A Neural Approach to Wind Speed Modeling

Jack Hammer, Luana Rampelotti, Wilhem Hector
MIT Mechanical Engineering
Cambridge, MA, United States

ABSTRACT

Modeling the long-term behavior of wind speed on a specific project site can streamline wind farm development worldwide. This investigation used machine learning to reconstruct long-term hourly wind speed data and to forecast a day ahead of fed data. 20 years of reanalysis and 1 year of measured wind data were used to train a generative model (CNN-LSTM) and two regression models (RBFN & CNN). They were deployed to forecast January 1, 2021, and reconstruct April 2024. Results show that forecasting is unfeasible beyond 10 hours due to error stack-up. On the other hand, the RBFN and CNN can reconstruct hourly data with a mean absolute percentage error of less than 20%. These findings are consistent with [1] and suggest that local long-term wind speed prediction is achievable using neural networks.

Keywords: Wind farm modeling, Weather prediction, Wind Speed

INTRODUCTION

Renewable energy development continues to be a critical strategy for agencies attempting to mitigate climate change and push society towards a more green state. Wind energy, in particular, represents a significant portion of the global renewable energy approach. With increasing investments and technological advancements, accurate wind speed prediction is a fundamental part of the successful development, design, and operation of modern wind farm infrastructure.

Wind speed variability, however, poses a major challenge to these developers. Traditional methods to make these predictions prove incredibly difficult and are limited by their inability to capture complex atmospheric dynamics and local meteorological variations. These methods rely heavily on numerical weather prediction (NWP) models that simplify atmospheric processes into deterministic equations. These approaches struggle with the chaotic nature of weather and often require vast computational resources. These limitations highlight the need for new methods and models to be implemented.

Machine learning algorithms have offered a new way to model and understand these complex dynamics. By using highly advanced interconnected models and massive datasets, great strides have been made in the extraction of intricate patterns that traditional statistical methods cannot discern. ML methods such as artificial neural networks (ANNs), support vector machines (SVMs), and more recently, hybrid models combining convolutional neural networks (CNNs) with long short-term memory (LSTM) layers as shown in [2], have demonstrated significant potential in accurately predicting wind speeds.

This research attempts to leverage reanalysis data and hybrid modeling techniques to predict future weather patterns and apply regression models to reconstruct long-term wind speed data. The study provides insight into the capabilities and limitations of simple neural network techniques as applied to wind speed prediction.

METHODS

We performed one generative and two regression tasks in this study. We first trained a hybrid Convolutional and Long Short Term Memory neural network (CNN-LSTM) on 20 years of atmospheric data and deployed it to forecast the month of January 2021. Then, we trained two regressors on 1 year of atmospheric data and deployed it to reconstruct the month of April 2024. The regressors were a 1D Convolutional Neural Network (CNN), and a Radial Basis Functional Neural Network (RBFN). While the training data for the generative model and the regressors have the same origin, it is important to highlight the specific processing steps undertaken before training each.

Data Processing

This investigation used environmental data obtained from the European Centre for Medium-Range Weather Forecasts (ECMWF) ERA5 reanalysis dataset. ERA5 is a

global atmospheric reanalysis product that combines measured data and estimated parameters to create coherent and consistent long-term datasets [3]. We used 4 ERA5 datasets corresponding to 4 stations of interest around our target site. Each of these datasets contained hourly environmental data including the wind speed, wind direction, temperature, pressure, and the date of collection. The ERA5 station data extends from January 2000 to April 2024. Our target site contains measured wind speed data from a meteorological mast controlled by a developer. However, the latter only had data from March 2023 to April 2024. Figure 1 showcases the relative locations of the ERA5 stations and the target site.

Figure 1: Geographical location of the 4 ERA5 stations and the target wind farm site. Each ERA5 dataset contains hourly environmental data including wind speed, wind direction, temperature, pressure, and timestamps. The target site only has measured wind speed data.

It is very hard to model atmospheric variables because of their chaotic nature. There are no defined relationships between the wind speed, pressure, and temperature at a given location. However, we hope that the neural networks can capture the nonlinearity of these datasets and learn from them. As weather patterns vary widely across seasons, our training data must be reflective of a typical meteorological year. The latter is a representative base year [4] of the site of interest. Figure 2 shows the distribution of the variables relative to each other.

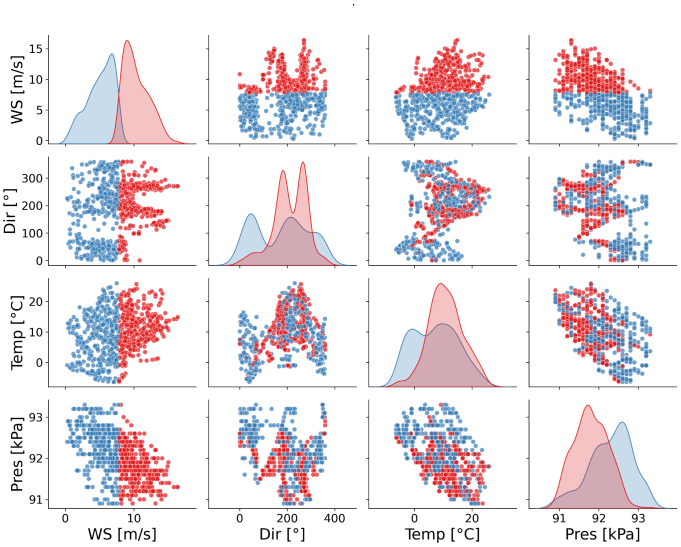

Figure 2: Pairplot showing the relative distribution of the Station 1 variables. The red dots represent values above the mean wind speed, while the blue ones show values below the mean wind speed of the site.

The above figure shows that the variables are roughly equally distributed around the mean wind speed. This suggests that our dataset is close to a typical meteorological year and good for training our models.

We used data from January 2000 to December 2015 from Station 1 to train our generative model. It was then validated using data from 2015 to 2020. Our goal was to test the feasibility of forecasting future wind speeds using previous wind speeds. We took inspiration from the rolling forecast method employed in this paper [5]. We sectioned our dataset into 48-hour sequences and used these as input data during training. As the model forecasts the hour ahead of each sequence, it compares it with the actual value and adjusts its weight. The input and target training data were of size (131447, 50, 4) and (52607, 4) respectively. We removed the timestamps column from the dataset. Figure 3 illustrates how the rolling forecasting technique allows us to move through time while training the CNN-LSTM model.

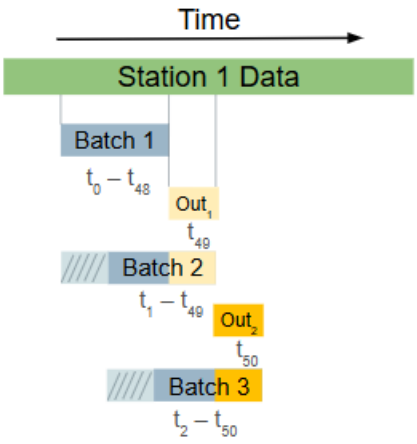

Figure 3: Rolling forecast technique applied to the Station 1 dataset. A sequence of 48 hours is used to forecast the 49th hour. The next sequence is created by concatenating the predicted value and removing the first hour of the batch.

We used data from March 2023 to March 2024 from all 4 stations to train the regression models. We concatenated them along the horizontal axis to create a vector of size (8016, 16). We purposefully opted for a 2-dimensional vector to effectively pass the data through a training loop without overly complex model architectures. The target values were measured by wind speed data of size (8016). Table 1 highlights the input and target used to train the regressors.

Table 1: Input and target datasets shape.

Station 1				...	Station 4				WS
WS	WDIR	Temp	Pres		WS	WDIR	Temp	Pres	

Performance metrics such as Mean Square Error (MSE), Mean Absolute Percentage (MAPE), Root Mean Squared Error (RMSE), and R-squared (R2) are calculated to evaluate the models' accuracy.

Additionally, the difference between actual and predicted values is plotted to assess the models' performance and identity biases.

CNN-LSTM Generative Model Architecture

The forecasting model constructed consists of a set of two 1D convolutional layers which are fed into a pooling layer. The data was subsequently pushed into a long short-term memory (LSTM) layer and finally into a fully connected layer to produce the final wind speed prediction. The choice of architecture was due to the convolutional layers being used to capture sophisticated spatial features of the data [6] while the LSTM layer allows the model to learn temporal patterns inherent to environmental data [7]. This model's training loop iterates through epochs, then calculates the loss (using Mean Squared Error) and updates the model's weights using the Adam optimizer. After each epoch, the model is evaluated on the testing data to track its performance. Early stopping was also implemented to prevent data overfitting. The trained model is then used to predict wind speed for the first 24 hours of 2021 using a rolling forecast approach. The rolling forecast method performed in this model starts with a time-period batch (of 48 hours) of real input data and uses this batch to attempt to predict the wind speed of the next hour. The next batch consists of both the first input real data plus the predicted hour, then the next hour is predicted considering this batch, and so on.

CNN Regressor Architecture

A convolutional neural network (CNN) was the first regressor tested. It contains a set of 1D convolutional layers that work by applying a sliding window (called a kernel or filter) across sequential data, performing convolution operations. The primary rationale behind this architecture was its ability to capture spatial aspects of the concatenated location data. The choice of this model showed to be relevant as the concatenated station data provides a solid basis for location-based patterns to emerge.

RBFN Regressor Architecture

The second regressor constructed was a standard radial basis function neural network (RBFN). The RBFN model consists of an input layer, one hidden layer – where the radial basis function is applied – and the output layer. The input layer receives features from the 4 ERA5 stations as input. The RBF Layer (the hidden layer) is responsible for transforming the input features into a higher-dimensional space using radial basis functions. The Linear Layer (output layer) combines the outputs of the radial basis functions to produce the final prediction using a linear activation function whose weights are learned during training. This model was selected because RBFNs naturally capture the data's spatial information [8], and introduce high nonlinearity which is nonnegotiable for environmental data that presents so much discrepancy.

Results and Discussion

Our trained CNN-LSTM model was deployed to forecast January 2021 of station 1. We gave it an initial sequence of 192 hours (December 24-December 31) and let it generate the first 24 hours in January using the batching technique previously outlined in Figure 3. We then compared this forecast with the actual values recorded by the ERA5 station 1. Figure 4 compares the actual hourly wind speeds with the forecasts made by the model.

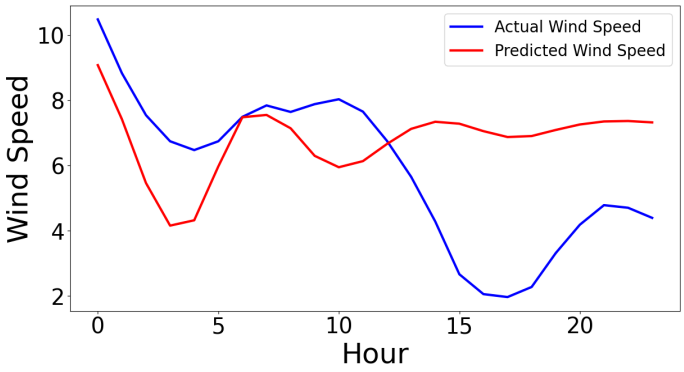

Figure 4: Hybrid CNN-LSTM forecasting the first 24 hours of wind speed in January 2021. The model is unable to accurately forecast the wind speed beyond 10 hours.

The prediction from the forecasting model appears to correlate with the actual data for the first 8-10 hours. However, the error explodes beyond this point. This is understandable given that our rolling forecasting technique uses the predicted values to forecast future wind speeds. The error from these compounds over time hinders the model's accuracy. Nonetheless, this result is significant given that the longest forecast reported in the literature is 24 hours ahead [1].

Figure 4 highlights the challenge of forecasting future atmospheric variables using previous ones. It is a nontrivial task because of the chaotic nature of the environment. Therefore, it is more viable to use regressors to reconstruct the hourly wind speed data locally. One can then assume that the future wind speed behavior will not be much different than the reconstruction behavior.

We deployed the two regressors in April 2024. The concatenated values from the four stations were given as input to predict the wind speed at the site. The predictions were then compared to the actual values. The 1D CNN regressor results are displayed in Figure 5 whereas the output from the RBFN is captured in Figure 6.

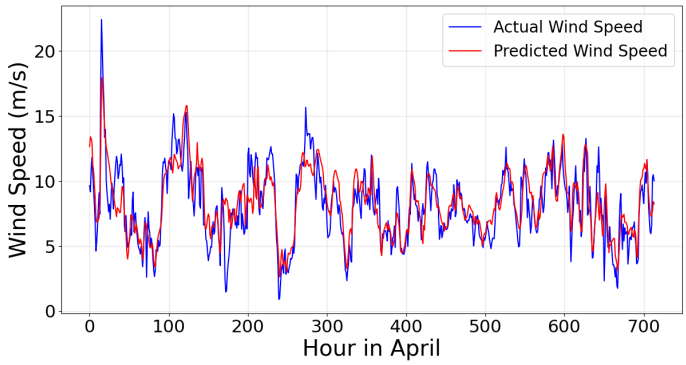

Figure 5: 1D CNN predictions for wind speed over April 2024 compared to the measured values. The trained model is successful at reconstructing the hourly wind speed data at the target site. Most of the temporal variation is captured.

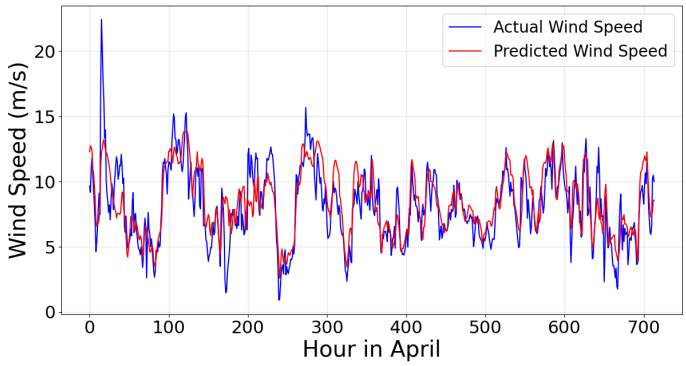

Figure 6: RBFN predictions for wind speed over April 2024 compared to the measured values. The trained model is successful at reconstructing the hourly wind speed data at the target site. Most of the temporal variation is captured.

The graphs above show that our regressors are capable of reconstructing the hourly wind speeds at our target site. They were able to capture the general trend as well as the temporal variations of April 2024. The performance metrics of each model are summarized in Table 2.

Table 2: Various Metrics for each model's performance.

	MSE [(m/s)^2]	R2	Mean Ratio	MAPE [%]	RMSE [m/s]
CNN-LSTM	7.46	0.41	1.152	64.4	2.73
RBFN	3.0	0.63	1.001	18	1.73

CNN	2.69	0.67	1.04	19	1.64

The predictions of the CNN-LSTM model were not accurate, achieving an R2 of 0.41. Meanwhile, the regressors show much more significant results. The RBFN predicted the mean wind speed with 0.1% error while the CNN had a 4% error. This is an important result as developers often care about the general wind resource before pursuing a potential site of interest. The RBFN also outperformed the CNN MAPE scores. On the other hand, CNN seems to be better at capturing the hourly variations in the data. Its R2 score and RMSE are 0.67 and 1.64 m/s – both better results than the RBFN

CONCLUSION

The study results have shown that the CNN-LSTM can only forecast the wind speed trend for up to 10 hours before the prediction becomes highly inaccurate. The error compounds over time and hinders future predictions in the batch. However, both regressors reconstruct their test data with a mean absolute percentage error of less than 20% – consistent with literature values. The RBFN outperforms the CNN in certain metrics. It predicts the site's mean wind speed with a 0.01% error while the latter had a 4% error. The CNN was better at capturing the temporal variation of the wind speed at the site with RMSE and R2 scores of 1.64 m/s and 0.67 respectively. These results highlight the ability of the regressors to reconstruct the long-wind speed behavior locally. As performance metrics vary depending on the model used, it is up to the user to determine which model best suits their needs.

Acknowledgments

The authors would like to express their sincere gratitude to Professor Ahmed, Lyle Regenwetter, and Noah Bagazinski for their invaluable feedback, guidance, and support throughout the development of this research project. Their insightful comments, constructive suggestions, and generous assistance were instrumental in shaping the direction and quality of this work.

References

[1] Alves, D., Mendonça, F., Mostafa, S. S., and Morgado-Dias, F., 2023, "The Potential of Machine Learning for Wind Speed and Direction Short-Term Forecasting: A Systematic Review," Computers, **12**(10), p. 206. https://doi.org/10.3390/computers12100206.

[2] Karaman, Ö. A., 2023, "Prediction of Wind Power with Machine Learning Models," Applied Sciences, **13**(20), p. 11455. https://doi.org/10.3390/app132011455.

[3] Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz‐Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., De Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J., 2020, "The ERA5 Global Reanalysis," Quart J Royal Meteoro Soc, **146**(730), pp. 1999–2049. https://doi.org/10.1002/qj.3803.

[4] Kalogirou, S. A., 2022, "Solar Thermal Systems: Components and Applications—Introduction," *Comprehensive Renewable Energy*, Elsevier, pp. 1–25. https://doi.org/10.1016/B978-0-12-819727-1.00001-7.

[5] Tang, C., Xu, Y., Bu, S., Xiao, S., Yuan, B., Cai, J., and Yu, Q., 2024, "Application of ARIMA-LSTM-CQP Time Rolling Window Multi-Factor Stock Selection Model in Quantitative Investment." https://doi.org/10.21203/rs.3.rs-3875083/v1.

[6] Nazir, A., He, J., Zhu, N., Qureshi, S. S., Qureshi, S. U., Ullah, F., Wajahat, A., and Pathan, M. S., 2024, "A Deep Learning-Based Novel Hybrid CNN-LSTM Architecture for Efficient Detection of Threats in the IoT Ecosystem," Ain Shams Engineering Journal, **15**(7), p. 102777. https://doi.org/10.1016/j.asej.2024.102777.

[7] Lindemann, B., Müller, T., Vietz, H., Jazdi, N., and Weyrich, M., 2021, "A Survey on Long Short-Term Memory Networks for Time Series Prediction," Procedia CIRP, **99**, pp. 650–655. https://doi.org/10.1016/j.procir.2021.03.088.

[8] He, Q., Shahabi, H., Shirzadi, A., Li, S., Chen, W., Wang, N., Chai, H., Bian, H., Ma, J., Chen, Y., Wang, X., Chapi, K., and Ahmad, B. B., 2019, "Landslide Spatial Modelling Using Novel Bivariate Statistical Based Naïve Bayes, RBF Classifier, and RBF Network Machine Learning Algorithms," Science of The Total Environment, **663**, pp. 1–15. https://doi.org/10.1016/j.scitotenv.2019.01.329.