Article

Not peer-reviewed version

A Comparative Study for Using Deep LSTMs and ARIMA for Imputing Missing Data for Wind Data in the Irish Sea

Gohar Shoukat , Abdollah Malekjafarian , Vikram Pakrashi *

Posted Date: 21 May 2024

doi: 10.20944/preprints202405.1179.v1

Keywords: Irish Offshore Wind; LSTM; Renewable Energy; ARIMA; Data Imputation

 doi:10.20944/preprints202405.1179.v1

Article

A Comparative Study for Using Deep LSTMs and ARIMA for Imputing Missing Data for Wind Data in the Irish Sea

Gohar Shoukat [1,†], Abdollah Malekjafarian [2] and Vikram Pakrashi [1,*]

[1] UCD Centre for Mechanics, Dynamical Systems and Risk Laboratory, School of Mechanical and Materials Engineering, University College Dublin, D04 V1W8 Dublin, Ireland 1; gohar.shoukat@ucdconnect.ie
[2] Structural Dynamics and Assessment Laboratory, School of Civil Engineering, University College Dublin, D04 V1W8 Dublin, Ireland
* Correspondence: vikram.pakrashi@ucd.ie
† Climate Risk Services, London, UK.

Abstract: Achieving Net Zero emissions target is driving up the need for offshore wind as an alternate energy source. Ireland has the capacity to produce upwards of 30GW of power through offshore wind alone. Wind resource assessments are extremely vital in determining the long term trends of a site. The instruments used for this often suffer breakdowns or miss readings which impacts the long term trend analysis. Wind time series data from Ireland's Marine Institute is used which is shown to have significant gaps in it's 20 years of service. Data imputation is necessary to fill in these gaps as accurately as possible. This paper compares LSTMs and ARIMA as two competing methodologies for data imputation. LSTMs are shown to be marginally superior to ARIMA imputation with a mean squared error of 0.45 compared to that of ARIMA of 0.60.

Keywords: Irish offshore wind; LSTM; renewable energy; ARIMA; data imputation

1. Introduction

Reducing carbon emissions to mitigate climate change and reducing dependence on fossil fuels are a priority for a lot of countries globally. Consequently, investments in the renewable energy sector, in particular, wind energy are experiencing a significant rise. The Global Wind Energy Council (GWEC) reports that an estimated 680 GW of wind energy are to be added to the global energy mix between 2023 and 2027 [1].

Ireland, in the Atlantic Ocean is quite well placed in Europe terms of renewable energy capacity [2]. Irish Government is aiming to generate around 7 GW through wind energy by 2030, however, estimates suggest that the wind potential in the North Atlantic around Ireland is such that between 30 and 70 GW of energy can be produced [3,4]. Ireland has entered a transition phase where it aims to phase out fossil based production and resort to more sustainable sources. Offshore wind is expected to be pivotal in that regard [3].

While Ireland is still in the early days of maritime planning, the recent designation of 8600 km^2 off the south-east coast [5] will provide ample opportunity to evaluate procedures and plans set in place by European Union(EU). This area alone will be used to setup offshore wind farm with power generation capacity of 900MW [5]. Additionally, the government also auctioned off four additional wind farms with a combined capacity of 9 GW and a total investment of €9bn. The winning bids included wind farms off the coast of Dublin (with a capacity of 850 MW) called Dublin Array, off the west coast near Galway (with a capacity of 450MW) named Sceirde Rocks, off the coasts of Dublin, Louthe and Meath (with a capacity of 500 MW) called North Irish Sea Array and finally, off the Wicklow Coast, just south of Dublin (with a capacity of 1450 MW) named Codling Wind Farm.

These wind farms, located on the west, south and east of Ireland aim to leverage Ireland's strategically advantageous location with regards to wind energy with the aim of exporting the excess energy into the EU [6]. However, a crucial factor in the siting of wind farm is the estimated annual energy production and how that value compares to that of energy produced by other sources [7]. Using

 doi:10.20944/preprints202405.1179.v1

accurate, long-term data to make these determinations is therefore critical. Data should be collected at the site of interest for at least two to three years after which questions about long-term annual variability and annual energy production can be gauged [7]. Wind turbines accumulate damage and fatigue overtime; it is therefore important they are designed to last for the duration of their service life [8] making it extremely vital that the long term complex loading data is available. This will prevent over or under designing the wind turbines.

Long-term wind resource analysis is conducted through historical data. Studies with historical data take more than three years worth of data into account when carrying out long-term analysis with most research groups leveraging at least 40 years of data [9–12]. However, historical data is difficult to acquire for a potential site since it would be impossible to know 40 years in advance about a potential wind site when renewable energy was still in its infancy.

Collecting data for potential wind sites for several years before deciding on whether this would be a good site can be logistically and financially infeasible. Therefore, data from nearby sites or third parties is regularly sourced to make assessments [13]. This historical data can be real observed data from a neighbouring site or numerically simulated. Kim and Kim [13] used data from Yeosu Airport to carry out pre-feasibility wind resource assessment for a 30MW wind farm. The determination that historical data from a site is sufficient to make long-term assumptions about another site needs rigorous analysis [7]. Nelson and Starcher [7] state that in order to use cross-site data to determine historical trends, the annual hourly linear correlation coefficient should at least be 0.90 between the reference site and off-site data. If the two sites do not show similar trends in wind speed and directionality or topography, the correlations would be weak.

Instruments used for measuring wind speed can occasionally suffer from breakdowns and stop recording. These breakdowns cause missing values which can extend from a few hours to days as is the case with the data collected by the offshore buoys of Marine Institute(MI) [14]. MI's buoys have been in service since early 2000 and as such provide excellent database for historical data [14] for Ireland's offshore met-ocean conditions. However, many of the buoys deployed record missing values. These missing values can be classified as missing completely at random (MCAR). This means that the pattern of missing values is completely random and does not depend on any information contained within the dataset [15]. The assumption for MCAR is that the probability of encountering missing values depends neither on the observed values nor on the unobserved values.

Reanalysis techniques such as those employed by Copernicus use data assimilation technique [11, 16] that rely on observed data. Reanalysis techniques provide the most comprehensive climate data at regular intervals over long time periods - often decades. The quality of reanalysis depends on the data assimilation system itself which in turn relies on the observed data. Data assimilation is the science of combining different sources of data to determine the state of a system as it evolves over time [17].

Missing data presents a serious challenge to data assimilation [18]. Therefore, imputation to improve assimilation is [19,20] imperative. Sareen et al. [21] argue that short term wind speed forecasts show poor results when they have a high number of missing values. When the time series is first imputed and then a bi-directional long term short memory neural network used to make predictions, the results show a higher degree of accuracy. Kaur et al. [22] similarly argue how artificial neural networks make improved predictions on avalanches with imputations.

Standard imputation techniques like averaging might not always be the most accurate as argued by Steffan et al. [23] where time series data is in question. This is because of the nature of the time series data where structural dependencies exist between future and past data. Steffan et al. used existing time-series packages in R to impute data and concluded that seasonal kalman filter and a linear interpolation on seasonal loess decomposed data were most effective.

Liu et al. [24] used Gaussian process regression (GPR) to develop short-term prediction models to impute wind speed time series. They compared this with mean substitution and k-nearest neighbour (KNN) and concluded that GPR outperforms them. Shukur et al. [25] developed a hybrid artificial neural network and auto-regressive model and reached similar conclusions. They showed that

 doi:10.20944/preprints202405.1179.v1

when the time-series is nonlinear, the hybrid machine learning technique produces better imputation results compared to linear regression, KNN and state space methods. Liao et al. [26] developed a model based on context encoders to handle highly non-linear data. To benchmark their model, they compared it against auto-encoder, K-means, k-nearest neighbor, back propagation neural network, cubic interpolation, and conditional generative adversarial network and concluded that context econding technique gives better results. Liu et al. [27] used a hybrid convolutional neural network with bi-directional recurrent neural network to impute spatio-temporal satellite based aerosol optical depth. Their model is reported to impute missing data with low error.

The present study targets wind time series around Ireland since Ireland is set to take off as a major wind energy producer. Additionally, the study makes use of MI buoys deployed and focuses on univariate time series data imputation. This is particularly important as the authors wanted to quantify the error in imputation in the absence of additional variables like sea-state conditions.

Bouys do not always have the sensors available to measure wave heights, currents and direction of waves. Wind speed is therefore the only data being collected. Thus, wind speed is the only variate considered for this data imputation study.

2. Data

The data being used for the study is sourced through from Marine Institute(MI) [14]. MI is an Irish State Agency with a focus on marine research and technological development. MI has a dedicated Oceanography Services Team [28] which operates and maintains devices for monitoring waves, currents, temperature and other properties.

2.1. Irish Marine Data Buoy Observation Network

This network of marine buoys is managed by the Marine Institute in a joint collaboration between Met Eireann and the UK Met Office. Their coordinates of position are given by Table 1. The data from the buoy network is used for weather forecasts, shipping bulletins, gale and swell warnings. It also as serves as a data source for general public information and research.

Table 1. Location of the five buoys deployed around Ireland for oceanographic and meteorological condition monitoring [29].

Buoy	Latitude (°N)	Longitude (°W)
M2	53.48	5.425
M3	51.2166	10.55
M4	54.9982	9.992154
M5	51.69	6.704
M6	53.07482	15.88135

2.2. M2

This study is based on the data obtained MI's M2 buoy. It is positioned in the east of Ireland, just off the coast of Dublin with coordinates shown in table 1. M2 is one of the six buoys deployed around Ireland by MI. This particular buoy is equipped to measure oceanographic and meteorological variables as shown in Table 2:

 doi:10.20944/preprints202405.1179.v1

Table 2. List of meteorological and oceanographic variables that the M2 buoy is equipped to measure.

Meteorological Variables	Oceanographic Variables
Atmospheric Pressure (mB)	Significant Wave Height (m)
Wind Speed (kn)	Wave Period (s)
Wind Direction (°)	Max Wave Height (m)
Max Gust (kn)	Max Wave Period (s)
Air Temperature (°C)	Mean Direction (°)
Relative Humidity (%)	Sea Temperature (°C)
-	Salinity

Under the forces of nature, the buoys often stop recording due to malfunction. However, not all instruments go down at the same time. Some instruments may still continue to function while others stop. In the case of M2, wind speed has approximately 7% missing data since it was first deployed in 2001. However, wave height for the same duration observes approximately 14% missing data. This also means that a multivariate analysis is not always possible since other meteorological and oceanographic variables for the same instance of time might not be available. These missing values occur at random and are not always normally distributed across the period of deployment.

The MI's records provide data for hourly average for the duration of deployment of the buoy. Each data point on the Figure 1 is an hourly average for wind speed. The data shows that the wind speed off the coast of Dublin has an average speed of 7.24m s^{-1}. A good representation of the missing values in the wind speed record can be seen in figure 1. Elongated period of missing values can be seen in the figure between 2011 and 2012 and then again from mid-2013 to end of 2014.

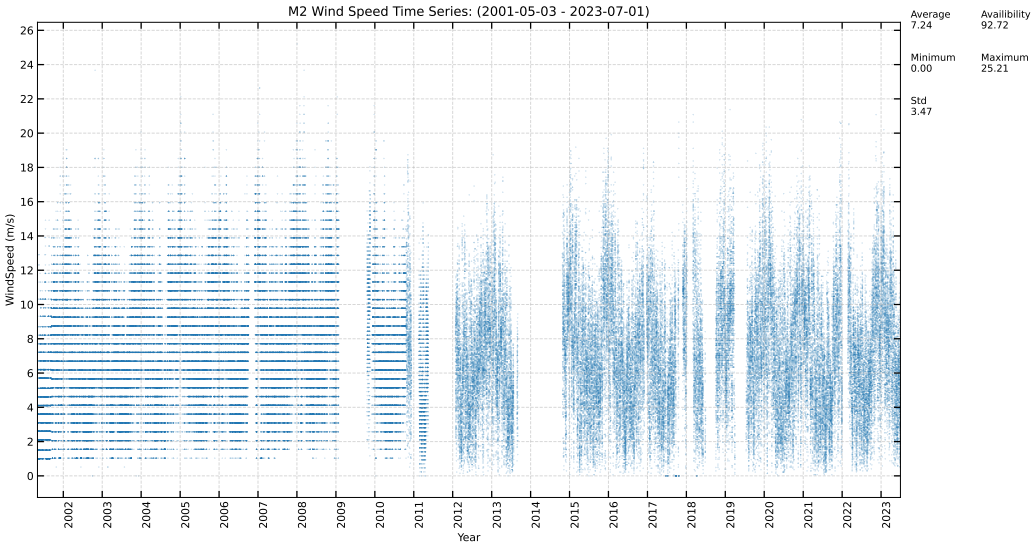

Figure 1. Wind speed time series for M2 buoy for the duration of deployment from 2001 to 2023. The gaps in the figure represent the missed values that the instruments on the bouy failed to record. The length of the gap signifies the number of readings missed. Summary statistics of the dataset are also visible in the top right corner. It is to be noted that the speeds measured prior to 2011 had a lower precision which resulted in same readings.

Figure 2 highlights the frequency of occurrence for the consecutive missing values observed within the dataset. For instance, for the M2 buoy, 939 times a reading was skipped for one hour, 110 times two consecutive readings were missed. Hourly missing values are the most common across the buoys. The frequency of occurrence for two hours or more sharply declines as can be seen in the Figure 2. This is primarily why the study focuses on hourly imputation of the time series data.

 doi:10.20944/preprints202405.1179.v1

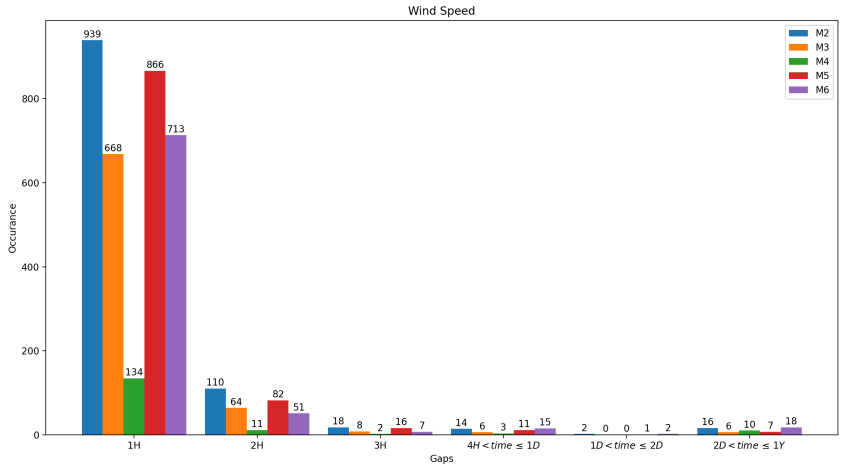

Figure 2. Bar chart showing the number of times each of the five buoys in the MI's network did not register a reading for the successive number of hours given by the x-axis. 'H' represents the hours.

The data is further investigated for any auto- and partial auto- correlations (ACF and PACF respectively). Figure 3 shows the result of the ACF for the first 720 hours after the buoy was deployed. The data shows high dependency on previous time steps with high seasonality across the first 30 days of operation. Only after the 450^{th} lag does the correlation go below the threshold of 0.05. Figure 4 gives the PACF of the time series for the same duration and that shows that the partial correlation goes below the threshold of 0.05 after the second lag. It does not show long term corelation between the

Separate investigations of the data using ACF showed that time windowing the series tends to reduce the seasonality. If the entire time series were to be taken into account, the ACF shows full dependence on previous values for the entirety of the time series as shown by the Figure A1.

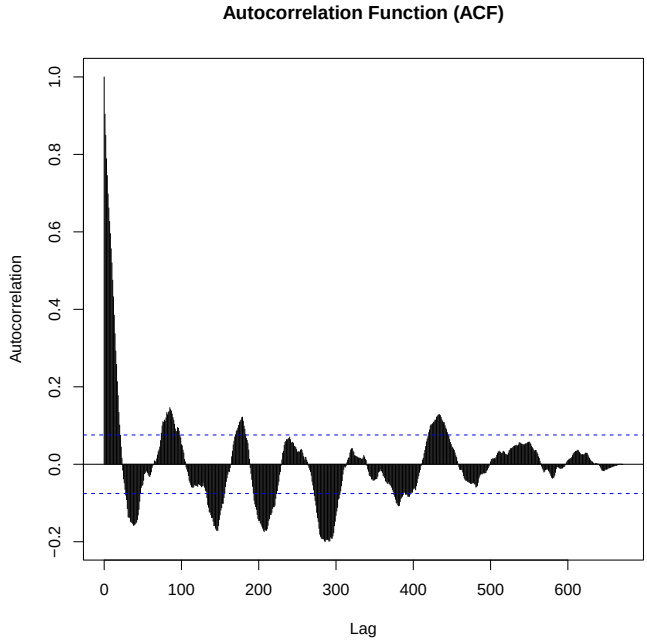

Figure 3. Auto-correlation function of the wind time series for 30 days. The time steps taken into consideration here are the first 30 days after buoys deployment to achieve maximum availability of data.

 doi:10.20944/preprints202405.1179.v1

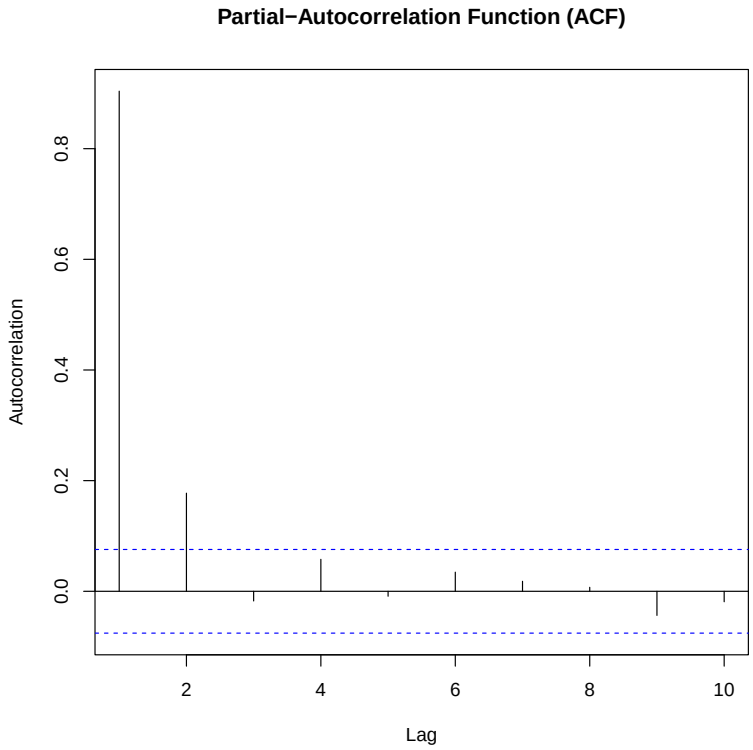

Figure 4. Partial auto-correlation function of the wind time series for 30 days. The time steps taken into consideration here are the first 30 days after buoys deployment to achieve maximum availability of data.

3. Methodology

While differencing the time series can reduce the seasonality and thus ACF, few problems arise while using the full time series for training any statistical technique for imputation or for that matter, forecasting.

It is evident that wind time series data from Figures 3 and A1 has long term trends and short term seasonality trends that need to be appropriately captured. ARIMA model is often termed as a linear time series analysis tool due to its inability to capture complex non-linear patterns in the dataset [30,31]. Khashei and Bijari [32] also highlighted the failure of ARIMA to capture the non-linear patters in the dataset and instead turned towards Artificial Neural Netowrks. Wang et al. [33] echoed the findings and concluded a hybrid ARIMA and metabolic grey model would instead be better to capture long term non-linearity trends.

However, Dong et al. [34] argued that many of the failures of ARIMA could be avoided by using a sliding window approach. They concluded that the lack of non-linearity capture by ARIMA is not a concern as long as the sliding window of training is carefully selected. Sheoran and Pasari. [35] reached the same conclusion that daily and weekly sliding windows with ARIMA outperform modelling the entire time series with conventional ARIMA.

LSTMs suffer from similar problems if not entirely the same. While they can retain information about complex non-linear patterns, they struggle with retaining this information over longer sequences. Miller and Hardt [36] considered RNNs and LSTMs as dynamic systems and concluded that LSTMs do not actually have long term memory. Tunnell and Harchaoui [37] applied LSTMs to music and language and reached the same conclusion that LSTMs struggle with fully representing the long memory effect in the input and can not generate long memory sequences from white noise inputs. Zhao et al. [38] while reaching the same conclusion as other researchers proposed a new definition for long-memory. They argued that because a time series is not inherently i.i.d, it violates the primary

 doi:10.20944/preprints202405.1179.v1

rule of an ANN that all inputs should be independent of each other. They go on to propose a new Memory-LSTM that attempts to retain long memory.

3.1. Training Dataset

Therefore, a 30 days period with a full lunar event is considered in this study instead of the entirety of the 20+ years of data available to avoid some of the problems discussed above. This study will attempt to highlight if LSTMs with their computational expense and hyperparameter tuning can outperform conventional ARIMA in a window of data for imputation of missing values.

Training data was created from this 30 day wind time series with complete hourly data. Artificial hourly gaps were introduced completely at random [39] in the dataset and the hourly availability of the dataset was reduced to 90%. Availability is calculated using equation 1. Consequently, the models will be trained on the 90% data and tested on 10% of unseen data.

Additionally, since the raw data showed auto-correlation up to 500 lags (Figure 3), a differenced time series was used.

$$Availability = \frac{No.\,of\ time\ steps\ with\ data}{Total\ no.\,of\ time\ steps} \times 100 \tag{1}$$

3.2. ARIMA

ARIMA model is fit onto the time series data. This is done through the R package *forecast* [40]. ARIMA combines autoregressive features with those of moving averages. An AR(1) - autoregressive order of one signifies that the current value is determined only by the immediate preceding value, while an AR(2) means that the current value is based on the previous two values. The moving average component on the other hand analyses data points by smoothing different subsets of the data set to remove the influence of outliers. MA(1) - moving average of order one truncates after a lag of one. This means that the auto-covariance function drops to zero after lag one. A MA(q) would therefore mean that the auto-covariance drops to zero after q lags. The I in ARIMA stands for Integrated which denotes the number of times the series has to be differenced to achieve stationarity. The Auto-regressive (AR) model of order p, denoted as AR(p), is given by the equation:

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \varepsilon_t$$

where:

- X_t is the value of the time series at time t,
- c is a constant term (often omitted),
- $\phi_1, \phi_2, \ldots, \phi_p$ are the parameters of the model,
- $X_{t-1}, X_{t-2}, \ldots, X_{t-p}$ are the lagged values of the time series,
- ε_t is the white noise error term at time t.

The Moving Average (MA) model of order q, denoted as MA(q), is given by the equation:

$$X_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

where:

- X_t is the value of the time series at time t,
- μ is the mean of the time series (often assumed to be zero),
- ε_t is the white noise error term at time t,
- $\theta_1, \theta_2, \ldots, \theta_q$ are the parameters of the model,
- $\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots, \varepsilon_{t-q}$ are the lagged error terms.

Auto-Arima, another one of the utilities within the package goes over several different combinations of the p, d and q parameters and settles on the ones with the lowest AIC.

 doi:10.20944/preprints202405.1179.v1

3.3. LSTM

LSTMs [41] is a form of RNN. A kind of RNN a architecture called Long Short-Term Memory (LSTM) was created to solve the vanishing gradient issue with conventional RNNs. LSTMs are extensively employed in a wide range of sequence generating and prediction tasks, including speech recognition, time series forecasting, natural language processing, and more. LSTMs are made with the intention of capturing long-term dependencies in sequential data by keeping a memory cell that has the capacity to hold data for extended periods of time.

An LSTM's memory cell is its central component. It enables LSTMs to forget or retain knowledge over time in a chosen manner. The network's "memory" can be compared to the state of the cell. To regulate the information flow via the memory cell, LSTMs employ gates. To regulate the information flow via the memory cell, LSTMs employs three different gates. The forget gate decides what information to discard from the cell state. It takes as input the concatenation of the current input and the previous hidden state, and its activation function is the sigmoid function. The input gate determines which new information to store in the cell state. It also takes as input the concatenation of the current input and the previous hidden state. The input gate uses the sigmoid function to regulate which values will be updated, and it also uses the hyperbolic tangent function to create a vector of new candidate values.

LSTMs can be further enhanced by modifying the network to be stateful [42]. This is particularly helpful in time series modelling where long term structures in data exist. A stateful LSTM is a type of RNN architecture that is capable of capturing long-term dependencies in sequential data while also maintaining an internal state or memory. Unlike its counterpart, the stateless LSTM, which resets its internal state after processing each sequence, the stateful LSTM retains its state across multiple sequences within a given batch of data. This allows the model to remember information from previous sequences and use it to make predictions or generate outputs for subsequent sequences.

The LSTM architecture used for this study is represented by Figure 5. It shows a deep neural network with five layers of LSTM, one FFN layer with eight neurons preceding a final layer of FFN connecting the output of LSTMs with the output layer. The activation function used for the penultimate layer is the RELU with a linear function used in the final layer before output. To reduce overfitting, dropouts ranging from 0 - 0.05 were employed. Several simulations for hyperparameter were performed to identify the optimum set of parameters.

Figure 5. High level architecture of the neural network used for training. x shows the input(s) and y shows the output which in this case is x_{t+1} if $x = x_t$.

3.4. Feature Selection

Feature selection for LSTMs was done through calculating PACF between the dataset. Figure 4 shows the PACF of the time series. For LSTM modelling, one time step will be used as the primary feature. Mathematically, this means that the feature chosen is x_{t-1} to predict or impute the x_t value. However, to safely conclude that this is indeed the right decision, models with up to two lags will also be used as primary features to predict the output state. Mathematically, up to two features will be used as inputs to the LSTM model: x_{t-2} and x_{t-1}.

3.5. Hyperparameter Tuning

The LSTMs were trained on several different configurations to identify the combination that produces the lowest MSE. The following parameters were adjusted and compared:

- Layers of LSTM
- Layers of FFN
- Regularisation for LSTM and FFN using dropout
- Stateful vs Stateless LSTMs

 doi:10.20944/preprints202405.1179.v1

4. Results

The parameters concluded for fitting ARIMA on the dataset were (2, 2, 0) for (p, d, q). It is worth noting that the lowest AIC was obtained with a second differenced model with AR(2). MSE for the imputed values for this model comes out to be 0.64.

Figure 6 shows the PDF of both the originally deleted values and the imputed values. It can be observed that the values are deleted at random with the highest number of values deleted near the middle. Coincidentally, the mean of the first four weeks of time series is 7.007m s^{-1}. Since this was a random deletion, values were deleted from both sides of the mean. The lowest recorded speed that was deleted was 2.05m s^{-1} and the highest was 12.37m s^{-1}.

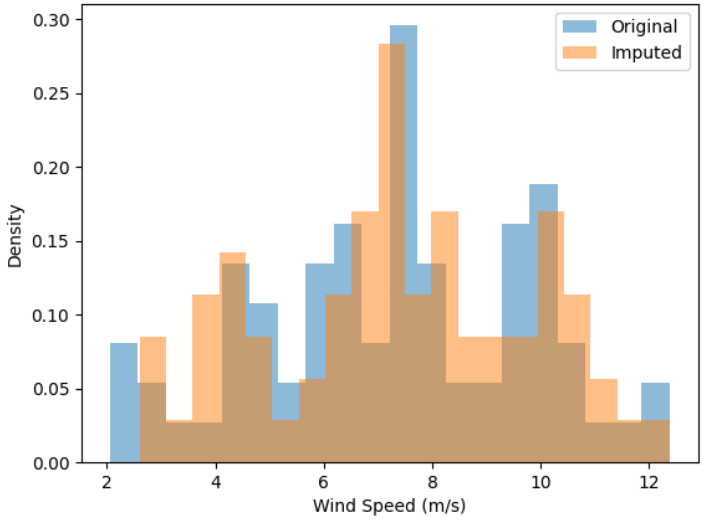

Figure 6. PDF of the 10% imputed values compared against the PDf of the original values using ARIMA as an imputation technique.

The ARIMA model's PDF is interesting. It shows, contrary to the PDF of the original series, the highest density of imputations are around the mean even though the spread of the actual values is quite even across the entire range of the series.

Figure 7 confirms this hypothesis as well. Missing values near the mean were imputed rather accurately as denoted by the line $y = x$. The nearer a data point is to this line, the lesser the error. Such that data points on the line show 100% accuracy and points above or below the line show inaccuracy with the vertical distance from the line showing the degree of inaccuracy. It is worth noting that extreme values either side of the mean show rather poor accuracy of imputation. The values below the median are over predicted and the values above the mean are under predicted. This is because of the inherent tendency of the statistical model to stay true of the mean of the time series.

 doi:10.20944/preprints202405.1179.v1

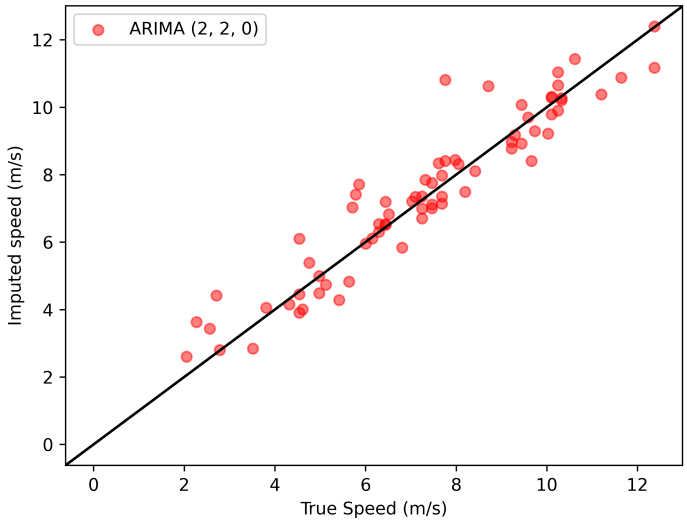

Figure 7. QQ comparison of the hourly imputed series against the original values using ARIMA as an imputation technique.

Figure 8 shows the PDF of the results obtained through LSTM imputation plotted in the backdrop of the true values. Interestingly, and unlike the ARIMA imputation, this is releatively more spread out across the range of the time series. While the ARIMA formed a strong aggregation with a higher density around mean values, LSTM imputation shows a far more equal spread out. This however, is observed only till 11 $\mathrm{m\,s^{-1}}$. Beyond this speed, LSTM fails to predict an equivalent magnitude. On the other hand, the lower extreme shows a much higher density of predictions. It can be argued that LSTM does a releatively good job in identifying the lower side of the range better than it does on the opposite end of the series.

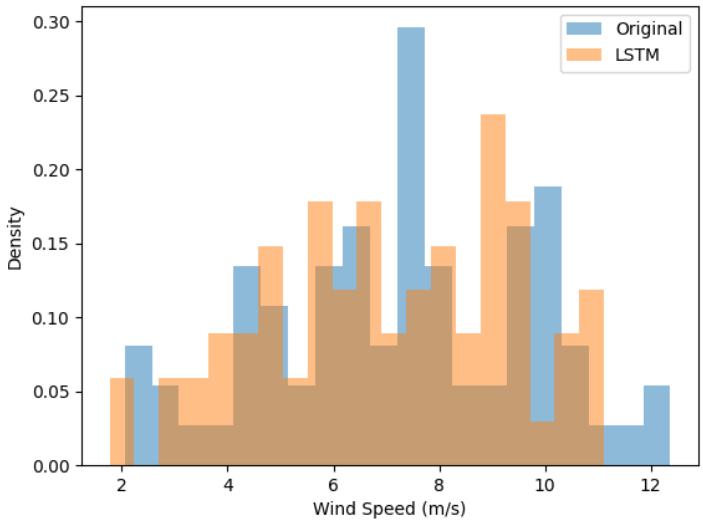

Figure 8. PDF of the 10% imputed values compared against the PDf of the original values using LSTM as imputation technique.

Figure 9 shows a visual comparison of the actual imputed values with the true values. The line $y = x$ shows where the true and predicted values are equal. The further away the marks are from the line, the greater the error. As with the PDF plot, near the median values, the cluster of points tends to

 doi:10.20944/preprints202405.1179.v1

stay very close to the line. On either side of the median though, there is a higher diversion. Similar to the results of ARIMA imputation in figure 7, the values below the mean are more often than not over predicted while the values above the mean are underpredicted. This is an observed trend for values above 11m s^{-1} quite regularly as all the values are under predicted and are significantly further away from true values as indicated by the line.

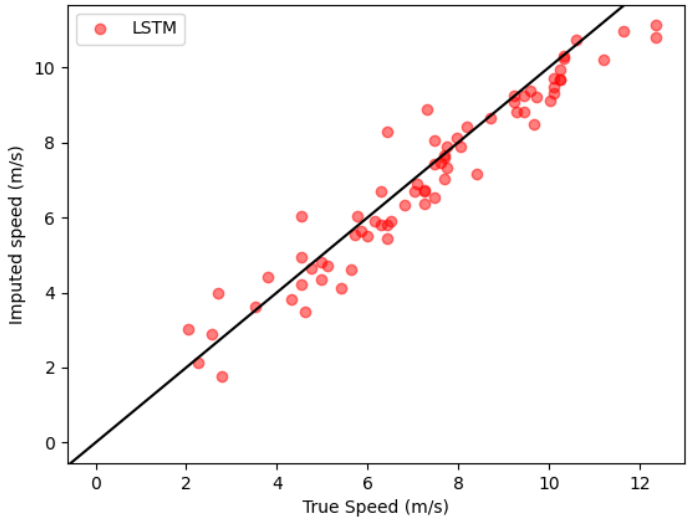

Figure 9. QQ comparison of the hourly imputed series against the original values using LSTM as an imputation technique.

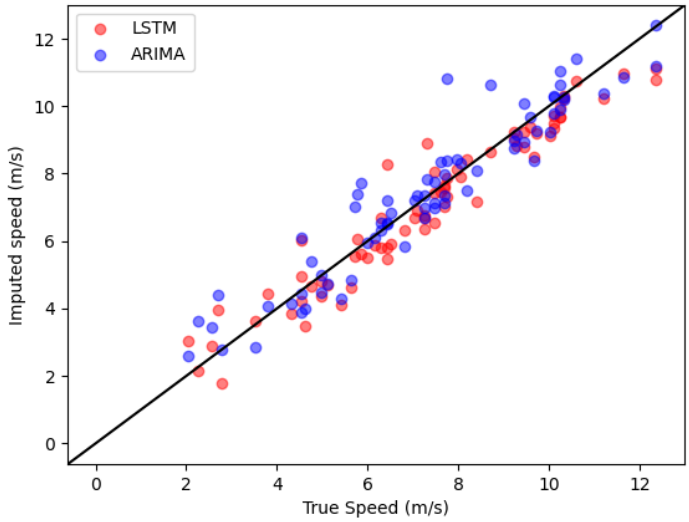

Figure 10. QQ plot of the hourly imputed series against the original values using LSTM and ARIMA as an imputation techqnique for comparing the two methods.

5. Conclusions

The goal of reaching Net Zero emissions has raised interest in offshore wind energy and established it as a vital alternative energy source. Ireland aims to produce over 7GW by 2030, with the capacity to produce more than 30 gigawatts of electricity from offshore wind alone. However, knowing the long-term feasibility of these sites requires precise evaluations of the wind resources.

 doi:10.20944/preprints202405.1179.v1

However, the buoys used to collect the data are not always reliable and can malfunction or provide inaccurate results, leaving large gaps in the data. We used 20 years of wind time series data from Ireland's Marine Institute for our investigation, and the results showed significant gaps in the dataset. Data imputation techniques were used to fill in these gaps as accurately as feasible in order to remedy this issue.

We investigated and compared the data imputation performance of ARIMA and LSTM approaches. LSTMs outperform ARIMA by a small margin, with a mean squared error of 0.45 as opposed to 0.60 for ARIMA. It can be argued that the long compute times and the hyperparameter tuning of LSTMs for a marginal improvement in impute accuracy is not the best way forward and therefore, ARIMA for hourly imputation should suffice. The PDF of the imputations through ARIMA and LSTMs are also inconclusive. While LSTM does reletively better on predicting the speed further away from mean, ARIMA registers a lower error for imputations near the mean. LSTMs then, can be preferred if the goal is to better capture the extreme left and right of the mean.

Author Contributions: Conceptualization, V.P. and G.S.; methodology, G.S., V.P; software, G.S.; validation, G.S.; formal analysis, G.S; investigation, G.S., V.P, A.M; resources, V.P., G.S; data curation, G.S, V.P; writing—original draft preparation, G.S; writing—review and editing, G.S, V.P, A.M; visualization, G.S; supervision, G.S, V.P, A.M; project administration, V.P., A.M; funding acquisition, V.P., G.S. All authors have read and agreed to the published version of the manuscript.

Funding: Vikram Pakrashi would like to acknowledge Science Foundation Ireland NexSys 21/SPP/3756, MaREI RC2302 2, Interreg Atlantic Area SiSDATA EAPA 0040 2022, Sustainable Energy Authority of Ireland Twinfarm RDD/604 and RemoteWind RDD/613.

Data Availability Statement: Data is available from the corresponding author on reasonable request.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

ANN	Artificial Neural Network
ARIMA	Auto-regressive Integrated Moving Average
MI	Marine Institute
ACF	Auto-correlation function
PACF	Partial Auto-correlation function
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
AIC	Akaike information criterion
i.i.d	Independent and identically distributed
PDF	Probability density function
FFN	Feed forward network

 doi:10.20944/preprints202405.1179.v1

Appendix A

Appendix A.1

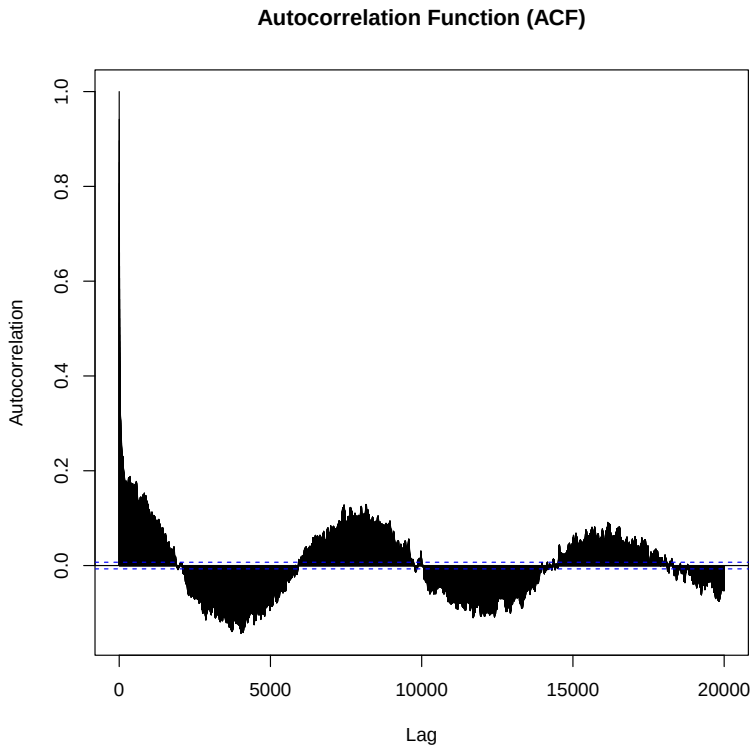

Figure A1. ACF calculated over the full time series for over 78 000 time steps spanning a period from 2001 to 2023.

References

1. Musial, W.; Spitsen, P.; Duffy, P.; Beiter, P.; Shields, M.; Mulas Hernando, D.; Hammond, R.; Marquis, M.; King, J.; Sathish, S. Offshore Wind Market Report: 2023 Edition. Technical report, National Renewable Energy Laboratory (NREL), Golden, CO (United States), 2023.
2. Gallagher, S.; Tiron, R.; Whelan, E.; Gleeson, E.; Dias, F.; McGrath, R. The nearshore wind and wave energy potential of Ireland: a high resolution assessment of availability and accessibility. *Renewable Energy* **2016**, *88*, 494–516.
3. Lefeuvre, E. The Wind That Shakes the Turbines: Analysis of Irish Energy Production and Sovereignty. *Irish Studies in International Affairs* **2023**.
4. Harry, M. Wind energy from Ireland's Atlantic coast could power almost 50m homes in Europe, says Ryan. *The Irish Times*.
5. of Ireland, G. South Coast Offshore Renewable Energy Designated Maritime Area Plan Proposal. Technical report, 2023.
6. Sharkey, F.; Honer, K.; Conlon, M.; Gaughan, K.; Robinson, E. The domestic and export market for large scale wave energy in Ireland and the economics of export transmission. 2013 48th International Universities' Power Engineering Conference (UPEC). IEEE, 2013, pp. 1–6.
7. Nelson, V. *Wind energy: renewable energy and the environment*; CRC press, 2009.
8. de N Santos, F.; D'Antuono, P.; Robbelein, K.; Noppe, N.; Weijtjens, W.; Devriendt, C. Long-term fatigue estimation on offshore wind turbines interface loads through loss function physics-guided learning of neural networks. *Renewable Energy* **2023**, *205*, 461–474.
9. Früh, W.G. Long-term wind resource and uncertainty estimation using wind records from Scotland as example. *Renewable Energy* **2013**, *50*, 1014–1026.

10. Jung, C.; Taubert, D.; Schindler, D. The temporal variability of global wind energy – Long-term trends and inter-annual variability. *Energy Conversion and Management* **2019**, *188*, 462–472. doi:https://doi.org/10.1016/j.enconman.2019.03.072.
11. Bonanno, R.; Viterbo, F.; Maurizio, R.G. Climate change impacts on wind power generation for the Italian peninsula. *Regional Environmental Change* **2023**, *23*, 15.
12. Kay, G.; Dunstone, N.J.; Maidens, A.; Scaife, A.A.; Smith, D.M.; Thornton, H.E.; Dawkins, L.; Belcher, S.E. Variability in North Sea wind energy and the potential for prolonged winter wind drought. *Atmospheric Science Letters* **2023**, p. e1158.
13. Kim, H.; Kim, B. Wind resource assessment and comparative economic analysis using AMOS data on a 30 MW wind farm at Yulchon district in Korea. *Renewable Energy* **2016**, *85*, 96–103. doi:https://doi.org/10.1016/j.renene.2015.06.039.
14. Marine Institute. https://www.marine.ie/site-area/about-us/about-us. Accessed: 2023-05-22.
15. Anil Jadhav, D.P.; Ramanathan, K. Comparison of Performance of Data Imputation Methods for Numeric Dataset. *Applied Artificial Intelligence* **2019**, *33*, 913–933, [https://doi.org/10.1080/08839514.2019.1637138]. doi:10.1080/08839514.2019.1637138.
16. de Rosnay, P.; Browne, P.; de Boisséson, E.; Fairbairn, D.; Hirahara, Y.; Ochi, K.; Schepers, D.; Weston, P.; Zuo, H.; Alonso-Balmaseda, M.; others. Coupled data assimilation at ECMWF: Current status, challenges and future developments. *Quarterly Journal of the Royal Meteorological Society* **2022**, *148*, 2672–2702.
17. Pandya, D.; Vachharajani, B.; Srivastava, R. A review of data assimilation techniques: Applications in engineering and agriculture. *Materials Today: Proceedings* **2022**, *62*, 7048–7052. International Conference on Additive Manufacturing and Advanced Materials (AM2), doi:https://doi.org/10.1016/j.matpr.2022.01.122.
18. Nijman, S.; Leeuwenberg, A.; Beekers, I.; Verkouter, I.; Jacobs, J.; Bots, M.; Asselbergs, F.; Moons, K.; Debray, T. Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review. *Journal of Clinical Epidemiology* **2022**, *142*, 218–229. doi:https://doi.org/10.1016/j.jclinepi.2021.11.023.
19. Wu, P.; Chang, X.; Yuan, W.; Sun, J.; Zhang, W.; Arcucci, R.; Guo, Y. Fast data assimilation (FDA): Data assimilation by machine learning for faster optimize model state. *Journal of Computational Science* **2021**, *51*, 101323. doi:https://doi.org/10.1016/j.jocs.2021.101323.
20. Ba, S.O.; Corpetti, T.; Chapron, B.; Fablet, R. Variational data assimilation for missing data interpolation in SST images. 2010 IEEE International Geoscience and Remote Sensing Symposium, 2010, pp. 264–267. doi:10.1109/IGARSS.2010.5649206.
21. Sareen, K.; Panigrahi, B.K.; Shikhola, T.; Sharma, R. An imputation and decomposition algorithms based integrated approach with bidirectional LSTM neural network for wind speed prediction. *Energy* **2023**, *278*, 127799.
22. Kaur, P.; Joshi, J.C.; Aggarwal, P. Estimation of missing weather variables using different data mining techniques for avalanche forecasting. *Natural Hazards* **2024**, pp. 1–24.
23. Moritz, S.; Sardá, A.; Bartz-Beielstein, T.; Zaefferer, M.; Stork, J. Comparison of different methods for univariate time series imputation in R. *arXiv preprint arXiv:1510.03924* **2015**.
24. Liu, T.; Wei, H.; Zhang, K. Wind power prediction with missing data using Gaussian process regression and multiple imputation. *Applied Soft Computing* **2018**, *71*, 905–916.
25. Shukur, O.B.; Lee, M.H. Imputation of missing values in daily wind speed data using hybrid AR-ANN method. *Modern Applied Science* **2015**, *9*, 1.
26. Liao, W.; Bak-Jensen, B.; Pillai, J.R.; Yang, D.; Wang, Y. Data-driven missing data imputation for wind farms using context encoder. *Journal of Modern Power Systems and Clean Energy* **2021**, *10*, 964–976.
27. Liu, N.; Li, Y.; Zang, Z.; Hu, Y.; Fang, X.; Lolli, S. A deep learning-based imputation method for missing gaps in satellite aerosol products by fusing numerical model data. *Atmospheric Environment* **2024**, p. 120440.
28. Marine Institute Oceanography. https://www.marine.ie/site-area/areas-activity/oceanography/oceanography. Accessed: 2023-05-22.
29. Marine Institute Buoy Locations. http://www.marine.ie/site-area/data-services/real-time-observations/irish-marine-data-buoy-observation-network. Accessed: 2023-05-22.
30. Beckers, S.; Blair, B. Non-parametric forecasting for conditional asset allocation. *Journal of Asset Management* **2002**, *3*, 213–228.
31. Pai, P.F.; Lin, C.S. A hybrid ARIMA and support vector machines model in stock price forecasting. *Omega* **2005**, *33*, 497–505.

 doi:10.20944/preprints202405.1179.v1

32. Khashei, M.; Bijari, M. An artificial neural network (p, d, q) model for timeseries forecasting. *Expert Systems with applications* **2010**, *37*, 479–489.
33. Wang, Q.; Li, S.; Li, R.; Ma, M. Forecasting US shale gas monthly production using a hybrid ARIMA and metabolic nonlinear grey model. *Energy* **2018**, *160*, 378–387.
34. Dong, H.; Guo, X.; Reichgelt, H.; Hu, R. Predictive power of ARIMA models in forecasting equity returns: a sliding window method. *Journal of Asset Management* **2020**, *21*, 549–566.
35. Sheoran, S.; Pasari, S. Efficacy and application of the window-sliding ARIMA for daily and weekly wind speed forecasting. *Journal of Renewable and Sustainable Energy* **2022**, *14*.
36. Miller, J.; Hardt, M. Stable recurrent models. *arXiv preprint arXiv:1805.10369* **2018**.
37. Greaves-Tunnell, A.; Harchaoui, Z. A statistical investigation of long memory in language and music. International Conference on Machine Learning. PMLR, 2019, pp. 2394–2403.
38. Zhao, J.; Huang, F.; Lv, J.; Duan, Y.; Qin, Z.; Li, G.; Tian, G. Do RNN and LSTM have long memory? International Conference on Machine Learning. PMLR, 2020, pp. 11365–11375.
39. Li, C. Little's test of missing completely at random. *The Stata Journal* **2013**, *13*, 795–809.
40. Hyndman, R.J.; Khandakar, Y. Automatic time series forecasting: the forecast package for R. *Journal of statistical software* **2008**, *27*, 1–22.
41. Patterson, J.; Gibson, A. *Deep learning: A practitioner's approach*; " O'Reilly Media, Inc.", 2017.
42. Gulli, A.; Pal, S. *Deep learning with Keras*; Packt Publishing Ltd, 2017.