Short–Term Electricity Load Forecasting Using the Temporal Fusion Transformer: Effect of Grid Hierarchies and Data Sources

3 authors:

Elena Giacomazzi
University of Bamberg
2 PUBLICATIONS   **26** CITATIONS

SEE PROFILE

Felix Haag
University of Bamberg
9 PUBLICATIONS   **53** CITATIONS

SEE PROFILE

Konstantin Hopf
University of Bamberg
40 PUBLICATIONS   **360** CITATIONS

SEE PROFILE

Short-Term Electricity Load Forecasting Using the Temporal Fusion Transformer: Effect of Grid Hierarchies and Data Sources

Elena Giacomazzi
University of Bamberg
Bamberg, Germany
elena@giacomazzi.de

Felix Haag
University of Bamberg
Bamberg, Germany
felix.haag@uni-bamberg.de

Konstantin Hopf
University of Bamberg
Bamberg, Germany
konstantin.hopf@uni-bamberg.de

ABSTRACT

Recent developments related to the energy transition pose particular challenges for distribution grids. Hence, precise load forecasts become more and more important for effective grid management. Novel modeling approaches such as the Transformer architecture, in particular the Temporal Fusion Transformer (TFT), have emerged as promising methods for time series forecasting. To date, just a handful of studies apply TFTs to electricity load forecasting problems, mostly considering only single datasets and a few covariates. Therefore, we examine the potential of the TFT architecture for hourly short-term load forecasting across different time horizons (day-ahead and week-ahead) and network levels (grid and substation level). We find that the TFT architecture does not offer higher predictive performance than a state-of-the-art LSTM model for day-ahead forecasting on the entire grid. However, the results display significant improvements for the TFT when applied at the substation level with a subsequent aggregation to the upper grid-level, resulting in a prediction error of 2.43% (MAPE) for the best-performing scenario. In addition, the TFT appears to offer remarkable improvements over the LSTM approach for week-ahead forecasting (yielding a predictive error of 2.52% (MAPE) at the lowest). We outline avenues for future research using the TFT approach for load forecasting, including the exploration of various grid levels (e.g., grid, substation, and household level).

KEYWORDS

Short-Term Load Forecasting, Artificial Neural Networks, Temporal Fusion Transformer (TFT), Long-Term Short-Term Memory (LSTM)

1 INTRODUCTION

Precise electricity load forecasts are key elements for planning and operating electrical power systems [18]. As utilities rely on such forecasts to purchase or generate electricity [6], the forecasting performance has a direct impact on their decision quality. With increasingly volatile electricity production and demand, electric load forecasting has recently become more difficult. Reasons include, among others, a more decentralized electricity generation, additional loads through sector coupling (heat pumps, electric vehicles, etc.) [13], changing behavioral patterns caused by the COVID-19 pandemic [37, 42], and the war in Ukraine.

Recent review studies on short-term electricity load forecasting [7, 13, 16, 18, 35] note that many research works examine the prediction of load at the level of complete energy systems (e.g., country or grid level). Another frequently investigated level is the household demand, given that smart meter data is increasingly available [13, 19, 36]. Yet, only a few studies have investigated intermediate levels in the low-voltage (distribution) grid for short-term load forecasting [13]. Within this forecasting domain, hierarchical forecasting turns out to be a promising approach, as it can model the topological distribution of load across the grid [2, 17].

While many approaches have been applied for short-term load forecasting, the most effective forecasting models currently employ variants of deep Artificial Neural Network (ANN) algorithms, such as Long-Term Short-Term Memorys (LSTMs) [31, 32, 41, 43]. LSTMs can handle sequences of data points well, yet, they have difficulties with learning patterns in long time series [3]. To remedy this drawback, Vaswani et al. [34] presented a new architecture called Transformer, which outperforms LSTMs in several sequential modeling problems like natural language processing, text generation, and machine translation [34].

Research has started to examine the Transformer approach for short-term electricity load forecasting [40] and also applied the Temporal Fusion Transformer (TFT), a Transformer variant for time series data, to short-term [21] and mid-term [26] time horizons with promising results. However, the studies on Transformers and TFTs currently investigate selective aspects and focus on suggesting new algorithmic variants rather than thoroughly testing existing approaches for various problem facets. Particularly, the use of benchmark datasets and the examination of forecasts on various grid levels are missing, although both are known limitations in the field of energy forecasting [18].

Our study addresses this research gap by conducting several experiments on the performance of the TFT in hourly electricity forecasting on the distribution grid. We vary time horizons (day-ahead and week-ahead), data sources (electricity consumption, calendar data, weather data, epidemic data), and network levels (grid and substation level). Before we present our evaluation approach in section 4 and analyze the results in section 5, we review current time series forecasting methods and related works in the electricity forecasting field.

2 BACKGROUND

Starting with the first studies on short-term load forecasting in the 1960s [14, 40], scholars have conducted intensive research within this field. Such research includes conventional statistical approaches (e.g., linear regression and Auto-Regressive Integrated Moving Average (ARIMA) models), but also Machine Learning (ML) methods

Table 1: Studies examining Transformer architectures for short-term electricity load forecasting

Ref.	Model	Place	Years	N	Cal.	Temp.	Level
[29]	TFT	PT	2011-14	1			Grid
[39]	Transf.+k-Means	FR	2006-10	1			Household
[40]	Transf.+k-Means	AU	2006-10	1			Grid
[27]	Transf.-variant	CN	-	1			Heat appl.
[21]	TFT+lin. reg.	VN	2014-21	1	x	x	Grid
[30]	Transf.	US	2004-08	20		x	Substation
[20]	Transf.-variant	SP	2016	1			Grid
[38]	Transf.-variant	AU	2006-10	1	x	x	Grid
[8]	Transf.	PA	2017-20	1	x	x	Grid
[26]	Transf., TFT, ITFT	CN	2016-17	1	x	x	Grid
		UK	2004-09	1	x	x	
This study	TFT	US	2004-08	20	x	x	Substation
study	TFT	DE	2019-22	70	x	x	Substation

such as fuzzy logic [24], and random forest [6]. In recent years, similar to other applications of forecasting, (deep) ANN approaches have gained prominence in the field of electricity load forecasting [6, 18]. Particularly LSTMs have proven to be a robust forecasting approach in several variations, as shown by studies based on data obtained from Scotland [43], Malaysia [10], the U.S. [41], and Great Britain, France, and Germany [1, 10].

Recent studies examined the Transformer architecture [34] for load forecasting. The TFT [28] in particular holds significant potential to boost the predictive performance, as it overcomes known limitations of both, the Transformer and the LSTM architecture for time series forecasting. For the application field of short-term load forecasting, we found[1] several recent studies (see Table 1) that evaluate Transformers and TFTs (and variants of them) with diverse sets of parameters and data. All of these studies indicate that the Transformer and TFT approaches outperform other methods for short-term load forecasting. However, we identify three issues that need further investigation.

First, almost all studies that we found propose a slightly different version of the Transformer or TFT architecture and test them with a single dataset (only one study [26] uses a second, publicly available dataset to demonstrate external validity). Hence, it remains unclear to what extent the reported performance results are generalizable or dataset-specific (an aspect that is also criticized in several review studies on short-term load forecasting [13, 18]). A comparison across different datasets would be helpful, although this requires significant effort and computational resources.

Second, several studies that we identified are quite selective regarding the input variables they consider. Some use electricity load data only [20, 27, 29, 39, 40], although the inclusion of exogenous variables such as weather or calendar data are known to improve forecast quality and should therefore be included [13, 23]. A detailed analysis of the performance of the TFT with different exogenous variables would benefit the assessment of the architecture's potential for load forecasting.

Third, the majority of studies focus on a single forecasting unit, e.g., the load of the whole grid [8, 20, 21, 26, 29, 38, 40], a single household [39] or a heating system [27]. This observation is also echoed in comprehensive review studies on short-term load forecasting [13], and energy forecasting [18]. Yet, forecasting on secondary levels of the distribution grid (e.g., substations or grid zones) is beneficial for grid operation and planning and has the potential to boost predictive performance, as hierarchical forecasts can model the topological distribution of load across the grid.

Our study addresses the outlined research gaps by using the TFT architecture for short-term load forecasting (day-ahead and week-ahead) on the grid and substation level while considering an acknowledged benchmark dataset.

3 FORECASTING METHOD FOR SHORT-TERM LOAD FORECASTING

We instantiate the TFT architecture to forecast the hourly electricity load based on past electricity consumption data and a variety of further data sources, as we illustrate in Figure 1. Thereby, our particular focus lies on the variation of the grid-level of the forecast (b). More specifically, we examine the TFT performance in the distribution grid using a single time series on the first grid-level and using the aggregation of multiple time series from the secondary grid-level (substation-level). The latter makes use of the TFT's and LSTM's capability to forecast several target variables for future time steps at the same time (also known as multi-horizon time series forecasting).
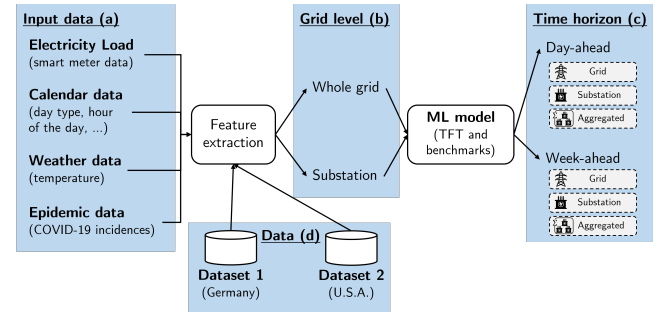
Figure 1: Simplified illustration of the approach; numbers indicate evaluation variations

For comparison, we also examine the predictive performance of LSTM and ARIMA models. By varying four elements of the forecasting setup, we analyze how well the TFT architecture performs under certain configurations. Thereby, we vary the input data (a) the time horizon (c) and test day-ahead and week-ahead forecasts, as well as two datasets (d).

Feature extraction. To obtain a multidimensional forecast, we consider four different data sources: Electric load, calendar, weather, and epidemic data. Table 2 provides an overview of all features employed, their value ranges, and their use in our study. We list further details on the data preparation in appendix A.2.

ANN model architectures. Our forecasting approach provides load forecasts on an hourly level (i.e., the next 24 hours *day-ahead*

---

[1]We searched Google Scholar, ACM digital library, and IEEE Xplore using the keywords "short-term electricity load forecasting" and "transformer".

and the next 168 hours *week-ahead*). All features are connected with the TFT using a separate Variable Selection Network (VSN) per input type. Weights are shared among VSN for past known, future known, and static variables, respectively. Table 2 lists these variable types. For benchmarking the predictive performance of the TFT, we use a linear ARIMA estimator and an LSTM architecture. Our analysis uses the Python package *darts* [15].

Grid-level forecast. We vary the levels for which we obtain forecasts. First, by considering the complete grid, which is a single time series of demand data. Second, by obtaining a substation-level forecast, which considers multiple time series for training and forecasts in each time step to predict demand data for each substation. To obtain a more precise forecast on the grid-level, we aggregate all substation-level forecasts—an approach that the literature describes as hierarchical load forecasting [17].

4 PERFORMANCE EVALUATION

We rely on two datasets to evaluate the performance of the TFT-based forecasting approach. The first stems from a local grid operator located in central Germany (*DE*) and covers a recent time frame (2019–2021). The second is a validation dataset, which is publicly available and stems from the Global Energy Forecasting Competition 2012 (GEFC'12) [17] (*US*). It comprises data from 20 grid zones in the U.S., which we consider as substations. The detailed processing of both datasets is described in appendix A.2.

For model training, we choose a time-wise 80/20 train-test split. For the *DE* dataset, the training set spans from Jan 1st, 2019, to May 23rd, 2021. The test set comprises the period from May 24th, 2021, until Dec 31st, 2021. For the day-ahead forecast, we choose all complete days (0h-23h), and for the week-ahead forecast, all complete weeks (Mo-So) in the test set. In total, we rely on 219 days and 28 weeks for the evaluation using the *DE* test set data. The training set of the *US* dataset spans from Jan 1st 2004 to March 14th 2007. The test set consists of the remaining data until December 31st, 2007. For the evaluation of the *US* dataset, the day-ahead test set consists of 291 complete days, and the week-ahead evaluation of 38 full weeks. We normalize all input features for both datasets to ensure unbiased model training [33].

We performed a random hyperparameter search [4] for those parameters for which we could not obtain meaningful values through reasoning. For the TFT, the parameters are: Number of neurons in

Table 2: Features with value ranges and horizon

Data	Feature	Range	Horizon[a]	Mean *DE*	Mean *US*
Load	Consumption	$\mathbb{R}$	P		
	(grid)			3,175.82	392,945.48
	(substation)			46.93	20,681.34
Calendar	Hour of the day	[0, 23]	F		
	Day of the week	[0, 6]	F		
	Day of the year	[1, 365]	F		
	Holiday/weekend	[0, 1]	F	0.31	0.31
Weather	Temperature	$\mathbb{R}$	F	10.48	14.46
Epidemic	Covid-19 incidence	$\mathbb{R}$	P	6	x
Other	Grid node ID[b]	$\mathbb{I}$	S		

a) F: Feature known in the future, P: Feature known up to the present, S: Static feature
b) only on substation-level prediction

the hidden layer, the number of LSTM layers, the number of attention heads, the dropout value, the batch size, and the size of the input window. For this purpose, we conducted a hyperparameter search using sweeps from the "Weights & Biases" platform [5]. The configurations for each sweep are based on the parameter bandwidth suggested in [29]. In addition, we varied the input window size k across the day-ahead forecast with $k \in [24, 48, 72, 168, 336, 672]$ and for the week-ahead forecast with $k \in [168, 336, 504, 672]$. We list the final parameter configurations of the best-performing models for each task in appendix (Table 4).

To evaluate the TFT, LSTM, and ARIMA models, we compare the predicted values p_t with the actual demand values y_t for each time step $t = 1, \ldots, N$ and assess the forecasting performance using the Root Mean Square Error (RMSE) as absolute and Mean Absolute Percentage Error (MAPE) and Symmetric Mean Absolute Percentage Error (SMAPE) as relative error metrics, which find regular use in earlier studies.

5 RESULTS

Within the scope of this paper, we present and discuss the main results of our analysis. Thereby, we analyze the predictive performance of the TFT against acARIMA and LSTM and compare the performance of our approach with earlier studies. A detailed list of the performance results of our approach can be found in Table 5 as part of the appendix.

5.1 Baseline comparison

For the analysis of the models, we focus on SMAPE as it allows for a comparison across datasets and grid-levels while expressing the relative performance results in relation to the actual and forecasted value. Similar to earlier studies, the ARIMA models display a relatively high error, which can be attributed to their limited capacity to generalize over long time series. Consequently, the LSTM and TFT clearly outperform the (more simplistic) statistical approach.

Overall, we obtain lower errors for the TFT than the LSTM models, yet, not for all configurations. We observe a clear superiority of the TFT with a larger forecasting horizon (week-ahead). We attribute this result to the stronger capability of the TFT architecture to learn patterns over longer time intervals. The TFT also performs better than the LSTM for a forecast on the substation level. For day-ahead forecasts and single time series forecasts, the LSTM approach is still a reasonable alternative.

With demand and calendar features (configuration II in Table 5), the TFT has an average performance of 3.98 MAPE, which is similar to what [21] and [38] report in their studies, although we have a simpler data processing without applying linear regression to the input features to estimate trends. As these studies do not provide pure TFT and LSTM estimates and only partially use public datasets, an appropriate comparison is not feasible. Lim et al. [29], who propose the TFT approach, conclude that the TFT results in a lower error than other approaches for time series forecasting. Using RMSE, MAPE, and SMAPE to review the forecasting error, we cannot confirm this result for the day-ahead, but for the week-ahead forecast. Other works applying the TFT to electricity forecasting do not report relative error metrics, which makes a reasonable comparison of the results unfeasible.

5.2 Forecasts on various grid-levels

The results show that the TFT's hierarchical forecast outperforms single time series forecasting regarding predictive performance on the grid-level: We observe a statistically significant difference for both approaches regarding the day-ahead $t(363.58) = 13.90$, $p < .001$, $d = 1.41$ (with MAPE 2.43%) and week-ahead $t(38.27) = 6.56$, $p < .001$, $d = 1.82$ (with MAPE 2.52%) forecasts using the DE dataset. We validate this result by performing the same tests on the US dataset, where we also find a significant difference with a large effect size for both approaches (day-ahead $t(550.99) = 10.60$, $p < .001$, $d = 0.89$, and week-ahead $t(71.14) = 3.84$, $p < .001$, $d = 0.88$). Hence, we conclude that the hierarchical forecast approach outperforms single time series forecasting on the grid-level. Additionally, the results display performance improvements for the LSTM architecture when the load is predicted and aggregated at the substation level—however, this observation does not hold for all predictive cases of this study.

6 DISCUSSION

Practical contribution. Our analysis demonstrated the potential of the TFT approach compared to a state-of-the-art approach (LSTM) and a simple estimator (ARIMA). In addition, we empirically illustrated the benefits that can be obtained through hierarchical load forecasting in the electric grid (reflecting the call for such analyses from recent review studies [13, 18]).

While we observed that the TFT approach is on par with (or only slightly better than) the LSTM approach on a day-ahead horizon, the TFT clearly outperformed the LSTM on a week-ahead horizon. Conversely, the TFT seems to be more costly to train regarding the computational effort because it has significantly more parameters. Therefore, practitioners need to balance a trade-off between more accurate methods in longer time spans and computational costs.

Limitations and future work. Our study is a starting point for a more in-depth evaluation of Transformer and TFT approaches in the domain of load forecasting. In summary, we identify six areas for future research:

First, we used weather observations as inputs for the forecasting period, which leads to an underestimation of the forecasting error [13]. In practice, only weather forecasts are available. Future studies should therefore include historical weather forecasts and quantify their impact on the models' forecasting quality.

Second, we included the Covid-19 incidence as a covariate for the TFT. However, the incidence data do not properly represent the lockdown periods. Hence, additional epidemic data might reflect time periods and their effect on the energy demand more precisely (e.g., by employing a binary feature that reflects lockdown periods).

Third, we only compared point estimates of the forecasting models in our study. However, probabilistic forecasting is a very promising area in load forecasting [13, 16]. Future studies may extend the TFT approach and assess its potential for probabilistic forecasting.

Fourth, for real-world applications, the runtime performance of the models and their explainability might be of major importance to electricity vendors. In some cases, higher explainability outweighs higher costs for training [6, 18, 25]. The TFT architecture contains an interpretable multi-head self-attention mechanism that enables feature importance-based explanations [29]. So far, this functionality has not been studied for the case at hand, although explainable ML offers detailed insights on model forecasts that can benefit decision-makers [12].

Fifth, our analysis has shown the potential of predictions on more granular network levels employing a subsequent aggregation. Future work should make use of increasingly available smart meter data to obtain household level predictions and their aggregations to enhance the forecasting performance.

Sixth, we integrated empirical load data mostly as is in our analysis. The body of forecasting literature has suggested several meaningful data preprocessing steps that improved the performance of less complex forecasting models, such as taking into account typical daily or weekly load profiles [22]. Considering that varying existing algorithms often result in only small changes in predictive performance, we encourage future research to focus on an in-depth evaluation of existing methods, more advanced feature engineering, and the evaluation of real-world problems with (multiple) benchmark datasets.

7 CONCLUSION

Current developments related to more volatile electricity production and demand challenge the management of the electric grid. Thus, precise load forecasts become more and more important. Recent forecasting literature has proposed the TFT architecture, which theoretically addresses known limitations of the LSTM and Transformer approach. To date, studies on the TFT approach to short-term load forecasting have been empirically inconclusive and neglect external validity.

Our study carries out several experiments using the TFT architecture and multiple datasets. The results show that the TFT architecture does not outperform a LSTM model for day-ahead forecasting for the entire grid. Yet, we find that the predictive performance of the TFT is higher when applied at the substation level in conjunction with a subsequent aggregation to upper grid-levels.

Our investigation opens avenues for future research on the TFT approach for short-term load forecasting. In particular, we would like to motivate other scholars to conduct further experiments, specifically with respect to different network levels of forecasting (e.g., grid, substation, household) and the explainability of the models used.

REFERENCES

[1] Aasim, S. N. Singh, and Abheejeet Mohapatra. 2021. Data driven day-ahead electrical load forecasting through repeated wavelet transform assisted SVM model. *Applied Soft Computing* 111 (2021), 107730. https://doi.org/10.1016/j.asoc.2021.107730

[2] George Athanasopoulos, Roman A Ahmed, and Rob J Hyndman. 2009. Hierarchical forecasts for Australian domestic tourism. *International Journal of Forecasting* 25, 1 (2009), 146–166.

[3] Yoshua Bengio, Patrice Y. Simard, and Paolo Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks* 5 2 (1994), 157–66.

[4] James Bergstra and Yoshua Bengio. 2012. Random search for hyper-parameter optimization. *Journal of machine learning research* 13, 2 (2012).

[5] Lukas Biewald. 2020. Experiment Tracking with Weights and Biases. https://www.wandb.com/ Software available from wandb.com.

[6] Stefan Borkovski, Stefan Petkoski, and Maja Erkechova. 2019. Electricity consumption forecasting using recurrent neural network: Electrical trade market study. *Innovations* (2019), 8.

[7] Mathieu Bourdeau, Xiao qiang Zhai, Elyes Nefzaoui, Xiaofeng Guo, and Patrice Chatellier. 2019. Modeling and forecasting building energy consumption: A review of data-driven techniques. *Sustainable Cities and Society* 48 (July 2019), 101533. https://doi.org/10.1016/j.scs.2019.101533

[8] Yang Cao, Zhenzhen Dang, Feng Wu, Xovee Xu, and Fan Zhou. 2022. Probabilistic Electricity Demand Forecasting with Transformer-Guided State Space Model. In *2022 IEEE 5th International Conference on Automation, Electronics and Electrical Engineering (AUTEEE)*. 964–969. https://doi.org/10.1109/AUTEEE56487.2022.9994294

[9] Richard E. Edwards, Joshua New, and Lynne E. Parker. 2012. Predicting future hourly residential electrical consumption: A machine learning case study. *Energy and Buildings* 49 (June 2012), 591–603. https://doi.org/10.1016/j.enbuild.2012.03.010

[10] Behnam Farsi, Manar Amayri, Nizar Bouguila, and Ursula Eicker. 2021. On Short-Term Load Forecasting Using Machine Learning Techniques and a Novel Parallel Deep LSTM-CNN Approach. *IEEE Access* 9 (2021), 31191–31212. https://doi.org/10.1109/ACCESS.2021.3060290

[11] Arne Groß, Antonia Lenders, Friedhelm Schwenker, Daniel A. Braun, and David Fischer. 2021. Comparison of short-term electrical load forecasting methods for different building types. *Energy Informatics* 4, 3 (Sept. 2021), 13. https://doi.org/10.1186/s42162-021-00172-6

[12] Felix Haag, Konstantin Hopf, Pedro Menelau Vasconcelos, and Thorsten Staake. 2022. Augmented Cross-Selling Through Explainable AI—A Case From Energy Retailing. In *ECIS 2022 Research Papers*. AIS electronic library, Timisoara, Romania. https://aisel.aisnet.org/ecis2022_rp/129

[13] Stephen Haben, Georgios Giasemidis, Florian Ziel, and Siddharth Arora. 2019. Short Term Load Forecasts of Low Voltage Demand and the Effects of Weather. *International Journal of Forecasting* 35, 4 (Oct. 2019), 1469–1484. https://doi.org/10.1016/j.ijforecast.2018.10.007 arXiv:1804.02955.

[14] G T Heinemann, D A Nordmian, and E C Plant. 1966. and Summer Loads- A Regression Analysis. *IEEE TRANSACTIONS ON POWER APPARATUS AND SYSTEMS* (1966), 11.

[15] Julien Herzen, Francesco Lässig, Samuele Giuliano Piazzetta, Thomas Neuer, Léo Tafti, Guillaume Raille, Tomas Van Pottelbergh, Marek Pasieka, Andrzej Skrodzki, Nicolas Huguenin, Maxime Dumonal, Jan Kościsz, Dennis Bader, Frédérick Gusset, Mounir Benheddi, Camila Williamson, Michal Kosinski, Matej Petrik, and Gaël Grosch. 2022. Darts: User-Friendly Modern Machine Learning for Time Series. *Journal of Machine Learning Research* 23, 124 (2022), 1–6. http://jmlr.org/papers/v23/21-1177.html

[16] Tao Hong and Shu Fan. 2016. Probabilistic electric load forecasting: A tutorial review. *International Journal of Forecasting* 32, 3 (July 2016), 914–938. https://doi.org/10.1016/j.ijforecast.2015.11.011

[17] Tao Hong, Pierre Pinson, and Shu Fan. 2014. Global Energy Forecasting Competition 2012. *International Journal of Forecasting* 30, 2 (April 2014), 357–363. https://doi.org/10.1016/j.ijforecast.2013.07.001

[18] Tao Hong, Pierre Pinson, Yi Wang, Rafał Weron, Dazhi Yang, and Hamidreza Zareipour. 2020. Energy Forecasting: A Review and Outlook. *IEEE Open Access Journal of Power and Energy* 7 (Oct. 2020). https://doi.org/10.1109/OAJPE.2020.3029979

[19] Konstantin Hopf. 2019. *Predictive Analytics for Energy Efficiency and Energy Retailing* (1 ed.). Contributions of the Faculty Information Systems and Applied Computer Sciences of the Otto-Friedrich-University Bamberg, Vol. 36. University of Bamberg, Bamberg. https://doi.org/10.20378/irbo-54833

[20] Shichao Huang, Jing Zhang, Yu He, Xiaofan Fu, Luqin Fan, Gang Yao, and Yongjun Wen. 2022. Short-Term Load Forecasting Based on the CEEMDAN-Sample Entropy-BPNN-Transformer. *Energies* 15, 10 (Jan. 2022), 3659. https://doi.org/10.3390/en15103659

[21] Pham Canh Huy, Nguyen Quoc Minh, Nguyen Dang Tien, and Tao Thi Quynh Anh. 2022. Short-Term Electricity Load Forecasting Based on Temporal Fusion Transformer Model. *IEEE Access* 10 (2022), 106296–106304. https://doi.org/10.1109/ACCESS.2022.3211941

[22] Boye A. Høverstad, Axel Tidemann, Helge Langseth, and Pinar Öztürk. 2015. Short-Term Load Forecasting With Seasonal Decomposition Using Evolution for Parameter Tuning. *IEEE Transactions on Smart Grid* 6, 4 (July 2015), 1904–1913. https://doi.org/10.1109/TSG.2015.2395822

[23] Rishee K. Jain, Kevin M. Smith, Patricia J. Culligan, and John E. Taylor. 2014. Forecasting energy consumption of multi-family residential buildings using support vector regression: Investigating the impact of temporal and spatial monitoring granularity on performance accuracy. *Applied Energy* 123 (June 2014), 168–178. https://doi.org/10.1016/j.apenergy.2014.02.057

[24] Ahsan Raza Khan, Anzar Mahmood, Awais Safdar, Zafar A Khan, Syed Bilal, and Naveed Ahmed Khan Javaid. 2015. Load Forecasting and Dynamic Pricing based Energy Management in Smart Grid-A Review. In *International Multi-topic Conference*.

[25] Jesus Lago, Fjo De Ridder, and Bart De Schutter. 2018. Forecasting spot electricity prices: Deep learning approaches and empirical comparison of traditional algorithms. *Applied Energy* 221 (July 2018), 386–405. https://doi.org/10.1016/j.apenergy.2018.02.069

[26] Dan Li, Ya Tan, Yuanhang Zhang, Shuwei Miao, and Shuai He. 2023. Probabilistic forecasting method for mid-term hourly load time series based on an improved temporal fusion transformer model. *International Journal of Electrical Power & Energy Systems* 146 (March 2023), 108743. https://doi.org/10.1016/j.ijepes.2022.108743

[27] Guangxia Li, Cheng Zhou, Ruiyu Li, and Jia Liu. 2022. Heat load forecasting for district water-heating system using locality-enhanced transformer encoder. In *Proceedings of the Thirteenth ACM International Conference on Future Energy Systems (e-Energy '22)*. Association for Computing Machinery, New York, NY, USA, 440–441. https://doi.org/10.1145/3538637.3538751

[28] Bryan Lim, Sercan Ö. Arık, Nicolas Loeff, and Tomas Pfister. 2021. Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting* 37, 4 (2021), 1748–1764. https://doi.org/10.1016/j.ijforecast.2021.03.012

[29] Bryan Lim, Sercan Ö. Arık, Nicolas Loeff, and Tomas Pfister. 2021. Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting* 37, 4 (Oct. 2021), 1748–1764. https://doi.org/10.1016/j.ijforecast.2021.03.012

[30] Alexandra L'Heureux, Katarina Grolinger, and Miriam A. M. Capretz. 2022. Transformer-Based Model for Electrical Load Forecasting. *Energies* 15, 14 (Jan. 2022), 4993. https://doi.org/10.3390/en15144993

[31] Sana Mujeeb, Nadeem Javaid, Manzoor Ilahi, Zahid Wadud, Farruh Ishmanov, and Muhammad Afzal. 2019. Deep Long Short-Term Memory: A New Price and Load Forecasting Scheme for Big Data in Smart Cities. *Sustainability* 11 (02 2019), 987. https://doi.org/10.3390/su11040987

[32] Md Jamal Ahmed Shohan, Md Omar Faruque, and Simon Y. Foo. 2022. Forecasting of Electric Load Using a Hybrid LSTM-Neural Prophet Model. *Energies* 15, 6 (March 2022), 2158. https://doi.org/10.3390/en15062158

[33] J. Sola and J. Sevilla. 1997. Importance of input data normalization for the application of neural networks to complex industrial problems. *IEEE Transactions on Nuclear Science* 44, 3 (1997), 1464–1468. https://doi.org/10.1109/23.589532

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *arXiv:1706.03762 [cs]* (Dec. 2017). http://arxiv.org/abs/1706.03762 arXiv: 1706.03762.

[35] Frederik vom Scheidt, Hana Medinová, Nicole Ludwig, Bent Richter, Philipp Staudt, and Christof Weinhardt. 2020. Data analytics in the electricity sector – A quantitative and qualitative literature review. *Energy and AI* 1 (2020), Article no: 100009. https://doi.org/10.1016/j.egyai.2020.100009

[36] Yi Wang, Qixin Chen, Tao Hong, and Chongqing Kang. 2019. Review of Smart Meter Data Analytics: Applications, Methodologies, and Challenges. *IEEE Transactions on Smart Grid* 10, 3 (May 2019), 3125–3148. https://doi.org/10.1109/TSG.2018.2818167 Conference Name: IEEE Transactions on Smart Grid.

[37] Le Wen, Basil Sharp, Kiti Suomalainen, Mingyue Selena Sheng, and Fengtao Guang. 2022. The impact of COVID-19 containment measures on changes in electricity demand. *Sustainable Energy, Grids and Networks* 29 (March 2022), 100571. https://doi.org/10.1016/j.segan.2021.100571

[38] Guangqi Zhang, Chuyuan Wei, Changfeng Jing, and Yanxue Wang. 2022. Short-Term Electrical Load Forecasting Based on Time Augmented Transformer. *International Journal of Computational Intelligence Systems* 15, 1 (Aug. 2022), 67. https://doi.org/10.1007/s44196-022-00128-y

[39] Junfeng Zhang, Hui Zhang, Song Ding, and Xiaoxiong Zhang. 2021. Power Consumption Predicting and Anomaly Detection Based on Transformer and K-Means. *Frontiers in Energy Research* 9 (Oct. 2021), 779587. https://doi.org/10.3389/fenrg.2021.779587

[40] Zezheng Zhao, Chunqiu Xia, Lian Chi, Xiaomin Chang, Wei Li, Ting Yang, and Albert Y. Zomaya. 2021. Short-Term Load Forecasting Based on the Transformer Model. *Information* 12, 12 (Dec. 2021), 516. https://doi.org/10.3390/info12120516

[41] Huiting Zheng, Jiabin Yuan, and Long Chen. 2017. Short-term load forecasting using EMD-LSTM neural networks with a Xgboost algorithm for feature importance evaluation. *Energies* 10, 8 (2017), 1168.

[42] Haiwang Zhong, Zhenfei Tan, Yiliu He, Le Xie, and Chongqing Kang. 2020. Implications of COVID-19 for the electricity industry: A comprehensive review. *CSEE Journal of Power and Energy Systems* 6, 3 (Sept. 2020), 489–495. https://doi.org/10.17775/CSEEJPES.2020.02500

[43] Mingzhe Zou, Duo Fang, Gareth Harrison, and Sasa Djokic. 2019. Weather Based Day-Ahead and Week-Ahead Load Forecasting using Deep Recurrent Neural Network. In *2019 IEEE 5th International forum on Research and Technology for Society and Industry (RTSI)*. IEEE, Florence, Italy, 341–346. https://doi.org/10.1109/RTSI.2019.8895580

A DATA PROCESSING

A.1 Datasets

The two real datasets that we select stem from two geographic regions, differ in the number of substations, the magnitude of load connected, and the timespans of the data. The severe differences in the datasets should ensure the external validity of our results.

US dataset. The dataset stems from the Global Energy Forecasting Competition 2012 (GEFC'12) [17] and comprises data from 20 grid zones in the U.S., which we consider as substations. For our analyses, we consider the years 2004-2007 of the dataset. The authors of the dataset [17] advise excluding two substations, #4 because of outages and #9 because it was covered by an industrial customer. We remove substation #9 but keep substation #4 and add the following data cleaning to all substations to handle potential outages on every substation. We identify extreme values (e.g., outages) per substation using the statistical quartiles $q_{0.25}$ and $q_{0.75}$ and remove values smaller than $q_{0.25} - 1.5 * (q_{0.27} - q_{0.25})$, known as the inter quantile range. We calculate the quartiles per substation. Only demand values for substation #4 fall under this criterion, and we remove 52 out of 39,576 (0.0013%) data points. We replace the removed measurements using a linear interpolation [11].

The dataset also contains temperature data of 11 weather stations in the U.S. but the connection of the weather stations to the zones is not given (the connection of the weather stations to the grid zones was part of the GEFC'12 challenge). Hence, for our analysis, we use the average temperature of all 11 weather stations per hour.

DE dataset. In addition to the public dataset, which does not contain detailed geographic references, we use a dataset from central Germany, which we obtained from a local distribution grid operator. The dataset consists of hourly smart meter data on the household level, covering the years 2019–2021 (36 months). Given that the first wave of the COVID-19 pandemic started in Germany in March 2020, the dataset contains 14 months of pre-pandemic electricity load and 22 months within the pandemic. In total, 9,455 households are connected to one of 70 substations in the distribution grid, where each substation serves between 8 and 447 households (M=135.07, SD=94.75).

We prepare the data on the level of each household and apply the following preparation steps. First, we remove households with unusually low consumption values. For this purpose, we exclude observations with a mean consumption less than 0.01 kWh or a total consumption less than 100 kWh. In total, we remove 598 households applying this criterion. Second, we harmonize time shift events (to and from daylight saving time) in spring and fall. For time shifts in the fall, where a single day has 25 hours, we exclude the extra hour. For the days with only 23 hours (i.e., time shift in spring), we linearly interpolate the missing value to harmonize the data into a 24-hour shape. Third, we remove all values from the meter readings that were labelled as "provisional", "defective", and "incorrect" and linearly interpolated the readings. In total, we replace 1,572,786 of such missing values out of 51,157,455,768 observations (0.0031%).

Finally, we aggregate the data on the level of the substations and on the grid-level for our analysis. We were also provided with the geographic location data for each substation, which we leverage to connect weather and epidemic data.

A.2 Features

From the *calendar data*, we extract the hour of the day, day of the week, day of the year, and a binary feature if a day is a national holiday or weekend day. We use the Python package *python-holidays*[2] to obtain the local holidays. As most of the calendar features have a cyclic pattern, we encode them cyclically combining sine and cosine, following [9].

One of the most common input features in demand forecasts are meteorological variables, in particular, temperature data [13, 22]. As *weather data*, we use the hourly temperature of the region obtained from the *Meteostat*[3] Python package, which uses, for example, data from the German Meteorological Service[4]. For cities and areas with no own weather station, we interpolate the temperature for the selected geographical point using the geographic reference and altitude. We assume that the temperature data is also available for the test data horizon and apply the measurements as a proxy for a weather forecast.

Finally, we consider a data source that we have not found to be used by earlier studies, namely *epidemic data*. This is feasible, as one of the datasets we include covers the beginning of the COVID-19 pandemic and thus accounts for multiple lockdowns in the area of the grid. This lead us to include the officially announced number of infected people in the area as a feature. Such data is, for example, published by the German Robert Koch Institute[5].

B HYPERPARAMETERS

We list the value ranges for our hypoerparameter search in Table 3 and the finally used parameters in Table 4.

Table 3: Value ranges hyperparameter tuning

TFT hyper-parameter	
att heads	[1, 4]
hidden size	[16, 32, 64]
dropout	[0.1, 0.3]
batch size	[32, 128]
LSTM layers	[1, 2, 4]
LSTM hyper-parameter	
batch size	[50, 10, 120, 150]
learning rate	[0.001, 0.01, 0.1]
dropout	[0.1, 0.2, 0.3]
LSTM layer	[1, 2, 4]
hidden size	[64, 128, 248, 496]

C DETAILED RESULTS

See our detailed evaluation results in Table 5.

---

[2]https://github.com/dr-prodigy/python-holidays
[3]https://meteostat.net/en/
[4]https://www.dwd.de/
[5]https://github.com/robert-koch-institut/SARS-CoV-2-Infektionen_in_Deutschland

Table 4: Final hyper-parameters

Horizon	Dataset	TFT						LSTM					
		att. heads	hidden size	LSTM layers	input size	dropout	batch size	hidden size	num layers	input size	dropout	batch size	learning rate
Grid-level forecast, consumption + calendar													
Day	DE	1	64	2	24	0.1	32	64	2	48	0.2	10	0.01
Week	DE	1	64	4	504	0.1	32	64	4	336	0.1	10	0.001
Day	US	1	64	2	336	0.3	32	496	1	48	0.2	10	0.001
Week	US	1	64	2	336	0.1	32	248	1	504	0.2	10	0.001
Grid-level forecast, consumption + weather + calendar													
Day	DE	4	64	2	672	0.3	32	64	4	168	0.1	10	0.01
Week	DE	4	64	4	672	0.1	32	128	1	504	0.3	10	0.01
Day	US	4	64	2	168	0.1	32	128	1	168	0.2	10	0.01
Week	US	1	64	4	168	0.1	32	64	1	672	0.2	10	0.01
Grid-level, consumption + weather + calendar + epidemic features													
Day	DE	4	32	1	48	0.1	32	x	x	x	x	x	x
Week	DE	4	64	2	168	0.1	32	x	x	x	x	x	x
Hierarchical/substation forecast, consumption + weather + calendar													
Day	DE	1	32	1	336	0.1	128	128	2	72	0.2	120	0.01
Week	DE	4	64	2	168	0.1	128	248	4	672	0.3	10	0.001
Day	US	4	64	2	72	0.1	31	128	2	48	0.1	10	0.001
Week	US	1	16	1	672	0.3	32	64	2	168	0.1	50	0.001
Hierarchical/substation forecast, consumption + weather + calendar + epidemic feature													
Day	DE	4	64	2	336	0.3	32	x	x	x	x	x	x
Week	DE	4	32	2	336	0.1	32	x	x	x	x	x	x

Table 5: Forecasting performance

Model	day-ahead						week-ahead					
	RMSE		MAPE		SMAPE		RMSE		MAPE		SMAPE	
I. Grid-level forecast, demand only (both datasets)												
ARIMA DE	191,376.30	(±73,874.24)	116.23	(±60.62)	183.82	(±22.69)	197,129.80	(±63,766.59)	115.25	(±42.65)	185.51	(±18.69)
ARIMA US	251,621.73	(±675,237.99)	56.99	(±161.73)	64.39	(±14.66)	868,061.20	(±2,431,013)	209.04	(±602.5)	80.56	(±35.12)
II. Grid-level forecast, demand + calendar (both datasets)												
LSTM DE	154.79	(±68.68)	3.94	(±1.52)	4.01	(±1.54)	190.73	(±48.37)	4.94	(±0.91)	4.88	(±0.84)
TFT DE	151.13	(±58.52)	3.98	(±1.25)	3.95	(±1.22)	167.12	(±54.37)	4.13	(±0.92)	4.18	(±0.96)
LSTM US	32,116.58	(±16,425.95)	6.25	(±2.76)	6.35	(±2.88)	50,107.86	(±18,577.51)	10.14	(±4.27)	9.79	(±3.67)
TFT US	30,977.48	(±15,061.03)	6.11	(±2.73)	6.19	(±2.82)	50,232.94	(±17,623.17)	9.98	(±3.23)	10.01	(±3.24)
III. Grid-level forecast, demand + weather + calendar												
LSTM DE	137.42	(±59.21)	3.52	(±1.21)	3.54	(±1.23)	157.66	(±39.60)	4.24	(±1.13)	4.18	(±1.05)
TFT DE	164.53	(±60.43)	4.22	(±1.34)	4.20	(±1.31)	151.41	(±52.1)	3.88	(±0.96)	3.83	(±0.9)
LSTM US	25,531.28	(±14,523.90)	4.89	(±2.35)	4.99	(±2.50)	57,549.48	(±24,345.73)	11.77	(±5.55)	11.26	(±4.99)
TFT US	22,825.27	(±10,190.03)	4.59	(±1.86)	4.70	(±1.98)	26,967.22	(±9,804.02)	5.22	(±1.80)	5.36	(±1.97)
IV. Hierarchical forecast, demand + weather + calendar												
LSTM DE	146.43	(±67.89)	3.65	(±1.42)	3.60	(±1.33)	337.97	(±119.81)	7.99	(±1.98)	7.56	(±1.75)
TFT DE	102.46	(±55.09)	2.55	(±1.06)	2.54	(±1.05)	102.75	(±29.58)	**2.5**	(±0.49)	2.52	(±0.5)
LSTM US	19,751.91	(±9,955.10)	3.88	(±2.02)	3.86	(±1.95)	32,064.27	(±10,200.25)	6.23	(±1.88)	6.39	(±2.03)
TFT US	15,712.65	(±8,763.61)	3.04	(±1.59)	3.09	(±1.65)	18,955.79	(±6,553.06)	3.76	(±1.42)	3.81	(±1.49)
V. Substation forecast, consumption + weather + calendar												
LSTM DE	7.26	(±5.82)	16.98	(±16.07)	28.70	(±48.55)	10.97	(±8.73)	27.58	(±29.90)	35.17	(±47.92)
TFT DE	4.52	(±2.78)	10.15	(±7.52)	23.38	(±49.36)	4.83	(±2.71)	10.56	(±7.65)	23.91	(±49.29)
LSTM US	1,518.69	(±1,592.79)	6.54	(±3.73)	6.45	(±3.50)	2,453.43	(±2,397.44)	10.04	(±3.62)	10.07	(±3.57)
TFT US	1,172.17	(±1,310.05)	4.81	(±2.61)	4.85	(±2.62)	1,527.51	(±1,456.43)	6.43	(±2.63)	6.38	(±2.51)
VI. Forecast with demand + weather + calendar + epidemic features												
TFT DE (grid-level)	169.59	(±63.46)	4.46	(±1.2)	4.48	(±1.2)	149.39	(±44.35)	3.89	(±0.61)	3.88	(±0.61)
TFT DE (hierarchical)	98.32	(±50.49)	2.43	(±0.88)	2.44	(±0.89)	100.59	(±26.72)	2.52	(±0.47)	2.51	(±0.46)
TFT DE (substation)	4.39	(±2.81)	9.86	(±7.68)	23.12	(±49.42)	4.84	(±2.74)	10.85	(±8.05)	23.99	(±49.27)