DO FOUNDATION MODELS GENERALIZE TO REAL-WORLD EV FLEETS? A 1.1M-DRIVE BENCHMARK

Patrick Ahrend[1,2] **Stefan Pfund**[2,3] **Felix Divo**[4]

[1]TU München [2]BMW Group [3]TU Braunschweig [4]TU Darmstadt

patrick.ahrend@tum.de, stefan.pfund@bmw.de, felix.divo@tu-darmstadt.de

ABSTRACT

Time series foundation models (TSFMs) promise general-purpose forecasting, yet their effectiveness on large-scale industrial data with physical heterogeneity remains underexplored. We benchmark three zero-shot TSFMs and three supervised models against a Cluster-Aware Mixture of Experts on 1.1 million real-world electric vehicle driving sequences. The Cluster-Aware approach routes inputs to specialized LSTM experts based on Soft Dynamic Time Warping clusters, reducing mean absolute error by 14.7% over a global LSTM baseline and outperforming all other models. Evaluation on an unseen vehicle model confirms robust transfer with only 8.1% increase in error. Our results suggest that while TSFMs deliver competitive zero-shot performance, domain-informed supervised specialization remains advantageous on this heterogeneous industrial dataset.

Track: Industry & Applications

1 INTRODUCTION

The thermal management system (TMS) of a battery electric vehicle (BEV) regulates the temperature of the battery, electric motor, and cabin, accounting for up to 33% of total energy consumption (Kang et al., 2017; Kiss et al., 2015). Accurately forecasting this consumption across heterogeneous real-world conditions, including varying ambient temperatures, trip lengths, driving styles, and transient versus stationary operating phases, is critical for optimizing vehicle range and efficiency, yet poses a substantial modeling challenge. Time series foundation models such as Chronos-2 (Ansari et al., 2025), TiRex (Auer et al., 2025), and FlowState (Graf et al., 2025) aim to provide strong zero-shot predictions across diverse domains, while supervised architectures like PatchTST (Nie et al., 2023), TiDE (Das et al., 2024), and Long Short-Term Memory (LSTM) modules (Hochreiter et al., 1997) offer competitive task-specific performance. In parallel, Mixture of Experts (MoE) architectures have shown that partitioning heterogeneous data into specialized sub-models can outperform monolithic approaches (Jacobs et al., 1991; Yuksel et al., 2012). However, evaluations of TSFMs on large-scale industrial time series with known physical heterogeneity remain scarce. **Contributions.** We benchmark these approaches on 110M timesteps from 1.1M real-world driving sequences from a global production BEV fleet (see Appendix A for details). To this end, we compare six state-of-the-art time series models against a Cluster-Aware MoE that routes inputs to specialized LSTM experts using time series clustering with learned residual corrections.

2 CLUSTER-AWARE MIXTURE OF EXPERTS

We partition driving sequences into $K = 6$ operating regimes using k-means with Soft Dynamic Time Warping (SoftDTW) (Cuturi et al., 2017) as the distance metric, which accounts for temporal distortions and avoids overreaction to localized spikes. Each cluster is assigned a dedicated LSTM expert. A lightweight gating mechanism routes inputs by computing SoftDTW distances to all cluster centroids, scaling them by learnable parameters, and combining them with residual logits from a two-layer MLP that processes pooled input features: $z_k = -\beta_k \, d(\boldsymbol{X}, \boldsymbol{C}_k) + r_k$, where $d(\boldsymbol{X}, \boldsymbol{C}_k)$ is the SoftDTW distance to centroid k, $\beta_k > 0$ is a learnable scale, and r_k is the residual correction. A softmax over the logits $\boldsymbol{z}$ produces the final expert weights. An auxiliary distance-margin loss

Table 1: **The Cluster-Aware MoE outperforms all foundation and other supervised baselines on 1.1M real-world driving sequences.** Reported are means and 5th–95th-percentile ranges over $n = 25$ runs (5 folds, 5 seeds) in grey. **Bold** highlights best scores.

	Approach	MAE in Wh	RMSE in Wh	sMAPE in %	WAPE in %	R^2
TSFMs	Chronos-2 (Ansari et al., 2025)	$1.672^{1.688}_{1.663}$	$3.556^{3.588}_{3.536}$	$15.860^{15.883}_{15.847}$	$16.534^{16.560}_{16.502}$	$0.803^{0.805}_{0.802}$
	FlowState (Graf et al., 2025)	$1.706^{1.793}_{1.633}$	$3.489^{3.686}_{3.291}$	$17.178^{17.858}_{16.498}$	$17.350^{17.531}_{17.168}$	$0.787^{0.794}_{0.780}$
	TiRex (Auer et al., 2025)	$1.580^{1.628}_{1.543}$	$3.338^{3.425}_{3.274}$	$15.403^{15.759}_{15.144}$	$15.683^{16.023}_{15.464}$	$0.821^{0.824}_{0.817}$
Supervised	PatchTST (Nie et al., 2023)	$2.627^{2.701}_{2.589}$	$4.800^{4.918}_{4.728}$	$27.597^{28.376}_{27.159}$	$27.975^{28.771}_{27.552}$	$0.524^{0.538}_{0.500}$
	TiDE (Das et al., 2024)	$1.536^{1.556}_{1.513}$	$3.221^{3.277}_{3.193}$	$16.537^{16.720}_{16.165}$	$16.307^{16.467}_{16.013}$	$0.788^{0.793}_{0.782}$
	LSTM (Hochreiter et al., 1997)	$1.756^{1.761}_{1.748}$	$3.936^{3.967}_{3.919}$	$15.557^{15.606}_{15.510}$	$14.569^{14.702}_{14.511}$	$0.900^{0.901}_{0.899}$
	Cluster-Aware MoE (ours)	$\mathbf{1.498}^{1.514}_{1.489}$	$\mathbf{3.031}^{3.099}_{3.006}$	$\mathbf{14.016}^{14.188}_{13.936}$	$\mathbf{12.429}^{12.554}_{12.351}$	$\mathbf{0.941}^{0.942}_{0.938}$

encourages decisive routing by penalizing samples that are not sufficiently close to their assigned centroid. Hyperparameter details for the MoE and for the baselines are provided in Appendix B.

3 Results

Main Findings. Table 1 summarizes the performance across all approaches. The Cluster-Aware MoE achieves the lowest MAE of 1.498 Wh, a 14.7% reduction over the LSTM baseline (1.756 Wh). This error represents only 1.3% of the physical maximum per timestep (116.67 Wh) and approximately 12% of the mean target value, with an R^2 of 0.941. Among the foundation models, TiRex (1.580 Wh) and Chronos-2 (1.672 Wh) deliver competitive zero-shot performance, while FlowState performs comparably to Chronos-2. PatchTST, despite fine-tuning, performs worst overall (2.627 Wh), suggesting that its patch-based tokenization struggles with the short context length and multivariate sensor dynamics of this domain. TiDE, also fine-tuned, achieves 1.536 Wh, the strongest among the baselines, yet still falls short of the Cluster-Aware MoE. **Out-of-Distribution Evaluation.** To assess whether this specialization comes at the cost of generalizability, we evaluated the model on 85,000 drives from an unseen vehicle model with systematically different thermal properties and collected exclusively in Scandinavian winter conditions. Despite substantial distribution shifts, including an inverted relationship in refrigerant temperature, the MAE increased by only 8.1% (from 1.498 to 1.619 Wh), with R^2 still reaching 0.823. This confirms that the learned representations capture transferable physical patterns rather than vehicle-specific artifacts.

4 Discussion

Our results indicate that, for large-scale industrial time series with known physical heterogeneity, a lightweight domain-informed structure can outperform both fine-tuned supervised models and zero-shot foundation models. The TMS dataset contains fundamentally distinct physical modes, ranging from minimal cooling demand in mild weather to intensive heating in sub-zero conditions, which the Cluster-Aware MoE explicitly separates, enabling each expert to specialize in a narrower distribution. This evaluation does not uncover a limitation of foundation models per se, but it indicates that large datasets with distinct operating regimes benefit from explicit regime separation — an inductive bias that current foundation model architectures do not incorporate. Beyond accuracy, cluster-aware routing provides interpretable operating-mode assignments consistent with known TMS regimes, providing valuable validation for industrial deployment. **Limitations.** This study focused on LSTM experts within a single industrial domain; alternative backbone architectures and other heterogeneous time series settings were not evaluated. The offline clustering step requires substantial computational resources (approximately 4 hours for $K = 6$), and the number of clusters is determined heuristically rather than via joint learning. **Outlook.** Future research directions include integrating cluster-aware expert layers into foundation models, replacing offline clustering with differentiable partitioning, and extending comparisons to domains such as manufacturing, energy grids, or aviation to clarify when domain-informed specialization offers the greatest benefit relative to general-purpose models.

ACKNOWLEDGMENTS

This work benefited from the German Federal Ministry for Economic Affairs and Energy (BMWE) project "EU-SAI: Souveräne KI für Europa" (grant number 13IPC040G).

REFERENCES

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, Mononito Goswami, Shubham Kapoor, Danielle C. Maddix, Pablo Guerron, Tony Hu, Junming Yin, Nick Erickson, Prateek Mutalik Desai, Hao Wang, Huzefa Rangwala, George Karypis, Yuyang Wang, and Michael Bohlke-Schneider. Chronos-2: From univariable to universal forecasting, 2025. URL https://arxiv.org/abs/2510.15821.

Andreas Auer, Patrick Podest, Daniel Klotz, Sebastian Böck, Günter Klambauer, and Sepp Hochreiter. Tirex: Zero-shot forecasting across long and short horizons with enhanced in-context learning. *Advances in neural information processing system (NeurIPS)*, 2025.

Marco Cuturi and Mathieu Blondel. Soft-DTW: a Differentiable Loss Function for Time-Series. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 894–903. publisher: PMLR, August 2017.

Abhimanyu Das, Weihao Kong, Andrew Leach, Shaan Mathur, Rajat Sen, and Rose Yu. Long-term forecasting with tide: Time-series dense encoder. *Transactions on Machine Learning Research (TMLR)*, 2024.

Lars Graf, Thomas Ortner, StanisŁ WoŁşniak, Angeliki Pantazi, et al. Flowstate: Sampling rate invariant time series forecasting. *Advances in neural information processing system (NeurIPS) Workshop on Recent Advances in Time Series Foundation Models*, 2025.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive Mixtures of Local Experts. *Neural Computation*, 3(1):79–87, 1991. doi: 10.1162/neco.1991.3.1.79.

Byung Ha Kang and Hyun Jin Lee. A review of recent research on automotive HVAC systems for EVs. *International Journal of Air-Conditioning and Refrigeration*, 2017. publisher: World Scientific.

Tibor Kiss, Jason Lustbader, and Daniel Leighton. Modeling of an electric vehicle thermal management system in MATLAB/Simulink. *National Renewable Energy Lab (NREL)*, 2015.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *International Conference on Learning Representations (ICLR)*, 2019.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations (ICLR)*, 2023.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing system (NeurIPS)*, 2019.

Romain Tavenard, Johann Faouzi, Gilles Vandewiele, Felix Divo, Guillaume Androz, Chester Holtz, Marie Payne, Roman Yurchak, Marc Rußwurm, Kushal Kolar, and Eli Woods. Tslearn, a machine learning toolkit for time series data. *Journal of Machine Learning Research*, 21(118):1–6, 2020. URL http://jmlr.org/papers/v21/20-091.html.

S. E. Yuksel, J. N. Wilson, and P. D. Gader. Twenty Years of Mixture of Experts. *IEEE Transactions on Neural Networks and Learning Systems*, 23(8):1177–1193, August 2012. ISSN 2162-237X, 2162-2388. doi: 10.1109/TNNLS.2012.2200299.

A DATASET

We use a dataset of 1.1 million driving sequences totaling 110M timesteps collected from a global production BEV (*BMW iX*) fleet via an integrated logging system. Each sequence is sampled at 30-second intervals and consists of 12 sensor signals, including ambient, battery, and cabin temperatures; vehicle speed; refrigerant pressure and temperature; and battery current. All models operate on a context window of $L{=}20$ timesteps (10 minutes) and predict the TMS energy consumption one step ahead, i.e., a sequence-to-one regression task. This context length is shared across all supervised baselines (PatchTST and TiDE) and foundation models (Chronos-2, TiRex, and FlowState). The fleet data spans diverse geographic regions and climatic conditions, introducing substantial heterogeneity in both driving behavior and TMS operating regimes.

Table 2: **A set of 12 distinct sensor signals was selected to predict the target variable, TMS Energy Consumption.** The table lists the final features and their corresponding units used for modeling.

Sensor Name	Unit
Driving Duration	s
Ambient Temperature	°C
Cabin Temperature	°C
Battery Temperature	°C
Rear Motor Temperature	°C
Vehicle Speed	km/h
Refrigerant Temperature	°C
Battery Current	A
Cooler Outlet Temperature	°C
Refrigerant Pressure	bar
Target Cabin Temperature	°C
Delta Vehicle Speed	km/h
TMS Energy Consumption	Wh

Features were selected through domain knowledge. Table 2 shows the final dataset and the corresponding units. The physical maximum energy consumption of the TMS per 30-second timestep is 116.67 Wh; the Cluster-Aware MoE's MAE of 1.498 Wh therefore corresponds to 1.3% of this physical maximum and approximately 12% of the mean target value.

Out-of-Distribution Dataset. The OOD dataset exhibits substantial distribution shifts from the original training data described above due to the different geographic region and car model *(BMW X1)*. Table 3 summarizes the distributional shift between datasets. The cold Scandinavian climate causes substantial reductions in ambient (-100%), battery (-63%), and cooler outlet (-91%) temperatures, while the refrigerant temperature increases by 257% due to inverted TMS operation (now heating instead of cooling). Driving behavior features remain comparable across both datasets.

Table 3: **Distributional comparison between iX training data and X1 OOD test set** (Scandinavia, $n = 85{,}000$ drives). Mean Diff. denotes the relative mean difference $(\mu_{\text{X1}} - \mu_{\text{iX}})/\mu_{\text{iX}} \times 100$ and percentile position indicates where the X1 median falls within the iX distribution.

Feature	Mean Diff. (%)	OOD at Training Pctl.
Driving Duration	+7	50.7
Ambient Temperature	-100	12.7
Cabin Temperature	-12	21.8
Battery Temperature	-63	9.8
Rear Motor Temperature	-37	17.6
Vehicle Speed	+9	57.8
Refrigerant Temperature	+257	100.0
Battery Current	-19	40.4
Cooler Outlet Temperature	-91	5.7
Refrigerant Pressure	-1	54.2
Target Cabin Temperature	-1	42.8
Delta Vehicle Speed	-24	37.5
TMS Energy Consumption	-10	52.2

B Implementation Details

LSTM Experts and Global Baseline. All LSTM models use 2 hidden layers with 128 units, dropout of 0.2, and Leaky ReLU activation. Training runs for 200 epochs with a batch size of 128, using the AdamW optimizer (Loshchilov et al., 2019) and early stopping with a patience of 5. We employ a one-cycle learning rate scheduler with an initial rate of 10^{-4}, a peak rate of 10^{-3}, and a 30% warmup. Gradient clipping is applied at 4. The gating MLP consists of two fully connected layers with GELU activation and dropout. The distance-margin parameter is set to $m = 1.0$. Hyperparameters were selected via random search over the space detailed in Table 4.

Table 4: **Hyperparameter search space for LSTM models.**

Hyperparameter	Explored Values
Learning rate	10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}
Batch size	32, 64, 128, 256, 2048, 4096
Epochs	50, 100, 200, 500, 1000, 5000
Optimizer	Adam, SGD, AdamW
Hidden size	64, 128, 256, 512
Number of layers	2, 3, 4
Dropout	0.1, 0.2, 0.3, 0.4
Activation	ReLU, LeakyReLU, Tanh, Sigmoid
Scheduler	OneCycle, Step, Plateau, Cosine, Exponential, None
Gradient clipping	2, 3, 4, 5, None

Supervised Baselines. PatchTST uses patch length 4, hidden dimension 64, 4 attention heads, 2 transformer layers, and FFN dimension 256. TiDE uses a hidden size of 128, 2 encoder layers, 2 decoder layers, and a decoder output dimension of 32. Both were trained with AdamW (weight decay 0.01) and ReduceLROnPlateau scheduling with early stopping (patience of 10 epochs). Learning rates were tuned over $\{10^{-3}, 5 \times 10^{-4}, 10^{-4}, 5 \times 10^{-5}\}$ on 20% subsampled data.

Foundation Models. Chronos-2, TiRex, and FlowState were evaluated in zero-shot mode with a context length of 20 timesteps. No fine-tuning or hyperparameter optimization was performed.

Infrastructure. All experiments were conducted on a virtual machine with 256 GB RAM and a single NVIDIA L4 GPU (16 GB VRAM), using PyTorch 2.6.0 (Paszke et al., 2019) with CUDA 12.4 and tslearn 0.6.3 (Tavenard et al., 2020) for clustering.