Evaluating the effectiveness of synthetic training data for day-ahead wind speed prediction in the Great Lakes

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 hypothesized that training long short-term memory (LSTM) neural networks to predict day-ahead wind speeds on synthetic wind data instead of observational wind data would increase accuracy since synthetic wind data is available over longer time spans than observational wind data. At an hourly sampling rate, wind patterns in synthetic wind data are similar to those in observational wind data, making synthetic wind data a viable substitution. We used observational data from the Lake Michigan Wind Assessment and synthetic data from the closest available site in the National Offshore Wind (NOW-23) Great Lakes dataset. We trained identical LSTM networks on both datasets, evaluating them using withheld observational data. While LSTM networks have been applied to wind speed prediction at a few selected sites, the networks' effectiveness when trained on synthetic data has not yet been tested. We also optimized additional parameters of the networks, further improving accuracy. The availability of a wind speed prediction model trained on synthetic data would 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 synthetic wind data to create a wind speed prediction model capable of making 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 LiDAR data from Lake Michigan (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). To validate the effectiveness of synthetic wind data for day-ahead wind speed prediction, we utilized observational data captured during the Lake Michigan Wind Assessment off the coast of Muskegon, Michigan (11).

In our study, we aim to create more versatile wind speed prediction models using synthetic wind data. We hypothesized that training long short-term memory (LSTM) neural networks to predict day-ahead wind speeds on synthetic wind data instead of observational wind data would increase accuracy. We used multiple experiments to optimize the parameters of the networks, further improving accuracy. To evaluate our hypothesis, we compared the accuracy of identical networks trained using synthetic and observational data on a sample of observational data withheld from training. Finally, we used statistical procedures to estimate the true mean difference in accuracy between the networks.

Results

We developed and tuned an LSTM network for time series prediction using data from the NOW-23 Great Lakes dataset at an elevation of 80 meters and the Lake Michigan Wind Assessment at 75 meters. Data from the 2013 Lake Michigan Wind Assessment was taken from a buoy approximately 10 kilometers from the eastern shoreline of Lake Michigan near Muskegon, Michigan, and data from the NOW-23 Great Lakes dataset was taken from the closest site to this buoy (**Figure 1**). The networks trained used the previous 24 hours of wind speed and direction data at a temporal resolution of 1 hour to predict wind speed in 24 hours at a given site. We trained LSTM networks on either the Lake Michigan Wind Assessment observational data or synthetic wind data from the NOW-23 Great Lakes dataset.

Networks trained on synthetic wind data used data from January 2000 to November 2013, and networks trained on observational wind data used data from April 2013 to November 2013. We evaluated all networks' mean absolute error (MAE) scores in meters per second on the final month of data from the Lake Michigan Wind Assessment, which was withheld from training. LSTM networks trained on synthetic data were significantly more effective than a persistence model, with the MAE score of 4.023 m/s achieved by a persistence model being 26% higher than the mean score of 2.995 m/s achieved by LSTM networks trained on synthetic wind data.

Through multiple experiments, we discovered that using more than ten years of data decreased network accuracy, while two or fewer years of data were not comprehensive enough to train an LSTM network effectively (**Figure 2**). Furthermore, we observed that larger increases in the number of epochs used to train networks on synthetic data had diminishing returns in accuracy and that additional epochs worsened accuracy in models trained on observational wind data (**Figures 3 and 4**). We also experimented with the batch size used to train networks on synthetic wind data, discovering that batch size had little effect on the accuracy of these networks but was significant in determining the accuracy of networks trained on observational wind data (**Figures 4 and 5**). Because of the limited amount of data available, networks trained on observational wind data often overfit to training data at higher numbers of epochs and require smaller batch sizes to better learn their datasets. Thus, we used 50 epochs and a batch size of 8 to train networks on observational and synthetic wind data for comparison since these parameters resulted in the lowest variance in MAE for both network types (**Figure 4**).

While the mean MAE of LSTM networks trained on synthetic wind data is lower than that of networks trained using observational data, it is not by a significant margin (**Figure 4**). To determine statistical significance, we recreated our experiment by retraining each network type 30 times to compile a distribution of MAE scores for both (**Figure 6**). Since we have distributions of 30 MAE scores per model, our data is sufficiently large and thus meets the conditions needed for a two-sample t-test. We will use the significance level $\alpha = 0.01$. Let μ_o be the true mean MAE of networks trained on observational wind data and μ_s be the true mean MAE of networks trained on synthetic wind data. Our null hypothesis is that $\mu_o - \mu_s = 0$, and our alternative hypothesis is that $\mu_o - \mu_s > 0$. We use $\overline{x_o} = 3.105\ m/s$ and $\overline{x_s} = 2.995\ m/s$ as the sample mean MAE of networks trained on observational and synthetic wind data, respectively. We also

use $s_o = 0.178\ m/s$ and $s_s = 0.105\ m/s$ as the standard error of MAE scores for networks trained on observational and synthetic wind data, respectively. A two-sample t-test yields a t-statistic $t = 2.915$ with a corresponding p-value of $p = 0.0027$. Since $0.0027 < \alpha = 0.01$, there is convincing evidence that LSTM networks trained on synthetic data are more accurate in day-ahead wind speed prediction than networks trained on observational data at the Muskegon site.

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 training on synthetic wind speed data could improve the accuracy of LSTM networks in day-ahead wind speed prediction. We tested LSTM networks using data from the Lake Michigan Wind Assessment, using statistical procedures to determine a significant difference in accuracy between the networks trained using synthetic and observational data. We concluded that there is convincing evidence that LSTM networks trained using synthetic data are more accurate than those trained using observational data. Furthermore, we optimized additional parameters of the networks through multiple experiments to improve network accuracy for both network types, ensuring that the networks are practical for use at a large scale.

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 approximately 17% across all models tested (8). Persistence models use the last known wind speed measurement to predict the next and are a typical benchmark for the performance of wind speed prediction models. When evaluated on observational data withheld from training, our LSTM networks trained on synthetic data had a mean MAE score approximately 26% lower than that obtained using a persistence model. 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 (12). 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 made in this study to further 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 higher accuracy in day-ahead prediction when using synthetic data. 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 synthetic 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 with LiDAR data from Lake Michigan (13). We used observational data from a buoy approximately 10 kilometers from the eastern shoreline of Lake Michigan in the 2013 Lake Michigan Wind Assessment near Muskegon, Michigan, at an elevation of 75 meters (11). We utilized both datasets at a temporal resolution of 1 hour. We trained LSTM neural networks on either synthetic or observational data to predict day-ahead wind speeds at the same location near the coast of Muskegon, Michigan (**Figure 1**).

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 our LSTM networks, we selected only wind speed in meters per second and wind direction in degrees as features. These observations were taken at either 75 or 80 meters, reflecting 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 (14).

Synthetic wind data from 2000 to 2013 for the NOW-23 Great Lakes dataset were retrieved from the National Renewable Energy Laboratory developer network API and concatenated for the Muskegon site using a script available on GitHub (15). Observational wind data from April 2013 to November 2013 for the Lake Michigan Wind Assessment were retrieved from the Atmosphere to Electrons website. The columns of the data were transformed to match those of the NOW-23 data and cleaned by replacing missing data values with the last known value from the column (15). Missing data made up approximately 5% of the data used for

testing. Synthetic and observational wind data were 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 for 50 epochs with a batch size of 8 and 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 1**). Multi-layer neural networks with dropout layers show better convergence when using the Adam optimizer, so it was chosen as the optimizer for training (16). 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 1**). Each network's MAE accuracy was calculated using functions from the Scikit-Learn library.

To optimize the accuracy of networks trained on synthetic data, we experimented with the number of epochs, years of data, and batch size used in training the networks on the Muskegon site. To evaluate the effect of altering these variables, we considered a network's MAE score when tested on a month of observational data from this site that had been withheld from training. We used the results of these experiments to inform the parameters of the networks trained on synthetic data we tested, which utilized ten years of data spanning from 2004 to 2013. We also compared networks trained on synthetic and observational data with various training parameters, which led us to train the networks we compared through statistical analysis with 50 epochs and a batch size of 8. We trained these LSTM networks on synthetic wind data 30 times and observational wind data 30 times to compile distributions of the MAE of each type of network.

References

1. Laurie, Carol. “Exploring Offshore Wind Energy Opportunities in the Great Lakes.” *National Renewable Energy Laboratory*, 26 June 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. Musial, Walt, et al. “New York Great Lakes Wind Energy Feasibility Study.” *New York State Energy Research and Development Authority*, December 2022, www.nyserda.ny.gov/All-Programs/Clean-Energy-Standard/Clean-Energy-Standard-Resources/Great-Lakes-Wind-Feasibility-Study
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. Standridge, Charles. *Lake Michigan Buoy near Muskegon, MI / Raw Data*. 2023, United States, https://doi.org/10.21947/1877893
12. 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
13. 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
14. 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
15. Wycoff, Alex, “windspeed: Wind Speed Prediction with Synthetic Data.” GitHub. github.com/alexwycoff/windspeed
16. 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

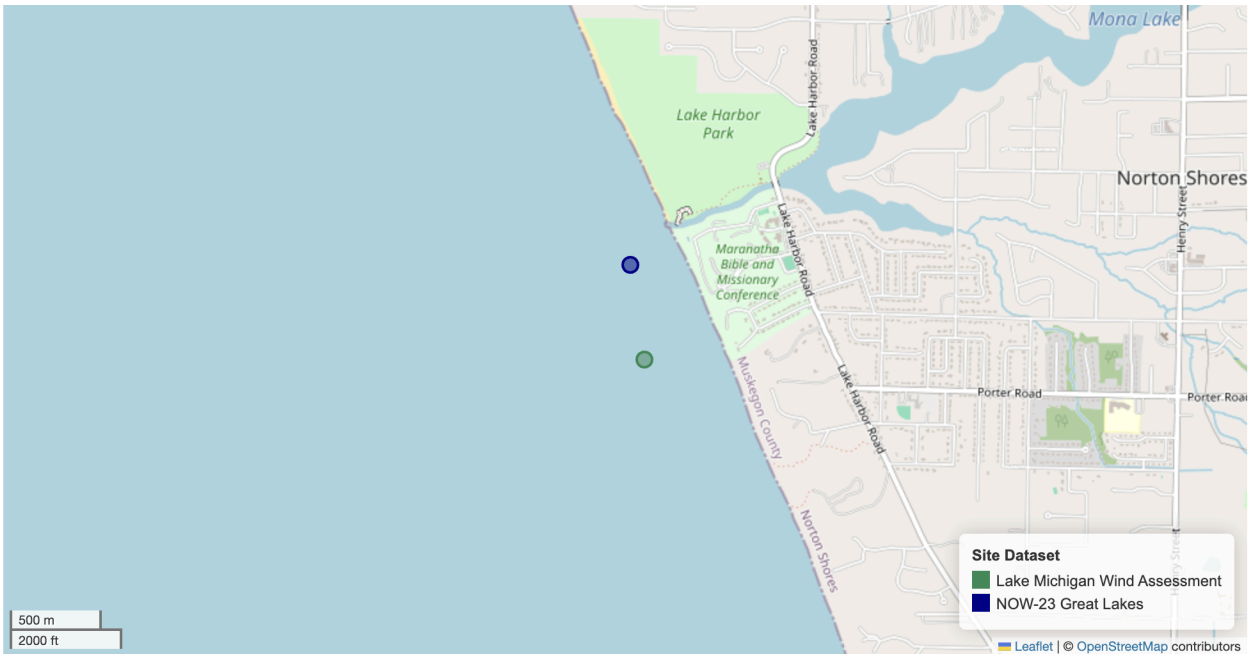

Figure 1: Map of Muskegon, Michigan, with site locations. The physical locations of the sites selected from the NOW-23 Great Lakes dataset and Lake Michigan Wind Assessment are displayed over a map of Muskegon, Michigan. Map data is available from openstreetmap.org under the Open Database License.

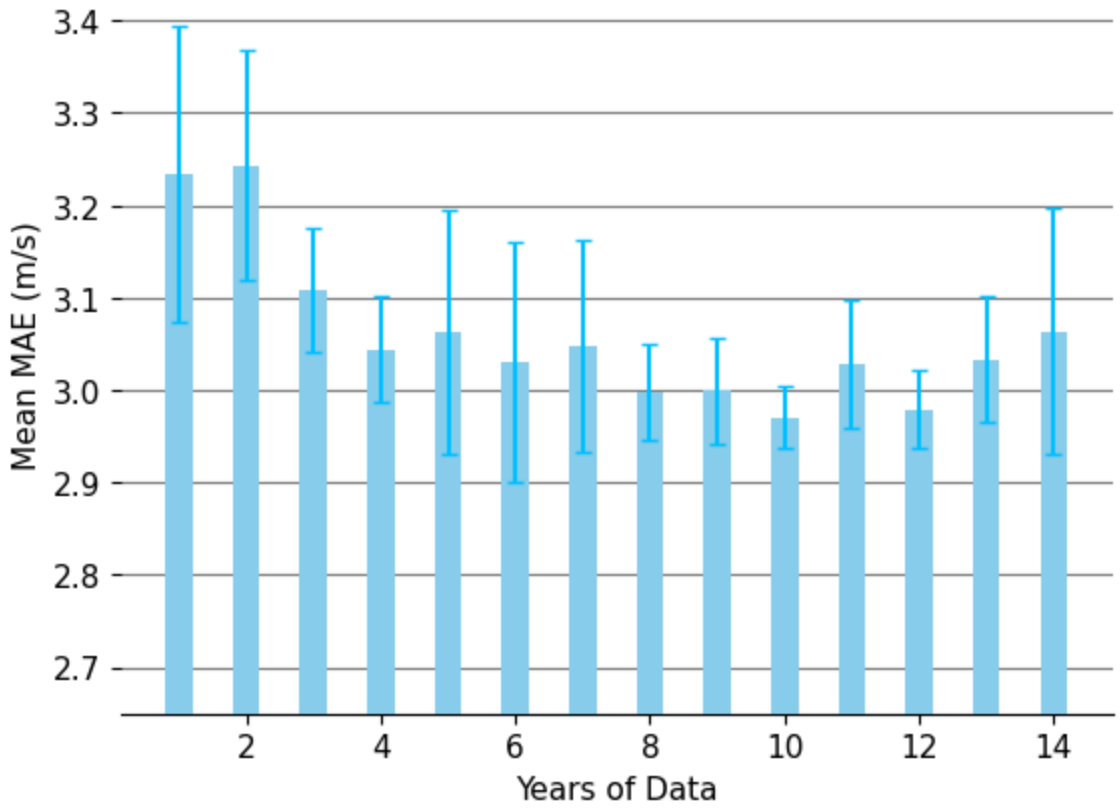

Figure 2: Effect of years of data on network MAE scores. We experimented with how the number of years of synthetic data used in training would affect networks' mean absolute error (MAE) scores on withheld observational data. For each number of years of data, 10 LSTM networks were trained for 50 epochs with a batch size of 128 using the most recent synthetic data for the Muskegon site. Error bars represent the standard deviation of the MAE scores of networks trained on data spanning a certain number of years.

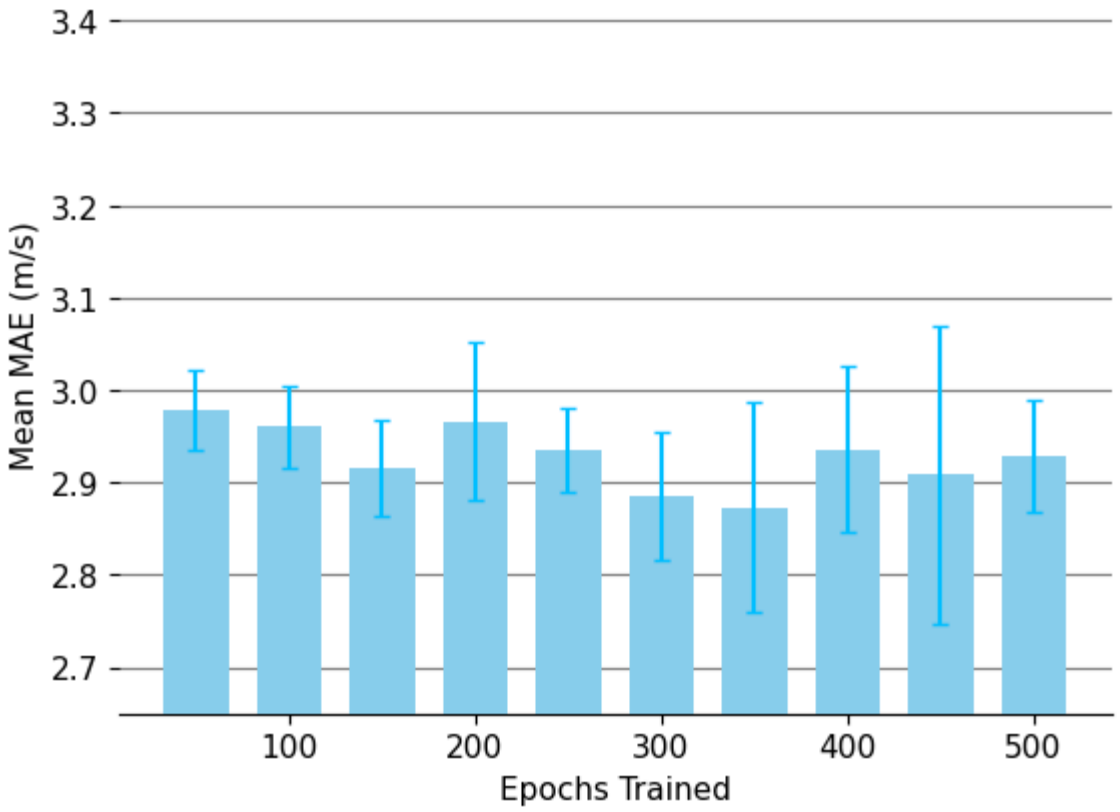

Figure 3: Effect of the number of epochs trained on network MAE scores. We experimented with how the number of epochs used to train a network on synthetic data would affect its mean absolute error (MAE) scores on withheld observational data. For each number of epochs, 10 LSTM networks were with a batch size of 128 on synthetic data from 2000 to 2013 for the Muskegon site. Error bars represent the standard deviation of the MAE scores of networks trained with a certain number of epochs.

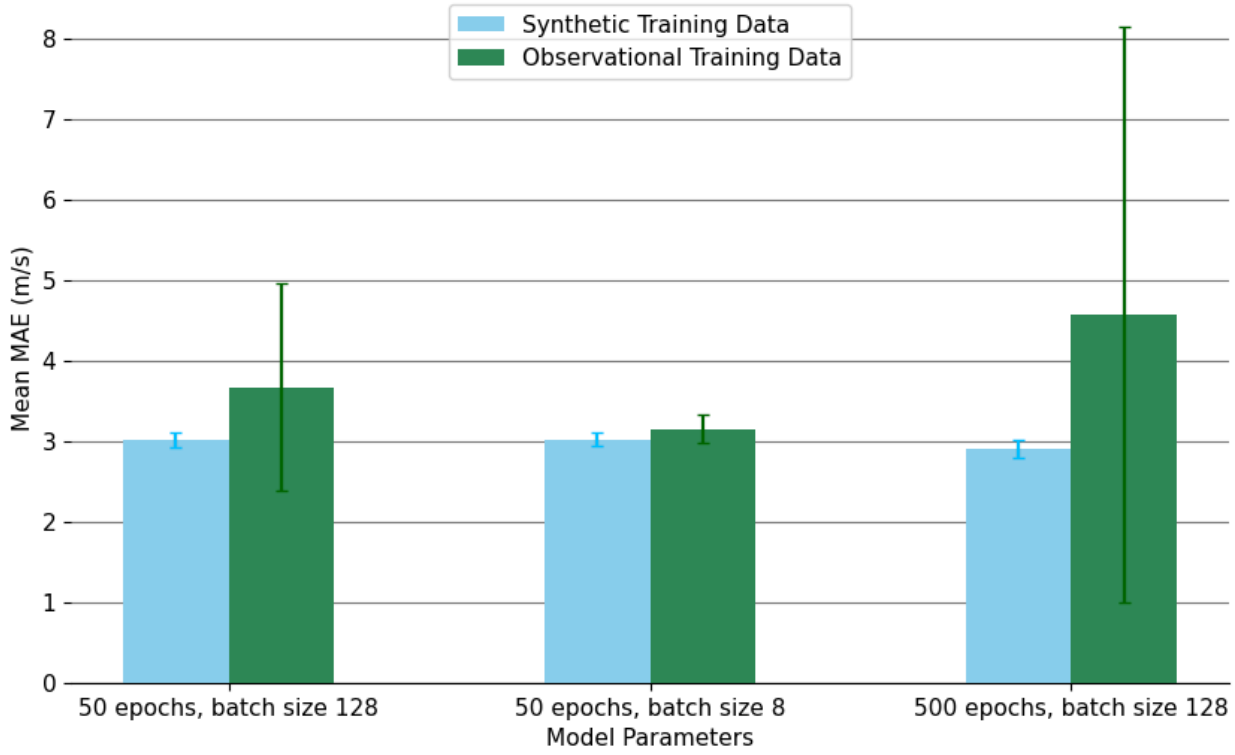

Figure 4: Comparison of MAE scores for models with varying parameters. We trained LSTM networks on synthetic or observational training data while varying the number of epochs and batch sizes used in training. For each unique network type, we trained 30 networks to account for random variability in the network training process. Error bars represent the standard deviation of the MAE scores of the 30 networks trained of each type.

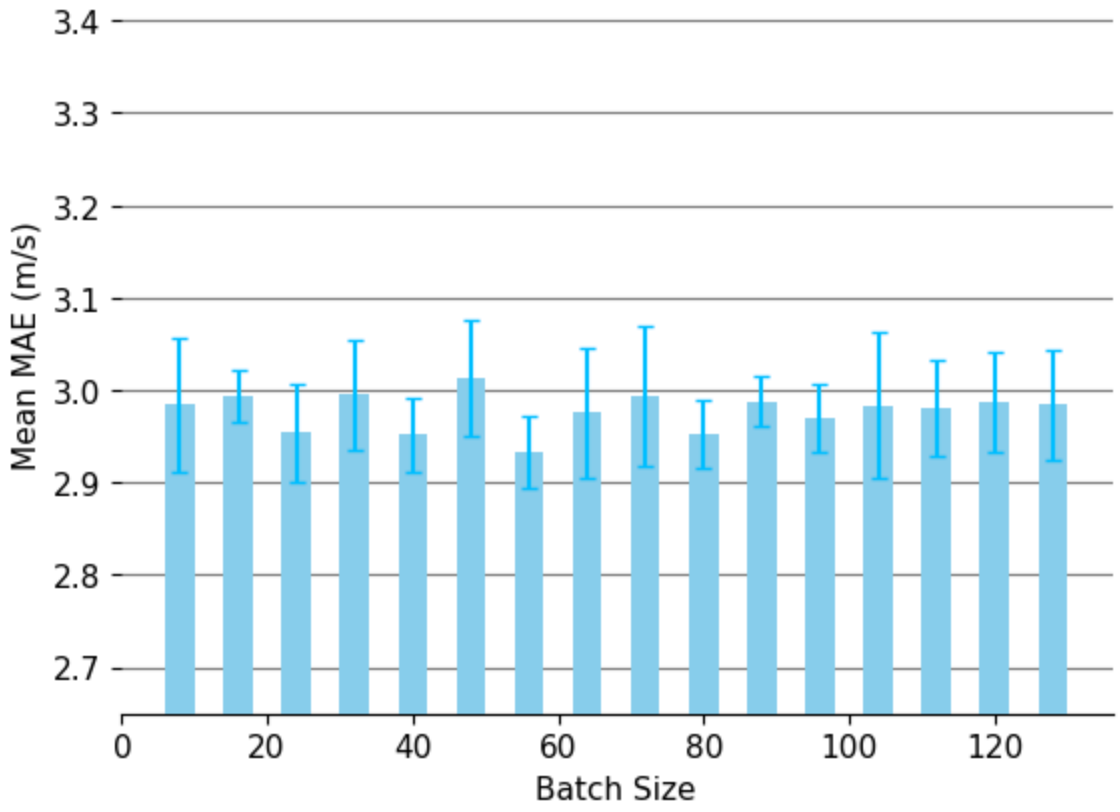

Figure 5: Effect of batch size on network MAE scores. We experimented with how the batch size used while training a network on synthetic data would affect its Mean Absolute Error (MAE) score on withheld observational data. For each batch size, 10 LSTM networks were trained for 50 epochs on synthetic data from 2000 to 2013 for the Muskegon site. Error bars represent the standard deviation of the MAE scores of networks trained with a specific batch size.

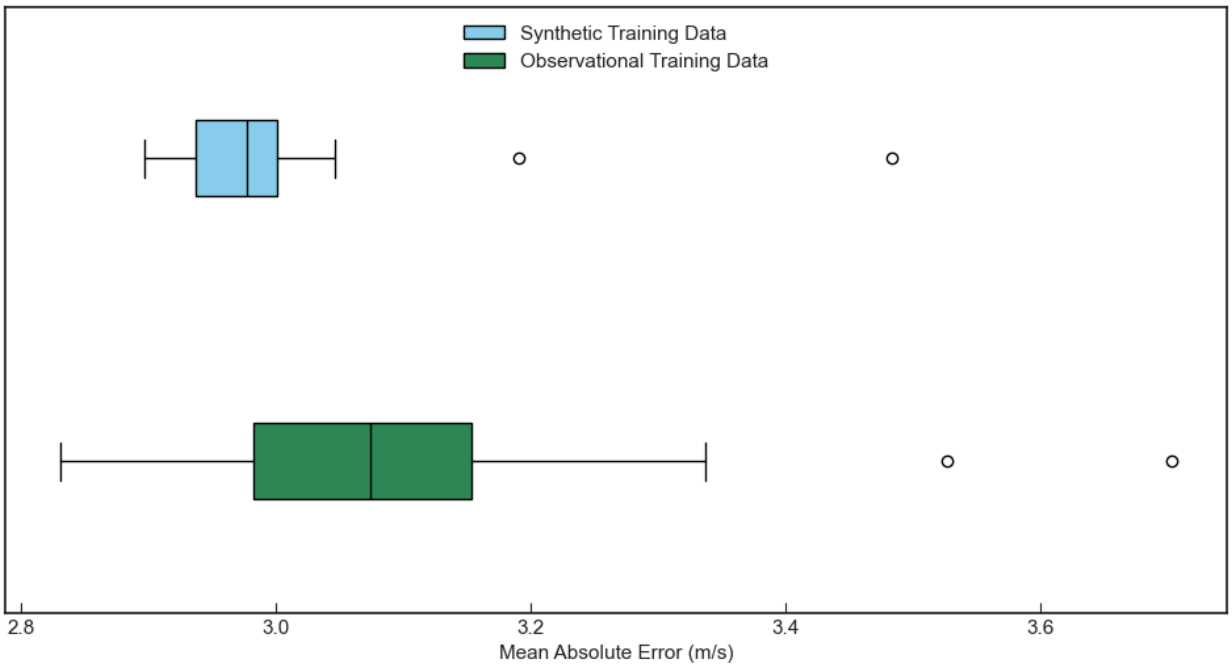

Figure 6: Box plots of distributions of MAE scores. The mean absolute error (MAE) scores on a month of observational data withheld from training were recorded for 30 LSTM networks trained on synthetic data and 30 identical LSTM networks trained on observational data. The distribution of these scores for both network types is shown here, with observational networks in green and synthetic networks in blue. Circles denote outliers in each distribution. Both network types were trained using 50 epochs and a batch size of 8.

Tables with Captions

Layer	Output Shape	Number of Parameters
LSTM	(None, 16)	1,216

Dropout	(None, 16)	0
Dense	(None, 8)	136
Dense	(None, 1)	9

Table 1: 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.