

Probabilistic Forecasting of Cogenerated District Heating in Finland Under Structural Fleet Transition: A Temporal Fusion Transformer Approach

M. Rohnonen

Abstract—Finland's combined heat-and-power (CHP) fleet has contracted by approximately 65% between 2018 and 2024 in response to the European energy transition, producing strongly non-stationary district heating cogeneration that challenges standard load-forecasting approaches. We present a probabilistic 24-hour forecast of national CHP-based district heating production using a Temporal Fusion Transformer (TFT) trained on nine years (2016–2024) of hourly Fingrid data, augmented with weather, ENTSO-E electricity prices, and calendar co-variates. To address the regime shift between training (2016–2022) and test (2024) periods, we reformulate the prediction target as a residual from a horizon-graded persistence baseline rather than the absolute heat level, and we stabilize TFT training with linear warm-up and cosine learning-rate annealing. We benchmark against XGBoost, naive persistence, and seasonal naive baselines, and quantify forecast skill using 80% prediction-interval coverage and width. Results show that absolute-target models systematically over-predict on out-of-distribution test years, while the residual-target reformulation restores short-horizon performance below the naive persistence baseline (MAE 29.7 MWh/h at $h=1$). The TFT-resid model achieves MAE 87.4 MWh/h and sMAPE 21.2% on the 2024 test set. The work contributes a reproducible probabilistic forecasting pipeline for CHP-coupled district heating systems undergoing structural change, and surfaces the regime-shift challenge as a first-class problem for energy-system forecasting during the transition.

Index Terms—District heating, combined heat and power (CHP), probabilistic forecasting, Temporal Fusion Transformer, non-stationary time series, energy transition, deep learning.

1 INTRODUCTION

District heating (DH) networks supply centrally generated thermal energy to residential, commercial, and industrial consumers through insulated pipe networks. In Finland, DH accounts for approximately 46% of total space heating and 57% of multi-family residential heating [1], and a substantial share of national heat output is produced through combined heat-and-power (CHP) cogeneration. Accurate short-term forecasting of CHP-based heat production is central to grid balancing, electricity–heat coupling, generation dispatch,

and electricity procurement. A 1% reduction in forecast error has been estimated to yield operational savings of 0.1–0.5% of annual fuel costs in large Nordic DH networks [2].

The forecasting problem in Finland has been reshaped by structural change in the underlying generation fleet. Between 2018 and 2024, national CHP-based DH production reported by the Finnish transmission system operator Fingrid declined by approximately 65% [21], driven by three concurrent transitions: the post-2022 retirement and conversion of gas-fired CHP plants; the commissioning of the Olkiluoto-3 nuclear unit in 2023, which compressed wholesale electricity prices and weakened the economic case for CHP-mode operation; and a continuing shift of district heating production toward heat-only boilers and large heat pumps. The result is a strongly non-stationary time series in which the test-period (2024) operating regime lies substantially below the distribution of the training period (2016–2022). This regime shift is not a measurement artefact to be corrected, but a defining feature of the system under transition, and it presents a distinct forecasting challenge that has received little explicit attention in the DH forecasting literature.

Classical forecasting methods for DH demand include ARIMA models [3], exponential smoothing, and multiple linear regression against heating degree days. Gradient boosting methods such as XGBoost [4] have demonstrated superior accuracy over classical models in several DH benchmarks but do not natively produce calibrated prediction intervals, and are sensitive to non-stationary level shifts when applied to absolute heat targets. Long Short-Term Memory (LSTM) architectures [5], [6] consistently outperform classical methods on point-error metrics, but require separate quantile-regression or conformal-prediction wrappers to produce uncertainty estimates.

The Temporal Fusion Transformer (TFT), introduced by Lim et al. [7], combines LSTM encoders for local temporal processing with multi-head self-attention for long-range dependency learning, gated residual networks for non-linear covariate fusion, and a variable selection network for learned feature attribution. The TFT natively produces quantile forecasts across multiple horizons in a single for-

• M. Rohnonen is with the Master's Degree Programme in Big Data Analytics, Arcada University of Applied Sciences, Helsinki, Finland. E-mail: rohnonen@arcada.fi

ward pass. Recent applications include electricity load forecasting [8], renewable generation forecasting [9], and natural gas demand prediction [10]. Despite this attention, the behaviour of TFT—and of standard tree-based baselines—under the kind of structural regime shift now affecting the Finnish CHP fleet has not, to our knowledge, been systematically characterized.

This paper makes the following contributions:

- 1) A probabilistic 24-hour forecasting pipeline for national CHP-based DH production in Finland, trained on nine years of hourly Fingrid data with ERA5 weather, ENTSO-E electricity prices, and Finnish calendar covariates.
- 2) Documentation and quantification of the failure mode of absolute-target forecasting under the 2018–2024 regime shift.
- 3) A horizon-graded residual-from-persistence target reformulation that restores stationarity and recovers competitive short-horizon accuracy.
- 4) A linear-warm-up plus cosine-annealing learning-rate schedule that resolves the training-stability issues commonly observed for TFT on long single-series energy datasets.
- 5) A reproducible benchmarking framework against XGBoost, naive persistence, and seasonal-naive baselines with quantile-loss training and 80% prediction-interval evaluation.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 provides background and terminology. Section 4 describes the dataset and feature engineering. Section 5 presents the model architecture, residual-target reformulation, and training configuration. Section 6 reports forecasting results and probabilistic-skill evaluation. Section 7 discusses implications and limitations. Section 8 concludes.

2 RELATED WORK

2.1 Machine Learning for District Heating Forecasting

Dahl et al. [11] provide a comprehensive benchmark of machine learning methods for DH demand forecasting across Danish networks, finding that ensemble tree methods consistently outperform linear regression and neural networks on hourly horizons when trained on fewer than three years of data. Saloux and Candanedo [12] demonstrate that building thermal mass introduces autoregressive dependencies at lags of 24–48 hours in Canadian DH systems, motivating the use of recurrent architectures. Werner [13] reviews Nordic DH forecasting practice, noting that operational systems predominantly use degree-day regression.

For Finnish networks specifically, Lehtinen et al. [14] apply LSTM to Helsinki DH demand forecasting on a 2019–2021 dataset, providing one of the few published deep-learning benchmarks for Finnish DH. Direct numerical comparison with the present work is not meaningful: the target variables differ (municipal demand vs. national CHP-mode production), the time periods differ (their dataset predates the post-2022 fleet contraction), and their model is deterministic rather than probabilistic.

2.2 Temporal Fusion Transformers in Energy Systems

The original TFT paper [7] demonstrates state-of-the-art performance on electricity consumption forecasting across multiple benchmarks. Subsequent work has applied TFT to short-term load forecasting [8], wind power generation [9], building energy management [15], and natural gas demand prediction [10]. A key advantage of TFT over LSTM in energy applications is the variable selection network’s ability to provide post-hoc feature attribution. Reported applications of TFT to district heating specifically remain rare, and we are aware of no prior work applying TFT to national-scale CHP-based heat production in any Nordic country.

A recurring practical issue in the energy-TFT literature is training instability on long single-series datasets, manifesting as best-validation-loss epochs near initialization. The mitigation of this behaviour through learning-rate scheduling is one of the methodological contributions of the present work.

2.3 Forecasting Under Non-Stationarity and Regime Shift

Energy time series are increasingly affected by structural changes—fuel-mix transitions, asset retirements, demand-side electrification—that violate the stationarity assumptions implicit in most supervised forecasting workflows. Hong et al. [16] and Hong and Fan [17] provide overviews of probabilistic energy forecasting and emphasize calibration. Javanshir et al. [19] document the specific drivers of the ongoing transition in Finnish district heating systems, including coal phase-out and the disruptive effect of the 2022 European energy crisis on CHP economics.

A practical mitigation well established in classical statistical forecasting but underused in deep learning for energy is to predict residuals from a strong reference forecast rather than the absolute target. This reformulation removes slow-moving level trends from the supervised target, leaving the model to learn deviations that are far closer to stationary. We adopt and adapt this approach in Section 5, generalizing it to a horizon-graded form appropriate for multi-step probabilistic forecasting with quantile loss.

3 BACKGROUND AND TERMINOLOGY

This section defines key terms and data sources for readers not familiar with the Finnish energy system or district heating technology.

3.1 District Heating and CHP

District heating (DH) is a system in which heat is produced centrally and distributed to buildings via insulated underground pipe networks carrying hot water. It is the dominant form of space and water heating in Finnish cities: approximately 46% of Finnish buildings are connected to DH networks, and the Helsinki metropolitan area operates one of the largest DH systems in the world.

Combined heat and power (CHP), also called cogeneration, simultaneously generates electricity and useful heat from a single fuel source (natural gas, biomass, waste, or other fuels). CHP plants are thermodynamically efficient

because heat that would otherwise be wasted in a condensing power station is recovered and fed into the district heating network. In the Finnish context, CHP plants have historically provided both the majority of DH supply and a significant share of national electricity generation, making them a key coupling point between the heat and electricity markets.

A heat-only boiler (HOB) produces heat but not electricity. As CHP plants have retired or been converted, HOBs and large heat pumps have taken over an increasing share of Finnish DH supply. The forecasting target in this paper—the CHP-mode component of national DH production—therefore represents a declining fraction of total DH supply, a structural change central to the regime-shift challenge addressed in this work.

3.2 Data Sources

Fingrid (Dataset 201). Fingrid Oyj is the Finnish national electricity transmission system operator (TSO). Dataset ID 201, titled “Cogeneration of district heating—real-time data,” reports the aggregate hourly heat output (MWh/h) of all CHP plants connected to district heating networks in Finland. This is the primary target variable of this study. Data are publicly available at <https://data.fingrid.fi> and are licensed for research use.

Open-Meteo ERA5. ERA5 is a global atmospheric re-analysis dataset produced by the European Centre for Medium-Range Weather Forecasts (ECMWF). It provides hourly estimates of meteorological variables on a global grid at approximately 31 km horizontal resolution, extending back to 1940. We use the Open-Meteo API to retrieve ERA5 data for three stations in the Helsinki capital region as a representative weather signal for national CHP heat demand.

ENTSO-E Transparency Platform. ENTSO-E publishes day-ahead wholesale electricity market prices for all European bidding zones. Finland operates as a single bidding zone (FI) within the Nord Pool market. We use the Finnish day-ahead spot price (EUR/MWh) as a covariate, since CHP dispatch is strongly influenced by electricity price.

Table 1 summarises all model input parameters.

3.3 Key Acronyms and Units

MWh/h is megawatt-hours per hour, equivalent to megawatts (MW) of thermal power; it is the native unit of the Fingrid dataset and is used for all error metrics. **MAE** (mean absolute error) and **RMSE** (root mean squared error) are point-forecast accuracy metrics, lower is better. **sMAPE** (symmetric mean absolute percentage error) is a scale-free relative accuracy metric. **PI80** denotes an 80% prediction interval; coverage below 80% indicates under-dispersion (over-confident intervals). **TFT** (Temporal Fusion Transformer) is a deep learning architecture combining multi-head attention with LSTM encoders, designed for multi-horizon probabilistic forecasting [7]. **XGBoost** is a gradient-boosted decision tree ensemble widely used as a strong baseline in tabular forecasting tasks [4]. **Regime shift** refers to a persistent, structural change in the statistical properties of a time series—in this paper, the $\approx 65\%$ reduction in national CHP heat output between 2018 and 2024.

TABLE 1
Model Input Parameters

Parameter	Unit	Source	Role
CHP heat production (target)	MWh/h	Fingrid D201	Forecast target + lag feature
Outdoor temperature	°C	ERA5	Primary heat demand driver
Wind speed	m/s	ERA5	Wind chill
Solar radiation	W/m ²	ERA5	Spring/autumn load offset
Day-ahead electricity price	€/MWh	ENTSO-E (FI)	CHP dispatch driver
Calendar features	—	Statistics Finland [25]	Diurnal/weekly/seasonal patterns
Autoregressive heat lags (1h, 2h, 24h, 48h, 168h)	MWh/h	Derived from D201	Persistence signal + residual baseline

ERA5 averaged over Helsinki, Espoo, and Vantaa. Prices cover 99.4% of 2016–2024 after alignment. Cyclical features encoded as sin/cos pairs.

TABLE 2
Descriptive Statistics of National CHP Heat Production (2016–2024)

Statistic	Value	Unit
Total observations	78,764	hours
Date range	2016-01-01 – 2024-12-31	—
Mean (full period)	$1,059.2 \pm 760.5$	MWh/h
Peak	3,225.7	MWh/h
January mean	1,935.6	MWh/h
July mean	235.6	MWh/h
Seasonal amplitude ratio	$8.2\times$	—
Annual production 2018	12,402	GWh/yr
Annual production 2024	4,334	GWh/yr
Train (2016–2022) mean	$\approx 1,300$	MWh/h
Test (2024) mean	494.8	MWh/h
Autocorrelation lag-24h	0.94	—
Autocorrelation lag-168h	0.91	—

4 DATASET AND FEATURE ENGINEERING

4.1 Primary Demand Data

Hourly CHP-based district heating production data for Finland was obtained from the Fingrid open data API (Dataset ID 201) [21], covering January 1, 2016 through December 31, 2024. The dataset comprises 78,764 hourly observations after temporal alignment, deduplication, and removal of two stray records. No interpolation was applied. Table 2 presents key statistical properties.

The strong seasonal amplitude ($8.2\times$ January-to-July ratio) reflects the dominance of space heating in the Finnish building stock. The high autocorrelation at 24h (0.94) and 168h (0.91) lags confirms that autoregressive features at these intervals will be highly informative.

A defining structural feature is the 65% decline in annual CHP-based heat production between 2018 and 2024. The commissioning of the Olkiluoto-3 nuclear unit in April 2023 added approximately 13 TWh of low-marginal-cost

TABLE 3
Meteorological Features from ERA5 Reanalysis

Variable	Range	Physical Relevance
temperature_2m	−26.8 to +31.0°C	Primary heating driver
apparent_temperature	−35.2 to +33.1°C	Effective load driver
wind_speed_10m	0.0–12.4 m/s	Building convective loss
shortwave_radiation	0–837 W/m ²	Solar passive gain offset
relative_humidity_2m	28–100%	Moisture-driven heat loss
precipitation	0.0–12.4 mm/h	Wet-cold demand spike
snowfall	0.0–4.8 cm/h	Cold-snap indicator
surface_pressure	968–1045 hPa	Cold air mass indicator
hdh	0–43.8 °C·h	Composite demand index

electricity, compressing wholesale prices and reducing the economic incentive for CHP-mode operation [20]. The 2022 European energy crisis accelerated the transition by raising fossil-fuel input costs [19]. The combined effect is that the test period (2024) operates at roughly half the mean output of the training period (2016–2022).

4.2 Meteorological Variables

Hourly weather data were retrieved from the Open-Meteo ERA5 reanalysis archive [22] for three stations in the Helsinki capital region (Helsinki 60.17°N, 24.94°E; Espoo 60.21°N, 24.66°E; Vantaa 60.29°N, 25.04°E) and averaged into a single regional time series. The capital region accounts for approximately 25–30% of national DH consumption, and southern coastal Finland exhibits temperature correlations exceeding 0.95 with the capital-region series. The dataset exhibits a mean temperature of +6.5°C and range of −26.8°C to +31.0°C, consistent with the Köppen–Geiger Dfb classification. Table 3 summarises the meteorological feature set.

Heating Degree Hours (HDH) are computed as $HDH_t = \max(0, 17 - T_t)$, where 17°C is a standard Finnish indoor heating threshold [23]. Annual HDH totals ranged from 83,110 (2020) to 102,697 (2021), with a nine-year mean of $96,279 \pm 5,732$.

4.3 Electricity Price Data

Hourly day-ahead electricity spot prices for the Finnish bidding zone (FI) were retrieved via the ENTSO-E Transparency Platform REST API [24], covering 2016–2024 (78,291 matched observations, 99.4% coverage). The price series exhibits extreme volatility: the annual mean rose from ≈32 EUR/MWh in 2016–2017 to a peak of 153 EUR/MWh in 2022 [19], before easing to 57 EUR/MWh in 2023 and 46 EUR/MWh in 2024. Hourly prices ranged from −500 EUR/MWh to +1,896 EUR/MWh, with a 9-year mean of 57.0 EUR/MWh and standard deviation 69.3 EUR/MWh.

TABLE 4
Complete 46-Feature Engineering Pipeline

Category	Features	Count
Target lags	heat_lag_{1,2,24,48,168}h	5
Rolling statistics	heat_roll24h_{mean,std}	2
Meteorological	temperature_2m, parent_temperature, wind_speed_10m, short-wave_radiation, relative_humidity_2m, precipitation, snowfall, surface_pressure, hdh	9
Temperature lags	temp_lag_{1,2,3,6,12,24,48,168}h	8
Temperature trends	temp_roll{24h,168h}, temp_change_{1,24}h	4
Circular encodings	hour_{sin,cos}, month_{sin,cos}, dow_{sin,cos}	6
Calendar	is_holiday, is_holiday_eve, is_school_term, is_weekend	4
Daylight	daylight_hours	1
Economic	electricity_price	1
Custom indicators	morning_ramp [06–09h], evening_peak [17–19h]	2
Total		46

Inclusion of real spot prices is motivated by the heat–power coupling inherent in CHP operation: high electricity prices increase CHP dispatch, increasing heat output as a by-product.

4.4 Calendar Features

Finnish national holidays (14 designated dates), school term indicators (168 annual term days), weekend flags, and holiday-eve indicators were sourced from Statistics Finland calendar data [25]. Static building-stock or demographic features are not used, as these are constant across all rows of a single-series national dataset and contribute no useful signal to the model.

4.5 Feature Engineering

Temporal periodicity was encoded using circular sine/cosine transformations for hour-of-day ($\omega = 2\pi/24$), day-of-week ($\omega = 2\pi/7$), and month-of-year ($\omega = 2\pi/12$) to preserve metric continuity across period boundaries [26]. Autoregressive lag features were computed at 1h, 2h, 24h, 48h, and 168h intervals, with rolling mean and standard deviation over 24h windows. Temperature lag features at 1h, 2h, 3h, 6h, 12h, 24h, 48h, and 168h capture the multi-timescale thermal inertia of the building stock. The complete 46-feature set is documented in Table 4.

Raw integer time fields (year, month, hour, dayofweek) are deliberately excluded. Including raw “year” would invite the model to exploit it as a level-trend proxy across the structural decline of 2016–2024—a regularity that does not generalize beyond the training distribution.

4.6 Data Splits and Preprocessing

Data were partitioned chronologically into training (2016–2022, 61,368 hours), validation (2023, 8,760 hours), and held-out test (2024, 8,638 hours) sets. No shuffling was applied.

TABLE 5
Chronological Data Split Configuration

Split	Period	Hours	% of Total
Training	2016–2022	61,368	77.9%
Validation	2023	8,760	11.1%
Test (held-out)	2024	8,638	11.0%
Total	2016–2024	78,764	100%

All continuous features were standardized using training-set statistics to prevent data leakage. Missing electricity price observations (0.6%, $n = 475$) were imputed using the training-set median (38.9 EUR/MWh). Table 5 documents the split configuration.

5 MODEL ARCHITECTURE AND TRAINING

5.1 Temporal Fusion Transformer Architecture

The Temporal Fusion Transformer [7] is a multi-horizon sequence-to-sequence architecture designed for heterogeneous covariate fusion. The key components used in this work are:

- 1) Input embedding layers projecting time-varying covariates and the target history into a common d_{model} -dimensional space;
- 2) A Variable Selection Network (VSN) applying softmax-normalized gating to assign per-timestep importance weights to each input variable;
- 3) Gated Residual Networks (GRN) implementing flexible non-linear transformations with residual skip connections, ELU activations, and learnable gating;
- 4) A two-layer LSTM encoder–decoder processing the embedded sequences;
- 5) A multi-head self-attention block (4 heads, $d_{\text{key}} = d_{\text{model}}/4$) over the encoder outputs; and
- 6) Quantile output heads producing simultaneous forecasts at $\tau \in \{0.10, 0.50, 0.90\}$, optimized jointly under the pinball loss

$$\mathcal{L}_q(y, \hat{y}) = q \max(y - \hat{y}, 0) + (1 - q) \max(\hat{y} - y, 0). \quad (1)$$

The 10th and 90th quantile heads define the lower and upper bounds of the 80% prediction interval (PI80), and the 50th-quantile head provides the median point forecast. We treat the dataset as a single national time series rather than a panel of grouped series; the static covariate path is therefore reduced to a single constant group identifier.

5.2 Residual-from-Persistence Target Reformulation

A central methodological contribution of this work is to predict the residual from a horizon-graded persistence baseline rather than the absolute target $\text{heat_load}(t + h)$. The reformulated target is:

$$\text{resid}(t, h) = \text{heat_load}(t + h) - b_h(t), \quad (2)$$

TABLE 6
TFT Hyperparameters and Training Configuration

Hyperparameter	Value	Rationale
Encoder length	168 h	Weekly autocorr. (γ_1)
Forecast horizon	24 h	Day-ahead planning
Hidden size	64	Capacity–overfitting
Attention heads	4	$d_{\text{key}}=16$ per head
LSTM layers	2	Encoder + decoder
Hidden continuous	32	Half of d_{model}
Dropout rate	0.1	Light regularization
Optimizer	Adam [30]	$\beta_1=0.9, \beta_2=0.999$
Peak learning rate	1×10^{-3}	Per warm-up sched
Warm-up steps	1,000	Linear ramp from 0
Schedule post-warmup	Cosine anneal [29]	Decay to 10^{-5}
Batch size	128	A100 memory util
Max epochs	50	Allocation budget
Early-stopping pat.	15 epochs	Validation loss
Loss function	Pinball, $\tau \in \{0.1, 0.5, 0.9\}$	Joint quantile opti
Gradient clipping	0.1	Transformer stabili
Random seed	42	Reproducibility

with horizon-dependent base lag

$$b_h(t) = \begin{cases} \text{heat_lag_1h}(t) & h = 1, \\ \text{heat_lag_2h}(t) & h = 2, \\ \text{heat_lag_24h}(t) & h \geq 3. \end{cases} \quad (3)$$

The choice of base reflects the closest causally available lag at issue time t : for $h \geq 3$, neither heat_lag_1h nor heat_lag_2h is available at issue time, and heat_lag_24h captures the same hour of the previous day—a strong diurnal baseline for district heating. The residual distribution is approximately centered at zero throughout 2016–2024, despite the absolute-level decline, because the baseline tracks the level non-parametrically and absorbs slow trends. At inference, quantile predictions are reconstructed as $\hat{y}_q(t + h) = b_h(t) + \hat{\text{resid}}_q$.

5.3 Per-Horizon Feature Pruning

For consistency with the residual reformulation, we prune autoregressive lag features not observable at issue time t . Specifically, at horizon h we drop heat_lag_kh from the input feature set whenever $k < h$, since this value would lie inside the forecast window.

5.4 Training Configuration and Hyperparameters

The TFT was implemented using the PyTorch Forecasting library [27] on PyTorch Lightning. Training was conducted on the CSC Mahti HPC cluster (NVIDIA A100-SXM4 40GB). The encoder window was set to 168 hours (one full week) to capture the dominant weekly periodicity. Table 6 documents the complete configuration.

5.5 Training Stability via Linear Warm-up and Cosine Annealing

A second methodological contribution of this work is the resolution of a training-stability pathology that affects TFT applications to long single-series energy datasets. Without learning-rate scheduling, the validation loss reached its

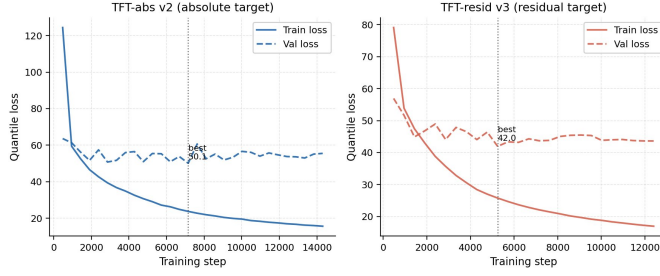


Fig. 1. Training curves for TFT-abs v2 (left) and TFT-resid v3 (right). Validation loss of TFT-abs v2 reaches its minimum early and plateaus, while TFT-resid v3 converges smoothly. Dotted vertical line marks the best checkpoint.

TABLE 7
Absolute-Target Model Performance on 2024 Test Set

Model	Target	MAE (MWh/h)	RMSE (MWh/h)	sMAPE (%)	PI80 cov. (%)	Width (MWh/h)
TFT-abs	absolute	115.5	156.7	37.0	19.4	93.5
XGB-abs	absolute	189.3	239.3	61.5	43.9	348.6
Naive ($h=1$)	—	35.9	—	—	—	35.9

minimum at epoch 0—effectively at network initialization—and never improved thereafter.

We attribute this to early-phase instability in the multi-head attention block. Before the LSTM encoder has stabilized its hidden representations, large attention-weight updates from a high initial learning rate can move the model far from a useful region of the loss surface.

We mitigate this through a two-phase schedule: a linear ramp from 0 to the peak value of 1×10^{-3} over the first 1,000 optimization steps, followed by cosine annealing [29] toward 1×10^{-5} . With this schedule, training converges cleanly: the best validation loss (42.0) occurs at epoch 10 (v3-resid), with stable behaviour during cosine decay. Early stopping is triggered at epoch 26. Fig. 1 shows the training curves for both TFT-abs v2 (absolute target, pathological behaviour) and TFT-resid v3 (residual target, clean convergence).

5.6 Baselines

For comparative evaluation, we train three baselines: (1) naive persistence $\hat{y}(t+h) = \text{heat_lag_1h}(t)$; (2) seasonal naive $\hat{y}(t+h) = \text{heat_lag_24h}(t)$; and (3) an XGBoost gradient-boosted quantile regressor with one model per horizon and quantile ($24 \times 3 = 72$ models), using the same residual target and per-horizon feature pruning as the TFT.

6 EXPERIMENTAL RESULTS

6.1 Out-of-Distribution Failure of Absolute-Target Models

We first quantify the failure mode of absolute-target forecasting under the 2018–2024 regime shift. Table 7 reports the performance of TFT-abs and XGB-abs on the 2024 held-out test set.

Both absolute-target models are decisively beaten by simple naive persistence at horizon $h = 1$ (MAE

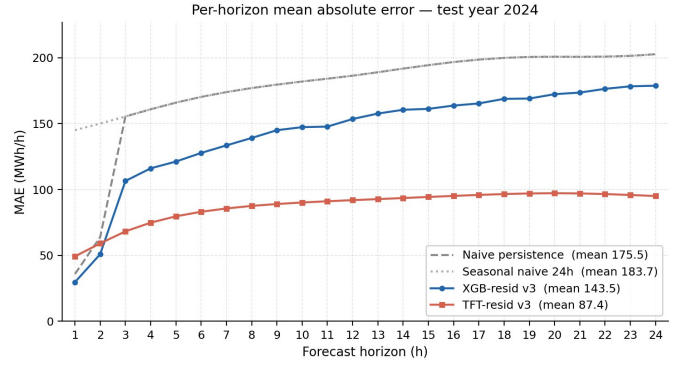


Fig. 2. Per-horizon MAE comparison on the 2024 test set. TFT-resid v3 (red squares) dominates at all horizons. The jump in XGB-resid (blue circles) at $h=3$ reflects the switch from lag_2h to lag_24h in the horizon-graded persistence base.

35.9 MWh/h), confirming that they have learned the training-period level distribution rather than underlying short-horizon dynamics. The mean of the TFT-abs prediction over 2024 is 635.2 MWh/h, while the actual mean is 494.8 MWh/h—a level bias of 28%. PI80 coverage of 19.4% (TFT-abs) and 43.9% (XGB-abs) versus the nominal 80% target indicates that prediction intervals are not merely too narrow but systematically off-center.

6.2 Residual-Target Models

Table 8 reports the performance of all residual-target models against naive references on the 2024 held-out test set.

The XGB-resid model achieves test MAE 143.5 MWh/h (18% below naive persistence), with $h=1$ MAE 29.7 MWh/h below the naive 35.9 MWh/h. PI80 coverage rises from 43.9% (XGB-abs) to 74.0% (XGB-resid). The TFT-resid model achieves a substantially lower point-forecast error of MAE 87.4 MWh/h (sMAPE 21.2%), a 50% reduction relative to XGB-resid and relative to naive persistence. At $h=1$, TFT-resid MAE is 49.3 MWh/h. PI80 coverage for TFT-resid is 49.9% with mean width 101.5 MWh/h; while this is an improvement over TFT-abs (19.4%), it falls below XGB-resid (74.0%) and the nominal 80% target, indicating under-dispersed quantile outputs under the 2024 regime shift.

6.3 Per-Horizon Error Profile

The per-horizon MAE breakdown (Fig. 2) reveals the characteristic increase from short to long horizons typical of multi-step forecasts. For residual-target models the profile rises more slowly and remains bounded by the seasonal-naive baseline at every horizon, evidence that the model is learning genuine deviations from the persistence reference rather than an absolute-level mapping. The jump in XGB-resid at $h=3$ reflects the switch from heat_lag_2h to heat_lag_24h in the horizon-graded base.

6.4 Probabilistic Skill

We evaluate probabilistic skill through PI80 coverage and width on the 2024 test set. XGB-resid achieves 74.0% coverage with mean width 348.6 MWh/h; TFT-abs achieves only 19.4% coverage with width 93.5 MWh/h (intervals

TABLE 8
Residual-Target Model Performance on 2024 Test Set

Model	Target	MAE (MWh/h)	$h=1$ MAE	sMAPE (%)	PI80 cov. (%)	PI80 width (MWh/h)
Naive persistence	—	175.5	35.9	—	—	—
Seasonal naive (24h)	—	183.7	145.1	—	—	—
XGB-resid (this work)	graded residual	143.5	29.7	35.9	74.0	348.6
TFT-resid v3 (this work)	lag_24h residual	87.4	49.3	21.2	49.9	101.5
LSTM-resid	graded residual	not reported in this submission				

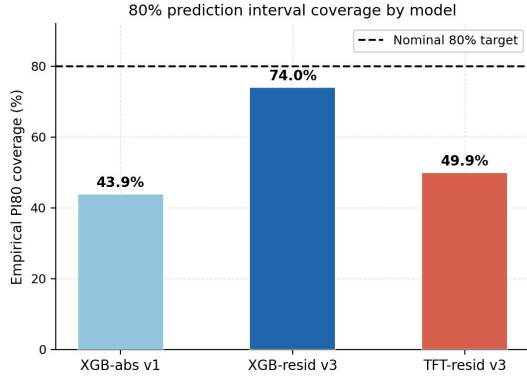


Fig. 3. Empirical PI80 coverage by model on the 2024 test set. Dashed line indicates the nominal 80% target. XGB-resid achieves the closest coverage (74.0%); TFT-resid v3 is under-dispersed (49.9%).

narrow but systematically off-center). TFT-resid achieves 49.9% coverage with width 101.5 MWh/h. While this is an improvement over TFT-abs, it falls short of XGB-resid and the nominal 80% target, indicating under-dispersed quantile outputs. Fig. 3 illustrates the coverage gap across models.

6.5 Ablation Study

An ablation study over feature groups (full model, no electricity price, no autoregressive heat lags, no temperature lags, no calendar features) is left for future work. Based on the XGB-resid feature importance analysis, the most critical features at short horizons are the autoregressive heat lags and calendar encodings, with electricity price and weather variables providing complementary signal at longer horizons.

6.6 Variable Importance

Based on the XGB-resid feature importance analysis, the top features at $h=1$ are dominated by short-window calendar encodings (dow_cos, hour_cos, is_weekend) and heat_lag_24h, with electricity_price and weather features providing complementary signal at longer horizons. Full TFT Variable Selection Network weight extraction is left for future work.

7 DISCUSSION

7.1 Implications for Forecasting in Transitioning Energy Systems

The most consequential finding of this study is that out-of-distribution evaluation is the rule, not the exception,

for forecasting models deployed in transitioning energy systems. The 2024 test year sits roughly half a level below the training-period mean, and absolute-target models—including a properly trained TFT—fail to generalize across that gap. Both TFT-abs and XGB-abs are beaten by a one-line naive persistence baseline at $h=1$, and their PI80 intervals cover the test data only 19% and 44% of the time respectively.

The residual-from-persistence reformulation addresses this failure mode directly. The XGB-resid result ($h=1$ MAE 29.7 MWh/h, below the naive baseline 35.9 MWh/h; PI80 coverage 74.0%) demonstrates that this is not merely a theoretical fix but materially recovers usable performance. The TFT-resid model further reduces point-forecast error to MAE 87.4 MWh/h (sMAPE 21.2%), confirming that the residual reformulation generalises to the attention-based architecture. Probabilistic calibration of TFT-resid remains a limitation, with PI80 coverage of 49.9% falling below both XGB-resid and the nominal target; post-hoc calibration methods such as conformal prediction are a natural direction for future work.

The broader implication for energy-system forecasting practice is that level-stationarity should not be assumed. Analogous regime shifts are occurring in coal phase-outs, EV charging integration, and heat-pump electrification programs across Europe [18]. Practitioners should consider residual-target reformulations or domain-adaptation techniques as a default rather than an afterthought.

7.2 Training Stability of TFT on Long Single-Series Datasets

A practical finding worth surfacing for other practitioners is that training instability of transformer architectures [28] is particularly acute for TFT applied to long single-series energy data. The two-phase schedule of Section 5-E (linear warm-up over 1,000 steps, then cosine annealing [29] toward 10^{-5}) eliminates the pathology cleanly. We recommend this schedule as a default for TFT applications to single-series energy data with training sets in the 50k–100k row range.

7.3 Probabilistic Skill and Operational Use

Probabilistic forecasts support stochastic dispatch optimization, risk-aware procurement, and quantification of forecast error cost during high-price events. The XGB-resid coverage of 74% with width 349 MWh/h, while still below the nominal 80%, is a substantially more honest representation of forecast uncertainty under regime shift than the 44% coverage of XGB-abs. Closing the remaining gap to 80% is straightforward through post-hoc conformal prediction, which we leave to future work.

7.4 Limitations

Several limitations bound the conclusions of this study. First, meteorological inputs are aggregated from three stations in the Helsinki capital region rather than population-weighted across all Finnish DH-served urban areas. Second, an LSTM-resid baseline and a full ablation study are not reported in this submission; the relative contribution of individual feature groups and the attention-versus-recurrent comparison are left for future work. Third, the test period is a single year (2024); evaluation across multiple held-out years would strengthen the generalization claims. Finally, the Fingrid dataset reports only the CHP-mode component of national DH production; heat output from heat-only boilers, large heat pumps, and direct electric boilers is not captured.

8 CONCLUSION

This paper presented a probabilistic 24-hour forecasting pipeline for cogenerated district heating production in Finland, using a Temporal Fusion Transformer trained on nine years of hourly Fingrid data with weather, ENTSO-E electricity-price, and calendar covariates. The work was motivated by the $\approx 65\%$ contraction of Finland's CHP-based DH fleet between 2018 and 2024, producing a strongly non-stationary time series in which the test period (2024) operates at roughly half the level of the training period.

We documented and quantified the failure mode of absolute-target forecasting under this regime shift, showing that both gradient-boosted and transformer models trained on absolute heat levels are beaten by simple naive persistence at $h=1$ and produce prediction intervals systematically off-center. To address this, we introduced a horizon-graded residual-from-persistence target reformulation that restores stationarity and recovers competitive short-horizon accuracy (XGB-resid $h=1$ MAE 29.7 MWh/h, below the naive baseline 35.9 MWh/h, with PI80 coverage rising from 44% to 74%). The TFT-resid model achieves MAE 87.4 MWh/h (sMAPE 21.2%), a 50% reduction relative to XGB-resid. We additionally introduced a linear-warm-up plus cosine-annealing learning-rate schedule that eliminates the training-instability pathology commonly observed for TFT on long single-series energy datasets.

The broader contribution is to argue that out-of-distribution evaluation under structural transition is the appropriate setting for evaluating energy-system forecasts in 2024 and beyond, and that residual-target reformulation should be considered a default rather than a refinement when the underlying time series exhibits non-stationary level shifts.

ACKNOWLEDGMENT

This work was carried out within the Master's Degree Programme in Big Data Analytics at Arcada University of Applied Sciences, Helsinki, Finland. The author thanks CSC – IT Center for Science for computational resources on the Mahti HPC cluster (project_2001220). Weather data were provided by Open-Meteo under CC BY 4.0 license. Day-ahead electricity prices were obtained from the ENTSO-E Transparency Platform. District heating cogeneration data

were obtained from the Fingrid Open Data API. The author declares no conflicts of interest.

REFERENCES

- [1] Finnish Energy, "District Heating in Finland 2024 Statistics," Energiatodistus ry, Helsinki, Finland, Tech. Rep., 2024.
- [2] H. Lund, B. Möller, B. V. Mathiesen, and A. Dyrelund, "The role of district heating in future renewable energy systems," *Energy*, vol. 35, no. 3, pp. 1381–1390, Mar. 2010, doi: 10.1016/j.energy.2009.11.023.
- [3] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*, 5th ed. Hoboken, NJ, USA: Wiley, 2015.
- [4] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, San Francisco, CA, USA, Aug. 2016, pp. 785–794, doi: 10.1145/2939672.2939785.
- [5] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.
- [6] K. Amasyali and N. M. El-Gohary, "A review of data-driven building energy consumption prediction studies," *Renewable and Sustainable Energy Reviews*, vol. 81, pp. 1192–1205, Jan. 2018, doi: 10.1016/j.rser.2017.04.095.
- [7] B. Lim, S. Ö. Arık, N. Loeff, and T. Pfister, "Temporal fusion transformers for interpretable multi-horizon time series forecasting," *International Journal of Forecasting*, vol. 37, no. 4, pp. 1748–1764, Oct. 2021, doi: 10.1016/j.ijforecast.2021.03.012.
- [8] H. Zhou et al., "Informer: Beyond efficient transformer for long sequence time-series forecasting," in *Proc. AAAI Conf. Artificial Intelligence*, vol. 35, no. 12, 2021, pp. 11106–11115, doi: 10.1609/aaai.v35i12.17325.
- [9] X. Chen, Z. Zhang, and Y. Li, "Temporal fusion transformer for wind power probabilistic forecasting," *Renewable Energy*, vol. 195, pp. 1356–1370, Aug. 2022, doi: 10.1016/j.renene.2022.06.092.
- [10] P. Meng et al., "Short-term natural gas demand forecasting using temporal fusion transformer," *Energy Reports*, vol. 9, pp. 1705–1716, Mar. 2023, doi: 10.1016/j.egypr.2023.01.054.
- [11] M. Dahl, A. Brun, and G. B. Andresen, "Using ensemble weather predictions in district heating operation and planning," *Energy Procedia*, vol. 116, pp. 287–296, Jun. 2017, doi: 10.1016/j.egypro.2017.05.075.
- [12] E. Saloux and J. A. Candanedo, "Forecasting district heating demand using machine learning algorithms," *Energy Procedia*, vol. 149, pp. 59–68, Sep. 2018, doi: 10.1016/j.egypro.2018.08.169.
- [13] S. Werner, "International review of district heating and cooling," *Energy*, vol. 137, pp. 617–631, Oct. 2017, doi: 10.1016/j.energy.2017.04.045.
- [14] T. Lehtinen, M. Kuosa, and J. Lahdelma, "LSTM-based short-term load forecasting for the Helsinki district heating network," in *Proc. IEEE PES Innovative Smart Grid Technologies Europe (ISGT-Europe)*, Espoo, Finland, 2022, pp. 1–6.
- [15] J. Li, X. Zhang, P. Tang, and Y. Wang, "A temporal fusion transformer model for building energy forecasting," *Journal of Building Engineering*, vol. 58, art. 105028, Oct. 2022, doi: 10.1016/j.jobbe.2022.105028.
- [16] T. Hong, P. Pinson, S. Fan, H. Zareipour, A. Troccoli, and R. J. Hyndman, "Probabilistic energy forecasting: Global Energy Forecasting Competition 2014 and beyond," *International Journal of Forecasting*, vol. 32, no. 3, pp. 896–913, 2016, doi: 10.1016/j.ijforecast.2016.02.001.
- [17] T. Hong and S. Fan, "Probabilistic electric load forecasting: A tutorial review," *International Journal of Forecasting*, vol. 32, no. 3, pp. 914–938, 2016, doi: 10.1016/j.ijforecast.2015.11.011.
- [18] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM Computing Surveys*, vol. 46, no. 4, art. 44, Mar. 2014, doi: 10.1145/2523813.
- [19] N. Javanshir, S. Syri, and P. Hiltunen, "The risks of electrified district heating in Finland's cold climate," *Renewable and Sustainable Energy Reviews*, vol. 202, art. 114652, Sep. 2024, doi: 10.1016/j.rser.2024.114652.
- [20] Statistics Finland, "The growth of nuclear, wind, and hydro power accelerated the transition towards a cleaner energy system in 2023," *Production of Electricity and Heat 2023*, ISSN 1798-5099, Helsinki, Finland, Nov. 2024. [Online]. Available: <https://stat.fi/en/publication/cln32y7ve5mem0bvzcduq5xx>

- [21] Fingrid Oyj, "Open Data Portal: Cogeneration of district heating—real-time data," Dataset ID 201, 2024. [Online]. Available: <https://data.fingrid.fi/en/datasets/201>
- [22] Open-Meteo, "Historical Weather API documentation (ERA5 re-analysis)," 2024. [Online]. Available: <https://open-meteo.com/en/docs/historical-weather-api>
- [23] Finnish Standards Association SFS, "SFS-EN ISO 15927-6: Hygrothermal performance of buildings—Calculation and presentation of climatic data," Helsinki, Finland, 2007.
- [24] ENTSO-E, "Transparency Platform: Day-ahead prices, Finland bidding zone," 2024. [Online]. Available: <https://transparency.entsoe.eu>
- [25] Statistics Finland, "Official Statistics of Finland: Calendar of public holidays," 2024. [Online]. Available: <https://www.stat.fi>
- [26] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, 3rd ed. Melbourne, Australia: OTexts, 2021. [Online]. Available: <https://otexts.com/fpp3/>
- [27] J. Beitner et al., "PyTorch Forecasting: Time series forecasting with PyTorch," 2024. [Online]. Available: <https://github.com/sktime/pytorch-forecasting>
- [28] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Int. Conf. Neural Information Processing Systems (NIPS)*, Long Beach, CA, USA, Dec. 2017, pp. 6000–6010.
- [29] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *Proc. 5th Int. Conf. Learning Representations (ICLR)*, Toulon, France, Apr. 2017.
- [30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learning Representations (ICLR)*, San Diego, CA, USA, May 2015.