see Figure 5a) reflects the structure of the inputs. Within the 72 temperature-forecast input columns, the largest coefficients form a near-diagonal band: each forecast output hour is associated primarily with the temperature forecast for the same hour and the adjacent hours $t - 1$ and $t + 1$. The division of this band into three visible segments results from the day boundaries in the forecast horizon. Additional off-diagonal coefficients link the midday temperature forecast of the first day to outputs on the second and third days, and the corresponding inputs of the second day to outputs on the third day. The final seven input columns represent the weekday block and encode a learned calendar correction. By contrast, the coefficients associated with the reference-day load and temperature are more dispersed and do not exhibit an equally clear temporal pattern.

Viewed over the forecast horizon (see Figure 5b), which shows the mean contribution of each block over the test season, the temperature forecast contributes most from late morning into the afternoon of each day and least at night, while the calendar correction counts less at night than by day. The reference-day load and the reference-day temperature contribute less.



(a) Coefficient matrix, with one row per forecast output hour and one column per input feature. Color indicates the coefficient associated with a one-standard-deviation change. Values are clipped symmetrically at the 99th percentile of the absolute coefficient values.



(b) Mean absolute contribution of each input block in MW at every output hour, averaged across the test season. Contributions are obtained by summing the absolute contributions of the inputs within each block. The gray curve shows the intercept, which corresponds to the forecast horizon profile when all standardized inputs are zero. It therefore represents a baseline level rather than an input contribution.

Figure 5: Interpretation of the selected ElasticNet model.

The learned weekday pattern. The seven columns of the weekday block encode a calendar correction. Figure 6 illustrates the relationship among weekdays based on the mean daily load profile for each weekday over the six training seasons; they serve as a reference for the correction shown in Figure 7. The working days show only small differences of a few MW, whereas Saturday and Sunday profiles remain below the mean profile across all days. This difference is most pronounced during the morning peak, which is substantially flatter on weekends.

Figure 7 shows the contribution of the first forecast day feature. It is obtained by applying the fitted standardization to the one-hot vector for the first forecast day d and multiplying the resulting vector by the coefficient columns of the weekday block. This contribution is added to the intercept. The resulting pattern in Figure 7 therefore resembles the final seven columns of the coefficient matrix in Figure 5a but is not identical to them as they are individual coefficients. Because setting one weekday in the one-hot encoding simultaneously determines the other six entries, an individual column cannot be interpreted in isolation. Only the contribution of the block as a whole is therefore meaningful. The visualized contributions in Figure 7 show how the model adjusts the reference-day-based forecast to the weekday structure of the forecast horizon. When the reference-day is a Saturday or Sunday, the subsequent working days are corrected upward toward the working-day profile. Conversely, forecast hours falling on a Saturday or Sunday are corrected downward. For transitions between working days, the correction remains close to zero.

7 Discussion

We discuss the implications of the results for a utility like the Stadtwerke Karlsruhe GmbH (Germany), how much model complexity is required to achieve the observed accuracy gains, and which aspects are not covered in the present work.

Operational implications of the results. All benchmark models beat the operational reference model; the best, TabPFN, lowers the MAE from 14.46 MW

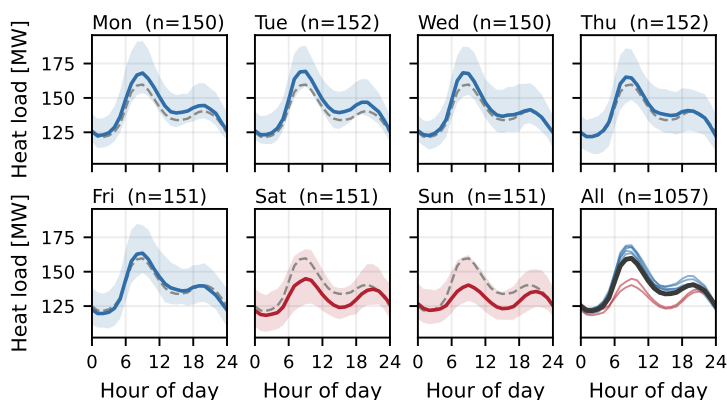
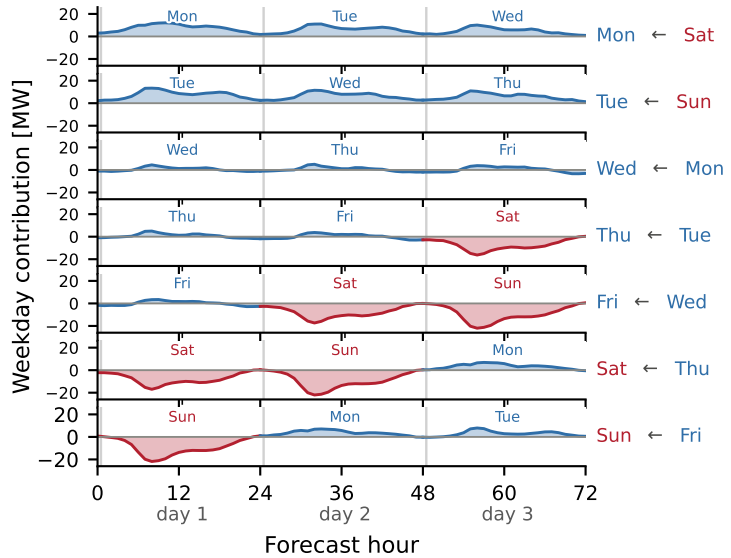


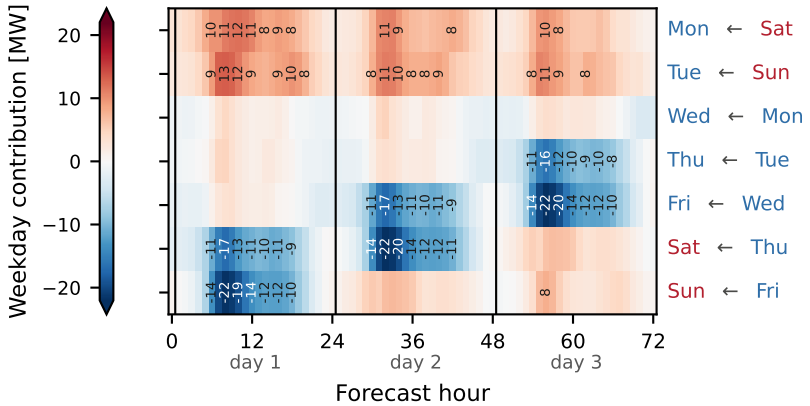
Figure 6: Mean daily load profile per weekday over the six heating seasons (2018/19–2023/24), public holidays excluded; n is the number of days per panel. Color marks the kind of day forecast, working day in blue and weekend in red, and is used the same way in Figure 7. The dashed gray curve, identical in all seven panels, is the mean over all days, so the gap between the curves is the weekday effect. The band spans the minimum and maximum of the per-season means of that weekday for all six seasons: for the working days this year-to-year spread exceeds the distance between the weekdays themselves. The last panel overlays the seven means and the all-day mean.

to 10.69 MW. Unlike the operational reference model, the benchmark models may draw on any hour of the input rather than on forecast hour t alone; this may allow them to capture temporal dependencies across the forecast horizon. The operational reference model is nonetheless not weak: it holds 12.5 MW to 14.9 MW across eight unseen seasons without ever being refitted and clearly beats the naive baselines. For the operator, this means that, on the test season, the forecast error could be reduced by about a quarter without additional data, since all benchmark models use only the inputs already available to the operational reference model. In particular, ElasticNet achieves almost the entire gain while its forecast remains a weighted sum of these inputs; it could therefore be evaluated with an effort comparable to that of the operational reference model and remain open to inspection by the operators who review the forecast.

The amount of complexity required to improve on the operational reference model is considerable. The leading models achieve similar error levels (Section 5). The remaining 2.4 % of the improvement between ElasticNet and



(a) One panel for each forecast day, corresponding to the rows in (b). Vertical lines mark the day boundaries, and each section is labeled with the weekday it covers.



(b) The same contribution shown as a heatmap in (a), with one row per forecast day $\leftarrow$ reference-day.

Figure 7: The weekday block of the ElasticNet in MW; blue working day, red weekend. The negative blocks follow Saturday and Sunday wherever they fall in the horizon. Both visualizations show how much is added to the intercept, that is, the mean profile (gray curve in Figure 5b), when the one-hot encoded weekday of the forecast day is set, with all other inputs at their training mean. Zero indicates no contribution from the weekday block relative to the intercept.

TabPFN costs the ability to read the forecast off the model, and the more complex gradient-boosting models and the MLP do not improve on the regularized linear models.

One possible explanation for the strong performance of the regularized linear models is that the input representation already provides the main determinants of the forecast: the reference-day load anchors the load level and profile, the temperature forecast describes the expected non-linear weather conditions, and the weekday block represents recurring calendar effects. Given a weather forecast as input, much of the remaining mapping may be approximately linear over the observed heating seasons. This interpretation remains a hypothesis and would require targeted ablation or nonlinear-dependence analyses for confirmation.

Beyond forecast accuracy, additionally, interpretability is important in the present operational context because forecasts inform dispatch, storage operation, and fuel procurement. ElasticNet supports model inspection through its sparse coefficient matrix, identifiable input blocks, and learned weekday correction. This provides a higher degree of model-level transparency than the nonlinear benchmark models, although it does not by itself guarantee that every individual forecast can be fully explained.

Limitations. Our findings are subject to several limitations. First, this work covers a single network run by a single operator. It therefore varies neither network size nor climate nor operating practice. Whether the results transfer to other district-heating networks also remains untested.

Second, the evaluation relies on a single test season of 164 days. How strongly a single winter can shift the results is illustrated by the operational reference model itself: evaluated across the eight seasons, its error ranges from 12.5 MW to 14.9 MW. Since the best-performing models differ by far less than this, their exact ranking should be interpreted with caution; evaluating all benchmark models on several seasons would make it more robust. In addition, the preconditions of the comparison between the operational reference model and the benchmark models are not identical, because only the latter are fitted on the training seasons. In the operational past, the parameters of the operational reference model were

also re-estimated, which yielded only marginal improvements. This suggests that the remaining gap stems from the structure of the gray-box model itself rather than from its calibration.

Third, the set of benchmark models is limited. We evaluate only two tabular foundation models, each in a single version. As described in Section 4.2, the two tabular foundation models are not fitted in the same way as the other benchmark models. Moreover, time-series foundation models are a natural candidate for this task but are absent from the comparison, because they do not fit our data setup, which holds every benchmark model to the same information base. Including them would require masking the issue day, providing a long historical context, and deciding whether to use measured or forecast temperature, or how to combine them. Further adding covariates for calendar features would be important; covariate-informed models such as Chronos-2 [1] offer a natural way to do so: Kreusel et al. [17] apply them to probabilistic building energy forecasting, and Kreusel et al. [16] evaluate time-series foundation models for cooling demand. Furthermore, the second requirement, explainability, could be addressed: Hertel et al. [8] show that covariate-informed time-series foundation models can provide explainable load forecasts.

Finally, the forecast setting itself is restricted. Because the information base is fixed by design, influences it does not contain, such as holidays, are only captured if they are added explicitly. Furthermore, all forecasts are point forecasts, whereas a probabilistic setting could be more natural for operational planning and may improve decisions. Moreover, it could be interesting to examine how errors in the temperature forecast propagate into errors of the heat load forecast.

8 Conclusion

District-heating forecasts support unit dispatch, storage operation, and fuel procurement, so they must be both accurate and interpretable. At the Stadtwerke Karlsruhe GmbH (Germany), a transparent gray-box model has provided these forecasts since 2017. In this paper, we benchmark the operational reference model against linear models, gradient-boosting models, a neural network, and

two tabular foundation models, all given the same inputs, to determine how much accuracy can be gained and at what cost in interpretability.

On a held-out test season of 164 days, every benchmark model outperforms the operational reference model, and the best, TabPFN, reduces the MAE from 14.46 MW to 10.69 MW. A regularized linear model, ElasticNet, reaches 10.78 MW and thus 97.6 % of this improvement while remaining interpretable: its coefficients reflect the input structure, and it learns a calendar correction.

Future work could pursue several directions. First, it could incorporate time-series foundation models, which would require adapting the data setup. Second, all models could be extended with additional inputs (e.g., holidays, more detailed weather variables) and reformulated in a probabilistic setting, for example via quantile regression or distributional outputs, to support uncertainty-aware decisions. Third, dedicated experiments could quantify how temperature-forecast errors propagate into heat load forecast errors. Finally, evaluating the benchmark models across multiple test seasons and replicating the analysis across networks, operators, and climates would provide stronger evidence regarding the robustness and external validity of the findings.

Acknowledgments

The authors thank the Helmholtz Association for its support under the program “Energy System Design” and its predecessor, the Joint Initiative “Energy System 2050 – A Contribution of the Research Field Energy”. In addition to the Helmholtz support, this work is also part of the research partnership “Forschungspartnerschaft Karlsruhe Energie” (FPS KA Energie), within which the four partners Karlsruhe Institute of Technology (KIT), the Stadtwerke Karlsruhe GmbH, the Stadtwerke Karlsruhe Netzservice GmbH, and the DVGW Research Center at the Engler-Bunte-Institut of KIT intensify and coordinate their collaboration on sustainable energy supply.

Declaration of Generative AI and AI-Assisted Technologies

During the preparation of this work, the authors used Claude Code (Anthropic), running the Claude Opus model, to assist with the development of the analysis and evaluation code and with structuring and language editing of the manuscript. No other generative AI tools were used. All scientific content, including the study design, the experimental protocol, the evaluation and the interpretation of the results, was conceived and verified by the authors. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

- [1] Abdul Fatir Ansari et al. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025. doi:10.48550/arXiv.2510.15821.
- [2] Maciej Bujalski, Paweł Madejski, and Krzysztof Fuzowski. Day-ahead heat load forecasting during the off-season in the district heating system using generalized additive model. *Energy and Buildings*, 278:112630, 2023. doi:10.1016/j.enbuild.2022.112630.
- [3] 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 (KDD)*, pages 785–794, 2016. doi:10.1145/2939672.2939785.
- [4] Magnus Dahl, Adam Brun, Oliver S. Kirsebom, and Gorm B. Andresen. Improving short-term heat load forecasts with calendar and holiday data. *Energies*, 11(7):1678, 2018. doi:10.3390/en11071678.
- [5] Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

- [6] Erik Dotzauer. Simple model for prediction of loads in district-heating systems. *Applied Energy*, 73(3-4):277–284, 2002. doi:10.1016/S0306-2619(02)00078-8.
- [7] Matthias Finkenrath, Till Faber, Fabian Behrens, and Stefan Leiprecht. Holistic modelling and optimisation of thermal load forecasting, heat generation and plant dispatch for a district heating network. *Energy*, 250:123666, 2022. doi:10.1016/j.energy.2022.123666.
- [8] Matthias Hertel, Alexandra Nikoltchovska, Sebastian Pütz, Benjamin Schäfer, Ralf Mikut, and Veit Hagenmeyer. Explainable load forecasting with covariate-informed time series foundation models. In *Proceedings of the 17th ACM International Conference on Future and Sustainable Energy Systems*, E-Energy '26, pages 612–626, New York, NY, USA, June 2026. Association for Computing Machinery. doi:10.1145/3744255.3811724.
- [9] Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. doi:10.1080/00401706.1970.10488634.
- [10] Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeyer, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025. doi:10.1038/s41586-024-08328-6.
- [11] Shi Bin Hoo, Samuel Müller, David Salinas, and Frank Hutter. The tabular foundation model TabPFN outperforms specialized time series forecasting models based on simple features. *arXiv preprint arXiv:2501.02945*, 2025. doi:10.48550/arXiv.2501.02945.
- [12] Pengmin Hua, Haichao Wang, Zichan Xie, and Risto Lahdelma. District heating load patterns and short-term forecasting for buildings and city level. *Energy*, 289:129866, 2024. doi:10.1016/j.energy.2023.129866.
- [13] 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 *Advances in Neural Information Processing Systems 30 (NeurIPS)*, pages 3146–3154, 2017.

- [14] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. arXiv:1412.6980.
- [15] Weihao Kong, Erez Louidor Ilan, Shuxin Nie, Taman Narayan, Rajat Sen, Yichen Zhou, Deqing Fu, Samet Oymak, and Abhimanyu Das. TabFM: A zero-shot foundation model for tabular data. *arXiv preprint arXiv:2609.37959*, 2026. doi:10.48550/arXiv.2609.37959.
- [16] Alexander Kreusel, Matthias Hertel, Moritz Noskiewicz, Heiko Maaß, Ralf Mikut, and Veit Hagenmeyer. Evaluating time-series foundation models for cooling demand forecasting with little data. In *Proceedings 35. Workshop Computational Intelligence: Berlin, 20.–21. November 2025*, page 301, 2025.
- [17] Alexander Kreusel, Matthias Hertel, Moritz Noskiewicz, Frederik Zahn, Heiko Maaß, Ralf Mikut, and Veit Hagenmeyer. Covariates are the key to accurate probabilistic building energy forecasting with time series foundation models. In *Proceedings of the 2026 ACM Sustainability Week, ACM Sustainability Week '26*, pages 172–182, New York, NY, USA, June 2026. Association for Computing Machinery. doi:10.1145/3765611.3815507.
- [18] Teresa Kurek, Artur Bielecki, Konrad Świrski, Konrad Wojdan, Michał Guzek, Jakub Białek, Rafał Brzozowski, and Rafał Serafin. Heat demand forecasting algorithm for a Warsaw district heating network. *Energy*, 217:119347, 2021. doi:10.1016/j.energy.2020.119347.
- [19] Stefan Leiprecht, Fabian Behrens, Till Faber, and Matthias Finkenrath. A comprehensive thermal load forecasting analysis based on machine learning algorithms. *Energy Reports*, 7:319–326, 2021. doi:10.1016/j.egyr.2021.08.140.
- [20] Henrik Aalborg Nielsen and Henrik Madsen. Predicting the heat consumption in district heating systems using meteorological forecasts. Technical Report IMM-REP-2000-16, Department of Mathematical Modelling, Technical University of Denmark, Lyngby, Denmark, 2000.

- [21] Henrik Aalborg Nielsen and Henrik Madsen. Modelling the heat consumption in district heating systems using a grey-box approach. *Energy and Buildings*, 38(1):63–71, 2006. doi:10.1016/j.enbuild.2005.05.002.
- [22] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [23] Primož Potočnik, Primož Škerl, and Edvard Govekar. Machine-learning-based multi-step heat demand forecasting in a district heating system. *Energy and Buildings*, 233:110673, 2021. doi:10.1016/j.enbuild.2020.110673.
- [24] Nicolas Renkamp. Short-term load forecasting in district heating networks. Master’s thesis, Karlsruher Institut für Technologie, 2016.
- [25] Nikolaos P. Sakkas and Roger Abang. Thermal load prediction of communal district heating systems by applying data-driven machine learning methods. *Energy Reports*, 8:1883–1895, 2022. doi:10.1016/j.egy.2021.12.082.
- [26] Ben Spoek, Karim K. Ben Hicham, Kai Derzsi, Philipp Althaus, Alexander Mitsos, and Dirk Müller. Systematic evaluation of TabPFN-TS for zero-shot probabilistic heat load forecasting in district heating networks. arXiv:2608.20024, 2026. doi:10.48550/arXiv.2608.20024.
- [27] Stadtwerke Karlsruhe GmbH. Umwelterklärung 2026 mit Klimareport und Energiebericht, May 2026.
- [28] Stadtwerke Karlsruhe GmbH. Zahlen, Daten und Fakten zur Fernwärme in Karlsruhe, 2026. Kennzahlen Stand 01.04.2026, Energieträgermix 2025.
- [29] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67(2):301–320, 2005. doi:10.1111/j.1467-9868.2005.00503.x.

The Benchmark Models in Detail

Table 3 specifies every benchmark model with its implementation and preprocessing, and Table 4 the search spaces and selected configurations.

Standardization. Each benchmark model is built as a scikit-learn pipeline consisting of a `StandardScaler` and the estimator [22], the scaler fitted on the training seasons alone and applied unchanged to the validation and test seasons. For the tree ensembles, a monotonic rescaling leaves the splits untouched; it is kept only so that the pipeline is identical across learners. The MLP alone also standardizes the 73 targets and inverts that transformation before any reported error is computed; all errors are reported in MW.

The two tabular foundation models. Both are applied without task-specific training: their weights stay frozen and the 1037 training days are supplied as in-context examples.² TabFM’s limit on context rows per ensemble member is disabled so that all training days are available; TabPFN has no such limit at this data set size.

² We avoid the term *zero-shot* here. The weights are never updated, but the models do see the 1037 labeled training days as context, which makes this in-context learning rather than forecasting without examples.

Table 3: Every benchmark model, together with the operational reference model and the naive baselines: implementation, preprocessing and source. *Feat.* and *Targ.* mark standardization of the 129 inputs and of the 73 targets, fitted on the training seasons only; the operational reference model and the naive baselines are not scaled.

Model	Implementation	Feat.	Targ.	Reference
<i>Naive baselines</i>				
Naive baseline: reference-day	take the reference-day as forecast	n/a	n/a	n/a
Naive baseline: previous week	repeat the same 73 h one week earlier	n/a	n/a	n/a
<i>Operational reference</i>				
Operational reference model	own	n/a	n/a	gray-box archetype
<i>Learned models</i>				
ElasticNet (tuned)	<code>ElasticNet</code>	✓	n/a	[22, 29]
Ridge (tuned)	<code>Ridge</code>	✓	n/a	[9, 22]
LightGBM (tuned)	<code>LGBMRegressor</code> $\times$ 73	✓	n/a	[13]
XGBoost (tuned)	<code>XGBRegressor</code> , multi-output tree	✓	n/a	[3]
MLP (tuned)	<code>MLPRegressor</code> , Adam	✓	✓	[14, 22]
Linear regression (OLS)	<code>LinearRegression</code>	✓	n/a	[22]
<i>Tabular foundation models</i> (frozen weights, training days as in-context examples)				
TabPFN (v3)	<code>tabpfn</code> , 73 sequential regressions	n/a	n/a	[10]
TabFM (v1.0) Ensemble n=32	<code>tabfm</code> , 73 sequential regressions	n/a	n/a	[15]

Table 4: Search spaces and the configuration selected by MAE on the 2024/25 validation season (Val. MAE). For the MLP, the grid search ranks the configurations by MAE on the standardized targets; the Val. MAE is given in MW. The selected models keep their fit on the training seasons and are not refitted on the validation season. Benchmark models not listed here use fixed settings.

Model	Search space (grid)	Selected	Val. MAE ↓
ElasticNet (tuned)	$\alpha \in \{0.001, 0.01, 0.05, 0.1, 0.3, 1.0\}$; $l1_ratio \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$	$\alpha = 0.1$; $l1_ratio = 0.9$	10.85
LightGBM (tuned)	$n_estimators \in \{100, 200, 400, 800\}$; $num_leaves \in \{15, 31, 63\}$; $learning_rate \in \{0.03, 0.05, 0.1\}$	$learning_rate = 0.03$; $n_estimators = 200$; $num_leaves = 15$	10.83
MLP (tuned)	$hidden_layer_sizes \in \{(100,), (128, 64), (256, 128), (256, 128, 64), (512, 256, 128)\}$; $\alpha \in \{0.0001, 0.001, 0.01\}$	$\alpha = 0.0001$; $hidden_layer_sizes = (100,)$	10.64
Ridge (tuned)	$\alpha \in \{0.1, 1.0, 3.0, 10.0, 30.0, 100.0, 300.0\}$	$\alpha = 30.0$	10.87
XGBoost (tuned)	$n_estimators \in \{100, 200, 400, 800\}$; $max_depth \in \{3, 5, 6\}$; $learning_rate \in \{0.03, 0.05, 0.1, 0.3\}$	$learning_rate = 0.05$; $max_depth = 3$; $n_estimators = 800$	11.21