Energy and AI 26 (2026) 100942

Contents lists available at ScienceDirect

Energy and AI

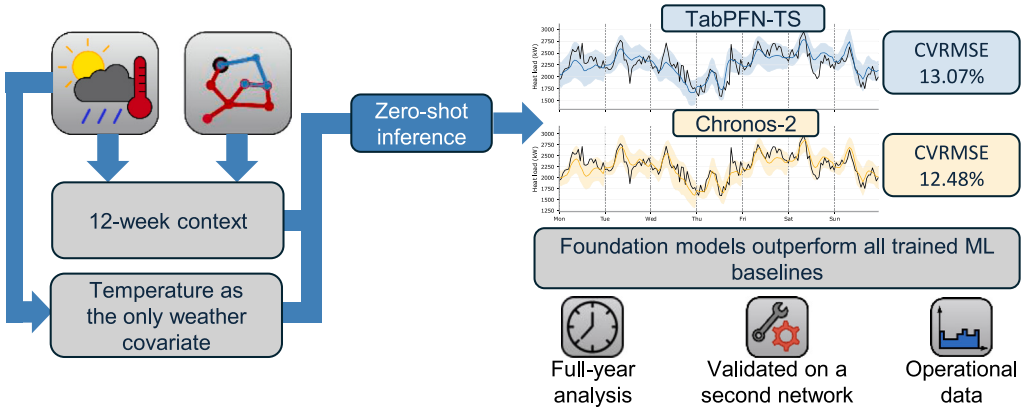
journal homepage: www.elsevier.com/locate/egyai

Systematic evaluation of TabPFN-TS and Chronos-2 for zero-shot heat load forecasting in district heating networks

Ben Spoek [a,*], Karim K. Ben Hicham [b], Kai Derzsi [a], Philipp Althaus [a], Alexander Mitsos [b], Dirk Müller [a]

[a] Chair of Energy Efficient Buildings and Indoor Climate, RWTH Aachen University, Mathieustr. 10, Aachen, 52074, Germany
[b] Process Systems Engineering, RWTH Aachen University, Forckenbeckstr. 51, Aachen, 52074, Germany

GRAPHICAL ABSTRACT



HIGHLIGHTS

- TabPFN-TS and Chronos-2 are systematically evaluated for heat load forecasting.
- Moderate context windows provide robust hourly forecasts for both models.
- Chronos-2 achieves the lowest deterministic and probabilistic forecast errors.
- TabPFN-TS remains competitive despite pretraining only on synthetic tabular data.
- The selected configuration transfers to an independent heating network.

ARTICLE INFO

Dataset link: https://github.com/benspoek/tsfm-heatload, https://doi.org/10.5281/zenodo.23101871

Keywords:
District heating

ABSTRACT

District heating energy hubs require reliable heat load forecasts for efficient operational scheduling. Forecasting models trained on historical data may require retraining as networks evolve. Zero-shot time-series foundation models and in-context forecasting therefore offer a promising alternative: they can adapt at inference time from recent observations rather than by repeated retraining. This study systematically evaluates TabPFN-TS and Chronos-2 for probabilistic heat load forecasting in two German district heating networks and

* Corresponding author.
E-mail addresses: ben.spoek@eonerc.rwth-aachen.de (B. Spoek), karim.benhicham@avt.rwth-aachen.de (K.K. Ben Hicham), kai.derzsi@eonerc.rwth-aachen.de (K. Derzsi), philipp.althaus@eonerc.rwth-aachen.de (P. Althaus), amitsos@alum.mit.edu (A. Mitsos), dmueller@eonerc.rwth-aachen.de (D. Müller).

https://doi.org/10.1016/j.egyai.2026.100942
Received 3 August 2026; Received in revised form 3 October 2026; Accepted 12 October 2026
Available online 24 October 2026
2666-5468/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
Open access funding enabled and organized by Projekt DEAL.

Heat load forecasting
In-context learning
Prior-data fitted networks
Time-series foundation models
Zero-shot forecasting
Probabilistic forecasting

compares them with trained baselines. Of particular interest is whether TabPFN-TS, whose underlying model is pretrained entirely on synthetic tabular rather than time-series data, can capture complex district heating dynamics. We analyze covariate choice, context length, temporal resolution, and forecast horizon on selected operating weeks, evaluate the selected configuration over the full year, and assess cross-network transfer. The principal benchmark assumes perfect weather forecasts; a separate sensitivity analysis uses retrospective weather predictions. Hourly 24-hour forecasting with a 12-week rolling context and ambient temperature provides a parsimonious configuration; longer context windows do not improve accuracy. Both TSFMs outperform all trained baselines in deterministic accuracy in the full-year benchmarks. Chronos-2 achieves the best deterministic scores, with TabPFN-TS remaining close: their CVRMSE values on the main data set are 12.48% and 13.07%, respectively. Chronos-2 also achieves lower continuous ranked probability scores in both networks, with TabPFN-TS remaining close. Finally, a TSFM-based Multi-Resolution Residual-Correction Forecaster combines an hourly base forecast with short-term high-resolution corrections. Relative to direct high-resolution forecasting, it generally reduces errors in total heat demand over 12-hour periods and recorded prediction times.

1. Introduction

To meet the climate targets of the European Union [1], the heat supply of buildings must be decarbonized. District heating networks (DHN) provide one promising pathway towards this goal [2] because they enable the integration of waste heat and renewable heat sources [3]. Realizing this potential requires coordinating heat sources and storage under fluctuating availability, making accurate heat-demand forecasts particularly important for cost- and emission-optimal operation [4]. Heat demand in DHNs is shaped by multiple superimposed drivers [5], including weather-dependent space-heating demand [6], comparatively stochastic domestic hot water consumption [7], and additional operational, daily, and seasonal variations [8].

DHNs evolve as buildings are connected or disconnected, renovated, equipped with additional technologies, or subject to changes in occupancy and usage. These changes may not be fully known to the operator and can alter the underlying heat-demand dynamics [9]. Data-driven forecasting methods can leverage operational observations, but conventional supervised machine-learning models typically require explicit training and retraining in response to such changes [10]. We hypothesize that in-context learning (ICL), which conditions predictions on supplied observations without updating model weights, can support adaptation to concept drift as a sliding context window continually incorporates recent observations. Whether such models can rapidly distinguish relevant post-change observations from older context remains an open question.

Recent advances in foundation models, models pretrained for reuse across different downstream tasks, are highly promising in this regard. TabPFN [11] and TabICL [12] achieve strong performance on tabular prediction benchmarks through ICL without task-specific training. The time-series extension TabPFN-TS [13] applies time-series featurization to TabPFN and is competitive with leading forecasting models [14]. Other foundation models, such as Chronos-2 [15], are pretrained directly for time-series forecasting. Both approaches enable zero-shot forecasting from historical observations without task-specific weight training. However, it remains unclear whether TabPFN-TS's synthetic tabular prior adequately captures the complex heat-demand dynamics of DHNs, or whether time-series-specific pretraining provides an advantage. We therefore systematically evaluate these two contrasting time-series foundation model (TSFM) approaches to probabilistic heat load forecasting. We compare their accuracy, predictive uncertainty and operational requirements, using classical statistical, machine-learning and deep-learning models as trained benchmarks.

The contributions of this work are as follows:

- We provide a systematic evaluation of two TSFMs, TabPFN-TS and Chronos-2, for probabilistic heat load forecasting in DHNs.
- We investigate how covariate choice and context length affect TSFM forecasts to identify configurations that achieve strong performance with limited input-data requirements.
- We compare hourly and 15-minute resolutions to assess whether TSFMs can capture higher-frequency heat load dynamics that are relevant for district heating operation.
- We evaluate forecast horizons from 4 h to one week to characterize horizon-dependent TSFM performance and its interaction with suitable context length.
- We compare both foundation models with trained forecasting baselines and evaluate cross-network transfer.
- We derive and evaluate a Multi-Resolution Residual-Correction (MRRC) Forecaster based on the diagnostic findings.

2. Related work

This section reviews established forecasting approaches for DHNs, the foundation-model concepts underlying TabPFN-TS and Chronos-2, recent energy-forecasting benchmarks, and residual-correction architectures relevant to the present study.

2.1. Heat load forecasting in district heating networks

Heat load forecasting methods range from physically informed models to statistical, machine-learning, and deep-learning approaches [16]. Physically motivated and gray-box methods incorporate knowledge of building and district-heating dynamics [17,18], while engineered load-shape indicators provide a more empirical representation of operational demand patterns [19]. Classical data-driven approaches formulate forecasting as supervised regression, including linear regression and autoregressive models with exogenous inputs [20,21].

Machine-learning methods extend these formulations through nonlinear relationships. Common examples include support vector regression [22], decision-tree models [23], and gradient-boosting methods such as XGBoost [24]. Neural approaches include feed-forward networks [25], nonlinear autoregressive models with exogenous inputs [26], convolutional-recurrent architectures [27], recurrent neural networks [28], attention-based long short-term memory (LSTM) networks [29], and Transformer-based models such as Informer [30]. Rather than replacing earlier methods, these model families provide increasingly flexible alternatives for representing heat-demand dynamics.

Most DHN load-forecasting models are fitted to a specific network and may require retraining as system dynamics change [31]. Adapting to concept drift remains challenging [32], as changes must be detected, suitable adaptation windows selected, and recurring concepts accounted for [33]. Such recurrence is particularly relevant in district-heating systems, where seasonal shifts in the relative dominance of space-heating and domestic-hot-water demand create recurring regimes [5]. These limitations motivate forecasting methods that can adapt through context rather than repeated parameter fitting. The following subsection introduces in-context learning for tabular prediction and its extension to time-series forecasting.

2.2. Tabular prediction with TabPFN and its extension to time-series data

TabPFN [11] is a foundation model for tabular prediction based on in-context learning. It is pretrained on a large collection of fully synthetic supervised-learning tasks and, at inference time, predicts query labels from labeled context examples. Unlike conventionally trained neural networks and gradient-boosted models, TabPFN requires no task-specific parameter fitting [11,12,34].

More formally, the model receives both labeled context examples $D_c = (X_c, y_c)$ and unlabeled query examples x_* and produces predictions y_* in a single forward pass. In this sense, the training data of the target task become context data for the pretrained model. The Bayesian interpretation is central to Prior-Data Fitted Networks (PFNs) [35]. Let ϕ denote a latent data-generating task or process. The posterior predictive distribution can be written as

$$p(y_* \mid x_*, D_c) = \int p(y_* \mid x_*, \phi)\, p(\phi \mid D_c)\, d\phi. \tag{1}$$

Using Bayes' rule,

$$p(\phi \mid D_c) = \frac{p(D_c \mid \phi) p(\phi)}{p(D_c)}, \tag{2}$$

so that

$$\begin{aligned} p(y_* \mid x_*, D_c) &= \frac{1}{p(D_c)} \int p(y_* \mid x_*, \phi)\, p(D_c \mid \phi)\, p(\phi)\, d\phi \\ &\propto \int p(y_* \mid x_*, \phi)\, p(D_c \mid \phi)\, p(\phi)\, d\phi. \end{aligned} \tag{3}$$

Computing this integral explicitly would generally be intractable. A PFN instead amortizes the computation: it is trained to map (D_c, x_*) directly to an approximation of the posterior predictive distribution in a forward pass. This approximation is Bayesian with respect to the synthetic pretraining prior over data-generating tasks, rather than an explicit posterior over the real-world target system. The practical consequence is that TabPFN removes the standard per-data-set gradient-based training loop in its default use [11,34]. No classical hyperparameter search is required for basic application, although inference-time ensembling, preprocessing variants, fine-tuning, or hyperparameter tuning can still be used to improve performance.

TabPFN-TS, as proposed by Hoo et al. [13], can be directly used for time series forecasting by reformulating a time series as a tabular regression problem. Each time step becomes a row in a table. The feature set includes temporal features such as a running time index, cyclic calendar features, a non-cyclic year feature, automatically extracted seasonal features, and optionally time-varying covariates whose future values are known at prediction time. The historical observations form the labeled context D_c, while the future time points form query rows with known features x_* and unknown targets y_*. TabPFN then outputs a probabilistic forecast for the full prediction horizon in a non-autoregressive forward pass. This is notable because the same study reports strong GIFT-Eval [36] performance for TabPFN-TS, although the underlying TabPFN model was pretrained only on synthetic non-temporal tabular data. The synthetic pretraining paradigm is important for the present application: since TabPFN is pretrained on synthetic supervised prediction tasks rather than on collections of real district-heating time series, the risk of direct or indirect pretraining-test overlap with historical district-heating backtesting data is eliminated. In addition, the probabilistic forecast can be interpreted as an approximation to the posterior predictive distribution under the learned synthetic prior. This makes TabPFN-TS particularly interesting for heat load forecasting, where reliable uncertainty estimates are operationally relevant and where point forecasts alone do not fully describe the stochastic demand process. Hoo et al. [13] report that TabPFN-TS can distinguish signal from noise, adapt its probabilistic predictions accordingly, incorporate covariates, and handle seasonal patterns, while trend recognition, especially for linear and exponential trends, remains a reported weakness.

2.3. Time-series forecasting with Chronos-2

Chronos-2, proposed by Ansari et al. [15], is a time-series foundation model capable of probabilistic predictions. The model uses normalization of the covariates and targets. Additionally, it applies patching where each patch represents several consecutive observations. Its transformer architecture combines temporal attention and group attention, which allows relationships across time and across targets and covariates to be captured. The model supports both past-only covariates and covariates with known future values. Chronos-2 is pretrained on time-series forecasting tasks using real and synthetic data. It differs from TabPFN-TS in its architecture and pretraining assumptions.

Since Chronos-2 was partially pretrained on real-world time-series data, potential train-test contamination between the pretraining corpus and the downstream evaluation data must be considered. Based on the disclosed Chronos-2 pretraining corpus, leakage of time-series training data into the 2024 heat load forecasting data sets for the DHNs in Munich or Flensburg appears unlikely. The real pretraining data sets listed in Appendix A of the Chronos-2 documentation are univariate time series from broad domains such as electricity, solar and wind generation, weather, transport, web, cloud operations, and simulated U.S. building energy, but no district-heating heat load data from Munich or Flensburg are reported [15]. Moreover, Chronos-2's multivariate and covariate-informed training tasks are stated to rely entirely on synthetic data.

2.4. Foundation-model benchmarks in energy forecasting

Recent work indicates that time-series foundation models can be useful in energy-forecasting applications. Meyer et al. [37] show that foundation models can provide performance comparable to approaches trained from scratch for household electricity-demand forecasts. Koch et al. [38] study thermal-energy forecasting and compare multisource transfer-learning strategies with time-series foundation models, including TimesFM, Toto, and Chronos-2. Chronos-2 achieves the strongest performance among the considered time-series foundation models. They further find that multisource transfer learning requires at least 16 building data sets before it outperforms time-series foundation models.

Comparative benchmarks including both TabPFN-TS and Chronos-2 in energy forecasting provide the closest context for this study: Hertel et al. [39] compare TabPFN-TS and Chronos-2 for load forecasting in electricity grids at transmission-system-operator level and use Shapley-based explanations to increase interpretability. In their study, Chronos-2 outperforms TabPFN-TS in both the univariate and covariate settings. Obermeier et al. [14] benchmark several time-series foundation models on energy-forecasting tasks, including Chronos-2 and TabPFN-TS in a covariate setting. The benchmark includes a data set from Flensburg's DHN as part of a heat load prediction subset. For this heat load subset and overall, covariate models outperform univariate models. Chronos-2 again outperforms TabPFN-TS overall and on the heat load subset, although TabPFN-TS obtains some wins over Chronos-2 in the covariate setting on other subcategories and several wins overall. These findings motivate evaluating both approaches under the same district-heating forecasting conditions.

TabPFN has also been successfully applied to energy-related regression tasks. Liu et al. [40] evaluate TabPFN for regression tasks in solar-energy meteorology and report that TabPFN outperforms multiple regression, artificial neural networks, Random Forest, and Gradient Boosting. However, they do not directly address TabPFN-TS as a time-series forecasting method.

Alkhulaifi et al. [41] show the importance of feature engineering for time-series forecasting with TabPFN. TabPFN-TS largely automates this step by embedding generic temporal and seasonal feature construction into the forecasting pipeline.

Ramachandran et al. [42] compare several time-series foundation models and classical machine-learning baselines with a method that

combines a continuous wavelet transform with LSTMs for heat-demand forecasting. Their proposed method outperforms the benchmark approaches. However, it is specifically engineered for the considered forecasting problem and is therefore less directly transferable than general-purpose time-series foundation model approaches.

Overall, the literature motivates a systematic comparison of TabPFN-TS and Chronos-2 for zero-shot probabilistic heat load forecasting. Open questions concern their application-specific configurations, deterministic and probabilistic performance, computational requirements, comparison with trained baselines, and transfer across DHNs with different characteristics. Their different pretraining approaches additionally allow us to examine how competitive a synthetic tabular prior remains against a model pretrained directly for time-series forecasting.

2.5. Residual-correction architectures for time-series foundation models

Residual correction of TSFM forecasts and the use of TSFMs as residual forecasters are emerging directions in time-series forecasting. Chitsaz and Haratizadeh [43] investigate TSFM-based residual correction for several regression and machine-learning models in stock-price forecasting. They compare this approach with a relational architecture in which a TSFM separately forecasts the stock price and auxiliary financial variables, and the concatenated forecasts are passed to a multivariate predictor. Residual correction reduces mean squared error (MSE) relative to the standalone multivariate baseline for all five model classes and outperforms the relational architecture in four of five comparisons. For price-direction classification, the relational architecture achieves a higher macro-averaged F1-score in three comparisons, residual correction in one, with one tie.

In contrast, Dai et al. [44] use a frozen TSFM to generate the base forecast. Their Online Residual Contextual Adaptation (ORCA) framework estimates an additive residual correction using a linear adapter that decomposes the concatenated historical input and base forecast into trend and seasonal components. The base and corrected forecasts are then combined using a Boltzmann softmax weighting mechanism based on recent prediction errors. Across eight data sets covering energy, finance, meteorology, and traffic, the authors report an average MSE reduction on every data set, with results aggregated over five TSFMs and three forecasting horizons.

Gao et al. [45] propose SMR^2, a TSFM wrapper that constructs forecasts using the same frozen TSFM with context data at different resolutions. A separate ridge-regression model predicts a residual correction for each resolution-specific forecast. The corrected forecasts are then combined using weights determined from validation performance. The authors report consistent improvements across several TSFMs for regional drought forecasting.

Related residual-correction schemes have also been investigated for Transformer-based forecasting models. Biswas et al. [46] train a base model on the original time series and a second model to forecast its residuals. The predicted residuals are scaled before being used to correct the base forecast. Both models use the same look-back length and prediction horizon. Their main iTransformer-based experiments show improved accuracy over the uncorrected model across all eight data sets from energy, transportation, meteorology, and finance.

These studies motivate adapting residual-correction architectures to the distinct heat-load dynamics and operational requirements of DHNs. In particular, they leave open how to combine longer-horizon planning forecasts with frequently updated, high-resolution residual corrections. Within the reviewed literature, using a TSFM to forecast the residuals of a TSFM prediction remains underexplored. The present study investigates this combination as an application-specific adaptation of residual correction.

3. Methodology

This section describes the forecasting models and rolling prediction setup, followed by the procedures for evaluating deterministic accuracy, probabilistic forecast quality and computational cost. It then defines the anomaly-screening criteria and the metric used to quantify interannual distributional change, before introducing the MRRC forecasting architecture.

3.1. Forecasting models and prediction setup

The forecasting workflow uses rolling historical context windows comprising past heat load and observed weather variables. Known future covariates comprise weather variables over the prediction horizon and calendar features. Unless stated otherwise, ambient temperature is the only weather variable. The prediction target is the future heat load. Both foundation models are evaluated without task-specific weight training and produce quantile forecasts. The median is used as the deterministic prediction, while the stored quantiles support probabilistic evaluation.

TabPFN-TS and Chronos-2 are evaluated on the same operational forecasting task to compare their different pretraining paradigms and inference architectures. Both models receive the same future covariate values over the prediction horizon and, where permitted by model constraints, the same rolling context windows. The maximum number of context points that can be passed to Chronos-2 is 8192 time steps. If a selected context window exceeds this limit, only the most recent 8192 time steps are passed as context to Chronos-2, and results are reported at the effective context length.

AutoGluon TimeSeries [10,47] is used as a benchmark suite for the selected hourly day-ahead forecasting task. The benchmark contains several model families. Persistence baselines are represented by Naive and Seasonal naive. Naive repeats the last observed heat load value over the prediction horizon. Seasonal naive uses a seasonal period of 24 h and therefore predicts each future hour from the corresponding observed value on the previous day. Classical statistical baselines include exponential smoothing (ETS), automatic ETS (AutoETS), and automatic autoregressive integrated moving average (AutoARIMA). Tabular machine-learning baselines include direct LightGBM (LGBM) and XGBoost models as well as a recursive LGBM model. Neural baselines include DeepAR, TemporalFusionTransformer (TFT), Feed-forward, and DLinear. AutoGluon's weighted ensemble is reported as AutoGluon Ensemble.

The predictors are configured with ambient temperature as a known covariate, so future ambient-temperature values are supplied over each prediction horizon. Models with native known-covariate support can use this information directly, including the direct and recursive tabular models, DeepAR, and TemporalFusionTransformer. Other models in the benchmark, such as the persistence, statistical, Feed-forward, and DLinear baselines, do not use the future covariate in the same way.

These AutoGluon benchmarks fitted once are not constrained to the same rolling context used for the selected foundation-model setup. Instead, each prediction call receives the observed target history from the start of 2023 up to the respective forecast start. Individual AutoGluon model classes then use this history according to their own defaults: statistical models may truncate long series internally, neural models use model-specific context lengths, and tabular models construct lag, time, and covariate features.

In the main experiments, future ambient temperature is provided as the realized temperature over the prediction horizon. This setting can be interpreted as a perfect-weather-forecast assumption: the heat load forecasting problem is isolated from errors in numerical weather prediction, and model behavior is evaluated conditional on known future weather.

All experiments were run on the RWTH Aachen University HPC cluster. Each job used one NVIDIA H100 GPU with 95,830 MiB memory,

8 CPU cores, and 64 GB RAM. Nodes are equipped with Intel Xeon Platinum 8468 CPUs, with two sockets and 48 physical cores per socket. The software environment used Python 3.12.3 and CUDA 12.6.3, with tabpfn-time-series 1.1.0, tabpfn 8.0.3, chronos-forecasting 2.2.2, and autogluon.timeseries 1.5.0.

3.2. Evaluation metrics and approach for model comparison

For experiments with overlapping forecast windows, metrics are computed after collapsing predictions to a single operational trajectory: for each target timestamp, the forecast with the latest forecast start at or before the target timestamp is retained. Thus each realization contributes once to the error metrics.

Point-forecast quality is assessed using the coefficient of variation of the root mean squared error (CVRMSE), the coefficient of determination (R^2), and mean absolute error (MAE). For n evaluation samples, let y_t denote the target value, $\hat{y}_t$ the corresponding point forecast, and $\bar{y} = \frac{1}{n}\sum_{t=1}^{n} y_t$ the mean target value of the evaluation set. In this study, y_t corresponds to the measured heat load.

The three metrics are defined as

$$\mathrm{CVRMSE} = 100 \cdot \frac{\sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(\hat{y}_t - y_t\right)^2}}{\bar{y}}, \tag{4}$$

$$R^2 = 1 - \frac{\sum_{t=1}^{n}\left(y_t - \hat{y}_t\right)^2}{\sum_{t=1}^{n}\left(y_t - \bar{y}\right)^2}, \tag{5}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{t=1}^{n}\left|\hat{y}_t - y_t\right|. \tag{6}$$

CVRMSE normalizes the root mean squared error by the mean target value, making the error comparable across operating regimes and between the two DHNs.

For longer-horizon planning, we evaluate the total heat demand over rolling 12-hour windows starting at hourly intervals. Let E_i and $\hat{E}_i$ denote the observed and predicted integrated heat demand, respectively, using the forecast available at anchor i, and let $\bar{E}$ be the mean observed integrated heat demand across the N evaluated windows. Energy CVRMSE (E-CVRMSE) and energy bias (E-bias) are defined as

$$\text{E-CVRMSE} = \frac{100}{\bar{E}}\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{E}_i - E_i)^2}, \tag{7}$$

$$\text{E-bias} = \frac{100}{N\bar{E}}\sum_{i=1}^{N}(\hat{E}_i - E_i). \tag{8}$$

Both metrics are expressed as percentages; negative E-bias indicates underprediction and positive E-bias indicates overprediction.

Probabilistic forecast quality is assessed from the stored quantile forecasts using the continuous ranked probability score (CRPS). Let F_t denote the predictive cumulative distribution function for time step t, where y_t is the observed heat load, and let $\mathbf{1}\{\cdot\}$ denote the indicator function, which equals one if the condition in braces is true and zero otherwise. The CRPS is defined as

$$\mathrm{CRPS} = \frac{1}{n}\sum_{t=1}^{n}\int_{-\infty}^{\infty}\left(F_t(z) - \mathbf{1}\{z \geq y_t\}\right)^2 dz. \tag{9}$$

Since the probabilistic forecasts are stored as discrete quantiles, CRPS is approximated by twice the trapezoidal integral of the pinball loss over the stored quantile levels from 0.01 to 0.99, averaged over the evaluated time steps. Lower CRPS values indicate better overall probabilistic forecast quality.

Central prediction intervals are additionally described by their absolute and relative widths. Relative interval width is the mean q90–q10 width divided by the mean observed heat load, expressed as a percentage.

Calibration is assessed using the probability integral transform (PIT), defined as

$$\mathrm{PIT}_t = F_t(y_t). \tag{10}$$

PIT values are approximated from the stored quantile forecasts. For calibrated continuous predictive distributions, PIT values are uniformly distributed. The histograms pool forecast horizons and dates.

For the full-year model comparison, daily forecast blocks are also used to compare the consistency of the algorithms over time. Each issued 24-hour forecast is treated as one block, and CVRMSE, R^2, and MAE are computed over the 24 hourly prediction points for each model and forecast start. Models are ranked within each block, using ascending ranks for CVRMSE and MAE and descending ranks for R^2; ties are assigned average ranks. The reported mean ranks therefore describe how consistently a model performs well across individual forecast days, not only its aggregate full-year error. Pairwise win-rate matrices and critical difference (CD) diagrams use daily CVRMSE comparisons. The CD diagrams use a nominal significance level of 0.05.

For full-year evaluations, uncertainty in aggregate metrics is quantified using 95% percentile confidence intervals from 10,000 paired circular block-bootstrap samples. Blocks comprise seven consecutive forecast days, retaining complete issued forecasts and using identical resampled blocks across models.

Computational cost is separated into model loading, estimated first-call overhead and steady-state inference. Steady-state inference time is calculated as the mean prediction-call duration from the second to the final call, excluding surrounding input preparation and prediction post-processing. Estimated first-call overhead is the first prediction-call duration minus the mean duration of subsequent calls and is reported separately from model loading.

To quantify computational cost, the mean steady-state prediction-call time and the operational real-time factor (RTF) are reported. The operational real-time factor is defined as the ratio between the total recorded prediction-call time and the total elapsed operating time over the evaluation period T_{eval},

$$\mathrm{RTF}_{\mathrm{op}} = \frac{\sum_i t_{\mathrm{comp},i}}{T_{\mathrm{eval}}}, \tag{11}$$

where $t_{\mathrm{comp},i}$ denotes the computation time of forecast i. This metric captures the computational burden of a forecasting scheme in an operational setting, accounting for both the cost of individual forecasts and their update frequency. Values below one indicate that the forecast is computed faster than real time.

3.3. Anomaly screening and exclusion

Potentially anomalous heat-load observations are identified using a model-independent screening proxy: nonpositive heat load or at least three consecutive exactly identical values at the data set's native resolution. These patterns can indicate maintenance, faults or signal outages, but do not establish their cause. As independently verified anomaly labels are unavailable, flagged observations are treated as suspected anomalies.

Weeks containing flagged observations are excluded from configuration selection to prevent brief anomalous periods from driving the configuration choices. The principal full-year benchmark retains the recorded heat-load series without anomaly-specific modification. This evaluates the selected configuration under continuous operational conditions and follows a data-driven setting in which no manual preprocessing of the target signal is applied for the final assessment.

3.4. Quantification of distributional change

To characterize interannual changes in the observed heat-load distribution, we compute the normalized Hellinger distance between consecutive years.

$$H(p,q) = \sqrt{\frac{1}{2}\sum_b\left(\sqrt{p_b} - \sqrt{q_b}\right)^2}, \tag{12}$$

where p_b and q_b are consecutive years' histogram probabilities for each of the 470 bins indexed by b. Histograms use all available raw 15-minute heat loads without anomaly exclusion and 470 shared equal-width bins determined on pooled 2020–2024 observations.

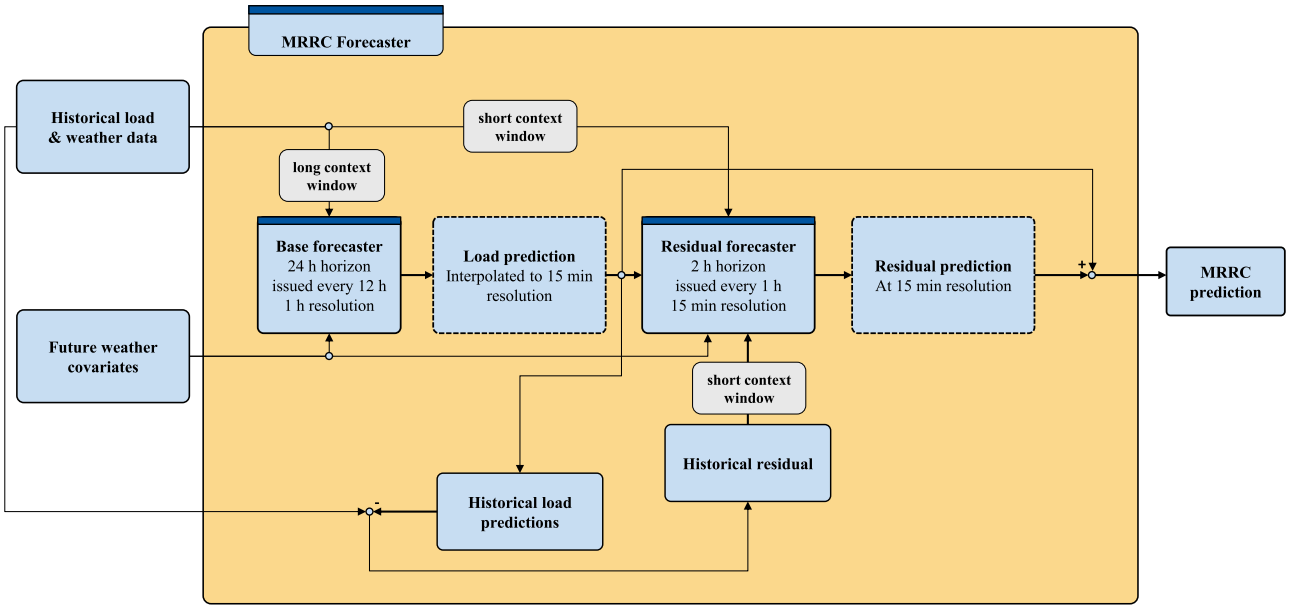


Fig. 1. Schematic structure of the Multi-Resolution Residual-Correction Forecaster.

3.5. Multi-resolution residual-correction forecaster

Building on the residual-correction approaches reviewed in Section 2.5 and our configuration findings, we develop the MRRC Forecaster as a follow-up extension of the evaluated forecasting approaches. It combines a long-horizon base forecast with a short-horizon residual correction at higher temporal resolution. The architecture is illustrated in Fig. 1.

The first component, referred to as the Base Forecaster, provides the general daily load trajectory potentially needed for scheduling and storage planning. Its point forecast is linearly interpolated to the higher temporal resolution of 15 min. The second component, referred to as the Residual Forecaster, forecasts the remaining high-frequency error of this baseline. The final point prediction is obtained by adding the predicted residual to the interpolated base forecast. The architecture aims to improve the representation of short-term dynamics while retaining a longer planning horizon. The MRRC Forecaster is evaluated here as a point-forecast architecture.

From a boosting perspective, the setup can be interpreted as a single residual-correction step. An initial model $F(x)$ produces the base prediction, while a second model $h(x)$ predicts the residuals $r = y - F(x)$. The final prediction is then given by $\hat{y} = F(x) + h(x)$. In the MRRC Forecaster, the Base Forecaster corresponds to $F(x)$, and the Residual Forecaster corresponds to $h(x)$.

Historical residual context is constructed by subtracting the interpolated base forecasts available at each historical timestamp from the observed high-resolution heat loads. The Residual Forecaster uses these residuals as its target context, with ambient temperature and the interpolated base prediction as covariates. Future inputs contain the corresponding temperature and base predictions.

MRRC's configuration, residual-context selection and evaluation are described in Section 4.6.

4. Case studies and experimental setup

This section introduces the Munich and Flensburg DHNs and their heat-load and weather data. Munich is a small, expanding mixed-use network, whereas Flensburg is substantially larger and comparatively stable. Munich's operational data are proprietary; the openly available Flensburg data support reproducible external validation. The section then presents the experimental design, supplementary analyses and MRRC configuration and evaluation.

4.1. District heating networks

The data used for model development and primary evaluation originate from a DHN serving a mixed-use urban district in Munich, Germany. At the end of the available measurement period, the network served 11 individual buildings, including large consumers such as hotels and office buildings. The building stock also includes retail, restaurants, cultural and educational facilities, and industrial uses.

Heat is supplied by a centralized energy hub, where the network heat load is measured. The network's supply line is approximately 400 m long. The design supply temperature is 80 °C, while the measured supply temperature averaged 84.44 °C with a standard deviation of 3.95 °C in 2024. The generation portfolio includes two combined heat and power units with a thermal capacity of 1111 kW each, a 1870 kW gas boiler, a 1284 kW high-temperature heat pump, and a 250 kW power-to-heat unit, supported by two 12 m^3 thermal buffer tanks. The combined nominal heat-generation capacity is 5.626 MW. The system is operated using a rule-based control strategy that prioritizes combined heat and power generation, while the heat pump, gas boiler, and power-to-heat unit provide supplementary heat depending on operating conditions.

The available Munich heat-load record spans 1 January 2020 to 31 January 2025 at a native temporal resolution of 15 min. Measured weather covariates are obtained from German Weather Service (DWD) stations and aligned with the 15-minute heat load data. Temperature, relative humidity, wind speed, and precipitation are taken from station 03379 München-Stadt [48]. As this station has no solar irradiation sensor, irradiation data are obtained from station 05404 Weihenstephan-Dürnast. In 2024, these data were 99.47% complete; remaining gaps were filled using surrounding DWD stations. In the primary benchmark, these weather variables are treated as known future inputs.

The measured heat-load signal contains occasional periods flagged by the screening procedure described in Section 3.3. After excluding flagged days, the maximum 15-minute load over the entire available data set was 4.929 MW, and the maximum hourly mean was 3.496 MW.

The Munich network combines consumers with diverse heat-demand characteristics, while its small number of connected buildings limits the smoothing of individual demand fluctuations through aggregation. Together, these features make it a challenging case for district heat-load forecasting. Ongoing expansion adds a further source

Table 1
Distributional-change indicators for the Munich network based on 15-minute heat-load observations. Hellinger distances compare each year's heat-load distribution with the preceding year; KSWIN drift-detection counts refer to changes detected within the indicated year.

Year	Hellinger distance H	KSWIN drift detections
2021	0.3921	65
2022	0.1601	26
2023	0.1827	27
2024	0.1583	31

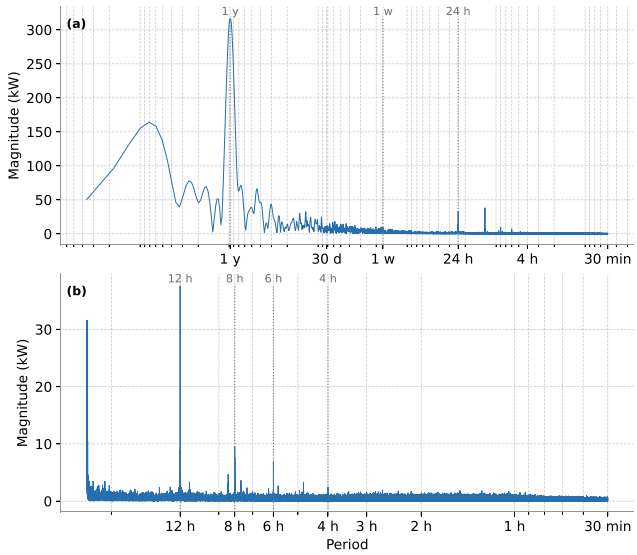


Fig. 2. Fourier magnitude spectra of the full available 15-minute heat load time series: (a) full period range and (b) periods up to one day. The x-axis is shown as period rather than frequency; marked periods indicate seasonal, weekly, daily, and sub-daily components.

of change and is accompanied by a substantial increase in annual heat demand, as shown in Supplementary Figure S1. Table 1 summarizes interannual differences and changes detected within individual years: the Hellinger distance compares each year's heat-load distribution with the preceding year, while the annual number of Kolmogorov–Smirnov Windowing (KSWIN) drift detections indicates how frequently distributional changes were signaled within that year.

To characterize the dominant time scales in the measured heat demand, a Fourier spectrum of the full available 15-minute heat load series is shown in Fig. 2. The very-long-period components above one year should not be interpreted as stationary cycles; they reflect the expansion of the Munich DHN and the associated increase in annual heat demand. Further details of this annual development are provided in Supplementary Figure S1. The spectrum also shows a dominant seasonal component close to one year, mainly attributable to space-heating demand, and a clear diurnal component around 24 h. Additional peaks at harmonics of the daily cycle, including 12 h, 8 h, 6 h, and 4 h, indicate that the diurnal load pattern contains substantial harmonic content rather than being well described by a single sinusoidal component. The pronounced 12 h component is consistent with two dominant demand periods per day, for example, morning and evening domestic-hot-water peaks. In contrast, no pronounced weekly peak is visible, suggesting that the aggregated DHN signal contains no strong strictly periodic seven-day component.

The data used in the transfer validation originate from the DHN of Flensburg, Germany. The network has developed historically from 1969 into a city-scale district heating system with more than 700 km of district heating pipelines and a connection rate of over 90% among private households. This is the same network analyzed by Obermeier et al. [14], discussed in Section 2.4.

Heat is supplied from a central heating plant connected to the DHN. As of 2024, the heat supply infrastructure comprises six generation units and two large thermal storage tanks [49]. The measured heat load signal reflects the aggregated demand supplied by this city-scale network. For external validation, we use the 2024 hourly heat-load observations from the publicly available 2020–2024 Flensburg district-heating data set [50]. Between 2023, the year supplying the training data for the trained baselines, and 2024, the deployment and evaluation year, Flensburg exhibited a smaller change in its annual heat-load distribution than Munich. Hellinger distances calculated between the annual distributions of hourly average heat load were 0.0780 for Flensburg and 0.1572 for Munich. Weather data for Flensburg are provided by DWD, using hourly air-temperature observations from station 01666 Glücksburg-Meierwik.

4.2. Configuration experiments and benchmark evaluation

The experimental design comprises configuration diagnostics on selected representative weeks, full-year benchmarks and external validation, and a multi-year evaluation of MRRC. The configuration diagnostics examine context length, temporal resolution, forecast horizon and weather-covariate choice for both foundation models, with additional model-specific diagnostics where appropriate. Both models are then evaluated over 2024 against trained AutoGluon baselines and on the Flensburg data set to assess cross-network transfer. Additional analyses assess sensitivity to anomaly exclusion, baseline retraining following detected concept drift and weather-forecast sensitivity, as described in Sections 4.3–4.5. The MRRC configuration and evaluation setup are presented in Section 4.6.

Although the native temporal resolution of the heat load data is 15 min, large thermal buffer tanks partially smooth the network demand. Consequently, this high temporal resolution is not strictly required for heat-generation scheduling. Selected experiments therefore also use hourly aggregated data to assess forecasting performance at a coarser, operationally relevant resolution.

Candidate weeks are restricted to complete Monday–Sunday weeks without observations flagged by the screening procedure in Section 3.3. From the remaining weeks, three operating regimes are selected based on ambient temperature: the coldest week represents winter operation, the hottest week represents summer operation, and the week with the largest within-week temperature variability represents transitional operation. The selected winter, transitional, and summer weeks in 2024 are January 15–21, April 22–28, and August 12–18, respectively. Their detailed ambient-temperature statistics are provided in Supplementary Table S1.

These representative weeks are used to study seasonal effects and to compare forecast horizons, context lengths, and temporal resolutions under characteristic operating conditions.

For the principal 2024 benchmarks, the AutoGluon predictors are trained on 2023 target values and evaluated on rolling 24-hour forecasts in 2024. For a transfer validation on a new data set, the AutoGluon models are trained from scratch on the corresponding 2023 data before being evaluated on 2024 data.

For the full-year day-ahead benchmarks, forecasts are issued every 24 elapsed hours, beginning at 00:00 Europe/Berlin on 1 January. This corresponds to 00:00 local time during standard time and 01:00 during daylight-saving time, with each forecast covering 24 elapsed hours. In the selected-week diagnostics, daily forecasts are instead anchored at local midnight at the beginning of each selected week.

4.3. Sensitivity to anomaly exclusion

To assess whether potentially anomalous targets affect the full-year benchmark conclusions, we repeat the evaluation after anomaly exclusion. We exclude complete 24-hour forecasts issued on local dates containing observations flagged by the screening procedure in Section 3.3.

4.4. Baseline retraining on detected concept drift

To test whether batch retraining triggered by active drift detection improves baseline performance, we use KSWIN, which achieved the highest detection accuracy in the benchmark study of Derzsi et al. [9]. We replay the Munich drift-detection timestamps supplied by the authors, generated using the KSWIN parameters specified for the medium-severity expansion scenario, referred to as EXP-M in Table B.1 of their study [9]. AutoGluon models initially fitted on 2020 are evaluated over 2021–2024 either as static models without retraining or with batch retraining from scratch after each KSWIN drift detection on the preceding year of observations, ending strictly before the detection time. At each forecast issuance, the latest eligible model is used; fitting is assumed complete before the next scheduled forecast. Each model receives observed history from its training-window start to the forecast issue time.

4.5. Weather-forecast sensitivity

To assess sensitivity to weather-forecast uncertainty, retrospective hourly temperature predictions for Munich were obtained from the Open-Meteo Single Runs API using ECMWF IFS HRES [51]. Each 24-hour heat-load forecast uses one coherent temperature trajectory from the preceding 12:00 UTC initialization. The archive contains IFS Cycle 49R1 hindcasts, which we use to assess sensitivity to coherent weather forecasts.

Forecast temperatures are available from 15 March 2024. Requiring a complete 12-week context and 24-hour horizon yields 208 matched daily forecast starts from 7 June to 31 December 2024. The realized-weather reference uses measured temperatures in both the context and future horizon. The forecast-weather variants use the same heat-load histories, forecast starts and targets, but incorporate measured and forecast temperatures as described below.

Measured and forecast temperatures are supplied as separate historical covariates; each historical forecast value comes from the trajectory selected for its corresponding daily forecast start. Chronos-2 receives only forecast temperature as a future covariate. Because the TabPFN-TS implementation requires future values for every covariate, both future columns are filled with the same forecast temperature. These configurations achieved the lowest point-estimate errors among the tested forecast-weather configurations, as shown in Supplementary Table S2.

4.6. MRRC configuration and evaluation

We evaluate the MRRC architecture described in Section 3.5 on the Munich data over 2021–2024 using both TabPFN-TS and Chronos-2 as backbones. The Base Forecaster operates at hourly resolution, and its predictions are linearly interpolated to the 15-minute grid used for residual construction and evaluation. Its forecast horizon is 24 h, with a new forecast issued every 12 h. Thus, at every point in time, a forecast for at least the next 12 h is available. The Residual Forecaster predicts the next two hours at 15-minute resolution and issues a new forecast every hour, ensuring that a high-resolution forecast for at least the next hour is available.

For the Base Forecaster, the context window is chosen based on the 24-hour forecast results described in Section 5.1. The base and direct high-resolution forecasts use a common 12-week context to avoid varying context length alongside resolution and update frequency.

Because residual prediction differs from direct heat-load forecasting, the context-window results from the main configuration experiments are not directly transferable to the Residual Forecaster. Residual contexts of 2, 4, 8, 24, 48, 168 and 336 h are therefore compared separately for each backbone on the held-out December 2020 selection period, preceding the 2021–2024 evaluation. Selection minimizes the combined MRRC forecast's CVRMSE. A seven-day residual context achieved the lowest CVRMSE among the tested context lengths for both TSFM backbones and was selected for evaluation.

We compare the MRRC Forecaster with two baseline setups. The first is the Base Forecaster linearly interpolated to 15-minute resolution without the residual correction. The second baseline is a direct high-frequency, high-resolution predictor: it uses 15-minute resolution, a 13-hour prediction horizon, and hourly re-issuing. This setup ensures that at least 12 h of forecast are available at every time step while matching the short-term resolution of the MRRC Forecaster. In the following, this setup is referred to as the High Frequency & High Resolution (HFHR) setup.

Two sets of forecast-quality metrics are reported to reflect the different operational requirements. The short-term metrics assess the latest available 15-minute operational trajectory. For the hourly updated high-resolution and MRRC setups, this corresponds to the first hour after each forecast update and therefore quantifies how well the most recent forecast tracks the realized high-resolution load profile. In contrast, the long-term metrics assess the forecast information available for planning decisions. This long-term evaluation uses the rolling 12-hour energy metric described in Section 3.2. For the 12-hour energy evaluation, the residual correction is applied during the first two hours; the remaining horizon uses the base forecast.

5. Results

This section presents configuration and seasonal analyses, followed by full-year benchmarks, external validation and probabilistic forecast quality. It then reports sensitivity to anomaly exclusion, KSWIN-triggered baseline retraining and weather-forecast sensitivity, before comparing MRRC with Base and HFHR forecasts across 2021–2024.

5.1. Configuration experiments

First, the required length of historical context data was investigated for the 24-hour forecast task. Context windows of 1 day, 1 week, 4 weeks, 12 weeks, and 1 year were evaluated for both 15-minute and hourly forecast resolutions. At hourly resolution, both models achieved their lowest CVRMSE with a 12-week context. At 15-minute resolution, the best context differed: 12 weeks for TabPFN-TS and 4 weeks for Chronos-2, as shown in Fig. 3. The R^2 and MAE results follow the same qualitative trend. For TabPFN-TS, the 12-week context achieved a CVRMSE of 11.42%, an R^2 of 0.970, and an MAE of 102.6 kW at hourly resolution; at 15-minute resolution, it achieved a CVRMSE of 15.71%, an R^2 of 0.945, and an MAE of 138.7 kW. Extending the recent context to one year did not further improve the TabPFN-TS point forecast at either resolution. In this experiment Chronos-2 slightly outperforms TabPFN-TS on every metric for all resolutions and context window lengths.

The probabilistic score follows the same main pattern as the deterministic metrics. For hourly forecasts, both models achieved their lowest CRPS with a 12-week context. At 15-minute resolution, the optimum was 12 weeks for TabPFN-TS and four weeks for Chronos-2. For TabPFN-TS, the 12-week context achieved a CRPS of 70.7 kW at hourly resolution and 97.3 kW at 15-minute resolution. Thus, the CRPS results support the choice of a moderate recent context window and do not indicate a probabilistic benefit from using the full one-year context either.

The corresponding steady-state prediction-call times are shown in the lower panel of Fig. 3. For the 12-week context, the mean time per forecast was 1.136 s and 1.665 s for TabPFN-TS at hourly and 15-minute resolution, respectively, compared to 0.022 s and 0.028 s for Chronos-2. Table 2 reports model loading and estimated first-call overhead alongside steady-state inference, showing shorter model-loading time for TabPFN-TS but lower first-call overhead and inference time for Chronos-2.

Table 2
Controlled TSFM timing for hourly 24-hour forecasts with a 12-week context. Values are mean ± standard deviation across five fresh processes.

Model	Model loading (s)	Estimated first-call overhead (s)	Steady-state inference (s/forecast)
TabPFN-TS	16.49 ± 3.45	7.47 ± 1.23	0.947 ± 0.021
Chronos-2	29.78 ± 6.14	0.260 ± 0.019	0.01809 ± 0.00018

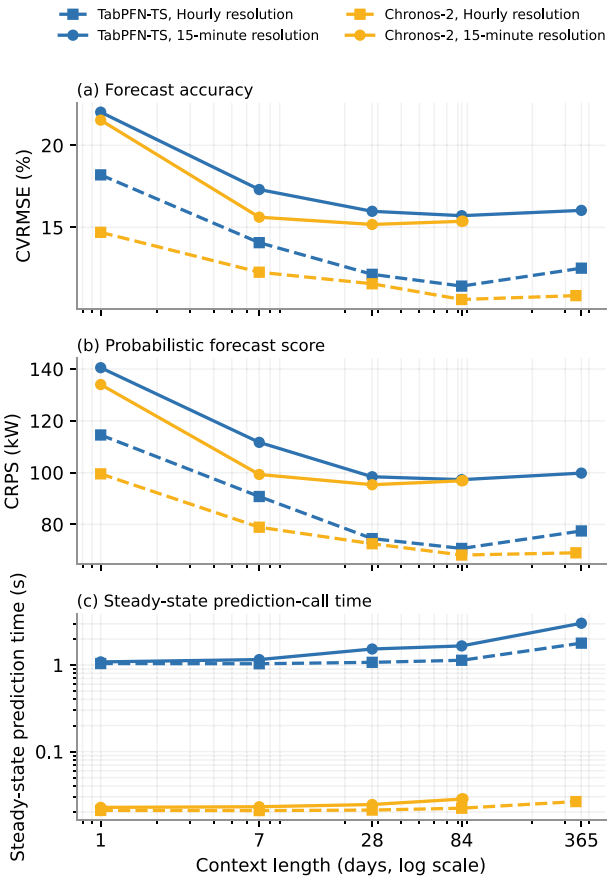


Fig. 3. Overall CVRMSE, CRPS, and mean steady-state prediction-call time for 24-hour heat load forecasts as a function of context length for the selected summer, winter, and transitional operating weeks.

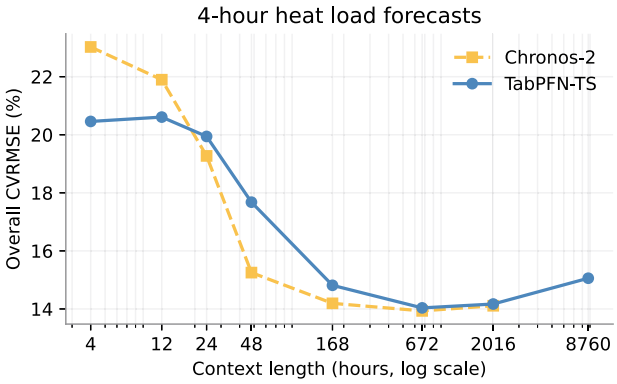


Fig. 4. Overall CVRMSE for 4-hour heat load forecasts at 15-minute resolution as a function of context length.

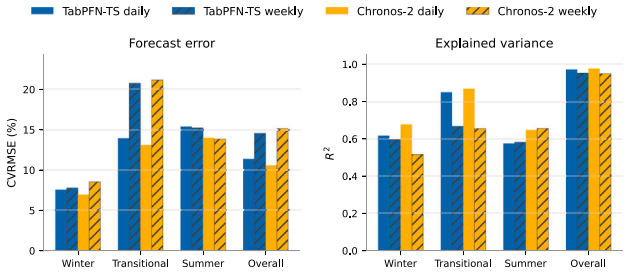


Fig. 5. CVRMSE and R^2 for daily 24-hour and weekly 168-hour forecasts at hourly resolution with a 12-week context window for TabPFN-TS and Chronos-2.

The influence of the forecast horizon was then evaluated separately. First, a shorter 4-hour horizon was evaluated at 15-minute resolution, with forecasts issued every 4 h. Fig. 4 shows how the context-window length affects this short-horizon task. When only very short context windows are available, namely 4 h or 12 h, TabPFN-TS outperforms Chronos-2. For medium context windows of 1 day, 2 days, and 1 week, Chronos-2 achieves lower CVRMSE values. For longer context windows of 4 weeks and 12 weeks, both models perform very similarly.

The shorter horizon also changes the context-length optimum for TabPFN-TS. Whereas the 24-hour forecast at 15-minute resolution performed best with a 12-week context window, the 4-hour forecast performed best with a 4-week context window. Compared with the best TabPFN-TS setup for 24-hour forecasts at 15-minute resolution, the best 4-hour setup reduced CVRMSE from 15.71% to 14.04%, corresponding to a decrease of 1.68 percentage points, or 10.7% relative to the 24-hour setup. This comparison reflects the combined effect of the shorter forecast horizon and the higher forecast-update frequency.

To separate the effect of the forecast horizon from the effect of the update frequency, an additional TabPFN-TS run issued 24-hour forecasts every 4 h and was evaluated with the non-overlapping latest-forecast-wins strategy defined in Section 3.2. With a 12-week context window, this setting achieved a CVRMSE of 14.17% and an R^2 of 0.955. Its CVRMSE is only 0.13 percentage points higher than the best 4-hour TabPFN-TS forecast, which used a 4-week context window. Thus, most of the improvement over the daily 24-hour 15-minute-resolution forecast is explained by the higher forecast-update frequency.

Finally, both models were evaluated for 168-hour forecasts at hourly resolution using a 12-week context. Moving from 24-hour forecasts issued daily to 168-hour forecasts issued weekly reduced accuracy for both models. On the selected operating weeks, TabPFN-TS achieved a slightly lower CVRMSE than Chronos-2, at 14.57% compared with 15.15%, with results by operating regime shown in Fig. 5. This reversal of the day-ahead ranking motivated an additional full-year evaluation to assess whether the advantage persisted across the year. The full-year results did not confirm the TabPFN-TS advantage: Chronos-2 achieved a lower weekly CVRMSE of 15.64%, compared with 17.06% for TabPFN-TS, as reported in Supplementary Table S3.

After the context-length, resolution, and horizon diagnostics, the sensitivity to weather-covariate choice was evaluated for the hourly 24-hour forecast setup with a 12-week context window. Table 3 reports the aggregate performance over all selected operating weeks.

Additional weather covariates produced small, model-dependent improvements on the selected weeks. Ambient temperature alone was retained as a parsimonious configuration. It is measured at many sites and is commonly available as a forecast variable, supporting transferability and operational simplicity. The mean prediction-call time increases by 2.4% for TabPFN-TS and 3.3% for Chronos-2 when moving from ambient temperature only to all weather covariates.

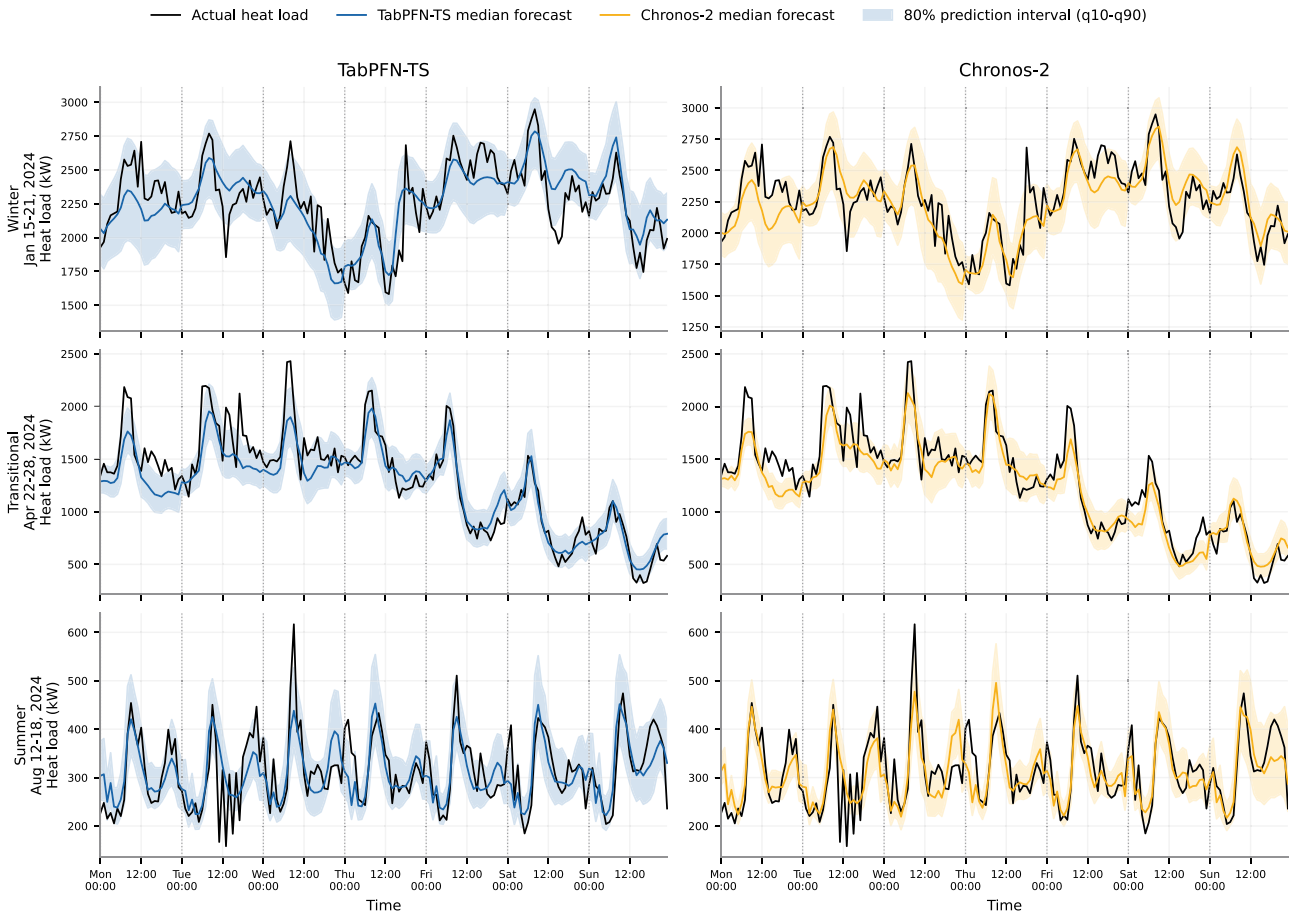


Fig. 6. Munich predictions for the selected 24-hour forecast setup at hourly resolution with a 12-week context window.

Table 3
Weather feature selection for 24-hour hourly forecasts with a 12-week context window for TabPFN-TS and Chronos-2.

Weather covariates	TabPFN-TS			Chronos-2		
	CVRMSE (%)	R^2	MAE (kW)	CVRMSE (%)	R^2	MAE (kW)
Amb. temp.	11.42	0.970	102.6	10.60	0.974	96.0
Amb. temp., relative humidity	11.43	0.970	102.2	10.62	0.974	96.7
Amb. temp., wind speed	11.09	0.972	98.8	10.56	0.975	96.0
Amb. temp., precipitation	11.16	0.972	100.1	10.59	0.975	96.0
Amb. temp., precip., wind sp.	11.25	0.971	99.9	10.38	0.976	94.2
Amb. temp., solar irradiation	11.33	0.971	101.5	10.47	0.975	95.3
All weather covariates	11.09	0.972	98.8	10.32	0.976	95.6

Motivated by the dominant annual spectral peak in the Munich heat-load data described in Section 4.1, we examine whether augmenting recent context with calendar-aligned observations from the previous year improves TabPFN-TS forecast accuracy.

Here, relevant observations denote historical data from a window around the current calendar date, shifted by one year. The recent-plus-relevant setup combines 6 weeks of recent context with a 6-week block from the same calendar period one year earlier, so that the total amount of context data matches the observed optimal 12-week recent-data context. According to the TabPFN-TS documentation, automatic temporal feature engineering must be disabled for this non-contiguous context configuration. Therefore, the recent-plus-relevant experiment used only calendar features and disabled the running-index and automatic seasonal features. This experiment was carried out only for TabPFN-TS, because Chronos-2 expects strictly regular time steps without gaps in the context.

No performance benefit was obtained from adding a 6-week block from the same calendar period one year earlier. Table 4 also shows that the default automatic temporal feature engineering reduced CVRMSE by 2.3% compared with the recent-only setup.

Table 4
Context-data selection for TabPFN-TS hourly 24-hour forecasts. Context lengths are in weeks; "auto feat." denotes automatic temporal features.

Setup	CVRMSE (%)	R^2	MAE (kW)
Recent 12 (auto feat.)	11.42	0.970	102.6
Recent 12	11.69	0.969	106.9
Recent 6 + Relevant 6	12.69	0.963	113.0

Taken together, the context-length, temporal-resolution, forecast-horizon, weather-covariate, and context-data diagnostics motivate using 24-hour-ahead forecasts at hourly resolution with a 12-week rolling context window and ambient temperature as the only weather covariate in the subsequent full-year and transfer analyses. Fig. 6 shows the resulting TabPFN-TS and Chronos-2 forecasts for the selected winter, transitional, and summer weeks.

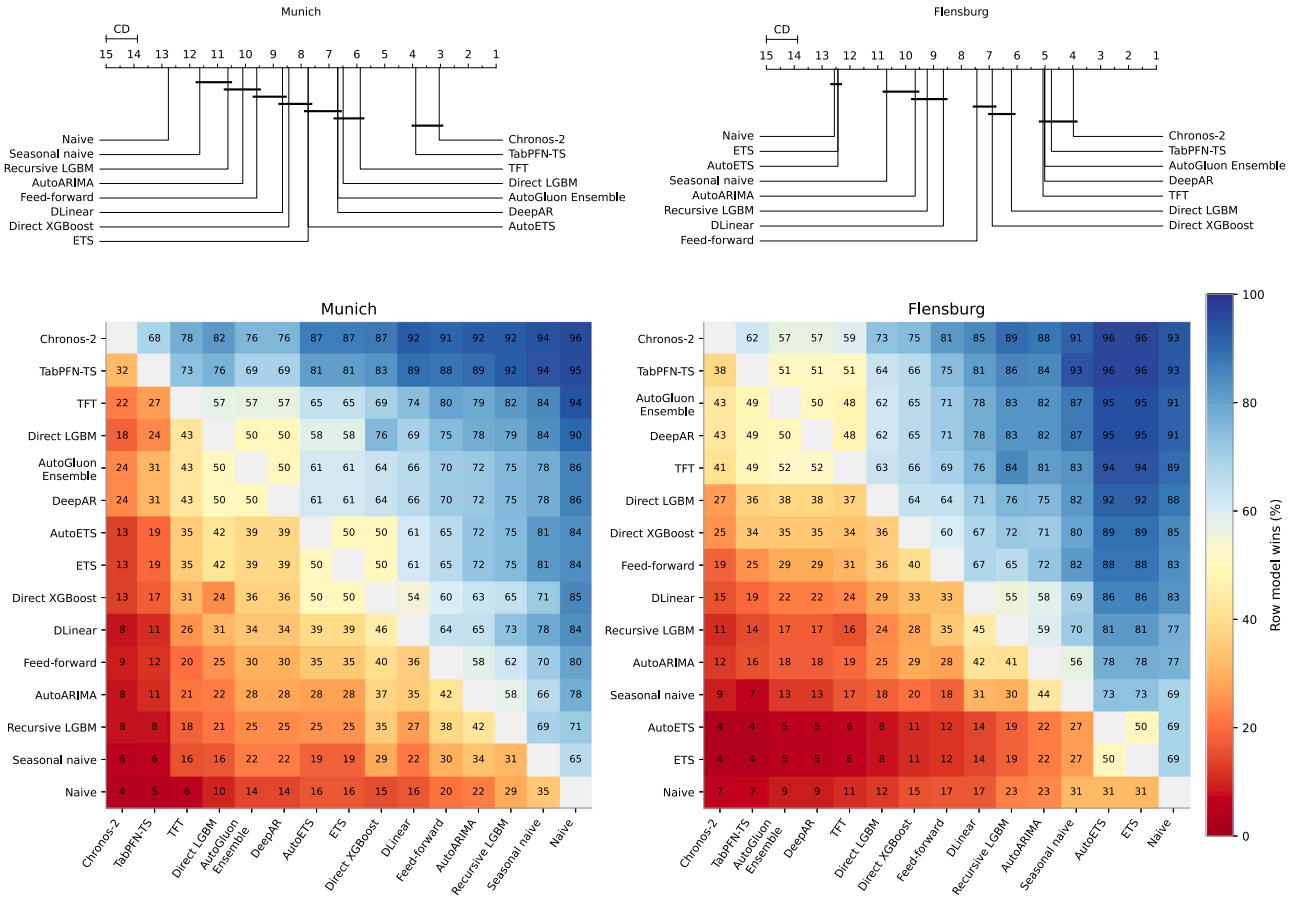


Fig. 7. Full-year model comparison based on daily CVRMSE. The upper panels show CD diagrams and the lower panels show pairwise win rates. Each heatmap cell gives the percentage of forecast days on which the row model outperforms the column model; exact ties count as half a win.

Table 5
CRPS and 80% prediction-interval widths for daily issued forecasts at hourly resolution with a 12-week context on the selected representative weeks.

Period	Model	CRPS (kW)	Width (kW)	Rel. w. (%)
Winter	TabPFN-TS	92.95	498.9	22.1
	Chronos-2	85.56	488.4	21.6
Transitional	TabPFN-TS	93.81	370.3	28.5
	Chronos-2	95.20	317.7	24.5
Summer	TabPFN-TS	25.42	114.8	36.7
	Chronos-2	23.78	93.9	30.0
Overall	TabPFN-TS	70.73	328.0	25.4
	Chronos-2	68.18	300.0	23.2

5.2. Seasonal effects on forecast accuracy and uncertainty

Fig. 5 shows that forecast performance depends strongly on the operating period. For daily 24-hour forecasts, the CVRMSE is lowest in winter, followed by the transitional period and summer. The R^2 score follows a different pattern: winter and summer are comparable, while the transitional week achieves the highest R^2. During the transitional week, the loss in accuracy when moving from daily to weekly forecasts is particularly visible.

The seasonal pattern is also reflected in the probabilistic forecasts. Table 5 reports CRPS and the absolute and relative widths of the 80% prediction interval. TabPFN-TS has higher CRPS than Chronos-2 in winter, in summer, and across the three weeks combined, indicating lower overall probabilistic forecast quality under this score. During the transitional week, however, TabPFN-TS achieves lower CRPS than Chronos-2, at 93.81 versus 95.20 kW. TabPFN-TS produces wider prediction intervals in all three weeks. For both models, the relative interval width is lowest in winter and highest in summer, with the transitional period in between. During the selected transitional week, both models show a systematic tendency to underestimate heat demand, particularly Chronos-2. The seasonal PIT histograms in Supplementary Figure S2 provide further detail.

5.3. Full-year benchmark and transfer validation

The results of the full-year simulations are reported in Table 6. The entries are sorted by aggregate CVRMSE within each network. On the Munich data set, both foundation models outperformed the trained AutoGluon baselines across the reported accuracy metrics, with Chronos-2 achieving the lowest aggregate error. To assess whether models perform consistently when transferred to a different data set, the rank-shift column reports the change in CVRMSE position in Flensburg relative to the Munich ordering. The relative error metrics were generally lower for Flensburg than for Munich, especially for the strongest models. Chronos-2, TabPFN-TS, and TFT remained the three strongest models in both networks, whereas the naive baselines stayed at the lower end of the ranking. The largest rank changes occurred for AutoETS and ETS, which lost five CVRMSE positions in Flensburg, and for Feed-forward, Recursive LGBM, and Seasonal naive, which gained three positions each. Chronos-2 again achieved the lowest aggregate error, while TabPFN-TS remained ahead of the trained AutoGluon models.

The critical-difference diagrams are shown in Fig. 7. For both data sets, TabPFN-TS lies within the critical difference of Chronos-2. On the Munich data set, the two time-series foundation models form the leading group and outperform all trained AutoGluon baselines in the

Table 6
Full-year benchmark and transfer-validation results for hourly 24-hour forecasts in 2024. Brackets denote 95% paired seven-day block-bootstrap confidence intervals.

Model	Rank shift	Aggregate metrics			Mean daily rank		
		CVRMSE (%)	R^2	MAE (kW)	CVRMSE	R^2	MAE
Munich							
Chronos-2		12.48 [10.88, 14.37]	0.962 [0.946, 0.972]	85.4 [75.9, 95.6]	3.04	3.01	3.11
TabPFN-TS		13.07 [11.49, 14.94]	0.959 [0.941, 0.969]	90.3 [79.9, 101.1]	3.89	3.86	3.83
TFT		14.61 [12.92, 16.61]	0.948 [0.928, 0.961]	102.1 [90.7, 114.3]	5.88	5.85	5.93
Direct LGBM		15.87 [14.25, 17.72]	0.939 [0.917, 0.953]	113.7 [99.1, 129.6]	6.49	6.42	6.49
DeepAR		16.54 [14.89, 18.35]	0.934 [0.910, 0.949]	117.2 [101.1, 134.1]	6.69	6.64	6.75
AutoGluon Ensemble		16.54 [14.89, 18.35]	0.934 [0.910, 0.949]	117.2 [101.1, 134.1]	6.69	6.64	6.75
Direct XGBoost		17.34 [15.59, 19.30]	0.927 [0.904, 0.942]	126.7 [110.6, 144.9]	8.44	8.39	8.61
AutoETS		18.43 [16.31, 20.90]	0.918 [0.884, 0.938]	132.8 [117.1, 148.9]	7.75	7.73	7.84
ETS		18.43 [16.31, 20.90]	0.918 [0.884, 0.938]	132.8 [117.1, 148.9]	7.75	7.73	7.84
DLinear		18.66 [16.58, 21.06]	0.916 [0.882, 0.937]	135.1 [118.6, 151.9]	8.67	8.69	8.75
Feed-forward		19.40 [17.33, 21.86]	0.909 [0.874, 0.931]	141.9 [124.5, 159.4]	9.59	9.63	9.67
AutoARIMA		20.48 [18.13, 23.24]	0.898 [0.857, 0.924]	148.3 [129.8, 167.0]	10.09	10.12	9.88
Recursive LGBM		21.07 [18.67, 23.91]	0.892 [0.848, 0.920]	152.0 [132.7, 171.7]	10.63	10.69	10.52
Seasonal naive		22.51 [19.94, 25.58]	0.877 [0.826, 0.909]	161.9 [141.1, 183.1]	11.64	11.72	11.52
Naive		23.66 [20.75, 27.22]	0.864 [0.803, 0.901]	172.1 [151.8, 193.8]	12.77	12.89	12.52
Flensburg							
Chronos-2		9.82 [8.77, 11.14]	0.970 [0.957, 0.977]	7913.6 [6828.9, 9058.9]	3.98	3.98	4.14
TabPFN-TS		10.37 [9.30, 11.71]	0.966 [0.952, 0.975]	8437.3 [7265.3, 9664.9]	4.77	4.77	4.60
TFT		10.69 [9.36, 12.27]	0.964 [0.948, 0.974]	8669.2 [7347.2, 10123.4]	5.07	5.07	5.06
DeepAR	↑ 1	10.72 [9.68, 11.91]	0.964 [0.950, 0.972]	8677.9 [7363.6, 10059.6]	5.01	5.01	5.12
AutoGluon Ensemble	↑ 1	10.72 [9.68, 11.91]	0.964 [0.950, 0.972]	8677.9 [7363.6, 10059.6]	5.01	5.01	5.12
Direct LGBM	↓ 2	11.12 [10.09, 12.27]	0.961 [0.947, 0.970]	9380.2 [8098.7, 10709.1]	6.20	6.20	6.50
Direct XGBoost		11.43 [10.39, 12.61]	0.959 [0.944, 0.968]	9716.2 [8396.1, 11066.6]	6.89	6.89	7.07
Feed-forward	↑ 3	13.05 [11.97, 14.35]	0.946 [0.926, 0.959]	10665.8 [9103.6, 12276.0]	7.44	7.44	7.53
DLinear	↑ 1	13.62 [12.59, 14.84]	0.941 [0.921, 0.954]	11200.1 [9568.3, 12896.1]	8.64	8.64	8.58
Recursive LGBM	↑ 3	13.75 [12.68, 14.92]	0.940 [0.921, 0.953]	11456.8 [9835.7, 13095.7]	9.23	9.23	9.22
Seasonal naive	↑ 3	15.53 [14.19, 17.12]	0.924 [0.896, 0.942]	12775.2 [10905.9, 14708.0]	10.68	10.68	10.40
AutoARIMA		15.80 [14.36, 17.54]	0.921 [0.890, 0.941]	12820.7 [10811.6, 14875.6]	9.66	9.66	9.55
AutoETS	↓ 5	21.21 [20.12, 22.30]	0.858 [0.813, 0.886]	18255.5 [15406.1, 21185.2]	12.43	12.43	12.38
ETS	↓ 5	21.21 [20.12, 22.30]	0.858 [0.813, 0.886]	18255.5 [15406.1, 21185.2]	12.43	12.43	12.38
Naive		22.39 [21.19, 23.57]	0.842 [0.791, 0.874]	18766.4 [15467.8, 22171.0]	12.56	12.56	12.34

Table 7
Annual mean CVRMSE across eight learned AutoGluon setups in Munich. Differences are KSWIN minus static; brackets denote 95% percentile confidence intervals.

Year	Static CVRMSE (%)	KSWIN CVRMSE (%)	ΔCVRMSE (percentage points)	Relative reduction (%)
2021	19.96 [18.24, 22.10]	19.09 [17.27, 21.34]	−0.88 [−1.42, −0.28]	4.4 [1.4, 7.2]
2022	19.55 [17.18, 22.68]	17.62 [14.96, 21.31]	−1.93 [−2.77, −1.02]	9.9 [4.8, 14.8]
2023	22.85 [20.40, 25.60]	19.67 [17.18, 22.73]	−3.19 [−4.08, −2.20]	13.9 [9.3, 18.1]
2024	20.69 [19.00, 22.49]	16.89 [15.09, 18.93]	−3.80 [−4.87, −2.65]	18.4 [12.7, 23.4]

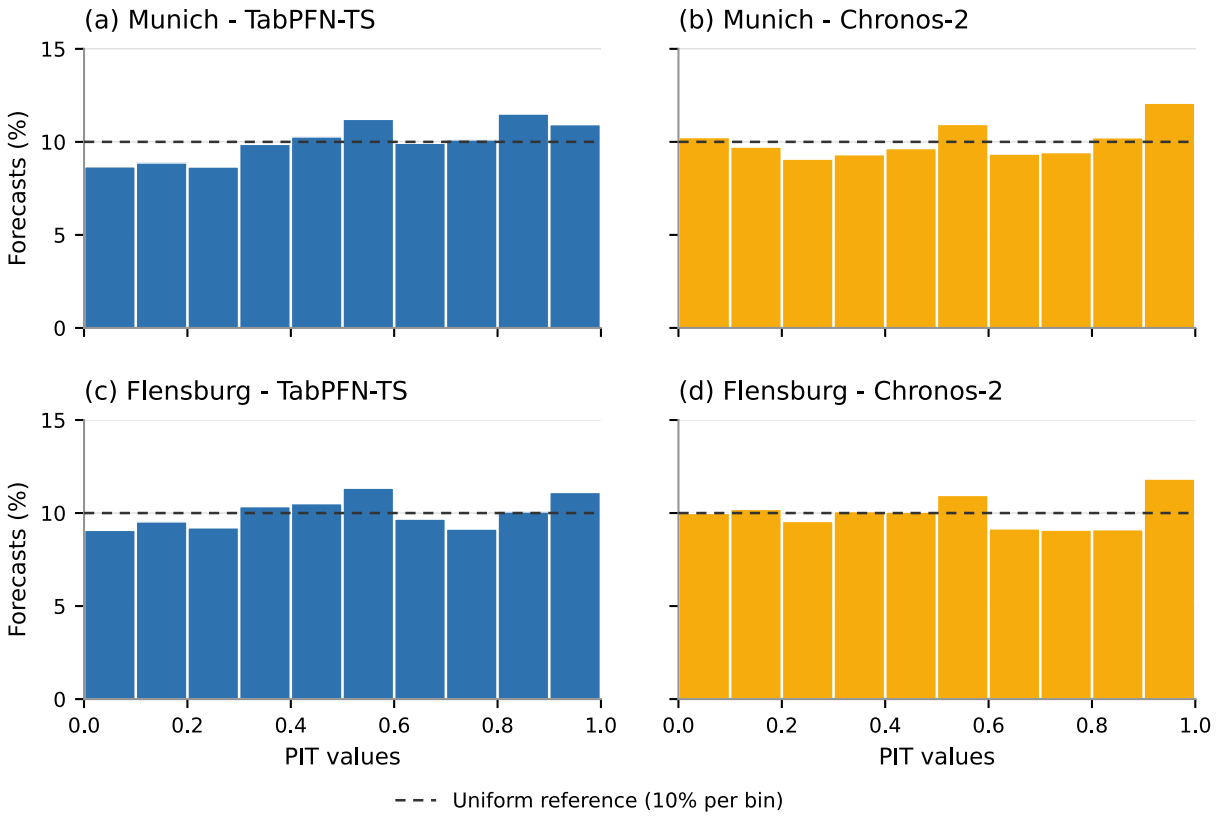


Fig. 8. Approximate PIT histograms for the principal 2024 hourly, 24-hour forecasting benchmark in Munich and Flensburg.

daily-rank comparison. In Flensburg, the leading group is broader: TFT, DeepAR, and AutoGluon Ensemble also lie within the critical difference of the foundation-model results. The lower panels of Fig. 7 show the corresponding pairwise daily CVRMSE win rates. Supplementary Table S8 reports initial AutoGluon fitting times.

5.4. Probabilistic forecast quality

Chronos-2 achieves lower full-year CRPS than TabPFN-TS in both networks: 60.49 versus 63.66 kW in Munich and 5.62 versus 5.94 MW in Flensburg, indicating better overall probabilistic forecast quality. Fig. 8 compares the approximate PIT distributions. Both models show deviations from uniformity. In Munich, TabPFN-TS shows a broader shift towards higher PIT values, suggesting a general tendency to underpredict heat load. Chronos-2 instead exhibits an excess concentrated in the highest PIT bin in both networks, indicating that its upper predictive tail does not adequately cover the observed loads.

5.5. Sensitivity to anomaly exclusion

We also investigated the influence of potentially anomalous target observations using the screening criteria described in Section 3.3. After excluding 24 Munich forecast days containing time steps identified by these criteria, 342 evaluation days remained. The complete aggregate CVRMSE ordering was unchanged, with Chronos-2, TabPFN-TS and TFT retaining the first three positions. Chronos-2 and TabPFN-TS also remained the leading group in the critical-difference comparison. CVRMSE and MAE decreased for 14 of the 15 models; the naive baseline was the exception. For TabPFN-TS and Chronos-2, the standard deviation of daily MAE decreased from 51.8 to 48.8 kW and from 49.4 to 45.0 kW, respectively. Both foundation models achieved lower CRPS after exclusion, with Chronos-2 retaining the lower score. No Flensburg day met the screening criteria. Supplementary Table S4 reports the detailed deterministic results in Panel A and paired changes after anomaly exclusion for all 15 models in Panel B. Full-year and anomaly-filtered CRPS values are reported in Supplementary Table S5.

5.6. Baseline retraining on detected concept drift

Table 7 compares static models and models undergoing KSWIN-triggered batch retraining over 2021–2024, averaging scores equally across the trained AutoGluon setups: TFT, DeepAR, DLinear, Feed-forward, Direct LGBM, Direct XGBoost, Recursive LGBM, and the AutoGluon Ensemble. Retraining improves their average performance in every year, with paired confidence intervals excluding zero. The relative CVRMSE reduction increases continuously from 4.4% in 2021 to 18.4% in 2024 as the deployment period moves further from the initial 2020 training period. MAE and R^2 show the same direction of improvement and increasing relative gains, as shown in Supplementary Table S6.

The benefit is model-dependent. Over the complete four-year period, Direct XGBoost and Direct LGBM show the largest CVRMSE reductions, from 30.03% to 19.31% and from 26.21% to 18.81%, respectively, whereas DLinear and Feed-forward deteriorate after retraining, as detailed in Supplementary Table S7. The annual results do not show a consistent relationship between the refitting benefit and either the Hellinger distance or the number of KSWIN drift detections in Table 1.

The baseline ranking also changes relative to the main 2024 benchmark trained on 2023 data: The static AutoGluon Ensemble ranks ahead of TFT over 2021–2024, while Direct LGBM falls behind several other baselines. Details can be found in Supplementary Table S7. This illustrates the dependence of trained-model performance on the training and deployment periods. Across the evaluated training setups and evaluation periods, Chronos-2 and TabPFN-TS retain the leading point-estimate performance, with and without KSWIN-triggered baseline retraining. The annual CVRMSE gap between the mean of the eight static trained-baseline setups and each TSFM increases monotonically

Table 8
Forecast quality with realized temperatures and coherent retrospective ECMWF temperature predictions for 208 matched Munich forecast starts from 7 June to 31 December 2024.

Model	Weather cov.	Point forecast			Probabilistic forecast
		CVRMSE (%)	R^2	MAE (kW)	CRPS (kW)
TabPFN-TS	Realized	11.89	0.974	72.2	50.9
	Predicted	12.77 (+7.4%)	0.971 (−0.4%)	77.8 (+7.7%)	55.0 (+8.1%)
Chronos-2	Realized	11.41	0.976	68.5	48.3
	Predicted	12.56 (+10.1%)	0.971 (−0.5%)	76.1 (+11.0%)	53.3 (+10.2%)

Table 9
Munich Base, HFHR and MRRC forecast performance pooled over 2021–2024. Brackets denote 95% percentile confidence intervals.

Setup	Short-term forecast			12-hour integrated heat demand		Computation
	CVRMSE (%)	R^2	MAE (kW)	E-CVRMSE (%)	E-bias (%)	RTF
TabPFN-TS						
Base	21.63	0.897	125.3	8.43	−0.99	2.24×10^{-5}
	[19.95, 23.47]	[0.880, 0.911]	[119.0, 132.1]	[7.79, 9.12]	[−1.38, −0.60]	
HFHR	21.13	0.902	115.4	8.78	−1.77	3.86×10^{-4}
	[19.37, 23.08]	[0.884, 0.916]	[109.6, 121.7]	[7.92, 9.74]	[−2.25, −1.32]	
MRRC	21.15	0.902	114.8	8.14	−1.09	2.70×10^{-4}
	[19.38, 23.10]	[0.884, 0.916]	[109.1, 120.9]	[7.52, 8.82]	[−1.44, −0.74]	
Chronos-2						
Base	21.24	0.901	121.9	7.88	−0.66	5.52×10^{-7}
	[19.55, 23.09]	[0.884, 0.914]	[115.6, 128.5]	[7.22, 8.61]	[−0.97, −0.35]	
HFHR	20.04	0.912	106.3	8.23	−0.18	7.72×10^{-6}
	[18.34, 21.87]	[0.896, 0.925]	[101.1, 111.9]	[7.33, 9.22]	[−0.47, 0.10]	
MRRC	20.21	0.910	106.9	7.61	−0.66	6.30×10^{-6}
	[18.44, 22.14]	[0.893, 0.924]	[101.7, 112.6]	[6.97, 8.33]	[−0.94, −0.38]	

from 2021 to 2024: from 4.73 to 7.62 percentage points for TabPFN-TS and from 5.42 to 8.21 percentage points for Chronos-2. With KSWIN-triggered retraining, the corresponding gaps remain within 3.31–3.86 and 3.66–4.54 percentage points, respectively, without a monotonic increase.

Retraining the complete AutoGluon model suite took an average of 179.72 s per refit. Supplementary Table S9 reports the cumulative refitting time divided by prediction-call time for each model.

5.7. Weather-forecast sensitivity

Table 8 summarizes the resulting change in point and probabilistic forecast quality. Both models are affected, with Chronos-2 showing slightly greater sensitivity in this experiment: its CVRMSE increases by 10.1%, compared with 7.4% for TabPFN-TS. The degradation is visible for both models in point-forecast accuracy and distributional scores. The paired 95% confidence intervals for changes in CVRMSE, MAE and R^2 exclude zero for both models. Details can be found in Supplementary Section S3. The standard deviation of daily MAE also increased from 42.0 to 46.2 kW for TabPFN-TS and from 40.0 to 46.7 kW for Chronos-2, indicating greater day-to-day variability in forecast error.

5.8. Multi-resolution residual-correction forecaster performance

Table 9 compares Base, HFHR and MRRC forecasts over the complete 2021–2024 evaluation period for both TSFM backbones.

Relative to the Base Forecaster, MRRC significantly reduces short-term CVRMSE and MAE, increases R^2, and reduces 12-hour E-CVRMSE for both backbones. The relative short-term CVRMSE reduction is larger for Chronos-2, at 4.9%, compared with 2.2% for TabPFN-TS. The relative reduction in 12-hour E-CVRMSE is approximately 3.4% for both.

Compared with direct HFHR forecasting, MRRC significantly reduces 12-hour E-CVRMSE by 7.3% for TabPFN-TS and 7.5% for Chronos-2. Its short-term benefits are less consistent: CVRMSE and R^2 differences are not statistically significant for either backbone, while MAE decreases slightly for TabPFN-TS and increases slightly for Chronos-2. Both MAE differences are statistically significant but below 1% in relative terms. Paired confidence intervals are reported in Supplementary Table S10.

MRRC also has lower recorded computational cost than HFHR, reducing RTF by 30% for TabPFN-TS and 18% for Chronos-2. The Base Forecaster remains the least expensive setup; MRRC requires approximately 12 and 11 times its recorded prediction time, respectively.

6. Discussion

Chronos-2 achieved stronger full-year deterministic performance and lower CRPS in both networks, consistent with the benefits of its time-series-specific architecture and pretraining. TabPFN-TS nevertheless remained within the critical-difference threshold of Chronos-2 in the daily CVRMSE-rank comparison in both networks. These results show that a model pretrained entirely on synthetic non-temporal tabular data can support competitive heat-load forecasting when combined with time-series featurization. Together with the solar-energy results of Liu et al. [40], this supports the applicability of synthetically pretrained TabPFN models across different energy-related tasks. Task- and domain-tailored priors may offer further improvements.

The context-length diagnostics in Section 5.1 show that forecast accuracy does not improve monotonically with context length: longer histories can yield poorer forecasts than shorter ones. Optimal context length therefore remains application-specific even for foundation models. This may reflect superimposed demand processes: space heating is strongly seasonal and weather-driven, whereas domestic hot water is more stochastic and exhibits different, weaker temporal patterns. Without explicit separation, long contexts may include regimes irrelevant to the current forecast and obscure rather than improve in-context prediction.

The results support the ability of both TSFMs to capture predictable components of higher-frequency heat-load dynamics at 15-minute resolution. In the four-year comparison, the direct HFHR setup achieved lower short-term CVRMSE than the interpolated hourly Base Forecaster for both backbones in every year. Nevertheless, the configuration experiments showed lower forecasting errors at hourly resolution. This

is likely due to temporal aggregation: hourly averaging smooths high-frequency dynamics that are difficult to predict at 15-minute resolution. One potential driver of these dynamics is domestic-hot-water draw-off behavior. The exact timing of individual draw-off events contains a stochastic component and is therefore inherently difficult to forecast [52], whereas broader temporal patterns in domestic-hot-water demand are easier to capture [53]. Additional sources of high-frequency variability may include hydraulic interactions between buildings [54], varying transport delays between the energy hub and consumers, delay-dependent heat losses, hysteresis-controlled pressure pumps, and other operational effects. The short-horizon experiments also revealed a setting in which TabPFN-TS outperformed Chronos-2 in deterministic accuracy: for the 4-hour forecast task at 15-minute resolution, TabPFN-TS performed better than Chronos-2 when only very short context windows of 4 or 12 h were available, as shown in Fig. 4. This is consistent with energy-demand forecasting results by Kim et al. [55], who report strong TabPFN performance under short-context conditions, and with the broader finding that TabPFN can deliver accurate predictions under low data availability [11].

Moving from daily day-ahead forecasts to weekly week-ahead forecasts reduced accuracy for both models, highlighting the importance of matching forecast horizon, update frequency and context length to the operational task.

The seasonal effects in Section 5.2 provide complementary evidence on how demand composition affects forecast accuracy and uncertainty. The lower winter daily CVRMSE in Fig. 5 is consistent with a larger share of weather-driven space-heating demand, creating a stronger and more predictable baseload. In summer, the smaller space-heating share and greater relative contribution of stochastic domestic hot water demand are expected to increase the relative forecast error. The higher R^2 during the transitional period results from greater variation in measured demand, allowing more variance to be explained despite the relative error not being lowest. The greater accuracy loss from daily to weekly forecasts in this period may arise because the longer horizon spans rapidly changing load and weather conditions, making multi-day extrapolation harder than in the more stable winter and summer periods. The relative interval widths in Table 5 similarly suggest that relative uncertainty rises as space-heating demand becomes less dominant and stochastic components gain importance. The transitional-week underprediction reported in Section 5.2, particularly for Chronos-2, may reflect difficulty adapting to changing operating conditions. One possible explanation is that the 12-week context window contains observations less representative of this regime.

The small, model-dependent improvements from additional weather covariates support ambient temperature as a parsimonious default for the evaluated forecasting task. This preference for a compact input set is consistent with Petković et al. [56], who favor compact weather-input sets over large sets of meteorological covariates for heat load prediction. Because the additional computational cost was small in the tested setting, including further reliably available weather covariates remains defensible when sufficiently accurate future values are available.

The transfer results in Section 5.3 and Table 6 should be interpreted in light of network scale. The lower relative errors in the Flensburg validation case may partly reflect aggregation: domestic hot water demand is comparatively stochastic at the individual-building level but becomes more predictable as more buildings are connected. Because Flensburg is a city-scale district-heating system whereas Munich is a smaller mixed-use district, stronger aggregation in Flensburg may smooth idiosyncratic consumption events and contribute to its lower relative forecasting errors.

Our literature review in Section 2.1 identified retraining under concept drift as a central challenge in heat load forecasting. The four-year comparison in Section 5.6 shows that the performance gap between static trained baselines and TSFMs widened over time, whereas the gap to baselines undergoing KSWIN-triggered batch retraining remained comparatively stable. This pattern supports our hypothesis that rolling-context in-context learning can accommodate changing system dynamics without task-specific weight updates. The remaining advantage over retrained baselines may reflect inherent model differences, the shorter 12-week context compared with a full training year that may retain more pre-change observations, and the TSFMs' ability to identify relevant dynamics within that context. Our experiments do not isolate the relative contributions of these factors.

As described in Section 5.1, adding calendar-aligned observations from the previous year did not improve forecasts. However, the 2023 and 2024 data span the ongoing expansion of the Munich district and associated demand growth. The added block may therefore represent a different system configuration. Calendar alignment also only approximates physical similarity; selecting context by ambient temperature or clustered operating conditions may be more meaningful. However, even temperature-matched historical observations may be less representative if the network configuration or operating conditions have changed. Seasonally relevant context may thus be more beneficial in stationary district-heating systems.

Using coherent ECMWF temperature hindcasts increased CVRMSE by 7.4% for TabPFN-TS and 10.1% for Chronos-2 relative to the perfect-weather reference, while CRPS increased by 8.1% and 10.2%, respectively. These moderate relative increases suggest that both models retain much of their predictive accuracy under imperfect weather inputs in the evaluated setting. The extent of deterioration will nevertheless depend on the quality of the weather predictions available in operation.

For both backbones, TSFM-based residual correction significantly improved short-term CVRMSE, MAE and 12-hour E-CVRMSE relative to the uncorrected Base Forecaster over 2021–2024. This demonstrates the feasibility of using TSFMs as residual forecasters for TSFM predictions in the evaluated multi-resolution setup. Compared with HFHR, MRRC improved predictions of integrated heat demand over 12 h while achieving broadly similar short-term accuracy. One possible explanation is its separation of forecasting time scales: the hourly Base Forecaster potentially captures slower variations associated with space heating, consistent with the dominant low-frequency patterns identified in the spectral analysis, while the Residual Forecaster addresses faster dynamics, including those associated with domestic hot water. Its computational savings arise from issuing expensive long-context base forecasts less frequently and using shorter-context residual forecasts for frequent updates. The MRRC architecture's benefit therefore depends on whether operational priorities favor short-term trajectory accuracy, longer-horizon heat-demand planning, or lower computational cost.

6.1. Limitations

Although the two case studies cover DHNs with substantially different scales and building stocks, both are located in Germany and are therefore subject to broadly similar seasonal weather effects. Additionally, configuration selection and MRRC evaluation are restricted to Munich only. Generalization beyond these case studies therefore remains to be established.

Configuration choices were evaluated on selected operating weeks rather than exhaustively optimized for every task. Results can depend on the evaluation period: TabPFN-TS's lower week-ahead CVRMSE on the selected operating weeks did not persist in the full-year evaluation. Findings from selected weeks therefore need not generalize to year-round performance.

The principal benchmark assumes perfect weather forecasts by using realized future temperatures. The weather sensitivity analysis uses retrospective hindcasts rather than verified forecasts delivered operationally in 2024. The sensitivity analysis of weather forecast uncertainty covers only 24-hour heat-load predictions. Its findings therefore cannot be directly extended to longer forecasting horizons, for which increasing weather forecast uncertainty may have a larger effect on heat-load prediction.

KSWIN-triggered batch retraining uses the preceding year of observations and trains the models from scratch. Consequently, observations preceding a detected change can remain in the training window; no explicit change-based selection of training data is applied.

Finally, the HFHR setup used as a benchmark in the MRRC experiments was not independently optimized for its 13-hour forecast horizon and increased update frequency.

7. Conclusion & future work

This work systematically evaluated TabPFN-TS and Chronos-2 for zero-shot probabilistic heat load forecasting in DHNs. The results demonstrate the suitability of both evaluated foundation models for this task, with Chronos-2 achieving the strongest overall forecast quality and TabPFN-TS also proving competitive. For hourly 24-hour forecasting, both models achieved their lowest error among the tested contexts with a 12-week window, whereas shorter contexts performed better in some high-resolution settings. Ambient temperature alone provided an effective weather-covariate set.

The experiments show that rolling-context foundation-model forecasting can perform well in an evolving DHN without repeated task-specific retraining. Both foundation models performed strongly in the Flensburg validation case and outperformed the trained AutoGluon baselines in the main full-year comparison. At the same time, the diagnostic experiments showed that configuration choices remain important: more context data, higher temporal resolution, and larger weather-covariate sets did not automatically improve performance.

The Multi-Resolution Residual-Correction Forecaster illustrates how residual correction can be adapted to DHN operational requirements. Compared with direct high-frequency, high-resolution forecasting, MRRC significantly reduced 12-hour E-CVRMSE for both backbones and required less recorded prediction time, while short-term accuracy remained broadly similar.

Future work on MRRC should address both probabilistic forecasting and further configuration and architecture development. A probabilistic extension requires representing uncertainty consistently across the hourly base forecast and quarter-hourly residual correction, including beyond the two-hour residual horizon. Additionally, configuration and architecture studies could vary the Residual Forecaster horizon, test multiple residual-correction stages, and compare combinations of TSFMs and classical machine-learning models as Base and Residual Forecasters. The separation of slow load trends and high-frequency deviations could also be evaluated in other energy-system forecasting tasks.

A second direction for future work is to adapt foundation models more closely to district-heating heat load forecasting by preconditioning them on real-world time series, simulation data, or synthetic data containing the relevant dynamics. Task-specific prior data or further weight adaptation may improve forecast accuracy. However, if target-network data are included in the prior or used for adaptation, robustness under concept drift must be evaluated separately. For the trained baselines, future work could investigate whether excluding pre-change training data or updating existing models through warm-starting or transfer learning improves adaptation and reduces retraining costs.

Finally, future work should extend forecasts beyond heat demand to mass flow rates as well as supply and return temperatures. The delivered heat rate is proportional to mass flow rate and temperature difference, while the efficiency and usability of sustainable sources such as heat pumps and waste-heat streams depend on the required temperature level. Spatially distributed multivariate forecasts within the network could further support decentralized heat-source integration by providing more detailed information for operational planning and control.

CRediT authorship contribution statement

Ben Spoek: Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Karim K. Ben Hicham:** Writing – review & editing, Validation, Software, Methodology, Conceptualization. **Kai Derzsi:** Writing – review & editing, Validation, Methodology, Formal analysis, Data curation. **Philipp Althaus:** Writing – review & editing, Supervision, Conceptualization. **Alexander Mitsos:** Writing – review & editing, Supervision, Project administration, Funding acquisition. **Dirk Müller:** Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of Generative AI and AI-Assisted Technologies in the Manuscript Preparation Process

During the preparation of this work, the authors used OpenAI Codex to assist with code review and editing, manuscript organization, language refinement, and consistency checks. The authors reviewed and edited the output as needed and take full responsibility for the content of the published article.

Funding

We gratefully acknowledge the financial support provided by the Federal Ministry for Economic Affairs and Energy (BMWE) under funding code 03EN3118. We acknowledge support of the Werner Siemens Foundation in the frame of the WSS Research Center “catalaix”. The funders had no role in the study design, data collection, analysis, interpretation, manuscript preparation, or decision to submit the article.

Acknowledgments and computational resources

Computations were performed with computing resources granted by RWTH Aachen University under project rwth2083.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.egyai.2026.100942.

Code & data availability

The code used for the experiments is available on GitHub at https://github.com/benspoek/tsfm-heatload and on Zenodo at https://doi.org/10.5281/zenodo.23101871. The Munich district heating data set used in this study is not publicly available due to third-party restrictions on the underlying operational data. The Flensburg district heating data set is a third-party data set and is publicly available [50].

References

[1] European Parliament and Council of the European Union. Regulation (EU) 2021/1119 of the European parliament and of the council of 30 june 2021 establishing the framework for achieving climate neutrality and amending regulations (EC) no 401/2009 and (EU) 2018/1999 (‘European climate law’). Official J Eur Union 2021;1–17, URL https://eur-lex.europa.eu/eli/reg/2021/1119/oj.

[2] Lund H, Werner S, Wiltshire R, Svendsen S, Thorsen JE, Hvelplund F, Mathiesen BV. 4Th generation district heating (4GDH): Integrating smart thermal grids into future sustainable energy systems. Energy 2014;68:1–11. http://dx.doi.org/10.1016/j.energy.2014.02.089.

[3] Jodeiri A, Goldsworthy M, Buffa S, Cozzini M. Role of sustainable heat sources in transition towards fourth generation district heating – a review. Renew Sustain Energy Rev 2022;158:112156. http://dx.doi.org/10.1016/j.rser.2022.112156.

[4] Vandermeulen A, Van Der Heijde B, Helsen L. Controlling district heating and cooling networks to unlock flexibility: A review. Energy 2018;151:103–15. http://dx.doi.org/10.1016/j.energy.2018.03.034.

[5] Leiria D, Johra H, Marszal-Pomianowska A, Pomianowski MZ. A methodology to estimate space heating and domestic hot water energy demand profile in residential buildings from low-resolution heat meter data. Energy 2023;263:125705. http://dx.doi.org/10.1016/j.energy.2022.125705.

[6] Wojdyga K. An influence of weather conditions on heat demand in district heating systems. Energy Build 2008;40(11):2009–14. http://dx.doi.org/10.1016/j.enbuild.2008.05.008.

[7] Ivanko D, Sørensen ÅL, Nord N. Splitting measurements of the total heat demand in a hotel into domestic hot water and space heating heat use. Energy 2021;219:119685. http://dx.doi.org/10.1016/j.energy.2020.119685.

[8] Dang LM, Lee S, Li Y, Oh C, Nguyen TN, Song H-K, Moon H. Daily and seasonal heat usage patterns analysis in heat networks. Sci Rep 2022;12(1):9165. http://dx.doi.org/10.1038/s41598-022-13030-6.

[9] Derzsi K, Klingebiel J, Müller D. Concept drift detection in district heating: A characterization framework and detector-centric benchmark. Energy AI 2026;25:100788. http://dx.doi.org/10.1016/j.egyai.2026.100788.

[10] Celik B, Vanschoren J. Adaptation strategies for automated machine learning on evolving data. IEEE Trans Pattern Anal Mach Intell 2021;43(9):3067–78. http://dx.doi.org/10.1109/TPAMI.2021.3062900.

[11] Hollmann N, Müller S, Purucker L, Krishnakumar A, Körfer M, Hoo SB, Schirrmeister RT, Hutter F. Accurate predictions on small data with a tabular foundation model. Nature 2025;637(8045):319–26. http://dx.doi.org/10.1038/s41586-024-08328-6.

[12] Qu J, Holzmüller D, Varoquaux G, Le Morvan M. Tabicl: A tabular foundation model for in-context learning on large data. In: Singh A, Fazel M, Hsu D, Lacoste-Julien S, Berkenkamp F, Maharaj T, Wagstaff K, Zhu J, editors. Proceedings of the 42nd international conference on machine learning. Proceedings of machine learning research, vol. 267, PMLR; 2025, p. 50817–47.

[13] Hoo SB, Müller S, Salinas D, Hutter F. From tables to time: extending tabpfn-v2 to time series forecasting. 2025, http://dx.doi.org/10.48550/arXiv.2501.02945, arXiv, arXiv preprint arXiv:2501.02945, URL https://arxiv.org/abs/2501.02945.

[14] Obermeier M, Pruckner M, Haselbeck F, Zeiselmair A. FETS benchmark: foundation models outperform dataset-specific machine learning in energy time series forecasting. 2026, http://dx.doi.org/10.48550/arXiv.2604.22328, arXiv, arXiv preprint arXiv:2604.22328, URL https://arxiv.org/abs/2604.22328.

[15] Ansari AF, Shchur O, Küken J, Auer A, Han B, Mercado P, Rangapuram SS, Shen H, Stella L, Zhang X, Goswami M, Kapoor S, Maddix DC, Guerron P, Hu T, Yin J, Erickson N, Desai PM, Wang H, Rangwala H, Karypis G, Wang Y, Bohlke-Schneider M. Chronos-2: from univariate to universal forecasting. 2025, http://dx.doi.org/10.48550/arXiv.2510.15821, arXiv, arXiv preprint arXiv:2510.15821, URL https://arxiv.org/abs/2510.15821.

[16] Ntakolia C, Anagnostis A, Moustakidis S, Karcanias N. Machine learning applied on the district heating and cooling sector: A review. Energy Syst 2022;13(1):1–30. http://dx.doi.org/10.1007/s12667-020-00405-9.

[17] Dotzauer E. Simple model for prediction of loads in district-heating systems. Appl Energy 2002;73(3–4):277–84. http://dx.doi.org/10.1016/S0306-2619(02)00078-8.

[18] Nielsen HA, Madsen H. Modelling the heat consumption in district heating systems using a grey-box approach. Energy Build 2006;38(1):63–71. http://dx.doi.org/10.1016/j.enbuild.2005.05.002.

[19] Gadd H, Werner S. Daily heat load variations in Swedish district heating systems. Appl Energy 2013;106:47–55. http://dx.doi.org/10.1016/j.apenergy.2013.01.030.

[20] Fang T, Lahdelma R. Evaluation of a multiple linear regression model and SARIMA model in forecasting heat demand for district heating system. Appl Energy 2016;179:544–52. http://dx.doi.org/10.1016/j.apenergy.2016.06.133.

[21] Dahl M, Brun A, Andresen GB. Using ensemble weather predictions in district heating operation and load forecasting. Appl Energy 2017;193:455–65. http://dx.doi.org/10.1016/j.apenergy.2017.02.066.

[22] Park TC, Kim US, Kim L-H, Jo BW, Yeo YK. Heat consumption forecasting using partial least squares, artificial neural network and support vector regression techniques in district heating systems. Korean J Chem Eng 2010;27(4):1063–71. http://dx.doi.org/10.1007/s11814-010-0220-9.

[23] Finkenrath M, Faber T, Behrens F, Leiprecht S. Holistic modelling and optimisation of thermal load forecasting, heat generation and plant dispatch for a district heating network. Energy 2022;250:123666. http://dx.doi.org/10.1016/j.energy.2022.123666.

[24] Runge J, Saloux E. A comparison of prediction and forecasting artificial intelligence models to estimate the future energy demand in a district heating system. Energy 2023;269:126661. http://dx.doi.org/10.1016/j.energy.2023.126661.

[25] Izadyar N, Ong HC, Shamshirband S, Ghadamian H, Tong CW. Intelligent forecasting of residential heating demand for the district heating system based on the monthly overall natural gas consumption. Energy Build 2015;104:208–14. http://dx.doi.org/10.1016/j.enbuild.2015.07.006.

[26] Powell KM, Sriprasad A, Cole WJ, Edgar TF. Heating, cooling, and electrical load forecasting for a large-scale district energy system. Energy 2014;74:877–85. http://dx.doi.org/10.1016/j.energy.2014.07.064.

[27] Song J, Zhang L, Xue G, Ma Y, Gao S, Jiang Q. Predicting hourly heating load in a district heating system based on a hybrid CNN-LSTM model. Energy Build 2021;243:110998. http://dx.doi.org/10.1016/j.enbuild.2021.110998.

[28] Kato K, Sakawa M, Ishimaru K, Ushiro S, Shibano T. Heat load prediction through recurrent neural network in district heating and cooling systems. In: 2008 IEEE international conference on systems, man and cybernetics. 2008, p. 1401–6. http://dx.doi.org/10.1109/ICSMC.2008.4811482.

[29] Xue P, Qi C, Li J, Kong L, Song X. Heating load prediction based on attention long short term memory: A case study of Xingtai. Energy 2020;203:117846. http://dx.doi.org/10.1016/j.energy.2020.117846.

[30] Gong M, Zhao Y, Sun J, Han C, Sun G, Yan B. Load forecasting of district heating system based on informer. Energy 2022;253:124179. http://dx.doi.org/10.1016/j.energy.2022.124179.

[31] Gama J, Žliobaitė I, Bifet A, Pechenizkiy M, Bouchachia A. A survey on concept drift adaptation. ACM Comput Surv 2014;46(4):1–37. http://dx.doi.org/10.1145/2523813.

[32] Lu J, Liu A, Dong F, Gu F, Gama J, Zhang G. Learning under concept drift: A Review. IEEE Trans Knowl Data Eng 2019;31(12):2346–63. http://dx.doi.org/10.1109/TKDE.2018.2876857.

[33] Webb GI, Hyde R, Cao H, Nguyen HL, Petitjean F. Characterizing concept drift. Data Min Knowl Discov 2016;30(4):964–94. http://dx.doi.org/10.1007/s10618-015-0448-4.

[34] Hollmann N, Müller S, Eggensperger K, Hutter F. TabPFN: A transformer that solves small tabular classification problems in a second. In: The eleventh international conference on learning representations. 2023, URL https://openreview.net/forum?id=cp5PvcI6w8_.

[35] Müller S, Hollmann N, Pineda Arango S, Grabocka J, Hutter F. Transformers can do Bayesian inference. In: The tenth international conference on learning representations. 2022, URL https://openreview.net/forum?id=KSugKcbNf9.

[36] Aksu T, Woo G, Liu J, Liu X, Liu C, Savarese S, Xiong C, Sahoo D. GIFT-Eval: A benchmark for general time series forecasting model evaluation. 2024, http://dx.doi.org/10.48550/arXiv.2410.10393, arXiv, arXiv preprint arXiv:2410.10393, URL https://arxiv.org/abs/2410.10393.

[37] Meyer M, Zapata Gonzalez DR, Kaltenpoth SB, Müller O. Benchmarking time series foundation models for short-term household electricity load forecasting. IEEE Access 2025;13:218141–53. http://dx.doi.org/10.1109/ACCESS.2025.3648056.

[38] Koch F, Raisch F, Tischler B. Thermal-GEMs: generalized models for building thermal dynamics. In: The 13th ACM international conference on systems for energy-efficient buildings, cities, and transportation. Banff, Canada: Association for Computing Machinery; 2026, p. 115–25. http://dx.doi.org/10.1145/3744256.3812565.

[39] Hertel M, Nikoltchovska A, Pütz S, Mikut R, Schäfer B, Hagenmeyer V. Explainable load forecasting with covariate-informed time series foundation models. In: Proceedings of the 17th ACM international conference on future and sustainable energy systems. Association for Computing Machinery; 2026, p. 612–26. http://dx.doi.org/10.1145/3744255.3811724.

[40] Liu B, Chen Y, Yang D. Evaluating TabPFN for regression tasks in solar energy meteorology. Sol Energy 2026;309:114472. http://dx.doi.org/10.1016/j.solener.2026.114472.

[41] Alkhulaifi N, Bowler AL, Pekaslan D, Watson NJ, Triguero I. AutoEnergy: An automated feature engineering algorithm for energy consumption forecasting with AutoML. Knowl-Based Syst 2025;329:114300. http://dx.doi.org/10.1016/j.knosys.2025.114300.

[42] Ramachandran A, Chatterjee S, Neergaard TFB, Oberndoerfer M, Maier A, Bayer S. A deep learning framework for heat demand forecasting using time-frequency representations of decomposed features. Energy AI 2026;24:100704. http://dx.doi.org/10.1016/j.egyai.2026.100704.

[43] Chitsaz F, Haratizadeh S. Bridging univariate foundation models and multivariate financial signals for stock price prediction. J Comput Secur 2025;12(1):57–64. http://dx.doi.org/10.22108/jcs.2025.144738.1160.

[44] Dai X, Liu Y, Xia H, Hu Y, Dong Z, Yang J-M, Xu Q. Learning the context of errors: Black-box online adaptation of time series foundation models. 2026, http://dx.doi.org/10.48550/arXiv.2606.14222, arXiv, arXiv preprint arXiv:2606.14222, URL https://arxiv.org/abs/2606.14222.

[45] Gao W, Li J, Liu L, Le TD, Liu J, Zhao Y, Chen Y. Lightweight wrappers for adapting time series foundation models to regional drought forecasting. In: Proceedings of the 32nd ACM SIGKDD conference on knowledge discovery and data mining v.2. ACM; 2026, p. 10957–68. http://dx.doi.org/10.1145/3770855.3818835, URL https://dl.acm.org/doi/10.1145/3770855.3818835.

[46] Biswas A, Kamal M, Krambroeckers R, Elahi MML, Momen S, Mohammed N, Rahman S. One step closer to ground truth: A multi-scale residual-aware representation learning pipeline for predicting time series data. 2026, http://dx.doi.org/10.48550/arXiv.2606.10678, arXiv, arXiv preprint arXiv:2606.10678, URL https://arxiv.org/abs/2606.10678.

[47] Shchur O, Turkmen AC, Erickson N, Shen H, Shirkov A, Hu T, Wang B. AutoGluon-TimeSeries: automl for probabilistic time series forecasting. In: Faust A, Garnett R, White C, Hutter F, Gardner JR, editors. Proceedings of the second international conference on automated machine learning. Proceedings of machine learning research, vol. 224, PMLR; 2023, p. 9/1–21.

[48] Deutscher Wetterdienst (DWD). 10-Minute station observation data for munich (ID 03379) for the period 2023–2024. 2026, URL https://opendata.dwd.de/climate_environment/CDC/observations_germany/climate/10_minutes/air_temperature/historical/[dataset].

[49] Stadtwerke Flensburg. Einzigartiges fernwärmenetz: Anschlussquote von über 90 %. 2024, URL https://www.stadtwerke-flensburg.de/unternehmen/nachhaltigkeit/transformationsplan/flensburger-fernwaerme.

[50] Freißmann J, Fritz M, Tuschy I, Stadtwerke Flensburg GmbH. Network data of the district heating system for the city of Flensburg from 2020–2024. Zenodo; 2025, http://dx.doi.org/10.5281/zenodo.17177421, [dataset].

[51] Zippenfenig P. Open-meteo.com weather API. Zenodo; 2023, http://dx.doi.org/10.5281/zenodo.7970649, [software].

[52] Fuentes E, Arce L, Salom J. A review of domestic hot water consumption profiles for application in systems and buildings energy performance analysis. Renew Sustain Energy Rev 2018;81:1530–47. http://dx.doi.org/10.1016/j.rser.2017.05.229.

[53] Maltais L-G, Gosselin L. Predictability analysis of domestic hot water consumption with neural networks: From single units to large residential buildings. Energy 2021;229:120658. http://dx.doi.org/10.1016/j.energy.2021.120658.

[54] Xu B, Fu L, Di H. Field investigation on consumer behavior and hydraulic performance of a district heating system in Tianjin, China. Build Env 2009;44(2):249–59. http://dx.doi.org/10.1016/j.buildenv.2008.03.002.

[55] Kim S, Lim J, Byun J, Kim J, Kim H. Data-efficient electricity consumption forecasting with a tabular foundation model. In: Proceedings of the 2025 international conference on information and communication technology convergence. IEEE; 2025, p. 489–93. http://dx.doi.org/10.1109/ICTC66702.2025.11389094.

[56] Petković D, Protić M, Shamshirband S, Akib S, Raos M, Marković D. Evaluation of the most influential parameters of heat load in district heating systems. Energy Build 2015;104:264–74. http://dx.doi.org/10.1016/j.enbuild.2015.06.074.