Real-E: A Foundation Benchmark for Advancing Robust and Generalizable Electricity Forecasting

Chen Shao*
Karlsruhe Institute of Technology
Karlsruhe, Germany
chen.shao2@kit.edu

Michael Färber
Karlsruhe Institute of Technology
Karlsruhe, Germany
michael.faerber@tu-dresden.de

Sebastian Pütz
Karlsruhe Institute of Technology
Karlsruhe, Germany
sebastian.puetz@kit.edu

Benjamin Schäfer
Karlsruhe Institute of Technology
Karlsruhe, Germany
benjamin.schaefer@kit.edu

Yue Wang*
Karlsruhe Institute of Technology
Karlsruhe, Germany
pg9697@partner.kit.edu

Tobias Käfer
Karlsruhe Institute of Technology
Karlsruhe, Germany
tobias.kaefer@kit.edu

Zhanbo Huang
Karlsruhe Institute of Technology
Karlsruhe, Germany
zhanbo.huang@rwth-aachen.de

Zhenyi Zhu
The Hong Kong University of Science
and Technology
Hong Kong, China
zzhubh@connect.ust.hk

Abstract

Energy forecasting is vital for grid reliability and operational efficiency. Although recent advances in time series forecasting have led to progress, existing benchmarks remain limited in spatial and temporal scope and lack multi-energy features. This raises concerns about their reliability and applicability in real-world deployment. To address this, we present the Real-E dataset, covering over 74 power stations across 30+ European countries over a 10-year span with rich metadata. Using Real-E, we conduct an extensive data analysis and benchmark over 20 baselines across various model types. We introduce a new metric to quantify shifts in correlation structures and show that existing methods struggle on our dataset, which exhibits more complex and non-stationary correlation dynamics. Our findings highlight key limitations of current methods and offer a strong empirical basis for building more robust forecasting models.

CCS Concepts

• **Computing methodologies → Machine learning algorithms**.

Keywords

Energy Forecasting, Multivariate Time Series Forecasting

---

*Both authors contributed equally to this research.

1 Introduction

Energy forecasting plays a critical role in modern power systems, supporting key tasks such as grid stability, energy planning and renewable energy integration. Recent advances in multivariate time forecasting (MTSF) have shown strong potential in modeling the complex spatiotemporal dependencies. Recurrent Neural Network (RNN) [8], Convolutional Neural Network (CNN) [2] and Transformers [7] primarily incorporate complex spatial and temporal dependencies using either an auto-regressive model or a global attention mechanism. Graph Neural Network (GNN)-based approaches incorporate spatial correlations by modeling spatial dynamics from inter-series correlations [9, 17, 18].

Despite the improved performance, these methods have been predominantly evaluated on small-scale datasets, such as electricity [1] and solar [8]. They focus solely on a *single generation type from a single country with limited temporal coverage*. However, modern power systems rely on an integrated mix of energy sources including wind, hydro, nuclear and coal that interact with one another. These interactions are further influenced by policy constraints and seasonal dynamics. Between such time-evolving correlation, each time series also exhibits seasonal patterns shaped by weather conditions, temperature fluctuations and geographic diversity. However, these complex data factors and their impact on the performance of current forecasting models have not yet been well studied.

Our work addresses these issues by first introducing **Real-E**, the largest electricity dataset to date, to support the development of robust and efficient methods for forecasting (Fig. 1). We conduct a thorough data analysis to compare the proposed approach with existing datasets and then perform extensive benchmark experiments. Based on the empirical results, we summarize key research challenges inherent in current approaches. Our main contributions include:
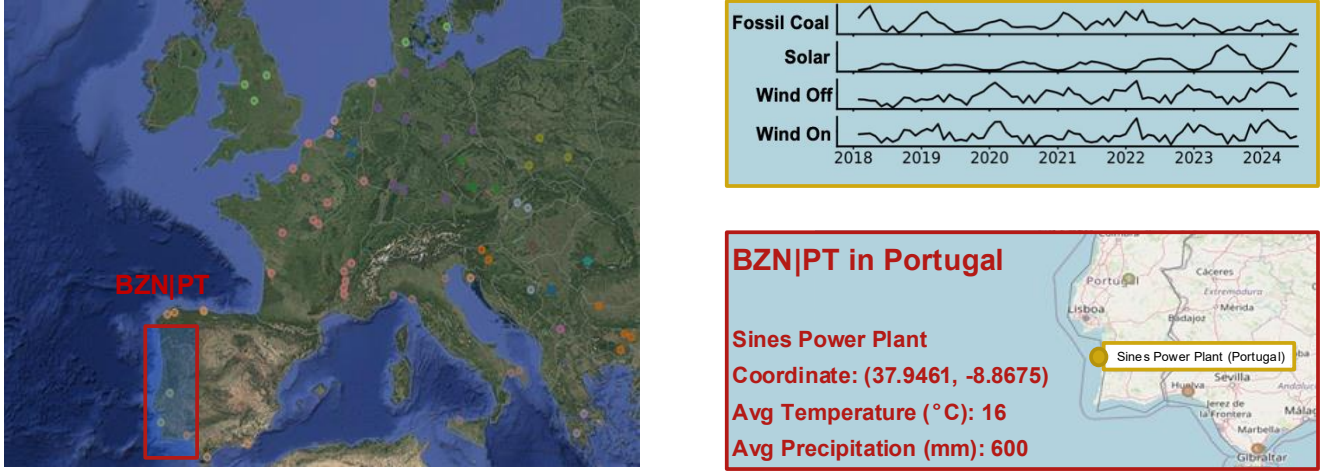
Figure 1: REAL-E is the largest electricity dataset to date, covering 10.18 million km^2 across 39 countries (colored dots) and spanning over 20 generation types (yellow boxes). Each power plant is accompanied by detailed metadata (red boxes). It offers a rigorous foundation for advancing multivariate electricity forecasting research.

(1) **Largest Real-world Electricity Dataset** The proposed REAL-E dataset offers unparalleled spatiotemporal breadth, encompassing 20 distinct energy types across 39 European countries over a period of up to 10 years. This dataset provides a foundational resource for advancing this field to address the challenges of generalizability and efficiency in real-world deployment.

(2) **Quantifying Temporal Correlation Evolution:** Based on our data analysis, we identify unique data characteristics of the proposed dataset: more complex time-varying complementary patterns among multiple energy sources. We introduce a metric to quantify this and compare it with existing datasets and demonstrate that our dataset exhibits significantly higher volatility and structural complexity in its correlation dynamics.

(3) **Comprehensive Benchmarking and Revealed Limitations.** We benchmarked over 20 approaches across diverse architectural families on REAL-E. Our empirical findings reveal that state-of-the-art Transformer-based and Spectral GNN models struggle to generalize under conditions of pronounced correlation dynamics. We make our dataset and benchmark publicly available at [Real-E Link] and [Benchmark Link].

2 Proposed REAL-E

We build REAL-E from the European Network of Transmission System Operators for Electricity (ENTSO-E) Transparency Platform [11] updated to June 2024. We excluded time series with substantial missing values to maximize the coverage of generation types and temporal alignment across energy types. Please refer to our repository for more details.

We compare REAL-E with existing benchmarks including ETT [22], Electricity [1] and Solar-Energy [8] from three perspectives: **(1) Realistic Setting:** REAL-E reflects multiple challenges in energy systems on a real-world scale. It covers the full electricity lifecycle (see Tab. 1), including *Generation* (Energy Production and production forecasts), *Transmission* (Power Transfer over borders between

areas), *Balancing* (Regulation Energy to ensure the electrical transmission grid is in balance), *Market* (Trade and Price) and *Load* (Data about power consumption). The data spans multiple energy sources such as wind, solar, hydro, thermal, nuclear and pumped storage. **(2) Extensive Temporal and Geographic Coverage:** As detailed in Tab. 1, REAL-E spans a full decade from *2014 to 2024* of multiple temporal resolutions ranging from 15-min to hourly records. In addition, REAL-E provides broad geographic representation across 39 European countries, including energy-specific correlation structures (Sec 3). **(3) Rich Operational Metadata:** Each time series is accompanied by contextual information. For example, the dataset includes spatial descriptors (e.g., coordinates or bidding zones) that enable geographical clustering, as well as system-level attributes (e.g., transmission distance, voltage level, or grid topology) that provide insights into the operational context of the measurements (see Fig. 1).

3 Data Analysis

We reveal the time-varying properties of energy forecasting across temporal and cross-energy dimensions. For data analysis, we use the dataset recording electricity production from different energy sources in Germany and France(highlighted in Tab. 1).

Seasonal Complementary Patterns. Figure 2 illustrates the annual trends of two representative energy sources, brown coal and solar power in Germany. These sources exhibit a clear seasonal complementarity: brown coal reaches its peak during the autumn and winter months, while solar power peaks in summer. Germany's overall electricity demand throughout the year displays a distinct bimodal pattern, with each peak predominantly supported by a different energy source. Such short and long-term time-evolving correlation continually challenges the prevailing short-window time series models, exposing their limitations in capturing long-range and evolving dynamics.

Cross Energy Time-Evolving Correlation. We generalize the analysis to encompass *all major energy sources* in Germany. Figure 3

Table 1: Comparing with existing datasets, REAL-E offers longer duration (Dur.), finer resolution (Res.), longer sequences (Len.), more energy categories (EC) and broader country coverage (C).

	Name	Dur.	Res.	Len.	EC	C
Existing Benchmark	Solar	1 y	10 min	52,560	1	1
	Electricity	3 y	1 hour	26,304	1	1
	ETTh1/ETTh2	20 mon	1 hour	14,400	1	1
	ETTm1/ETTm2	20 mon	15 min	57,600	1	1
REAL-E						
Generation	Actual-ByType	9.5 y	15 min	>330k	20	39
	Actual-ByUnit	9.5 y	1 hour	8.7k	20	39
	Renewables-Forecast	9.5 y	15 min	>330k	3	39
	Capacity-Annual	9.5 y	1 year	10	20	39
Load	Actual	9.5 y	15 min	>330k	20	39
	Forecast-WeekAhead	9.5 y	1 day	3.4k	20	39
Market	Price-QuarterHourly	9.5 y	15 min	>330k	20	39
	Price-Hourly	9.5 y	1 hour	8.7k	20	39
Transmission	Capacity-Forecast	9.5 y	1 hour	8.7k	20	39
	Flow-Actual	9.5 y	1 hour	8.7k	20	39
Balancing	Energy-Activated	9.5 y	15 min	>330k	20	39
	System-Imbalance	9.5 y	1 hour	8.7k	–	39

reveals dynamic shifts in the inter-dependencies between multiple energy resources. These evolving correlation structures reflect fundamental changes in the internal coordination of the energy system. As the figure highlights, we observe a clear intensification of correlations in the upper-left block informed by biomass, brown coal and gas.

Quantify Time-Evolving Correlation To characterize such a time-evolving correlation pattern, we propose two metrics as follows:

Temporal Graph Volatility (TGV) quantifies the structural variation in the adjacency matrix A between two adjacent time steps t via the Frobenius norm. A higher value indicates more frequent structural transitions.

$$TGV = \sum_{t=0}^{T-1} \|A_{t+1} - A_t\|_F = \sum_{t=1}^{T} \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}^{t+1} - a_{ij}^t|^2}. \quad (1)$$

Graph Spectral Divergence (GSD) C_f indicates periods of structural volatility via spectral distances. It identifies correlation regime shifts, such as policy events or grid reconfigurations. We compute the Laplacian spectral distances of matrices as follows:

$$GSD = \|\lambda(L_t) - \lambda(L_{t-1})\|_2, L_t = D_t - A_t \quad (2)$$

where $D_t = \text{diag}(A_t \mathbf{1})$, A_t are the degree and adjacent matrix. The spectral topology shift is measured as the Euclidean distance between the sorted Laplacian eigenvalues.

Metrics in Practice We report average metric values across all time steps on existing benchmarks. For the existing datasets, Temporal Graph Volatility (TGV) achieves an average of 0.5822, approaching a maximum of 0.6245, while Graph Spectral Divergence (GSD) averages 1.6298 and peaks at 1.8027. In contrast, our dataset attains higher values, with TGV averaging 1.03335 and peaking at 1.0969, and GSD averaging 2.3237 and peaking at 2.7289. For empirical analysis, we focus on Germany and France as representative cases.

(a) **(b)**
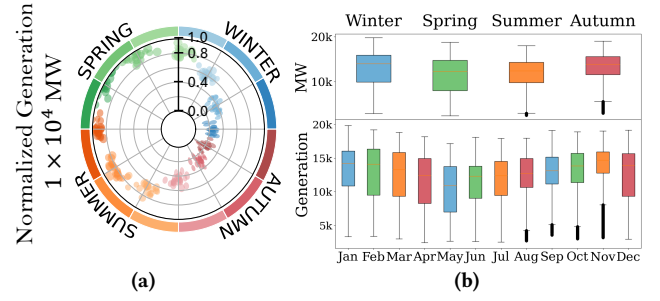
Figure 2: *We use a polar plot (a) and a boxplot (b) to illustrate seasonal generation patterns of two complementary energy sources in Germany: solar and brown coal. (a) Solar generation shows a clear seasonal trend, with higher output in spring and summer, peaking in June and reaching its minimum in December — primarily driven by solar irradiance. (b) In contrast, brown coal generation peaks in autumn and winter and drops in warmer months, reflecting increased demand during colder periods.*
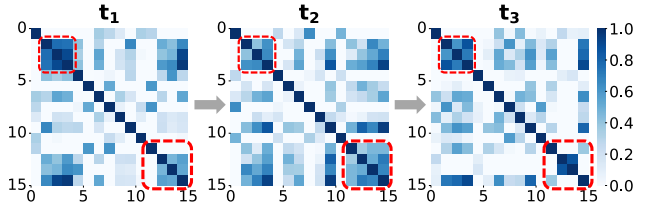
Figure 3: We present a series of Pearson correlation matrices across consecutive time windows. The bottom-right block highlights the strong, mutually inclusive correlations among lignite, gas and hard coal generation.

4 Experiments

In this section, we compare extensive approaches on both REAL-E and public benchmark, aiming to answer the following research questions: **RQ1:** How well do existing forecasting models generalize to the more complex, real-world structure of REAL-E? **RQ2:** Among different architectural paradigms, which models are most effective for forecasting under REAL-E's dynamic spatial-temporal structure?

Dataset & Metrics. We choose a subset of proposed datasets, REAL-E (*Germany, France*) and two public datasets (electricity and solar). It records electricity production in Germany and France in all generation types for all available years. We use the consistent data splits (70%/20%/10%) for all models and report *Mean Absolute Error (MAE)* and *Root Mean Squared Error (RMSE)* of forecasting horizons with 12 time steps on five random seeds with enabled early stop. Due to space constraints, we provide the details of data preprocessing and splitting in Link.

Hyperparameter Tuning. We performed a grid search over a wide range of hyperparameter values for all reported models. Due to limited space, we only report a subset of the tuned hyperparameters, which includes but not limited to: learning rate $10^{(-2\sim-4)}$, hidden embedding size $2^{8\sim10}$, batch size $2^{4\sim9}$, number of encoder and decoder layers $1 \sim 3$, number of attention heads $2 \sim 8$, kernel

sizes $3 \sim 7$ and rolling window sizes $12 \times (1 \sim 3)$. For the electricity, solar and ETT datasets, we report the best results obtained from [16, 19].

Baselines. We evaluate 20 baselines reported in Tab. 2, which includes three classical statistical methods [3, 4, 13], three MLP-based approaches [10, 14, 20] and several RNN/CNN-based architectures [2, 8, 12, 21]. We also include three advanced paradigms: 1) Transformer-based models [7, 15, 23, 24], which optimize temporal modeling via optimized attention mechanisms; 2) Spectral GNN models [5, 6, 19], which leverage nonlinear filtering in the spectral domain to capture spatial correlations; 3) Spatial graph-based approaches [9, 17, 18], which rely on localized message passing mechanisms to dynamically model and adapt to spatial dependencies.

Table 2: We report the average MAE and RMSE evaluated on REAL-E (subset) and compare them against existing benchmark datasets (Electricity and Solar). The * results are obtained from [16, 19]. Best model per benchmark highlighted in gray.

Model	Electricity 0.5391		Solar 0.6245		Germany 1.0969		France 0.9698	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Spatial GNN								
GWaveNet [18]	0.094	0.140	0.183	0.238	**0.013**	**0.028**	**0.012**	**0.025**
MTGNN [17]	0.077	0.113	0.151	0.207	**0.016**	**0.034**	**0.012**	**0.023**
TPGNN* [9]	**0.055**	**0.080**	**0.123**	0.214	**0.099**	**0.173**	**0.089**	**0.158**
Spectral GNN								
FourierGNN [19]	**0.051**	**0.077**	**0.120**	**0.162**	0.110	0.186	0.096	0.164
StemGNN* [5]	0.070	0.101	0.176	0.222	0.179	0.285	**0.148**	0.206
TGGC [6]	0.086	0.127	0.184	0.231	0.467	0.524	0.341	0.483
Transformer-based								
Autoformer [15]	**0.056**	**0.083**	**0.150**	**0.193**	0.204	0.376	0.165	0.263
FEDformer [24]	**0.055**	**0.081**	**0.139**	**0.182**	0.271	0.396	0.220	0.291
Informer [23]	0.070	0.119	0.151	0.199	0.283	0.324	0.137	0.217
Reformer [7]	0.078	0.129	0.234	0.292	0.297	0.361	0.141	0.233
RNN/CNN-based								
SFM [21]	0.086	0.129	0.161	0.283	0.184	0.245	0.156	0.238
TCN [2]	0.057	0.083	0.176	0.222	0.187	0.287	0.172	0.260
LSTNet [8]	0.075	0.138	0.148	**0.200**	0.193	0.346	0.177	0.263
DeepGLO [12]	0.090	0.131	0.178	0.400	0.264	0.372	0.181	0.246
MLP-based								
TimeMixer [14]	0.091	0.147	0.166	0.211	0.181	0.314	0.167	0.279
N-Beats [10]	0.067	0.126	0.162	0.203	0.189	0.325	0.174	0.300
DLinear* [20]	0.058	0.092	0.257	0.313	0.266	0.368	0.196	0.259
Statistic Methods								
VAR [13]	0.096	0.155	0.175	0.222	0.243	0.381	0.177	0.260
ARIMA [3]	0.107	0.169	0.185	0.244	0.275	0.319	0.164	0.281
S-ARIMA [4]	0.101	0.163	0.184	0.234	0.268	0.379	0.158	0.273

RQ1: Ranking Cross-Dataset Generalization Table 2 reports the average MAE and RMSE for a forecasting horizon of 12 time steps of partial results. Other detailed results and information can be found at our Github. We observed that a consistent trend emerges; many models fail to maintain performance across datasets with varying temporal dynamics and correlation complexity. For instance, the average MAE of the 4 Transformer-based models deteriorated from 0.117 to 0.215, reflecting an 85.4% increase in error. Spectral GNN, a method that consistently outperforms other approaches in existing benchmarks (highlighted in the Electricity and Solar datasets), experiences a performance degradation of approximately 16.20%.

RQ2: Rank for the Performance On the proposed dataset, the top three performing models are predominantly from the Spatial GNN category. Its design includes dynamic graph modeling that demonstrates the greatest robustness, consistently delivering stronger generalization across diverse datasets. This superiority can likely be attributed to the rich spatial correlations and temporal patterns present in the data, which enable the graph-based neurons to be effectively trained. In contrast, this advantage is less evident on traditional datasets. While Transformer-based models and certain Spectral GNNs perform comparably on earlier benchmarks, they fall short against Spatial GNN architectures on the proposed dataset—even under carefully tuned parameter settings.

Summary Our findings empirically reveal the limitation of the current transformer-based forecasting model: global attention mechanism struggles to generalize in the presence of rapidly shifting dependencies, which is crucial in real-world model deployment, while Spatial GNNs consistently outperform other models by explicitly modeling correlations through graph structures and dynamically adapting to evolving spatial patterns.

5 Conclusions

We evaluated the generalization ability of advanced models on real-world electricity datasets with challenging structural complexity. To support reproducibility, we have released all datasets and code under the Creative Commons Attribution 4.0 License (CC-BY 4.0), fully compliant with open data standards. Our work exposes current methodological limitations and underscores the urgent need for designs to effectively capture time-evolving dependencies in electricity forecasting. We believe our dataset lay a solid foundation for addressing this challenge and offer a viable direction for future research.

6 GenAI Usage Disclosure

In the paper writing, generative AI tools were used to assist with language polishing and LaTeX formatting. In algorithm development, we use AI to support debugging suggestions. The AI tools were not used to generate original scientific content, original implementation, or influence the interpretation of results. Any text generated or suggested by AI has been carefully reviewed and edited by the authors to ensure precision and clarity.

References

[1] Arthur Asuncion, David Newman, et al. 2007. UCI machine learning repository.
[2] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271* (2018).
[3] George E. P. Box and Gwilym M. Jenkins. 1970. *Time Series Analysis: Forecasting and Control.* Holden-Day, San Francisco.
[4] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung. 2015. *Time Series Analysis: Forecasting and Control.* Wiley.
[5] Liangzu Cao, Yixian Chen, and Junbo Li. 2020. Spectral temporal graph neural network for multivariate time-series forecasting. *Advances in Neural Information Processing Systems (NeurIPS)* (2020).
[6] Ming Jin, Guangsi Shi, Yuan-Fang Li, Bo Xiong, Tian Zhou, Flora D. Salim, Liang Zhao, Lingfei Wu, Qingsong Wen, and Shirui Pan. 2023. Towards Expressive Spectral-Temporal Graph Neural Networks for Time Series Forecasting. *arXiv preprint arXiv:2305.06587* (2023). https://doi.org/10.48550/arXiv.2305.06587 arXiv:2305.06587 [cs.LG] Revised version submitted 23 Feb 2025.
[7] Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The Efficient Transformer. In *International Conference on Learning Representations.*
[8] Guokun Lai, Wei-Cheng Chang, and Yiming et al. Yang. 2018. Modeling long- and short-term temporal patterns with deep neural networks. *ACM SIGIR* (2018), 95–104.

[9] Yijing Liu, Qinxian Liu, Jian-Wei Zhang, Haozhe Feng, Zhongwei Wang, Zihan Zhou, and Wei Chen. 2022. Multivariate Time-Series Forecasting with Temporal Polynomial Graph Neural Networks. In *Advances in Neural Information Processing Systems*, Vol. 36. https://papers.neurips.cc/paper/2022/file/7b102c908e9404dd040599c65db4ce3e-Paper-Conference.pdf

[10] Boris N. Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. 2020. N-BEATS: Neural Basis Expansion Analysis for Interpretable Time Series Forecasting. In *International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=r1ecqn4YwB

[11] ENTSO-E Transparency Platform. 2024. ENTSO-E Electricity Data Repository. https://transparency.entsoe.eu

[12] Rajat Sen, Haekyu Yu, and Inderjit S. Dhillon. 2019. Think Globally, Act Locally: A Deep Neural Network Approach to High-Dimensional Time Series Forecasting. In *Advances in Neural Information Processing Systems*, Vol. 32.

[13] Christopher A. Sims. 1980. Macroeconomics and Reality. *Econometrica* 48, 1 (1980), 1–48.

[14] Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y. Zhang, and Jun Zhou. 2024. TimeMixer: Decomposable Multiscale Mixing for Time Series Forecasting. In *International Conference on Learning Representations (ICLR)*. Poster.

[15] Haixu Wu, Jiehui Xu, and Jianmin Wang. 2021. Autoformer: Decomposition transformers with autocorrelation for long-term series forecasting. *Advances in Neural Information Processing Systems (NeurIPS)* (2021).

[16] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. 2023. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. *International Conference on Learning Representations (ICLR)* (2023).

[17] Yingheng Wu, Kristjan Greenewald, Jonathan Hertrich, et al. 2020. Connecting the dots: Multivariate time series forecasting with graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems* (2020).

[18] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. 2019. Graph wavenet for deep spatial-temporal graph modeling. In *Proceedings of International Joint Conference on Artificial Intelligence*. 1907–1913.

[19] Kun Yi, Qi Zhang, Wei Fan, Hui He, Liang Hu, Pengyang Wang, Ning An, Longbing Cao, and Zhendong Niu. 2023. FourierGNN: Rethinking Multivariate Time Series Forecasting from a Pure Graph Perspective. In *Thirty-seventh Conference on Neural Information Processing Systems*.

[20] Ailing Zeng, Wenqian Zhang, Yuxuan Zheng, Yixuan Xu, Jiajun Yan, Weinan Xu, Yong Zhang, and Xiaotian Ma. 2023. DLinear: A Baseline Algorithm for Time Series Forecasting. *arXiv preprint arXiv:2301.07021* (2023).

[21] Liheng Zhang, Charu C. Aggarwal, and Guo-Jun Qi. 2017. Stock Price Prediction via Discovering Multi-Frequency Trading Patterns. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, 2141–2149. https://doi.org/10.1145/3097983.3098061

[22] Haoyi Zhou, Shanghang Zhang, Jie Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. 2020. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 12 (2020), 11106–11115.

[23] H. et al. Zhou. 2021. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of AAAI Conference on Artificial Intelligence*. 11106–11115.

[24] Tian Zhou, Ziqing Ma, Qingsong Wen, Ruixin Jin, Jianyuan Zhu, Xinyu Zhang, and Liang Sun Wang. 2022. FEDformer: Frequency Enhanced Decomposed Transformer for Long-term Series Forecasting. In *International Conference on Machine Learning (ICML)*. 11214–11225.