

MoE-TFT: Mixture-of-Experts Enhanced Temporal Fusion Transformer for Time Series Forecasting

Saifullah Burero

Free University of Bozen-Bolzano

Joint work with: Anton Dignös · Michele Linardi · Johann Gamper

DAWAK 2026 · DEXA

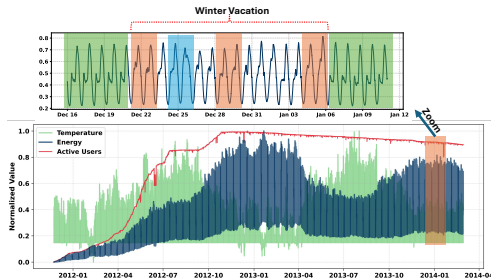
Time Series Data

- ▶ **Time Series Data:** A sequence of observations of an underlying phenomenon (e.g., electrical consumption, weather conditions) recorded over regular or irregular time intervals.
- ▶ **Applications of time series data**
 - ▶ **Forecasting (our focus):** electricity demand forecasting, predictive maintenance
 - ▶ **Anomaly detection:** detecting unexpected appliance consumption (e.g., equipment failure)
 - ▶ **Classification:** classifying fixed vs. dynamic tariff users
- ▶ **Complexity:** Analyzing time series data without contextual information is challenging, as external factors can cause simultaneous shifts in consumption patterns.
 - ▶ **Example:** When a large group of active consumers suddenly switch off their HVAC units during a holiday or vacation period, the total grid consumption drops instantly, deviating completely from historical temporal trends.

Why Context Matters in Time Series

Real-world phenomena (e.g., electrical consumption) are affected by **external factors**, modelling them improves both **forecast accuracy** and **interpretability**.

- **Seasonality**: winter heating increases demand
- **Time of day**: morning/evening peaks, night lows
- **Holidays & Vacations**: different usage patterns during mid-day than weekdays & school breaks
- **Weather**: cold days increase heating demand
- **Consumers**: new users change consumption trends



Context is heterogeneous

The context may come from diverse sources (e.g., sensors, operation logs) that lacks homogeneity. Integrating them with time series data requires careful handling.

Context Is Heterogeneous in Its Time Dimension

We distinguish four time dimensions present in most cases.^{1 2}

Static

Time-invariant properties
e.g., consumer segment, building type

Time Series

A value at every time step
e.g., temperature, humidity

Events

Instantaneous occurrences
e.g., public holidays

Intervals

Spans with a start and end
e.g., school vacations

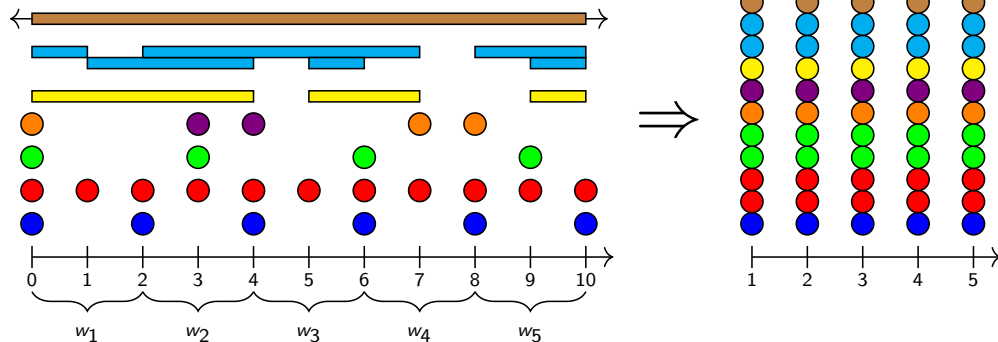
Requirement: ML/DL models require **homogeneous (multivariate)** input data.

¹Burero et al., *Blending Contextual Data with Heterogeneous Time Dimensions for Improved Time Series Analysis* (2026).

²Burero, *Integrating Heterogeneous Contextual Data for Enhanced Time Series Analysis* (2025).

Context Is Heterogeneous in Its Time Dimension

Creating multivariate time series from context



Approach [1]: Given granularity w :

- 1 Align using aggregation / imputation.
- 2 Compute lag information based on the time dimension.
- 3 Transformation introduce **multivariate heterogeneity**.
- 4 Modeling resulting time series requires specialized architecture.

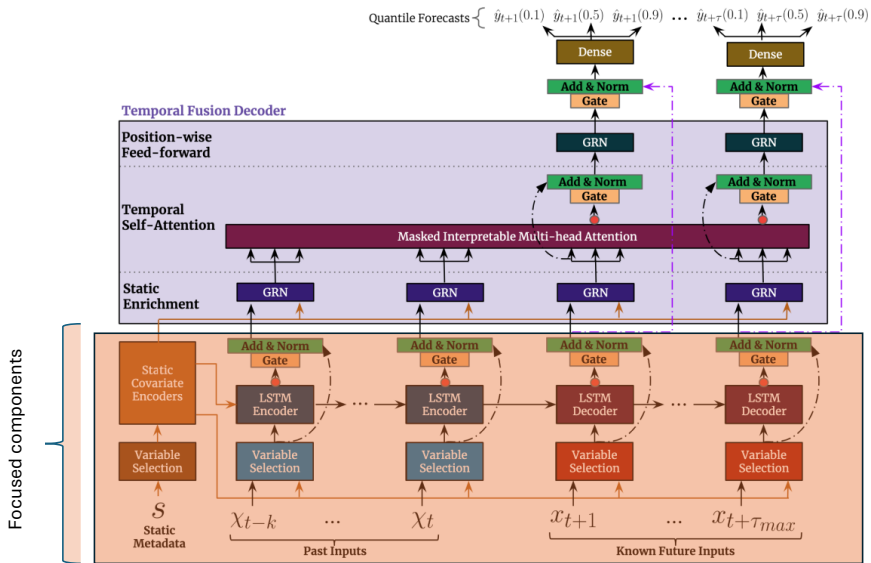
Related Work

Temporal Fusion Transformer (TFT) enables interpretable multi-horizon probabilistic forecasting.³

- ▶ **Multi-horizon & Probabilistic Forecasting:** Quantile regression approach for generating probabilistic forecasts over multiple future time steps.
- ▶ **Specialized Components for Heterogeneity:** Static covariate encoder for static data and LSTM encoder-decoder for temporal patterns.
- ▶ **Interpretability:** Built-in mechanisms to identify important features and time steps influencing forecasts.

³Lim et al., *Temporal Fusion Transformers for Interpretable Multi-Horizon Time Series Forecasting* (2021).

Temporal Fusion Transformer [2]

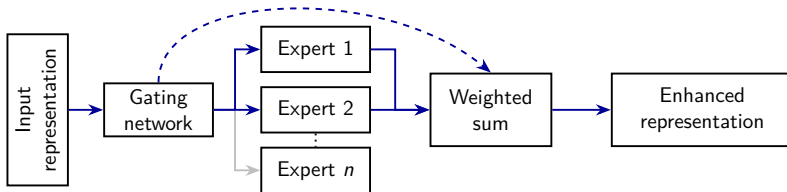


Temporal Fusion Transformer (TFT)

Attention-based multi-horizon architecture with dynamic variable selection [2].

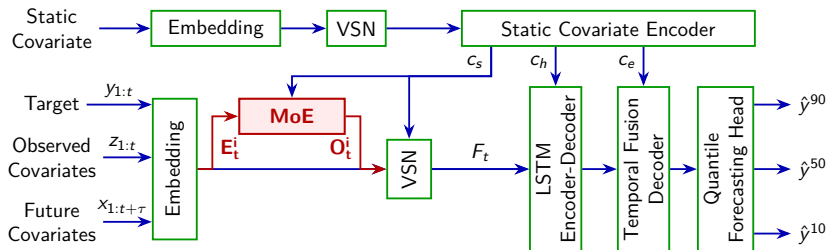
- ▶ **Unified Inputs:** Handles static covariates, known-future inputs, and observed past time series within a single forecasting model.
- ▶ **Built-in Interpretability:** Variable-selection networks and interpretable multi-head attention identify important variables and time steps influencing each forecast.
- ▶ **Gap:** TFT assumes contextual information can be aligned as homogeneous covariates along the forecasting timeline. However, real-world context consists of heterogeneous temporal dimensions:
 - ▶ **Static:** Time-invariant characteristics (e.g., consumer type, building characteristics).
 - ▶ **Time Series:** Continuously observed signals (e.g., temperature, humidity).
 - ▶ **Events and Intervals:** Sparse temporal information (e.g., holidays, vacations) represented by active/inactive states, with different frequencies, durations, overlaps, and irregular occurrences.

Mixture of Experts (MoE)



- **Specialisation:** Each expert learns a distinct regime or sub-population instead of a single averaged model [3].
- **Dynamic Gating:** A learned gating network assigns weights and routes inputs to relevant experts [4].
- **Expert Selection:** Soft (sum of all experts) vs Hard expert (top- k experts).
- **MoE Architecture:** Linear (learns linear projection) and non-linear (relu for non-linearity).
- **Architecture Variants:** Four variants, [MoELS](#), [MoENLS](#), [MoELH](#), [MoENLH](#)
- **End-to-End Training:** Experts and gating network are jointly trained as a single model.
- **Our Contribution:** Integrating MoE into TFT to enhance contextual input representation.

MoE-TFT: Experts Inside the TFT Pipeline



- **Where MoE enters:** MoE block enriches the embedded target, observed covariates, and future covariates before variable selection.
- **Objective:** multivariate heterogeneity are handled together, with TFT's interpretability preserved.

Use Case & Results (1/3): UK Electricity Consumption

5,567

consumers

aggregated by the UK Acorn
socio-economic classification

7

consumer groups

daily, per-group aggregated demand

The four time context dimensions:

- **Constant:** consumer group (UK Acorn segment)
- **Time series:** temperature & humidity (secondary series)
- **Events:** bank holidays
- **Intervals:** school vacation periods

Models comparison :

- TFT
- DeepAR
- MQRNN
- MoE-TFT variants: [MoELS](#), [MoENLS](#), [MoELH](#), [MoENLH](#)

Results (2 / 3) — Quantile Risk

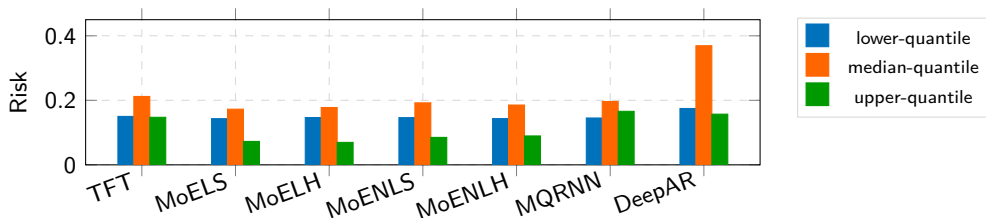


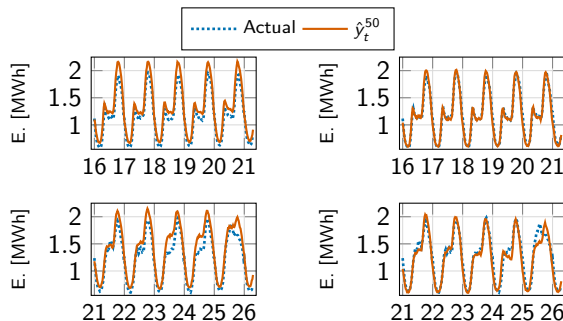
Figure: Performance comparison across quantile risk metrics (lower is better).

P10 (lower) · P50 (median) · P90 (upper) quantile risk across models (lower is better).

Takeaway

MoE variants cut upper-quantile (P90) risk sharply versus TFT and DeepAR, while holding lower- and median-quantile risk.

Results (3 / 3) — Forecast vs. Actual



Reading the panels

- **Top-weekdays:** both track the daily patterns, MoE-TFT follows the actual curve more tightly.
- **Bottom-vacations:** TFT drifts during the vacation interval, MoE-TFT stays closer to actual demand.

Where it wins

Largest gains appear in regime-shifting, i.e, vacation periods, multivariate-heterogeneity case MoE targets.

References I

- [1] S. Burero.
Integrating heterogeneous contextual data for enhanced time series analysis.
In *EDBT/ICDT Workshops*, volume 3946 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2025.
- [2] B. Lim, S. Ö. Arik, N. Loeff, and T. Pfister.
Temporal fusion transformers for interpretable multi-horizon time series forecasting.
International Journal of Forecasting, 37(4):1748–1764, 2021.
- [3] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton.
Adaptive mixtures of local experts.
Neural Computation, 3(1):79–87, 1991.
- [4] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean.
Outrageously large neural networks: The sparsely-gated mixture-of-experts layer.
In *International Conference on Learning Representations (ICLR)*, 2017.

Summary & Future Work

Summary

- **Context is heterogeneous in time:** four dimensions (constant, time series, events, intervals) must share one grid, yet keep multivariate heterogeneity.
- **TFT unifies inputs and stays interpretable:** but shared attention alone leaves multivariate heterogeneity between factors unresolved.
- **MoE-TFT closes the gap:** expert specialisation with dynamic gating: one interpretable, end-to-end model.
- **Validated on UK electricity demand:** better-calibrated tails and closer vacation-period tracking than TFT.

Summary & Future Work

Summary

- **Context is heterogeneous in time:** four dimensions (constant, time series, events, intervals) must share one grid, yet keep multivariate heterogeneity.
- **TFT unifies inputs and stays interpretable:** but shared attention alone leaves multivariate heterogeneity between factors unresolved.
- **MoE-TFT closes the gap:** expert specialisation with dynamic gating: one interpretable, end-to-end model.
- **Validated on UK electricity demand:** better-calibrated tails and closer vacation-period tracking than TFT.

Future Work

- **Router Interpretability:** analyze expert weights to understand how different experts learn and adapt to different regimes.
- **Benchmarking:** evaluate our approach on diverse datasets from multiple domains.
- **Multi-horizon Analysis:** investigate how the loss varies across different forecasting horizons.

Summary & Future Work

Summary

- **Context is heterogeneous in time:** four dimensions (constant, time series, events, intervals) must share one grid, yet keep multivariate heterogeneity.
- **TFT unifies inputs and stays interpretable:** but shared attention alone leaves multivariate heterogeneity between factors unresolved.
- **MoE-TFT closes the gap:** expert specialisation with dynamic gating: one interpretable, end-to-end model.
- **Validated on UK electricity demand:** better-calibrated tails and closer vacation-period tracking than TFT.

Future Work

- **Router Interpretability:** analyze expert weights to understand how different experts learn and adapt to different regimes.
- **Benchmarking:** evaluate our approach on diverse datasets from multiple domains.
- **Multi-horizon Analysis:** investigate how the loss varies across different forecasting horizons.

Thank you

saifullah.burero@student.unibz.it