Proceedings of the IISE Annual Conference & Expo 2025
XXX, XXX, XXX

Investigation of the Impacts of Training Sample Sizes and Features on Wind Speed Prediction Accuracy of Long Short-Term Memory Method

Abstract ID: 6652

Mariee Cruz Mendoza
Texas A&M University-Kingsville, Kingsville, TX, USA

Hua Li
Texas A&M University-Kingsville, Kingsville, TX, USA

Kaylee Xu
Flour Bluff High School, Flour Bluff, TX, USA

Reyna Xiao
Veteran Memorial High School, Corpus Christi, TX, USA

Jinyi Wang
Flour Bluff High School, Flour Bluff, TX, USA

Abstract

The accuracy of wind speed prediction is critical to wind farm development and operation, including site and turbine selection, energy production forecasting, and grid management. The increasing utilization of wind energy has required scientists to improve the prediction accuracy of the current approaches to maximize wind energy generation and utilization. In the field of wind energy, machine and deep learning methods have shown advantages in accurately forecasting data, such as wind speed and wind direction. This study focuses on evaluating the capabilities of Long Short-Term Memory (LSTM) models to consistently predict short-term wind speeds using different amounts of training data to test the model's generalizability and robustness. An LSTM model was trained using different sample sizes of training data and features, while the testing accuracies of the trained LSTM models were used to evaluate the models and compared to investigate the impacts of training data sizes and features. In addition, data from three locations with varying terrain and weather conditions was used to evaluate the performance of the trained LSTM model. The model was trained on data from each selected location and tested on the others to assess its predictive accuracy and generalizability across different geographic settings.

Keywords
Machine Learning, Prediction, Texas, Wind Speed

1. Introduction

In the field of wind energy, accurate forecasting and resource allocation are crucial for maximizing power output, managing grid stability, and reducing operational costs at wind farms. Nevertheless, the inherent variability of wind presents a significant challenge for wind farm operators due to its complex nature and influence of several factors such as atmospheric processes, local topography, and fine-scale temporal and spatial variations. Nowadays, researchers classify the forecasts of wind speed based on the time scale, mostly in short-term forecasting (minutes to hours), medium-term forecasting (hours to days), and long-term forecasting (weeks to years). The rise in wind energy developments around the world has made short-term wind forecasting an essential tool for transmission system operators and power traders, particularly in regions with significant wind power integration to balance the supply and demand of energy [1]. Over the past five decades, researchers have dedicated considerable effort to enhancing the accuracy of wind speed and direction predictions at wind sites. Various approaches have been employed to improve wind forecasting, including physical methods, statistical models, and artificial intelligence (AI) algorithms [1]. In

recent years, AI-based models have gained attention across diverse fields of study, particularly due to their high accuracy in prediction, classification, and anomaly detection tasks [2]. Currently, the Long-Short Term Memory (LSTM) method has become one of the most prominent algorithms for prediction across several fields, due to its capacity to handle sequential data over time and the advantages it has shown in the analysis of patterns, future predictions, and real-time monitoring [3].

The goal of this study is to evaluate whether LSTM models can consistently predict next hour wind speed across different geographic locations using different amounts of historical wind speed training data and a small number of randomly selected features, thereby testing the model's generalizability and robustness. The research method involved training an LSTM model with different sample sizes of training data and features to assess the impact of these elements on prediction accuracy. This approach aims to contribute to the ongoing efforts to improve wind forecasting techniques, ultimately enhancing the efficiency and reliability of wind energy production.

2. Background

Despite the significant growth of the wind energy sector, the forecasting of wind conditions is considered as one of the most challenging tasks in the field as wind farms require accurate forecasting of available resources to maximize energy generation. To advance research in clean energy and enhance the efficiency of wind power resources, it is important to continue exploring prediction methods aimed at improving the forecasting of wind energy sites, especially with the rise of machine and deep learning models. In recent years, several algorithms have been utilized to predict wind speed and wind direction, LSTM networks being one of the most studied ones due to their effectiveness in modeling sequential data [1,4,5]. In addition, researchers have also focused on exploring the capabilities of LSTMs with different variations, such as using LSTMs to predict wind speed profiles over complex terrains [6], or the investigation of strategies to balance computational efficiency and prediction accuracy in wind-induced response predictions, experimenting with various modifications during the model training process [6]. For this study, the impacts of training size and meteorological factors in the performance of next-hour predictions with LSTM models were further explored. Meteorological data for three different cities in Texas was employed to evaluate the performance of the LSTM model when varying training input size and features. This study selected the cities of Corpus Christi, El Paso, and Lubbock to analyze various factors that may impact the accuracy of wind speed predictions. These factors included sample sizes, meteorological features, and terrain characteristics, which can influence the training and testing of the machine learning algorithm. Table 1 summarizes some of the surface terrain characteristics for the selected cities in this study.

Table 1: Surface Terrain Characteristics for Three Cities in Texas [7].

City	Characteristics
Corpus Christi	-Flat terrain, nearly level. -Includes barrier islands, salt grass marshes, bays, and estuaries. -Elevation: 0 – 400. ft above sea level.
El Paso	-Most complex region, located in extreme western Texas to the Pecos River. -Diverse habitats: desert valleys, plateaus, and wooded mountain slope. -Elevation: 2500 – 5200 ft. above sea level.
Lubbock	-Flat terrains with a mix of shortgrass prairie and desert scrubland. -Subject to extended droughts. -Elevation: 3000 – 4500 ft. above sea level.

3. Methodology

LSTM networks are a type of Recurrent Neural Networks (RNN) that have the ability to handle and analyze sequential data effectively in a variety of application domains, such as language modeling, speech-to-text transcription, and prediction [8]. LSTMs work by handling sequential data input by processing it through multiple time steps. During each step, the network's internal states are automatically updated based on the current input and information from the previous state. LSTMs employ a gating mechanism to learn which information should be retained, discarded, or passed on to the next step. This system of gates allows LSTMs to selectively regulate the flow of information into and out of their memory cells [9].

To conduct this study, historical meteorological data for three selected locations between 1998 and 2022 were obtained from the National Solar Radiation Database (NSRDB). The datasets included hourly measurements for various meteorological variables, such as diffuse horizontal irradiance (DHI), direct normal irradiance (DNI), global horizontal irradiance (GHI), wind speed, relative humidity, temperature, pressure, and wind direction. To study the effects of training sample sizes and features on wind speed prediction accuracy using LSTMs, a model using the previous hour's values for several features to forecast the next hour was performed in MATLAB. Moreover, four distinct analyses were conducted: 1) Training with one week of data from a single city, followed by testing with one week of data from the same city; 2) Training with one month of data from a single city, followed by testing with one week of data from the same city; 3) Training on multiple features while testing on the same city; and 4) Training with data from one city and testing with data from a different city. For each analysis performed at each location, a random year was selected from the available data for training and testing the model. Likewise, all models were trained using 24 hours of data from randomly selected dates throughout the year organized sequentially. The accuracy of the analyses was quantified using the Root Mean Square Error (RMSE), which measures the differences between predicted and actual values, along with the standard deviation to assess the variability in the model's predictions.

4. Results & Discussion

The accuracy of short-term wind predictions can be influenced by the size and diversity of the historical datasets employed during the training of the model. The amount of available training data is critical to the model's effectiveness as larger datasets can enhance performance, while models trained on data from various locations and accounting for multiple factors can have a substantial impact on overall performance. Figure 1 shows distribution of data used in this study. A summary of the results of the four analyses is presented below.

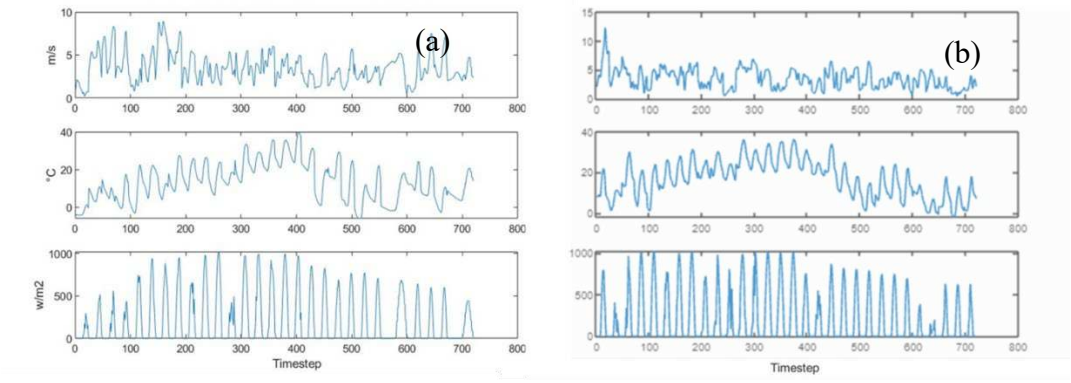

Figure 1: Time Series Graphs for Wind Speed, Temperature, and GHI: (a) El Paso and (b) Lubbock, Texas.

4.1 Training with data from one week from a single city, testing performed in one week from the same city.

The initial analysis evaluated the model's generalization ability by assessing its performance on data from different weeks within the same city. The model was trained using wind speed data from seven randomly selected days throughout the year, arranged sequentially for each location in this study. This training dataset comprised 168 observations, representing seven days of hourly data. The model was then tested with another 168 observations from a continuous seven-day period, beginning on a randomly selected date, all containing hourly measurements. The underlying hypothesis was that training the model on a randomly selected week of data would expose it to wind speed patterns influenced by varying atmospheric conditions, enabling it to generalize to another week within the same city. However, the training dataset size could also impact prediction accuracy. The results showed that the model effectively predicted wind speed for continuous seven-day periods, with RMSE values remaining consistent across tested cities, averaging 0.4035. Table 2 presents the average RMSE values after running the analysis 10 times.

Table 2: RMSE values for training and testing with one-week data from the same city.

City	Corpus Christi	El Paso	Lubbock	Total
Average RMSE	0.3298	0.3948	0.4858	0.4035
Standard Deviation	0.0628	0.0687	0.1279	0.0864

4.2 Training with data from one month from a single city, testing performed in one week from the same city.
The modification of the code for the second version of the model involved using a larger volume of training data to assess the model's ability to generalize more accurately with extended training periods, compared to shorter durations like one week. The training dataset comprised 720 observations corresponding to 30 days of hourly data. Subsequently, the model underwent testing with an additional 168 observations, representing 7 consecutive days of hourly data commencing on a randomly selected date, all containing hourly measurements. The initial hypothesis for this analysis was that training the model on longer periods of historical wind speed data, such as one month, would lead to more accurate forecasts compared to training on shorter time periods. Moreover, the model could benefit from diverse datasets that encompass varying wind speed conditions and patterns, offering more comprehensive insights than seven days of data. Based on the obtained RMSE values, the initial hypothesis was validated, as using a larger dataset of thirty days of wind speed data, instead of seven, for model training led to improved accuracy. Overall, the three cities showed lower RMSE values in this analysis, with the city of Corpus Christi performing better than the cities of El Paso and Lubbock. Table 3 presents the average RMSE values after running the analysis 10 times and the standard deviation for these values, while Figure 2 shows the difference between predicted values and actual values.

Table 3: RMSE values for training with data from one month from a single city, testing performed in one week from the same city.

City	Corpus Christi	El Paso	Lubbock	Total
Average RMSE	0.2715	0.3595	0.3915	0.3408
Standard Deviation	0.0479	0.0568	0.0647	0.0565

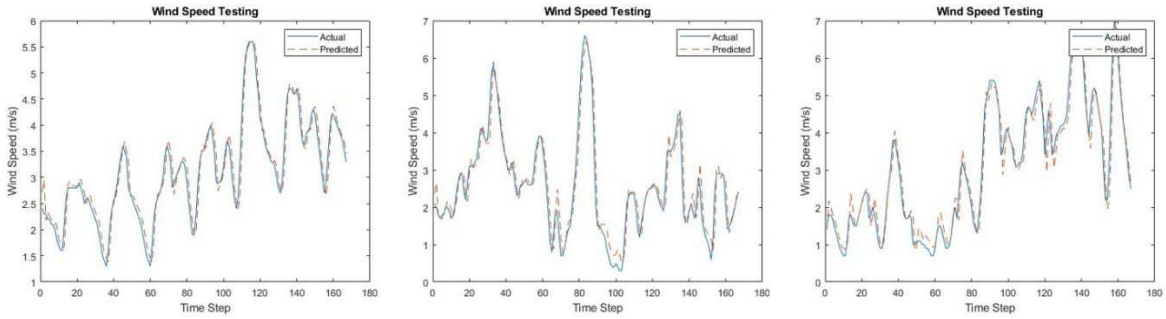

Figure 2: Wind Speed Predictions for Analysis 2: Corpus Christi, El Paso, and Lubbock.

4.3 Training on Multiple Features and Testing on the Same City.
The third analysis incorporated additional factors to train the model, as opposed to solely utilizing wind speed. This approach would allow the combination of meteorological factors available from the datasets collected from the NSRDB. The variables that were considered for the training of the model for this specific analysis were randomly selected, resulting in the selection of wind speed, temperature, and GHI for the performance of this analysis. Following a similar approach for the selection of the training data as analysis 2, thirty complete days (720 observations) of these data factors were randomly selected for model training. Subsequently, the model was evaluated using a continuous week of wind speed data (168 observations). The purpose was to assess the efficacy of incorporating additional features in predicting wind speed and to determine whether the model could generate accurate wind speed predictions when multiple features were utilized, as opposed to considering wind speed as the sole feature.

Table 4: RMSE values Training on Multiple Features and Testing on a Single Feature in the Same City.

City	Corpus Christi	El Paso	Lubbock	Total
Average RMSE	0.8090	0.8839	0.9648	0.8788
Standard Deviation	0.2019	0.1569	0.3824	0.2471

The results obtained as shown in Tables 3 and 4 indicated that the addition of more features to the model did not improve accuracy, suggesting that utilizing wind speed independently could produce reliable predictions and would

be sufficient for the model to learn patterns and make predictions. The higher RMSE values resulting from the inclusion of additional features suggested that the selected combination of features, such as wind temperature and GHI, may not have been optimal, potentially including weak features for predictions of wind speed. Future improvements to the model could involve selecting a different combination of features for training that more effectively captures the complex relationships between the variables. Figure 3 shows the predication results.

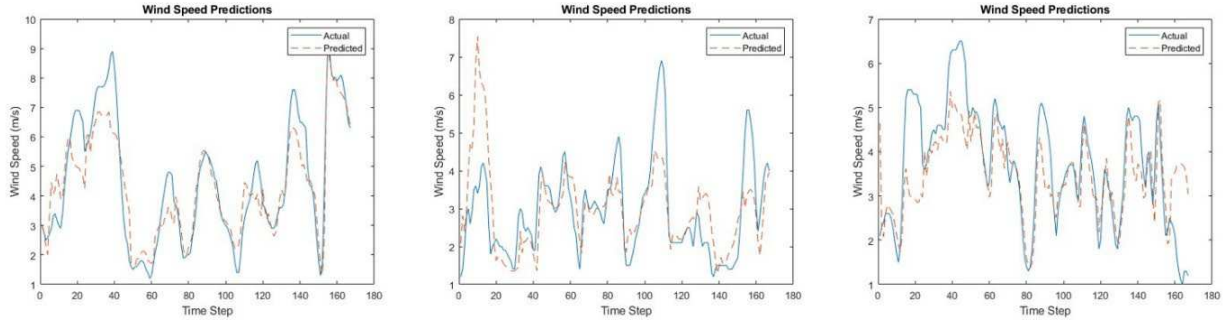

Figure 3: Wind Speed Predictions for Analysis 3: Corpus Christi, El Paso, and Lubbock.

4.4 Training with data from One City, Testing with data from a Different City.

The final analysis evaluated the model's ability to generalize across diverse locations with varying surface and terrain conditions, influenced by different weather patterns, to predict wind speed in another city. This assessment aimed to determine whether the model could identify patterns and capture physical relationships relevant to the cities selected for the study. The model was trained using data from Corpus Christi and tested with data from El Paso and Lubbock. Similarly, a model trained with data from El Paso was tested in Corpus Christi and Lubbock, and a model trained with data from Lubbock was tested in Corpus Christi and El Paso. This code followed a similar approach to the third analysis, utilizing monthly data for training and incorporating additional factors to enhance the model's performance. The results of this analysis are presented in Table 5 and Figure 4. While the model demonstrated reasonable generalization when applied to data from different cities, variations in elevation, topography, and local weather patterns can significantly affect the accuracy of the predictions.

Table 5: RMSE values for training with data from One City, Testing with data from a Different City.

City (Training)	**City (Testing)**		
	Corpus Christi	**El Paso**	**Lubbock**
Corpus Christi	-	1.5671	1.4089
El Paso	1.1898	-	1.5501
Lubbock	0.7344	0.9142	-

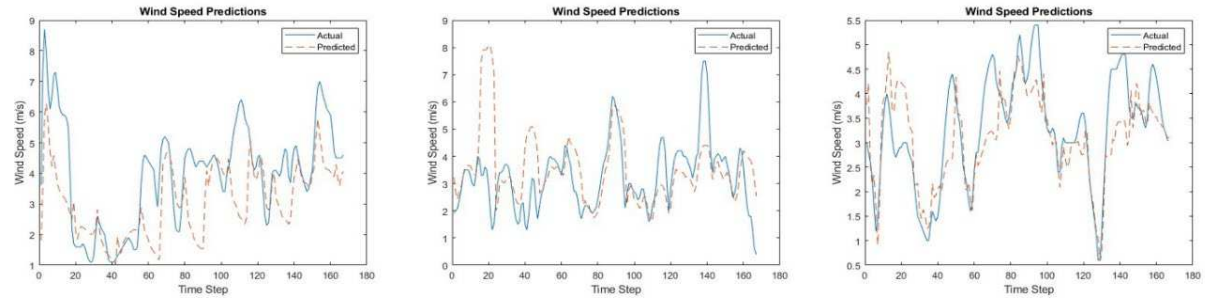

Figure 4: Wind Speed Predictions for Analysis 4 – Training in Corpus Christi and Testing in El Paso; Training in El Paso and Testing in Lubbock; Training in Lubbock and Testing in El Paso.

5. Conclusion and Final Thoughts

Machine learning algorithms have revolutionized data analysis by identifying patterns within datasets, enabling the development of models that generalize effectively and generate accurate predictions for unseen data. Similarly, time series forecasting using LSTMs have become a powerful tool for predicting time-dependent variables by capturing temporal dependencies in sequential data and using past observations to predict future values. In this study, four different analyses using LSTM models were performed to evaluate the impact that varying the size and features from the testing and training datasets have when predicting wind speed. Additionally, the performance of wind prediction models across different surface and terrain conditions was examined to assess their accuracy when being trained with weekly and monthly data. The results indicate that models trained on monthly data performed better than those trained on weekly data. A potential reason for this is that monthly datasets capture more valuable and informative patterns, such as seasonal trends or long-term behaviors, which may not be discernible in shorter training periods. Moreover, the increased number of observations in monthly data contributed to improved model performance. It was also observed that adding additional features did not improve model accuracy. This may be attributed to the inclusion of suboptimal, randomly selected variables, such as temperature and GHI, which may not be the most effective predictors for short-term wind speed forecasting with an LSTM model. Moreover, the incorporation of more features may have increased the model's complexity, thereby requiring a larger volume of training data for the effective learning of the model. Likewise, it was observed that when the model is trained using meteorological data from one city and tested on data from another, the RMSE values indicate that the model performs reasonably well. However, differences in terrain conditions between cities may impact performance, suggesting that model accuracy could be further improved with localized training data. Future improvements will involve selecting a more relevant combination of features that better capture the complex relationships between variables and the forecasting of wind speed at different time-horizons, which will be explored in future research.

.

Acknowledgements

The authors are thankful for the support from the U.S. Department of Agriculture (Award 2022-77040-37631) and the U.S. National Science Foundation (award 2244523). Any opinions, findings, or recommendations expressed were created by the authors and not reviewed by nor necessarily reflect the view of the USDA and NSF.

References

[1] Xie, A., Yang, H., Chen, J., Sheng, L., & Zhang, Q. (2021). A short-term wind speed forecasting model based on a multi-variable long short-term memory network. Atmosphere, 12(5), 651.

[2] Nasteski, V. (2017). An overview of the supervised machine learning methods. Horizons. b, 4(51-62), 56.

[3] Deb, C., Zhang, F., Yang, J., Lee, S. E., & Shah, K. W. (2017). A review on time series forecasting techniques for building energy consumption. Renewable and Sustainable Energy Reviews, 74, 902-924.

[4] Prema, V., Sarkar, S., Rao, K. U., & Umesh, A. (2019). LSTM based Deep Learning model for accurate wind speed prediction. Data Sci. Mach. Learn, 1, 6-11.

[5] Leme Beu, C. M., & Landulfo, E. (2024). Machine-learning-based estimate of the wind speed over complex terrain using the long short-term memory (LSTM) recurrent neural network. Wind Energy Science, 9(6), 1431-1450.[6] Li,

[6] L., Huang, X., Chen, S., Wu, T., Mei, L., Long, W., & Xiao, Y. (2023). Study on strategies for reducing training samples for accurate estimation of wind-induced structural response of LSTM networks. Journal of Wind Engineering and Industrial Aerodynamics, 238, 105421.

[7] Texas Parks and Wildlife Department. "Texas Ecoregions." Texas Parks and Wildlife Department, https://tpwd.texas.gov/education/hunter-education/online-course/wildlife-conservation/texas-ecoregions. Accessed 8 Feb. 2025.

[8] Sherstinsky, A. (2020). Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Physica D: Nonlinear Phenomena, 404, 132306.

[9] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.