**Comparing the effectiveness of LSTM and persistence models for wind speed prediction in the Great Lakes region**

**Alexander Wycoff1, Jonathan Stelman2**

1 Eden Prairie High School, Eden Prairie, Minnesota

2 Learning Sciences & Innovation,Renaissance Learning, Brooklyn, New York

**Student Authors**

**Alexander Wycoff – High School**

**SUMMARY**

With an estimated offshore potential wind energy capacity of 575 gigawatts, the Great Lakes region is a promising area for future wind energy development. Electric utilities rely on accurate day-ahead wind energy forecasts, mainly informed by predicted wind speed, to account for the variability of wind energy production. Hence, accurate wind speed prediction models are crucial to integrating wind energy reliably in the Great Lakes region. We hypothesize that long short-term memory (LSTM) neural networks trained on synthetic data will be more accurate than persistence models in predicting day-ahead wind speeds in the Great Lakes region due to their ability to learn short-term trends effectively. While LSTM networks have been applied to wind speed prediction at a few selected sites in the past, we aim to use newly available synthetic wind speed data from the 2023 National Offshore Wind (NOW-23) Great Lakes dataset to create a more versatile model. We selected 100 random sites from the NOW-23 Great Lakes dataset at an elevation of 80 meters and trained an LSTM network on each site. This allowed us to extrapolate our results to the entirety of the dataset, demonstrating our model’s potential for use at a large scale. We also optimized our model, significantly reducing network training time while preserving accuracy. The availability of a wind speed prediction model trained on synthetic data will reduce reliance on historical observational data at future sites of wind energy infrastructure, allowing utilities to swiftly adapt accurate prediction methods to new sites.

**INTRODUCTION**

The Great Lakes have an estimated offshore wind energy potential of 575 gigawatts (1). With most sites throughout the Great Lakes reaching annual average wind speeds of 9 meters per second or greater, the Great Lakes region has significant opportunities for offshore wind energy development (2). While it does face unique challenges, offshore wind energy development in the Great Lakes has also been deemed technically and economically feasible in some areas (3). When wind energy development begins at larger scales in the Great Lakes, electric utilities will need to be able to integrate this new source of energy effectively.

Electric utilities in the United States utilize energy forecasts to schedule daily electricity production from various sources. While wind energy production is variable by nature, accurate wind energy forecasts can assist utilities in integrating wind energy reliably (4). Wind speed forecasts assist in predicting the energy that wind farms will produce. Thus, more accurate wind speed predictions would allow utilities to improve wind energy production estimates, enhancing the reliability of wind energy. Previous projects to improve the forecasting of wind speed production have demonstrated that improvements in predictive wind speed models can reduce wind energy overprediction and underprediction, directly corresponding to a decrease in the excess costs incurred by electric utilities (4).

LSTM networks are recurrent neural networks that are particularly well-suited to modeling short- and long-term dependencies in time series data due to their minimal error propagation while performing multi-step ahead predictions (5). Accordingly, researchers have already validated the efficacy of LSTM networks in wind speed prediction. A comparison of previous studies found that deep learning approaches have surpassed traditional methods regarding the accuracy of wind speed predictions; of these approaches, LSTM networks performed the best despite representing a relatively small portion of the studies examined, indicating the potential for future research involving LSTM networks (6).

Previous studies have utilized multiple variations of LSTM networks to predict time series data involving wind speed (7–8). While these studies produced models that could accurately predict wind speed in a time series, the locations, data, and forecasting periods involved in each study varied dramatically. One study forecasted wind speeds a day in advance, while the other focused on short-term timeframes less suitable for day-ahead wind energy prediction. Additionally, both studies were confined to relatively small areas in which wind speed data had been collected, limiting both the spatial diversity and comparability of the results. While these studies focused on areas where wind speed data is historically available, our study used newly available synthetic wind speed data to create a more versatile wind speed prediction model capable of making accurate predictions over the entire Great Lakes region.

We leveraged synthetic data from the NOW-23 Great Lakes dataset generated through the Weather Research & Forecasting program and validated using observational data (9). Synthetic data is available at a higher spatial resolution than what is available purely observationally and has been confirmed to realistically represent observational wind speed data over larger timescales (10). Additionally, previous studies have used synthetic data to improve sub-hourly wind speed predictions (10). Still, whether similar improvements can be achieved over extended forecasting periods remains to be seen.

In our study, we aim to create more versatile wind speed prediction models using synthetic wind speed data. We hypothesized that long short-term memory (LSTM) neural networks trained on synthetic data are more accurate than a persistence model in predicting day-ahead wind speeds in the Great Lakes region. We trained and tested LSTM networks on data from a random selection of sites from the NOW-23 Great Lakes dataset at an elevation of 80 meters. We then used multiple experiments to optimize the parameters of the networks, significantly reducing training time while preserving accuracy. To evaluate our hypothesis, we compared our model to a persistence model, which uses the last known wind speed measurement to predict the next and is a typical benchmark for the performance of wind speed prediction models. Finally, we used statistical inference to extrapolate our results to the entirety of the Great Lakes region.

**RESULTS**

We developed and tuned an LSTM architecture for time series prediction using the NOW-23 Great Lakes dataset data at an elevation of 80 meters. The model used the previous 24 hours of wind speed, direction, and turbulent kinetic energy data at a temporal resolution of 1 hour to predict the wind speed in 24 hours at a given site. We trained 100 LSTM networks—one for each site randomly selected from the dataset—on four years of time series data. We used data spanning 2015 to 2020 to reduce the time required to train networks while preserving enough data to do so effectively. We also used 50 epochs to train the networks to minimize training time while ensuring accuracy. The mean absolute error (MAE) scores in meters per second achieved by testing these models on data from a single year withheld from training were then compiled into a distribution.

Within the sample of randomly selected sites, the LSTM model was significantly more effective than the persistence model, with the mean MAE score of 3.845 m/s achieved by the persistence model being 28% higher than the 3.003 m/s score achieved by the LSTM model (**Table 1**). While the variation of the MAE scores of the two models was relatively similar, the persistence model had noticeably more outliers, with eight compared to the LSTM model’s five (**Figure 1**).

            Through multiple experiments, we discovered that using significantly more than 5 years of data had diminishing returns in model accuracy, while less than two years of data was not comprehensive enough to train an LSTM model effectively (**Figure 2**). Furthermore, we observed that model accuracy quickly converged as the number of epochs increased, with any number greater than 50 yielding diminishing improvements in accuracy (**Figure 3**). However, these experiments were limited to a single site, so validating them by comparing LSTM models with varying parameters was necessary. While LSTM models trained with more epochs or years of data had minor improvements in average root mean squared error (RMSE) scores, all LSTM models trained and tested on all selected sites had almost no difference in mean and median MAE scores (**Table 1**). Despite this roughly equal performance, the model that used 50 epochs and 5 years of data had a significantly faster training time per network than the other two LSTM models, making it more viable for large-scale wind speed prediction (**Figure 4**).

In our sample, the mean MAE of the LSTM models tested was much lower than that of the persistence models tested (**Figure 1**). We will use a paired t-test to determine if this observation holds across the 388,080 total sites in the NOW-23 Great Lakes dataset at an elevation of 80 meters. Since we had a large, random, independent sample of sites from the NOW-23 Great Lakes dataset, the conditions for inference are met. Let be the true mean difference in MAE scores between the persistence and LSTM models for all sites in the NOW-23 Great Lakes dataset at an elevation of 80 meters. Our null hypothesis is that and our alternative hypothesis is that . Using as the sample mean difference and as the standard error of the differences in MAE scores of the LSTM and persistence models (n=100) tested, a paired t-test yields a p-value less than 0.01. Thus, there is convincing evidence of a true mean difference in MAE scores between LSTM and persistence models over the entire NOW-23 Great Lakes dataset at an elevation of 80 meters.

**DISCUSSION**

Our study aimed to utilize synthetic data from the NOW-23 Great Lakes dataset and LSTM networks to create a practical and versatile wind speed prediction model. Specifically, we examined whether LSTM networks could accurately predict wind speed one day in advance throughout the Great Lakes region. We tested our model on a random sample of 100 sites from the NOW-23 Great Lakes dataset at an elevation of 80 meters, inferring our results to the entirety of the dataset at this elevation. We concluded that there is convincing evidence that LSTM neural networks are more accurate than a persistence model in predicting wind speed in the Great Lakes region. Furthermore, we optimized our model’s training time and parameters through multiple experiments while preserving its accuracy, ensuring that the model is practical for use at a large scale.

Additionally, the LSTM networks used in this study showed significant improvement over persistence models within our random sample. The LSTM networks trained across all 100 selected sites had a mean MAE score approximately 28% lower than that obtained using persistence models (**Figure 1**). Previous studies have approached wind speed prediction using machine learning, deep learning, and artificial intelligence (6). Yet, these approaches have typically been confined to areas where wind speed data is historically available, limiting the extent to which they can be applied. A previous study focused on day-ahead wind speed prediction with LSTM networks reported a maximum improvement in MAE over a persistence model of only 17% across all models tested (8). However, the data this study used was set at an unrealistic elevation for wind turbines of 20 meters and covered only four sites, so its comparability is limited (8).

While commercial wind energy is not yet produced in the Great Lakes, researchers have proposed pathways to bring it within the next decade (2). Wind energy development in the Great Lakes would assist states in meeting their clean energy goals and provide economic benefits to nearby population centers (1). However, the variability of wind energy production can make its integration burdensome, as electric utilities must adapt to changes between forecasted and realized wind energy (11). More accurate day-ahead wind speed predictions could help utilities account for this variability, improving the reliability of wind energy production. Furthermore, sites viable for wind energy production that lack historical observational data can utilize wind speed prediction models trained on synthetic data, which can be further tuned as observational data becomes available.

Considering the data available, our study is still limited in scale. While we considered day-ahead predictions with an hourly sampling rate in this study, future studies could train additional models at different prediction timescales and compare models trained on data sampled at different rates. We also consider only one region from the NOW-23 dataset. Further studies on this topic may also seek to extrapolate results to the additional areas of the NOW-23 dataset. Nevertheless, by its nature, synthetic data is limited in its realism to observational data. While it realistically represents observational data at the timescale used in this study, it is not a perfect indicator of actual wind features (10). Additionally, future research could utilize the optimizations to network training time made in this study to improve the accuracy of the LSTM model presented.

Creating more accurate and versatile wind speed prediction models can help offset the variability of wind energy production. Our work contributes to the trend of utilizing deep learning for wind speed prediction, demonstrating that LSTM networks can achieve high accuracy over a spatially diverse range of sites. Future wind energy infrastructure in the Great Lakes region will benefit from the greater availability of accurate wind prediction models, encouraging further development.

**MATERIALS AND METHODS**

We used data from the National Renewable Energy Laboratory’s NOW-23 Great Lakes dataset simulated at an elevation of 80 meters, which was generated using the Weather Research & Forecasting program and validated using lidar data from Lake Michigan (12). We trained LSTM neural networks on this data to predict wind speed 24 hours in advance at 100 randomly selected locations from the Great Lakes portion of the dataset at a temporal resolution of 1 hour. The locations used in the study were randomly selected from the 388,080 total locations using a random number generator in the dataset and are roughly evenly dispersed throughout the Great Lakes region (**Figure 5**).

The use of weather variables other than wind features in wind speed forecasting is generally not associated with improvements in prediction accuracy (6). Hence, to train the LSTM networks, we selected only wind speed, measured in meters per second; wind direction, measured in degrees; and turbulent kinetic energy, measured in joules per kilogram, as independent variables. These observations were taken at an elevation of 80 meters, which reflects the height of most turbines in the United States. As of 2018, the average hub height for turbines in the United States was about 88 meters (13).

Time series data for wind speed, direction, and turbulent kinetic energy from 2000 to 2020 inclusive were retrieved from the National Renewable Energy Laboratory developer network API and concatenated for each of the 100 randomly selected sites using a script available on GitHub (14). This process resulted in a time series with the three chosen variables spanning 2000 to 2020 for each site. To prepare it for use in training LSTM networks, this time series data was then split into training and testing groups and normalized using min-max normalization.

We used TensorFlow, a machine learning library, and Keras, a deep learning library, run on Jupyter Notebook in Python 3.11 to train the networks used in this study. Each network was trained using a batch size of 128 and was also composed of the same architecture, which utilized an LSTM layer, a dropout layer to reduce overfitting to training data, and two densely connected layers (**Table 2**). Multi-layer neural networks with dropout layers show better convergence when using the Adam optimizer, so it was chosen as the optimizer for training (15). Each network layer containing weights was regularized using L2 regularization to reduce overfitting further. In all, each network contained about 1,400 total parameters (**Table 2**). Each network’s MAE and RMSE accuracy were calculated using functions from the Scikit-Learn library.

To optimize the training time for each network, we experimented with the number of epochs and the years of training data used while training the networks on a site in northern Minnesota (**Figures 2 and 3**). To evaluate the effect of altering these variables, we considered a network's MAE when tested on data from this site that had been withheld from training. A train-test split of 80:20 was used, so the size of the testing data was one-fourth that of the training data for each network. We used the results of these experiments to inform the parameters of our final model, which utilized 5 years of data and 50 epochs. We then compared the distribution of MAE scores of this model to that of a persistence model, which used the last known wind speed value as its prediction, on testing data for all 100 randomly selected sites used in this study (**Figure 1**).

**ACKNOWLEDGMENTS**

We greatly appreciate Renaissance Learning, which provided valuable mentorship and insight from multiple professionals throughout the research process.

**REFERENCES**

1. Laurie, Carol. “Exploring Offshore Wind Energy Opportunities in the Great Lakes.” National Renewable Energy Laboratory, 2023, www.nrel.gov/news/program/2023/exploring-offshore-wind-energy-opportunities-in-the-great-lakes.html
2. Musial, Walter, et al. “Great Lakes Wind Energy Challenges and Opportunities Assessment.” National Renewable Energy Laboratory, 2023, https://doi.org/10.2172/1968585
3. New York State Energy Research and Development Authority (NYSERDA). 2022. “New York Great Lakes Wind Energy Feasibility Study,” NYSERDA Report Number 22-12. Prepared by the National Renewable Energy Laboratory, Advisian Worley Group, and Brattle Group/Pterra Consulting. www.nyserda.ny.gov/About/Publications
4. Turner, David D., et al. "Evaluating the Economic Impacts of Improvements to the High-Resolution Rapid Refresh (HRRR) Numerical Weather Prediction Model." *Bulletin of the American Meteorological Society*, vol. 103, no. 2, 2022, pp. E198-E211, https://doi.org/10.1175/BAMS-D-20-0099.1
5. Lindemann, Benjamin, et al. “A survey on long short-term memory networks for time series prediction.” *Procedia CIRP*, vol. 99, 2021, pp. 650-655, https://doi.org/10.1016/j.procir.2021.03.088
6. Alves, Décio, et al. “The Potential of Machine Learning for Wind Speed and Direction Short-Term Forecasting: A Systematic Review.” *Computers*, vol. 12, no. 10, 2023, https://doi.org/10.3390/computers12100206
7. Neshat, Mehdi, et al. “A deep learning-based evolutionary model for short-term wind speed forecasting: A case study of the Lillgrund offshore wind farm.” *Energy Conversion and Management*, vol. 236, 2021, https://doi.org/10.1016/j.enconman.2021.114002
8. Araya, I. A., et al. “A Multi-Scale Model based on the Long Short-Term Memory for day ahead hourly wind speed forecasting.” *Pattern Recognition Letters*, vol. 136, 2019, pp. 333–340, https://doi.org/10.1016/j.patrec.2019.10.011
9. Bodini, Nicola, et al. *2023 National Offshore Wind data set (NOW-23)*. National Renewable Energy Laboratory, 2020, United States, https://doi.org/10.25984/1821404
10. Perr-Sauer, J., et al. “Short-term wind forecasting using statistical models with a fully observable wind flow.” *Journal of Physics: Conference Series*, vol. 1452, 2020, https://doi.org/10.1088/1742-6596/1452/1/012083
11. Konda, Srikanth Reddy, et al. "Dynamic Energy Balancing Cost Model for Day Ahead Markets With Uncertain Wind Energy and Generation Contingency Under Demand Response." *IEEE Transactions on Industry Applications*, vol. 54, no. 5, pp. 4908-4916, 2018, https://doi.org/10.1109/TIA.2018.2844363
12. Bodini, N., et al. “The 2023 National Offshore Wind Data Set (NOW-23).” *Earth System Science Data*, 2023, https://doi.org/10.5194/essd-2023-490
13. Lantz, Eric J., et al. “Increasing Wind Turbine Tower Heights: Opportunities and Challenges.” National Renewable Energy Laboratory, 2019, United States, https://doi.org/10.2172/1515397
14. github.com/alexwycoff/windspeed
15. Kingma, Diederik P., and Jimmy Ba. “Adam: A Method for Stochastic Optimization.” *arXiv preprint*, 2014, https://doi.org/10.48550/arXiv.1412.6980

**Figures and Figure Captions**

A graph with a green and blue box

Description automatically generated

**Figure 1: Box plots of the distribution of MAE scores.** For each site in the study (n=100), the mean absolute error (MAE) score on testing data was recorded for a persistence model and trained LSTM network. The distribution of these scores for each model is shown here, with persistence models in green and LSTM networks in blue. Black circles denote outliers in each distribution.

A graph with blue dots

Description automatically generated

**Figure 2: Effect of years of data on network MAE scores.**  LSTM networks were trained for 100 epochs using the most recent data for the northern Minnesota site for varying years of data. The data was split into training and testing groups, representing 80% and 20%, respectively. For each year, the network’s mean absolute error (MAE) on testing data was recorded and used to fit a logarithmic function.

A graph of a graph with blue dots

Description automatically generated with medium confidence

**Figure 3: Effect of the number of epochs trained on network MAE scores.** We experimented with how the number of epochs used to train a network would affect its mean absolute error (MAE) scores on test data from 2020. LSTM networks were trained using various numbers of epochs on data from 2015 to 2019 for the northern Minnesota site. For each number of epochs, the network’s MAE score on testing data was recorded and fit to a logarithmic function.

A graph of data with blue bars

Description automatically generated with medium confidence

**Figure 4: Average training time for an individual network with varying parameters.** We trained networks on all 100 sites selected in this study with each of the three network parameter setups given. The size of each error bar represents the standard deviation in the time to train one network with each setup.

A map of the united states

Description automatically generated

**Figure 5:** **Map of the Great Lakes region with site locations.** The physical locations of the 100 randomly selected sites from the NOW-23 Great Lakes dataset are displayed over a map of the Great Lakes region.

**Tables with Captions**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model and Parameters** | **Mean MAE** | **Median MAE** | **Mean RMSE** | **Median RMSE** |
| LSTM: 100 epochs, 20 years | 3.013 | 2.936 | 3.857 | 3.753 |
| LSTM: 100 epochs, 5 years | 2.992 | 2.933 | 3.914 | 3.835 |
| LSTM: 50 epochs, 5 years | 3.003 | 2.924 | 3.934 | 3.879 |
| Persistence: 20 years | 3.885 | 3.790 | 4.956 | 4.828 |
| Persistence: 5 years | 3.845 | 3.782 | 4.936 | 4.849 |

**Table 1: Model performance statistics.** We trained three LSTM models with varying epochs and years of data. LSTM models utilized 20 or 5 years of total data with an 80:20 train-test split. Both persistence and LSTM models were then evaluated using testing data for all 100 sites selected in this study, with each model's mean and median mean absolute error (MAE) and root mean square error (RMSE) in meters per second being recorded.

|  |  |  |
| --- | --- | --- |
| **Layer** | **Output Shape** | **Number of Parameters** |
| LSTM | (None, 16) | 1,280 |
| Dropout | (None, 16) | 0 |
| Dense | (None, 8) | 136 |
| Dense | (None, 1) | 9 |

**Table 2: Keras LSTM network architecture.** The architecture of the LSTM network used in this study as given by Keras. The parameters are the totals of the weights and biases associated with each layer. An output shape of (None, 16) indicates that one or more lists of length 16 are passed as output from a layer.