Wind Energy Prediction Guided by Multiple-Location Weather Forecasts

Charalampos Symeonidis and Nikos Nikolaidis

Artificial Intelligence and Information Analysis Lab, Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece
{charsyme,nnik}@csd.auth.gr

Abstract. In recent years, electricity generated from renewable energy sources has become a significant contributor to power supply systems over the world. Wind is one of the most important renewable energy sources, thus accurate wind energy prediction is a vital component of the management and operation of electric grids. This paper proposes a novel method for wind energy forecasting, which relies on a novel variant of the scaled-dot product attention mechanism, for exploring relations between the generated energy and a set of multiple-location weather forecasts/measurements. The conducted experimental evaluation on a dataset consisting of the hourly generated wind energy in Greece along with hourly weather forecasts for 18 different locations, demonstrated that the proposed approach outperforms competitive methods.

Keywords: Wind energy prediction · Renewable energy · Scaled-dot product attention

1 Introduction

Electricity generated from renewable energy sources, has been proven an effective solution against the energy shortage and the environmental pollution caused by conventional (e.g. fossil fuels) energy production methods. Wind energy is one of the most important renewable energy sources. However, wind energy is a highly fluctuating resource, mainly due to the respective unpredictable nature of weather conditions, mainly wind speed and direction. Accurate wind energy prediction is vital for lowering the impact of uncertainty, thus achieving a smoother integration of the respective energy sources (wind farms/parks) into the grid.

Most approaches for wind energy generation prediction can be classified based on either the applied methodology or the time horizon of the prediction [4]. Based on the predictive horizon, the methods are usually classified into the following four categories:

- Very short-term (up to 30 minutes) forecasting
- Short-term (30 minutes to 6 hours) forecasting
- Medium-term (6 hours to 1 day) forecasting
- Long-term (1 day to a month) forecasting

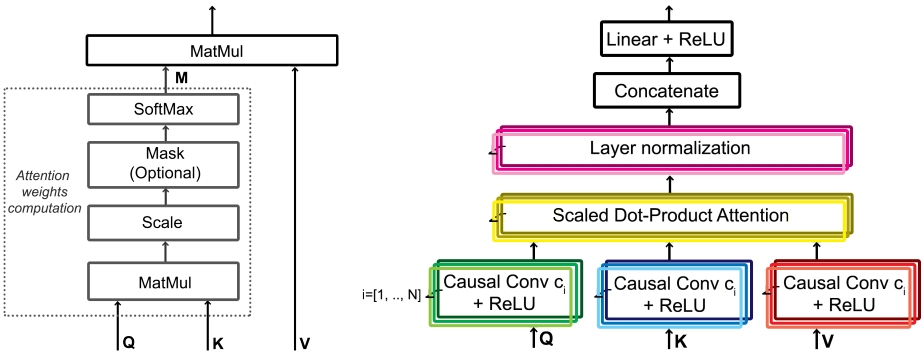
Regarding the applied methodology, wind energy prediction methods are categorized as physical or statistical. The first, explore the physical relations

between the wind speed, climate conditions, topological information and the energy generated from the corresponding wind power plant. Usually, physical models [2] [19] rely on numerical weather prediction models (NWP) that simulate atmospheric physics by utilizing boundary conditions and physical laws, in order to determine wind speed. The predicted wind speed is then used along with the related wind turbine power curve, usually provided by the turbine manufacturer, in order to predict the generated wind energy. Physical models are generally suitable for long-term wind energy forecasting, but their short-term precision remains low.

Statistical models/approaches [18], [6] are more appropriate for short-term wind energy prediction compared to physical models. Their aim is not to describe the physical steps involved in the wind power conversion process, but to directly obtain wind energy predictions, by exploring statistical relations between historical wind energy data and other relevant input data. A sub-class of statistical models are Deep Learning (DL) based methods. In recent years, several DL-based methods, including approaches utilizing convolutional neural networks (CNNs) [17] [20], autoencoders [16], recurrent neural networks (RNNs) [10] and spatio-temporal attention-based networks [11], have been proposed as suitable solutions for wind energy forecasting.

In [17], the authors mapped data collected from wind turbines into a grid space, which they called scene. The scene time series is a multi-channel image, which represents the spatio-temporal characteristics of wind in a certain area and time. Therefore, they developed a DL model based on CNN to extract features from these images, in order to predict the generated energy. The results showed that the proposed model achieves better accuracy than other existing methods. The authors in [11] proposed a sequence-to-sequence model for multi-step-ahead wind power forecasting, namely prediction of multiple future wind power values. The model architecture consists of two groups of Gated Recurrent Units (GRU) blocks, which work as encoder and decoder. The authors proposed the Attention-based GRU (AGRU) for embedding the task of correlating different forecasting steps by hidden activations of GRU blocks. The AGRU model achieved top performance against other competitive wind energy forecasting methods. In [13], the authors modified the N-BEATS [12] model towards making it suitable for the f wind energy forecasting task and proposed a loss function capable to confront the issue of forecast bias. The method, mostly evaluated on very-short term wind energy prediction datasets, was able to compete against other state-of-the-art approaches and even outperform them in terms of accuracy in most cases. In [20], the authors proposed a DL-based architecture, based on Temporal Convolutional Networks (TCNs) for short-term wind energy prediction. An experimental evaluation of the method on a dataset consisting of 5000 hourly wind power and meteorological data samples collected from a single wind energy power plant, showed promising results against the competitive methods. In [16] the authors proposed an architecture named SIRAE (Staked Independently Recurrent Auto-Encoder), suitable for ultra-short wind energy forecasting. According to the authors, this approach can accommodate a large volume of data in an effi-

cient manner while also overcoming the effects of random changes in the natural environment. To verify the effectiveness and stability of SIRAE, two comparative experiments, in which it outperformed several popular models, were conducted. In [10], the authors proposed DWT_LSTM, a short-term wind energy forecasting method based on Discrete Wavelet Transform (DWT) and Long Short-Term Memory (LSTM) networks. The method adopts a divide and conquer strategy, in which DWT is used to decompose original wind power data into sub-signals, while several independent LSTMs are employed to approximate the temporal dynamic behaviors of these sub-signals. The proposed method achieved top prediction accuracy rates against other state-of-the-art methods.



(a) Scaled Dot-Product Attention

(b) Multi-kernel convolutional scaled dot-product attention.

Fig. 1: (a): Scaled Dot-Product Attention, (b): the novel Multi-kernel convolutional scaled dot-product attention. c_i denotes the i-th convolutional kernel size, employed in the temporal domain whereas N denotes the number of convolutional kernels.

The method proposed in this paper relies on the scaled dot-product attention mechanism, initially proposed in [15]. Several methods [9] in the relevant literature have applied this mechanism in time-series forecasting for exploring temporal dependencies. More recently, spatio-temporal attention networks [3] have been introduced to wind energy forecasting, aiming to predict the generated energy of multiple, spatially neighboring, wind farms. Compared to methods in the relevant literature, our approach provides the following contributions:

– Utilizes past and future wind-related weather measurements/forecasts from multiple locations, aiming to explore temporal patterns between the time instances in the past and prediction windows. In addition, the method is able to explore pseudo-spatial relations between the energy generation location/region and the multiple locations of the weather measurements/forecasts, aiming to find how the weather in each of the locations for which weather data are available affects the energy generation prediction in the region under study. To achieve this, the method doesn't rely on any spatial information (e.g., geographic coordinates, or geographic distances) as input.

- Proposes a variant of scaled dot-product attention, which employs *causal convolutions* of multiple kernel sizes, for exploring context-based similarities, instead of point-based similarities, as proposed in [15]. To the best of our knowledge, this approach is novel. Indeed, although a similar approach has been proposed in [8], the corresponding authors employed *Causal Convolutions* with single sized kernels.
- Achieves top results, compared to SoA time-series forecasting methods, on a dataset suitable for wind energy generation prediction in Greece at hour-level resolution.

2 Proposed Method

2.1 Problem Statement and Notations

The problem of wind energy forecasting that is addressed in this paper can be formulated as:

$$\hat{\mathbf{E}^f} = g(\mathbf{E}^h, \mathbf{W}^h, \mathbf{W}^f) \tag{1}$$

In this equation $\mathbf{E}^h \in \mathbb{R}^{1\times H\times 1}$ corresponds to the past/history (h: history) wind energy measurements of a single region or power plant, H being the size of the past time window. Moreover, $\mathbf{W}^h \in \mathbb{R}^{B\times H\times D_{w^h}}$ corresponds to past weather measurements which are provided for B distinct locations or regions and D_{w^h} is the number of input weather variables, for the past. Also, $\mathbf{W}^f \in \mathbb{R}^{B\times F\times D_{w^f}}$ corresponds to weather forecasts (predictions in the future), where F is the size of the future time window and D_{w^f} is the number of input weather variables, for the future. Finally, $\hat{\mathbf{E}}^f \in \mathbb{R}^{1\times F\times 1}$ corresponds to the wind energy predictions that are generated by the method for the region of interest.

In short, given past energy measurements for a region or location and wind-related weather data from B distinct locations, our aim is to find how the energy generation is related to the weather on each of the B locations. Once those pseudo-spatial relations are estimated, wind energy predictions can be obtained by exploring temporal patterns between the past weather measurements and weather forecasts.

Adopting the typical attention mechanism [14], the single-time step prediction $\hat{e}^f$, namely one of the elements of $\hat{\mathbf{E}}^f = [\hat{e}_1^f, ..., \hat{e}_F^f]$ can be defined as:

$$\begin{gathered}\hat{e}^f = \mathbf{C}^T \sum_{j=1}^{H} \alpha_j \mathbf{e}_j^{h,r} \\ \text{where } \sum_{j=0}^{H} \alpha_j = 1\end{gathered} \tag{2}$$

In the above formulas, $\boldsymbol{\alpha} \in \mathbb{R}^{1\times H}$ corresponds to the attention weights, $\mathbf{e}_j^{h,r} \in \mathbb{R}^{1\times D_{e^{h,r}}}$ corresponds to the hidden representations (r: representations) of past energy measurements at the j^{th} time instance in the past. $D_{e^{h,r}}$ corresponds to size each hidden representation. Moreover, $\mathbf{C} \in \mathbb{R}^{D_{e^{h,r}}\times 1}$ are learnable parameters of a linear operator. In this formulation, wind energy is predicted based

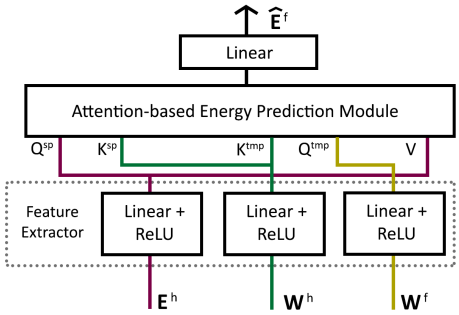


Fig. 2: The architecture of the proposed wind energy prediction method.

on the temporal patterns imposed by attention weights $\boldsymbol{\alpha}$, between the time step being predicted and past energy measurements (more specifically their internal representations) within the respective temporal window. It shall be noted that a multi-time step prediction formulation would involve a matrix $\mathbf{A} \in \mathbb{R}^{1\times F\times H}$ rather than $\boldsymbol{\alpha}$. Our objective is to explore, the previously described, pseudo-spatial and temporal relations between $\mathbf{E}^h$, $\mathbf{W}^h$ and $\mathbf{W}^f$ in order to efficiently approximate $\mathbf{A}$.

2.2 Multi-Kernel Convolutional Scaled Dot-Product Attention

The Scaled Dot-Product Attention, was presented in [15] and formulated as follows:

$$Attention(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathbf{M}\mathbf{V}, \tag{3}$$

where

$$\text{where}\quad \mathbf{M} = softmax(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{D_K}}) \tag{4}$$

$\mathbf{Q} \in \mathbb{R}^{N_Q\times D_Q}$, $\mathbf{K} \in \mathbb{R}^{N_K\times D_Q}$ and $\mathbf{V} \in \mathbb{R}^{N_K\times D_V}$ are the queries, keys and values respectively. Queries and keys have a dimension of D_K, while values have a dimension of D_V. N_Q is the number of queries while N_K is the number of keys and values. An illustration of the mechanism is depicted in Fig. 1a. Multi-head attention was also proposed in [15], allowing various attention mechanisms, including scaled dot-product attention, to run in parallel. To this end, instead of performing a single attention computation on queries, keys, and values of size D_L, the authors proposed their transformation with N independently learned linear projections. The attention computation is then performed, in parallel, on those N projected queries, keys, and values. More specifically, the multi-head attention module can be formulated as:

$$Multihead(\mathbf{Q},\mathbf{K},\mathbf{V}) = [\mathbf{p}_1,...,\mathbf{p}_N]\mathbf{S}^O, \tag{5}$$

$$\mathbf{p}_i = Attention(\mathbf{Q}\mathbf{S}_i^Q, \mathbf{K}\mathbf{S}_i^K, \mathbf{V}\mathbf{S}_i^V). \tag{6}$$

In this formulation, $\mathbf{S}_i^Q \in \mathbb{R}^{D_L\times D_K}$, $\mathbf{S}_i^K \in \mathbb{R}^{D_L\times D_K}$, $\mathbf{S}_i^V \in \mathbb{R}^{D_L\times D_V}$, $\mathbf{S}_i^O \in \mathbb{R}^{ND_V\times D_L}$ are projection parameter matrices, N is the number of heads, $D_K = D_V = \frac{D_L}{N}$, and the operator [...] implies concatenation.

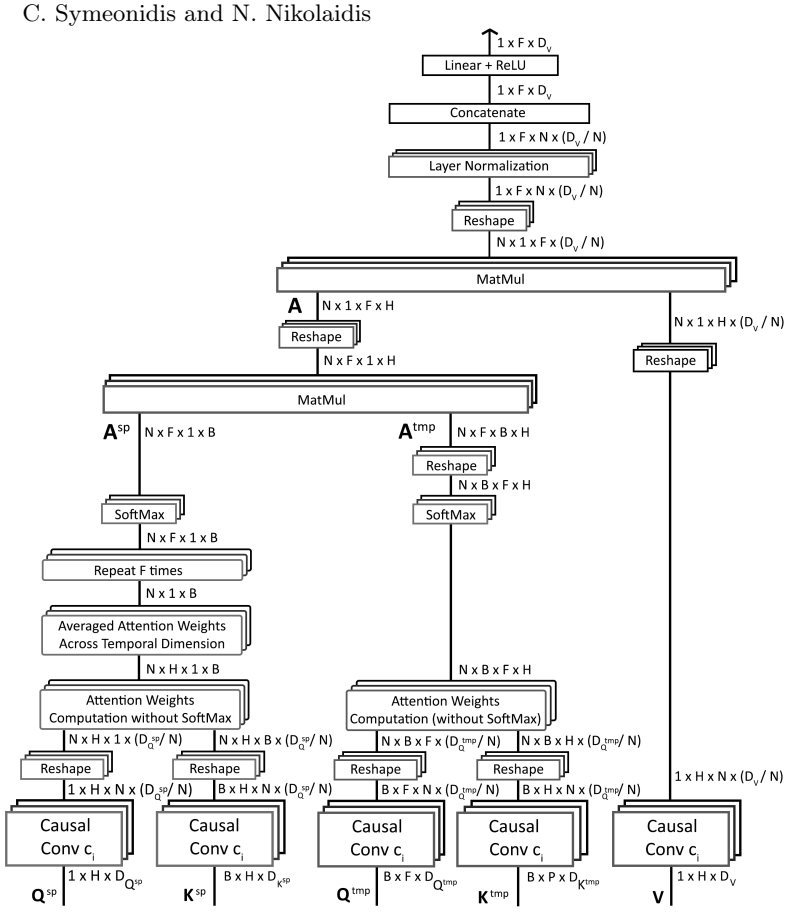


Fig. 3: The architecture of the core attention-based energy prediction module.

On the original formulation, the scaled dot-product attention was designed to explore point-wise similarities between queries and keys. However, in most time-series analysis tasks, information regarding the surrounding context of observed points is vital for exploring patterns among the series. The authors in [8], were able to employ *causal convolutions* of kernel size c to transform inputs into queries and keys. Thus, local context was exploited in the query-key matching, improving the way temporal patterns among the corresponding time series are explored. The authors experimented with various values of c in order to find the optimal one. To avoid selecting a specific kernel size, as well as for allowing the method to detect patterns in various kernel sizes, we propose the multi-kernel convolutional scaled dot-product attention. In our formulation, *causal convolutions* with N different kernel sizes are applied on $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$, resulting in N heads. The scaled dot-product attention is computed separately for each head. Layer normalization is then applied to the output of each head. Finally, the outputs are concatenated and projected, resulting in the final values, as depicted in Fig. 1b.

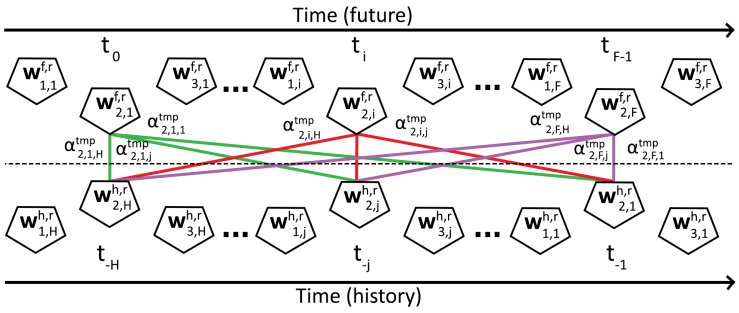


Fig. 4: The temporal attention mechanism captures correlations in weather forecasts between the time instances in the prediction window and the history (past) window. In this example, the number of weather forecasts/measurements B is set to 3.

2.3 Model Architecture

The overall architecture of our proposed method is depicted on Fig 2. The method receives as input $\mathbf{E}^h$, $\mathbf{W}^h$ and $\mathbf{W}^f$, and process those modalities through linear layers with the Rectified Linear Unit (ReLU) as activation function. Then, the hidden representations of all modality are fed into the Attention-based Energy Prediction module. Its architecture is depicted on Fig 3. The module is motivated by the typical attention mechanism, defined in Eq. 2, utilizing the multi-kernel convolutional scaled dot-product attention, previously described in Section 2.2. Its aim is to generate future energy representations, based on (i) temporal relations within past and future weather predictions, (ii) pseudo-spatial relations between the region of wind energy prediction and the locations of the weather forecasts. The temporal relations are imposed by $\mathbf{A}^{tmp} \in \mathbb{R}^{N\times B\times F\times H}$. The formulation of $\mathbf{A}^{tmp}$ involves a query-key matching of $\mathbf{W}^{f,r}$ and $\mathbf{W}^{h,r}$. An illustration of the described temporal attention mechanism is depicted on Fig. 4. The pseudo-spatial relations are imposed by $\mathbf{A}^{sp} \in \mathbb{R}^{N\times F\times 1\times B}$ and in the query-key matching $\mathbf{E}^{h,r}$ and $\mathbf{W}^{h,r}$ are involved. Illustrations of the described pseudo-spatial relations are depicted on Fig. 5. The final attention weights $\mathbf{A}$ can be defined as:

$$\mathbf{A} = u(\mathbf{A}^{sp} \otimes u(\mathbf{A}^{tmp})) \tag{7}$$

where $\mathbf{A} \in \mathbb{R}^{N\times 1\times F\times H}$ and $u(.)$ denotes a tensor reshaping function. In particular, element $a_{n,1,j,l}$ of $\mathbf{A}$ is computed as:

$$\begin{gathered} a_{n,1,l,j} = \sum_{i=1}^{B} \alpha^{sp}_{n,l,1,i} \cdot \alpha^{tmp}_{n,i,l,j} \\ \text{where } \sum_{i=1}^{B} \alpha^{sp}_{n,l,1,i} = 1,\ \sum_{j=1}^{H} \alpha^{tmp}_{n,i,l,j} = 1,\ 1 \le l \le F,\ 1 \le j \le H,\ 1 \le n \le N \end{gathered} \tag{8}$$

The output of the attention-based module are the hidden representations of the wind energy values $\hat{\mathbf{E}}^{f,r}$. Finally, a lineal layer is applied for generating the wind energy predictions.

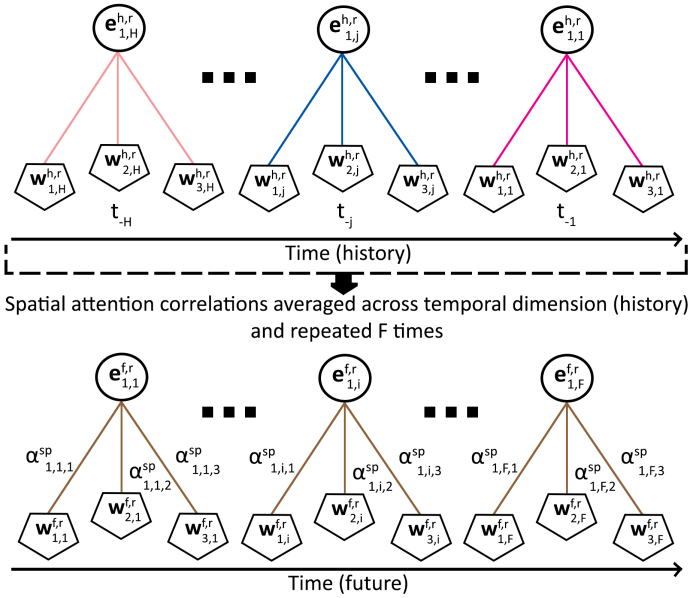


Fig. 5: The pseudo-spatial attention mechanism captures correlations between the generated energy and the multiple-location weather forecasts in the history window. In this example, the number B of locations for which weather forecasts/measurements are available is set to 3.

3 Experimental Evaluation

3.1 Dataset Description

The dataset employed in the experimental evaluation was initially proposed in [7]. It consists of (i) hourly wind energy generation data for Greece (the entire country), collected by the European Network of Transmission System Operators for Electricity[1], and (ii) hourly wind-related weather data, which correspond to 18 separate locations in Greece, retrieved by the Storm Glass weather API [2]. The weather data consist of forecasts regarding the wind speed, wind direction and gust. The dataset spans the period 2017-2020. The training set contains data for the period 2017-2019, while the data from the final year form the test set.

The provided wind energy generation values are not normalized/standardized and no information is provided for wind energy generation bounds within regular time intervals (e.g. per annum). Tab 1 depicts the large differences between various statistics of the generated energy at annual level. This is indeed a common real-world issue, since the number of wind stations/turbines of a region changes over time (usually increases due to the installation of new ones, as is obviously the case for Greece) and no information regarding this number is provided at

[1] https://transparency.entsoe.eu
[2] ttps://stormglass.io/

country level in regular intervals. A method employed to predict power generation under these circumstances must have a high generalization ability, and be able to overcome such significant data distribution shifts. Being fair to the corresponding dataset split, wind energy data used as input were scaled explicitly based on the minimum and maximum energy values of the training set. However, the metrics used in the evaluation were computed on output data (predictions) that were re-scaled on the min/max values of the overall dataset.

Table 1: Statistics derived from wind energy generation data (in MW) for Greece in the period 2017-2020.

Year	**Mean**	**Std.**	**Median**	**Max.**
2017	482.312	336.052	369.0	1702.0
2018	554.554	384.459	466.0	1695.0
2019	662.747	457.254	545.0	2107.0
2020	849.010	595.726	696.0	2630.0

Based on this dataset, two evaluation/benchmarking scenarios were formed. The first scenario assumes a forecast horizon of one hour and historic (past) data availability of up to 120 hours. Weather data are available for both input (past measurements, 120 measurements) and target (future forecasts) windows. Past energy production measurements for 72 hours are provided as input, starting 48 hours prior to the target period. This 48-hour gap in past energy data is due to the fact that measurements are not released immediately by the transmission system operators, i.e. it reflects the real situation. The second scenario assumes a 24-hour forecast horizon, in 1-hour intervals and data availability of up to 384 hours. In a similar fashion to the first scenario, weather data are available for both input (past measurements, 384 measurements) and target (future forecasts, 24 values) windows. Past energy production measurements for 336 hours are provided as input, starting 48 hours prior to the target period (48-hour gap).

3.2 Baseline Methods

Three SoA time series forecasting methods were trained and evaluated on each of the described wind energy prediction scenarios. The first employed method is N-BEATS [12], implemented by [5]. Compared to the original method, the implemented model can receive as input both historic wind energy measurements, as well as historic wind-related weather data for the corresponding time instances. This is accomplished by flattening the model inputs to a 1-D series.

The performance of a deterministic implementation of DeepTCN [1] was also evaluated in our scenaria. In addition to the historic wind energy measurements, we provide as input the historical wind-related weather data, since the method allows the use of past covariates. Finally, an implementations of TFT [9] was also employed in our experimental evaluation. Information regarding the type of input covariates of each method is provided in Table 2. In particular, TFT and our method are the only ones incorporating future weather forecasts as input.

Aiming to achieve a fair comparison, the weather data from all 18 locations, as well as weather data corresponding formed as an aggregated weather forecast from those 18 locations, were fed as input covariates to the three baseline methods. Furthermore, all four methods, including ours, were trained incorporating the 48-hour gap within the scenario specific prediction window (i.e., in the first scenario the methods were trained using a 49 prediction window). However, the predictions corresponding to the 48-hour gap were excluded during the evaluation process.

All methods, including our proposed method, were trained of 8 epochs. The learning rate was initially set to 5×10^{-4} and it was decreased twice by multiplying it with 0.1 at epochs 4 and 6, respectively. Regarding our proposed method, the number of kernels N in the multi-kernel convolutional scaled dot-product attention was set to 6, using 1, 3, 5, 9, 13 and 17 sized kernels. In addition, $D_{e^{h,r}}$, $D_{w^{h,r}}$, $D_{w^{f,r}}$ and $D_{e^{f,r}}$ were set to 66.

Table 2: Covariates used as input for each of the compared methods.

Method	**Covariates**	
	Past Weather Measurements	**Future Weather Forecasts**
N-BEATS [12]	✓	
DeepTCN [1]	✓	
TFT [9]	✓	✓
Ours	✓	✓

3.3 Experimental Results

This subsection presents the results of the forecasting experiments for each method along with a commentary on the findings. To be consistent with the literature [10] [20], Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) were used to measure the performance of the models. Each experiment was executed four times, and the mean value and standard deviation are reported.

Table 3: MAE and RMSE values for the two evaluation scenarios.

Method	**Scenario 1**		**Scenario 2**	
	MAE	**RMSE**	**MAE**	**RMSE**
N-BEATS	0.201 ± 0.003	0.262 ± 0.006	0.202 ± 0.002	0.262 ± 0.003
DeepTCN	0.189 ± 0.008	0.243 ± 0.017	0.226 ± 0.027	0.301 ± 0.030
TFT	0.113 ± 0.008	0.154 ± 0.012	0.118 ± 0.008	0.159 ± 0.012
Ours	**0.103 ± 0.002**	**0.139 ± 0.003**	**0.085 ± 0.001**	**0.112 ± 0.003**

Table 3 shows the MAE and RMSE of all compared methods for the two wind energy prediction scenarios. In both scenarios, methods which employ future weather forecasts as input covariates, i.e. TFT and the proposed method, demonstrate significant performance gains. In both scenarios our proposed method

achieved top results, compared to the three baseline methods. In particular, more significant results were attained in the second scenario achieving mean MAE and RMSE, among 4 experiments, of 0.085 and 0.112, respectively.

It is worth noting that the performance of our proposed method was better in the second scenario, compared to the first, in all metrics. This behaviour is exactly the opposite compared to the rest of the methods, where their performances downgraded in the second scenario. The improved performance of the proposed method, on a scenario in which data from a larger temporal window were used as input, highlights that the implemented temporal attention-based mechanism is able to effectively capture relations between distant samples within the sequences. Future work will focus on conducting more experiments, in respect to the size of input and prediction windows, as well as to extend the method, aiming to process and predict wind energy time-series from multiple stations or regions.

4 Conclusions

Energy generation from wind exhibits inherent uncertainties due to its intermittent nature. The accurate wind energy prediction can assist its integration, operation and management within the electric grids. This paper proposes a novel wind energy forecasting method, which relies on a novel variant of the scaled-dot product attention mechanism, for exploring relations between the generated energy and a set of multiple-location weather forecasts/measurements. The results of the conducted preliminary experimental evaluation against SoA time-series forecasting methods on a dataset consisting of the hourly generated wind energy in Greece, highlighted the potential of the proposed method.

Acknowledgements This work is co-financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH - CREATE - INNOVATE (project code: T2EDK-03048).

References

1. Chen, Y., Kang, Y., Chen, Y., Wang, Z.: Probabilistic forecasting with temporal convolutional neural network. Neurocomputing **399**, 491–501 (2020)
2. Focken, U., Lange, M., Waldl, H.P.: Previento-a wind power prediction system with an innovative upscaling algorithm. In: Proc. of the European Wind Energy Conference. vol. 276 (2001)
3. Fu, X., Gao, F., Wu, J., Wei, X., Duan, F.: Spatiotemporal attention networks for wind power forecasting. In: 2019 International Conference on Data Mining Workshops (ICDMW). pp. 149–154. IEEE (2019)
4. Hanifi, S., Liu, X., Linand, Z., Lotfian, S.: A critical review of wind power forecasting methods—past, present and future. Energies **13** (2020)

5. Herzen, J., Lässig, F., Piazzetta, S.G., Neuer, T., Tafti, L., Raille, G., Van Pottelbergh, T., Pasieka, M., Skrodzki, A., Huguenin, N.: Darts: User-friendly modern machine learning for time series. Journal of Machine Learning Research **23**(124), 1–6 (2022)
6. Hodge, B.M., Zeiler, A., Brooks, D., Blau, G., Pekny, J., Reklatis, G.: Improved wind power forecasting with arima models. Computer Aided Chemical Engineering **29**, 1789–1793 (2011)
7. Kouloumpris, E., Tsoumakas, G.: Short-term renewable energy forecasting in greece using prophet decomposition and tree-based ensembles. In: Database and Expert Systems Applications-DEXA 2021 Workshops: BIOKDD, IWCFS, MLKgraphs, AI-CARES, ProTime, AISys 2021, Virtual Event, September 27–30, 2021, Proceedings. p. 227. Springer Nature (2021)
8. Li, S., Jin, X., Xuan, Y., Zhou, X., Chen, W., Wang, Y.X., Yan, X.: Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. In: Advances in Neural Information Processing Systems. vol. 32 (2019)
9. Lim, B., Arık, S.O., Loeff, N., Pfister, T.: Temporal fusion transformers for interpretable multi-horizon time series forecasting. International Journal of Forecasting **37**, 1748–1764 (2021)
10. Liu, Y., Guan, L., Hou, C., Han, H., Liu, Z., Sun, Y., Zheng, M.: Wind power short-term prediction based on lstm and discrete wavelet transform. Applied Sciences **9**(6) (2019)
11. Niu, Z., Yu, Z., Tang, W., Wu, Q., Reformat, M.: Wind power forecasting using attention-based gated recurrent unit network. Energy **196**, 117081 (2020)
12. Oreshkin, B.N., Carpov, D., Chapados, N., Bengio, Y.: N-BEATS: neural basis expansion analysis for interpretable time series forecasting. In: Proc. of the Inter. Conf. on Learning Representations (2020)
13. Putz, D., Gumhalter, M., Auer, H.: A novel approach to multi-horizon wind power forecasting based on deep neural architecture. Renewable Energy **178**, 494–505 (2021)
14. Shih, S.Y., Sun, F.K., Lee, H.y.: Temporal pattern attention for multivariate time series forecasting. Machine Learning **108**, 1421–1441 (2019)
15. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Proc. of the Inter. Conf. on Neural Information Processing Systems (NIPS) (2017)
16. Wang, L., Tao, R., Hu, H., Zeng, Y.R.: Effective wind power prediction using novel deep learning network: Stacked independently recurrent autoencoder. Renewable Energy **164**, 642–655 (2021)
17. Yu, R., Liu, Z., Li, X., Lu, W., Ma, D., Yu, M., Wang, J., Li, B.: Scene learning: Deep convolutional networks for wind power prediction by embedding turbines into grid space. Applied Energy **238**, 249–257 (2019)
18. Zhang, J., Yan, J., Infield, D., Liu, Y., Lien, F.S.: Short-term forecasting and uncertainty analysis of wind turbine power based on long short-term memory network and gaussian mixture model. Applied Energy **241**, 229–244 (2019)
19. Zhao, J., Guo, Z.H., Su, Z.Y., Zhao, Z.Y., Xiao, X., Liu, F.: An improved multi-step forecasting model based on wrf ensembles and creative fuzzy systems for wind speed. Applied Energy **162**, 808–826 (2016)
20. Zhu, R., Liao, W., Wang, Y.: Short-term prediction for wind power based on temporal convolutional network. Energy Reports **6**, 424–429 (2020)