Intelligent digital twin – machine learning system for real-time wind turbine wind speed and power generation forecasting

Eamonn Tuton[1,2,*], *Xinhui* Ma[1], and *Nina* Dethlefs[1]

[1]University of Hull, School of Computer Science, HU6 7RX, Hull, United Kingdom

[2]Aura Centre for Doctoral Training in Offshore Wind Energy and the Environment, University of Hull, HU6 7RX, Hull, United Kingdom

Abstract. Wind power is a key pillar in efforts to decarbonise energy production. However, variability in wind speed and resultant wind turbine power generation poses a challenge for power grid integration. Digital Twin (DT) technology provides intelligent service systems, combining real-time monitoring, predictive capabilities and communication technologies. Current DT research for wind turbine power generation has focused on providing wind speed and power generation predictions reliant on Supervisory Control and Data Acquisition (SCADA) sensors, with predictions often limited to the timeframe of datasets. This research looks to expand on this, utilising a novel framework for an intelligent DT system powered by k-Nearest Neighbour (kNN) regression models to upscale live wind speed forecasts to higher wind turbine hub-height and then forecast power generation. As there is no live link to a wind turbine, the framework is referred to as a "Simulated Digital Twin" (SimTwin). 2019-2020 SCADA and wind speed data are used to evaluate this, demonstrating that the method provides suitable predictions. Furthermore, full deployment of the SimTwin framework is demonstrated using live wind speed forecasts. This may prove useful for operators by reducing reliance on SCADA systems and provides a research and development tool where live data is limited.

1 Introduction

Wind power is of particular interest in efforts to reduce greenhouse gas emissions [1] given its technological readiness, relatively low environmental footprint, and abundant availability [2]. Electricity is generated via the conversion of kinetic energy contained within the wind, governed by the wind power equation [3]:

$$P=0.5C_P\rho AV^3 \qquad (1)$$

Where P is power generated, C_P is a wind turbine's power coefficient, ρ is air density, A the rotor wind-swept area, and V is wind speed.

Given the fluctuating nature of the wind [4], wind turbine power output can be highly variable. This variability and the resultant difficulty in forecasting future power generation pose several challenges in power grid integration whilst ensuring stability [5]. This has led to a number of proposed solutions including electrical interconnectors [6], energy storage systems [7] and improved demand prediction [8].

Another potential mitigating measure is predicting wind turbine power generation. This allows better wind farm [9] and grid network management, reducing the need for generation reserves [5], as well as enabling other solutions such hydrogen energy storage [10]. Given that power generation is dependent on the wind speed cubed, this is frequently seen as the most important, and therefore the most used, input parameter for calculating power generation [11]. Manufacture power curves provide one method of doing so, detailing anticipated power output for a given wind speed. However, these can be overly optimistic and are often based on ideal conditions [12].

Supervisory Control Data Acquisition (SCADA) is commonly used in wind energy [13], providing wind turbines, wind farms, and associated equipment the ability to report their operational status [14]. Solutions using SCADA have become increasingly popular in fault diagnosis and prediction in wind turbines [15, 16, 17]. SCADA often details power output, wind speed, and other associated metrics, allowing its use for power generation predictions.

SCADA data is typically recorded at 10-minute intervals and provides a number of potentially useful measurements including wind speed, wind direction, power generation, voltages and component temperatures [18] making it ideal for data-driven methods utilising machine learning. These have the benefit of not requiring domain knowledge, with the potential for model improvement over time [19]. Additionally, the use of large datasets allows many different aspects to be considered [20], providing a good way to discover complex relations between data. However, it may be difficult to predict extreme conditions due to limitations in observations and whilst correlation may be determined, this will not give causality for what occurs [19].

[*] Corresponding author: e.tuton-2021@hull.ac.uk

Given the availability of SCADA datasets, numerous publications have undertaken wind turbine power generation-related research making use of this resource. Lin et al. [21] utilised isolation forest and deep learning techniques alongside high-frequency SCADA data to derive an improved technique for outlier detection, improving the accuracy of power generation predictions. Delago and Fahim [22] devised a Long Short-Term Memory (LSTM) powered data analysis framework, capable of predicting and visualising SCADA data and associated power generation prediction.

An important factor in the prediction of energy generation from a wind turbine is knowledge of forecasted wind speed, with a number of machine learning techniques also at the forefront of this. Lv and Wang [23] utilised deep learning within a newly proposed combined model to forecast wind speed which followed a “decomposition-optimisation-forecasting” principle. Hur [24] developed a 2-stage method for short-term wind speed prediction utilising an extended Kalman filter for estimation, with a combination of extrapolation and a double-layer perceptron (DLP) feedforward neural network for prediction.

Whilst methods such as those outlined above have seen successful academic implementation, it has been noted that these can often be overly complicated [25] increasing the difficulty of real-world deployment and use. As such, this has inspired research using simpler methods of wind speed and power generation. k-nearest neighbour (kNN) methods have been shown to be successful, producing robust wind speed and power generation predictions [9, 25], with results achievable that are comparable to more intensive methods [26].

A wide range of definitions as to what constitutes a DT have arisen, spurring attempts to consolidate definitions and better specify what constitutes a DT [27, 28, 29]. These suggest that at its most basic form, a DT consists of a physical asset, a virtual model of the asset, and a bidirectional link between them. This link is often considered “live”, providing real-time updates to the model, as well as enabling changes to the asset as a result of changes in the model. These changes vary from direct actions undertaken by the model to indirect actions resulting from operator decisions.

DTs provide increased integration between physical and virtual spaces by combining DTs with sensors, machine learning and Internet of Things (IoT)-based technologies [30]. DT technology offers several benefits including remote monitoring and operations from anywhere at any time [31], access to real-time monitoring data useful for decision-making [28], and the provision of continuously updating predictions of the future state of an asset [27]. These allow for improved decision-making and planning [27], optimisation of activities [32] and automation [31].

The use of DTs in the wind energy sector has seen increasing popularity. Numerous frameworks have been proposed, particularly for operations and maintenance purposes, for both onshore and offshore wind turbines [33, 34, 35].

However, there has been limited research regarding power generation-focused DTs. Fahim et al. [9] proposed a 5G DT platform powered by Microsoft Azure infrastructure. This looked to provide a framework for real-time monitoring and prediction of power generation, claiming to be the first to do so. This was tested utilising SCADA data for a turbine located at an onshore wind farm in the Yalova region of Turkey. Machine learning, in the form of a deep learning Temporal Convolution Network (TCN) and non-parametric k-nearest neighbour (kNN) regression, were also used to predict wind speed and power generation respectively.

Fahim et al. [9] were able to provide wind speed and power generation predictions rivalling other machine learning techniques. However, there was a lack of explanation as to how a DT would be deployed, either in the field or during testing, and power generation predictions were limited to 2018 (the extent of the SCADA dataset available). Additionally, the use of machine learning did not appear fully integrated with the DT. Rather the results presented appeared to utilise the historical SCADA dataset but did not provide evidence of being used in the context of the DT framework proposed.

Kim et al. [36] proposed a physics-based DT to overcome the reliance on historical SCADA data for model training. The approach provided promising results for a test floating wind turbine, though focused on “live” updates on power generation as opposed to longer-term predictions and requires continuous sensor readings in order to provide predictions. Additionally, the proposed physics-based approach requires a thorough understanding of the wind turbine and local environment in question, with a complete change in model required depending on wind turbine model and location.

As such, this research looks to contribute a DT with live power generation forecast capabilities, as opposed to being restricted to the timeframe of a given dataset and SCADA availability. This is achieved via the development of a machine learning model to predict wind turbine hub-height wind speeds based on live weather forecast data provided via an Application Programme Interface (API). Predicted hub-height wind speed is then used with a machine learning power generation prediction model to provide a power generation forecast. By incorporating this into a fully-functioning DT framework, a fully realised, deployable, power generation forecasting DT is delivered, capable of providing continuously updating live predictions for a wind turbine.

The proposed DT does not maintain a direct link to a wind turbine as might be expected from DT definitions highlighted, instead using live weather forecast data. It is considered that the DT proposed offers a simulation of the anticipated live and future state of the wind turbine in question and as such, the proposed framework and its deployment are referred to as a “Simulated Twin” (SimTwin) for the remainder of the paper.

2 Methodology

2.1. Data

Wind speed and power generation data were derived from historical SCADA data [37] for a wind turbine at

Kelmarsh Wind Farm near Haselbach, Northamptonshire comprising 6 wind turbines. Table 1 provides wind turbine data details. Historical wind speed forecasts for the 2-year 2019-2020 period and live wind speed forecasts were also used, sourced from OpenWeather [38].

Table 1. Wind turbine data overview [37]

Estimated location	52°24'2.30"N (latitude), 0°56'49.38"W (longitude)
Rated power generation	2.05 Mega Watt (MW)
Hub-Height	78.5m
Turbine type	Senvion MM92
Start/End of dataset	01/01/2019 – 31/12/2020
SCADA frequency	10-Minutes

2.2. Predictive models

kNN regression is a non-parametric pattern recognition method [39] that has seen popularity due to its successful use for time series forecasting whilst remaining relatively simple [40], with the technique seeing use for forecasting both wind speed and wind turbine power generation [41, 42]. kNN utilises the average of nearby observations to provide an estimated value [43] and distance is used to decide if observations are considered nearby, given by k [44]. An optimum k value is imperative. Smaller values can lead to over-fitting, with greater values potentially resulting in worse performance [43].

kNN regression is considered a suitable initial method for the development of the predictive models developed. Time-series SCADA and wind speed datasets were used to develop and implement the predictive capabilities of the SimTwin, with kNN highly compatible with time-series data. Additionally, the simple setup and relatively quick calculation time make it useful for continuous live updates. As previously highlighted, kNN-based research has produced robust power generation and wind speed prediction results [25, 9]. Taking into consideration the low computational and user requirements, kNN-based methods have demonstrated results comparable to more intensive methods, such as Long Short-Term Memory (LSTM), Support Vector Regression (SVR) and Bagging Regression (BR) [Mehr, 2021].

kNN regression was undertaken utilising a Minkowski distance measurement, given by [45]:

$$D=(\sum(|u_i-v_i|)^p)^{\frac{1}{p}} \qquad (2)$$

Where D is Minkowski distance, u is input array, v is output array and p is the order of the norm of the difference $||u-v||_p$ [45].

The use of kNN regression to predict hub-height wind speed (PW_{HUB}) from lower level wind speed (W_{LOW}) and predicted power generation (PG) from predicted hub-height wind speed is given by the following functions respectively:

$$F(W_{LOW}) \rightarrow PW_{HUB} \qquad (3)$$

$$F(PW_{HUB}) \rightarrow PG \qquad (4)$$

Wind speed prediction results have also been compared to alternative methods to kNN regression considered appropriate for wind speed forecasting. These have been chosen for their ability to provide relatively quick updates as would be required by the SimTwin to give continuously updating results. The Wind Speed Power Law (WSPL) is a physical law that can be used to estimate upscaled wind speeds utilising a reference height with a known wind speed given by the equation [46]:

$$U_h=U_g(\frac{Z_h}{Z_g})^{\alpha} \qquad (5)$$

Where U_h is target wind speed at height Z_h, U_g is known wind speed at height Z_g and α is the power law exponent. A value of 0.143 is typically adopted for α under neutral stability conditions.

Decision Tree Regression (DTR) has also been used, which breaks data into small sub-groups [9]. Extreme Gradient Boosting (XGBoost) regression is a gradient-boosted decision tree, which combines weaker models, providing a unified stronger model [47].

2.3. SimTwin framework

Fig. 1 highlights the high-level architecture used to implement the power generation forecasting SimTwin.

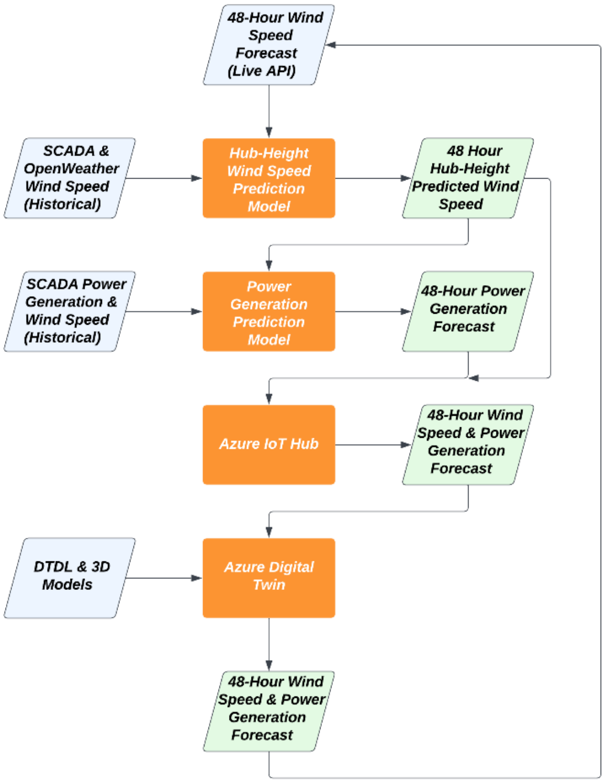

48-Hour Wind Speed Forecast (Live API)
SCADA & OpenWeather Wind Speed (Historical)
Hub-Height Wind Speed Prediction Model
48 Hour Hub-Height Predicted Wind Speed
SCADA Power Generation & Wind Speed (Historical)
Power Generation Prediction Model
48-Hour Power Generation Forecast
Azure IoT Hub
48-Hour Wind Speed & Power Generation Forecast
DTDL & 3D Models
Azure Digital Twin
48-Hour Wind Speed & Power Generation Forecast

Fig. 1. High-level SimTwin framework overview

2.3.1 Hub-Height wind speed prediction

Power generation forecasting requires the prediction of future wind speed, necessitating the use of weather forecasting services. Wind speed predictions for the wind turbine location were sourced from OpenWeather's One Call API 3.0 [48]. This provides hourly forecasts for a 48-hour period, including wind speed. However, as this is given at a different altitude than the hub-height of the wind turbine this needed to be converted.

A hub-height wind speed prediction model was derived using hub-height wind speed from the SCADA dataset and historical wind speed data sourced from OpenWeather [49] for the same timeframe (see equation 3). Historical lower-level wind speed data from OpenWeather is given for hourly periods and as such the 10-minute SCADA dataset was averaged to give an hourly value. Historic wind speed data was used as an input to kNN regression with the higher hub-height wind speed as the desired target. The kNN regressor looks to predict the anticipated hub-height wind speed for a given low-level wind speed by utilising the average of nearby wind speeds for a given point, as identified by k. Grid search was used to calculate the optimal k value for the model, with values between 1 and 100 trialled. In doing so, a kNN model was derived capable of taking lower-height wind speed and upscaling it to hub-height.

Live weather forecasts from the One Call API 3.0 were accessed using an API call to the OpenWeather One Call API 3.0 web address. Upon calling, 48-hour, hourly, wind speed dataset is provided in JSON format. This dataset is then inputted into the kNN wind speed upscaling model and anticipated correlating hub-height wind speed generated, providing a 48-hour prediction of hub-height wind speeds. This was undertaken at frequent repeating periods, providing continuous hub-height wind speed predictions based on the latest weather forecast.

2.3.2 Power generation forecast

Wind speed at hub-height and actual power generation from the SCADA dataset was used to train a kNN regression model capable of predicting power generation for a given hub-height wind speed (see equation 4). Data was averaged so to give an hourly value, reflecting the hourly values used for hub-height wind speed prediction and power generation forecasting. Grid search was also used, with values between 1 and 100 trialled.

The 48-hour hub-height wind speed predictions were used in conjunction with the power generation prediction model, giving a 48-hour power generation forecast for the wind turbine. This was set to occur at frequently repeating periods, providing continuously updated power generation forecasts.

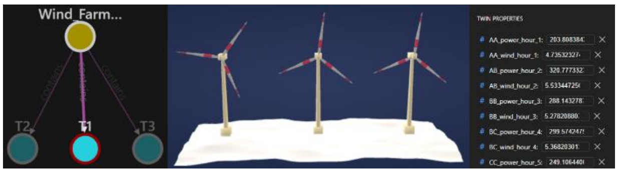
2.3.3 Exportation to Azure IoT Hub

The hub-height wind speed and power generation forecast are converted to JSON ("UTF-8" encoding and content type "application/json") and exported to Azure IoT Hub. This is then continuously updated upon receiving power forecast data.

2.3.4 Azure Digital Twin model display

Azure Digital Twin requires models to be defined using Digital Twin Definition Language (DTDL). This enables the establishment of relationships to allow for grouping of different components of a DT, allowing a better understanding of how these may interact with each other [50]. In this case, official documentation [50,51] has been used to generate a simple wind farm model comprising 3 wind turbines, one of which is used to represent the wind turbine tested. A 3D model [52] has also been imported for added visualisation.

An Azure Function, adopted from official documentation [53], decodes and sends relevant data from IoT Hub to Azure Digital Twin which is then displayed. The Azure Function continuously sends relevant data upon the arrival of new data to IoT Hub, allowing the DT model to display the most recent hub-height wind speed and power generation forecasts. Fig. 2 shows the wind turbine in Azure Digital Twin graph layout and the 3D wind turbine model. T1 refers to the wind turbine tested. It is noted that this is not intended to reflect the makeup of Kelmarsh wind farm but rather act as a general representation.



AA_power_hour_1: 203.808384
AA_wind_hour_1: 4.73532327
AB_power_hour_2: 320.777332
AB_wind_hour_2: 5.53344725
BB_power_hour_3: 288.143278
BB_wind_hour_3: 5.27820880
BC_power_hour_4: 299.574247
BC_wind_hour_4: 5.36820301
CC_power_hour_5: 249.106440

Fig. 2. Azure Digital Twin graph and 3D model overview

2.4. Model performance metrics

Root Mean Square Error (RMSE) was used to measure model performance, allowing comparisons with similar research. This considered a suitable performance evaluation metric for wind power as it assigns additional weight between large differences between actual and predicted values compared to smaller differences [54]. RMSE was calculated so to give the difference between predicted and actual values for wind speed and power generation. A lower RMSE value is considered better and is given by the equation [55]:

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(e_i-\bar{e})^2} \tag{6}$$

Where e_i is an actual value and $\bar{e}$ is a predicted value.

2.5. Power forecast SimTwin testing parameters

All tests undertaken are highlighted in Table 2. To test the deployment of the SimTwin framework, 12 months of power generation and wind speed SCADA data, as well as 12 months of historical wind speed forecasts were used to generate the power generation prediction model and hub-height wind speed prediction model. A wind speed forecast for the future 48-hour period of 16:00 on 21/06/2023 to 15:00 on 23/06/2023 was sourced from One Call API 3.0, upscaled to hub-height, and fed into the power generation prediction model, thus giving a power generation forecast for the wind turbine during this period.

Table 2. Model testing periods

Test	Training period	Test period
Q1 2019 predictions	01/01/2019 – 31/03/2019	01/04/2019 – 07/04/2019
Q2 2019 predictions	01/04/2019 – 30/06/2019	01/07/2019 – 07/07/2019
Q3 2019 predictions	01/07/2019 – 30/09/2019	01/10/2019 – 07/10/2019
Q4 2019 predictions	01/10/2019 – 24/12/2019	25/12/2019 – 31/12/2019
Q1 2020 predictions	01/01/2019 – 31/03/2019	01/04/2020 – 07/04/2020
Q2 2020 predictions	01/04/2019 – 30/06/2019	01/07/2020 – 07/07/2020
Q3 2020 predictions	01/07/2019 – 30/09/2019	01/10/2020 – 07/10/2020
Q4 2020 predictions	01/10/2019 – 24/12/2019	25/12/2020 – 31/12/2020
Future predictions	2019 (all)	21/06/2023 – 23/06/2023

Given that no actual wind turbine data is available for this period, the validity of the hub-height wind speed prediction and power generation prediction kNN models was calculated by producing alternative models. These are capable of making predictions during the SCADA dataset timeframe, allowing comparisons with actual output during this period. This was achieved by using historical OpenWeather wind speed data for both training the hub-height wind speed prediction model and for calculating power generation. Model training was undertaken using quarterly data, with the following week used to test the models. Q4 2019/2020 data was reduced by 1 week to allow a test period within the 2020 timeframe of the SCADA dataset.

3 Results

Results are split into 3 sections; The first section provides 2019 weekly results for the hub-height wind speed predictions based on quarterly 2019 training data, with the actual SCADA-derived hub-height wind speed for the same period provided for comparison. The kNN regression method adopted is also compared to WSPL, DTR and XGBoost-based approaches. kNN regression-derived weekly results for 2020 based on quarterly 2019 data are also provided.

The second section highlights the 2019 and 2020 weekly power generation forecast results using the kNN regression-derived predicted hub-height wind speeds and actual SCADA-derived hub-height wind speeds as inputs. These results have been presented alongside SCADA-derived wind turbine power generation for comparison.

The third demonstrates the deployment of the SimTwin for the future period of 16:00 on 21/06/2023 to 15:00 on 23/06/2023.

3.1. Quarterly hub-height wind speed results

Table 3 outlines measured performance in the form of RMSE for the 2019 test periods for each quarter, performed for the range of methods considered to be suitable for wind speed prediction. The k value used for kNN-based predictions is also presented.

Table 3. 2019 wind speed prediction comparative methods

Method	Q1 2019	Q2 2019	Q3 2019	Q4 2019
kNN (k value)	81	62	75	41
kNN (RMSE)	1.31	1.13	1.16	1.31
WSPL (RMSE)	1.36	1.28	1.27	1.45
XGBoost (RMSE)	1.32	1.14	1.17	1.31
DTR (RMSE)	1.64	1.3	1.27	1.52

Table 4 outlines the measured performance and k value used for the 2020 test periods for each quarter, performed using kNN regression for wind speed prediction.

Table 4. 2020 wind speed prediction performance values

Value	Q1 2020	Q2 2020	Q3 2020	Q4 2020
k value	81	62	75	41
kNN (RMSE)	1.13	1.17	1.34	1.99

Fig. 3 to Fig. 6 outline the 2020 week-long wind speed predictions based on 2019 quarterly training data. Predicted wind speed is given in meters per second (m/s).

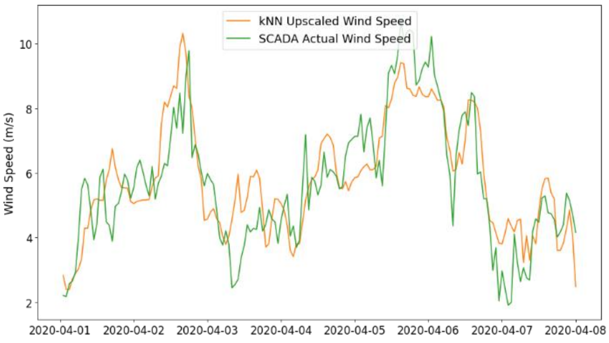


Fig. 3. 1-week 2020 wind speed prediction (Q1 2019 training)

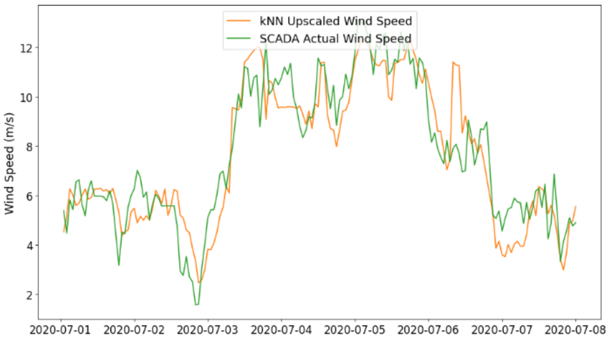


Fig. 4. 1-week wind speed prediction (Q2 2019 training)

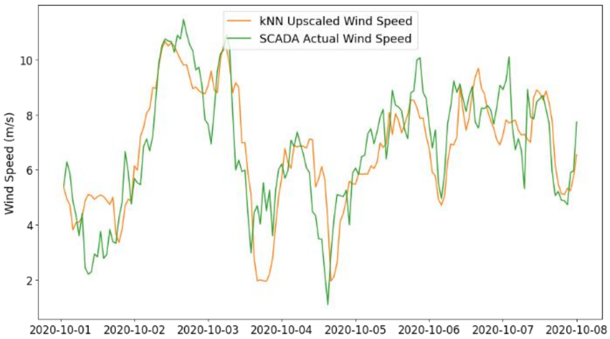


Fig. 5. 1-week wind speed prediction (Q3 2019 training)

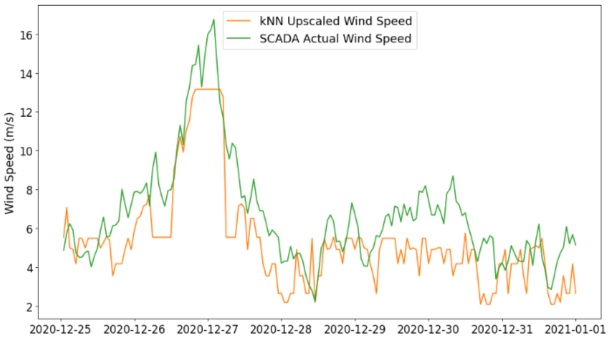


Fig. 6. 1-week wind speed prediction (Q4 2019 training)

3.2. Quarterly power generation results

Table 5 outlines measured performance and k value used for the 2019 test periods for each quarter, performed using kNN regression for power generation prediction based on predicted and SCADA-derived hub-height wind speeds.

Table 5. 2019 Power generation prediction performance values

Value	Q1 2019	Q2 2019	Q3 2019	Q4 2019
k value	38	30	19	97
Predicted wind speed	261.36	169.14	236.86	382.08
SCADA wind speed	57.20	35.96	45.86	63.33

Table 6 outlines measured performance and k value used for the 2020 test periods for each quarter, performed using kNN regression for power generation prediction based on predicted and SCADA-derived hub-height wind speeds.

Table 6. 2020 Power generation prediction performance values

Value	Q1 2020	Q2 2020	Q3 2020	Q4 2020
k value	38	30	19	97
Predicted wind speed	239.67	268.36	323.48	434.63
SCADA wind speed	49.10	90.06	59.46	253.12

Fig. 7 to Fig. 10 outline the 2020 week-long power generation predictions based on 2019 quarterly training data. Predicted power generation is given in kilowatts (kW).

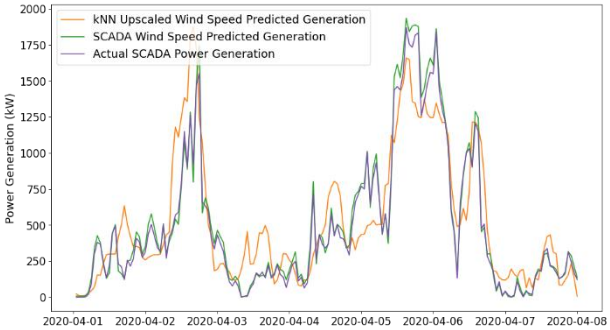


Fig. 7. 1-week 2020 power prediction (Q1 2019 training)

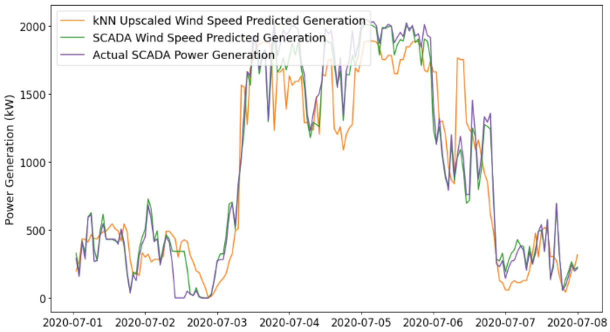


Fig. 8. 1-week 2020 power prediction (Q2 2019 training)

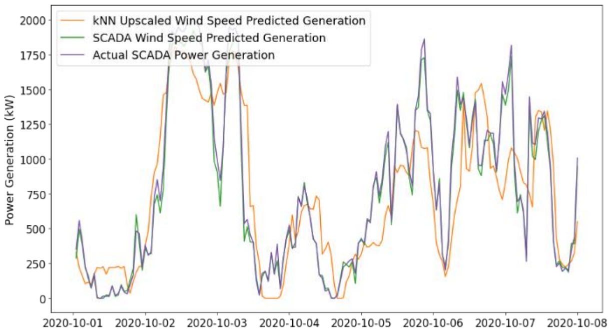


Fig. 9. 1-week 2020 power prediction (Q3 2019 training)

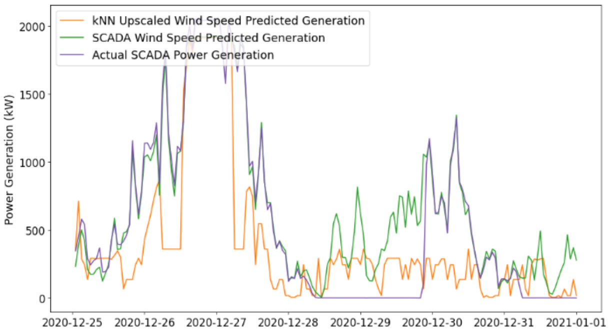


Fig. 10. 1-week 2020 power prediction (Q4 2019 training)

3.3. Live SimTwin result

Fig. 11 and Fig. 12 show the graphical output of the forecasted hub-height wind speed and power generation calculated for a future 48-hour period, giving a graphical demonstration of the live deployment of the SimTwin as would be displayed in the Azure Digital Twin (see Fig. 2).

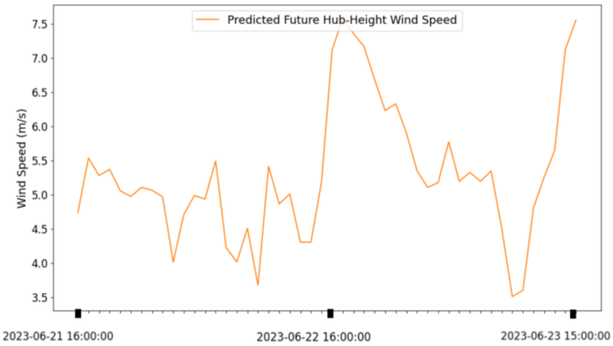


Fig. 11. “Live” hub-height wind speed forecasting

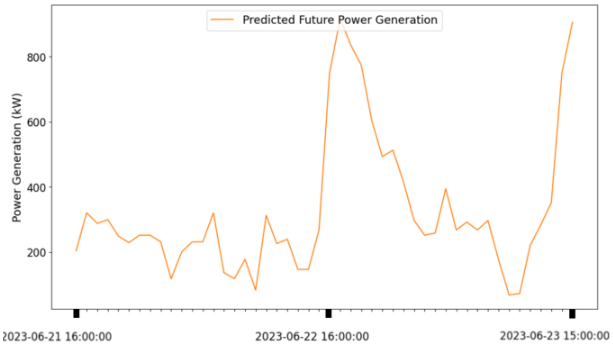


Fig. 12. “Live” power generation forecasting

4 Discussion

It has been demonstrated that the SimTwin framework proposed is deployable and capable of providing live and updating forecasts of future hub-height wind speed and power generation for a wind turbine. This may prove useful for operators by reducing reliance on SCADA systems and associated physical wind turbine sensors. Additionally, as a DT necessitates the need for live data, this provides a useful tool for DT development where live and historical data may be limited.

The testing results show that the kNN hub-height wind speed model was capable of upscaling historical OpenWeather low-level wind speed hub-height forecasts to a reasonable accuracy across all quarters tested barring Q4 2020, though still picked up on the general trend in wind speed fluctuations. The kNN regression approach taken produced lower RMSE results than WSPL and DTR. XGBoost results are generally considered on a similar level to that of kNN regression whilst requiring significantly greater model tuning and additional training time. As such, the kNN approach presented is considered the most appropriate for the development of an easy to deploy, responsive, DT system.

The kNN power generation model showed relatively impressive results when using SCADA-derived hub-height wind speed for the week tests undertaken for Q1, Q2 and Q3 of 2020. Q4 proved more challenging to predict, resulting in lower accuracy, however this is anticipated to be due to a partial shutdown of the wind turbine when no generation occurred. This demonstrates that the kNN power generation prediction model is generally capable of providing accurate power generation results given an optimum hub-height wind speed input.

When predicting power generation using predicted hub-height wind speeds, the model also demonstrated reasonably accurate results, capturing the general trend of power generation. The RMSE for this was higher than utilising the SCADA-derived wind speeds, due to the error already introduced when upscaling the predicted hub-height wind speeds.

It was also demonstrated that the proposed framework can produce reasonable predictions for both short and long timescales, thus increasing its potential usefulness.

Table 7 highlights select results from Fahim et al. [9]. Week-long quarterly hub-height wind speed predictions were undertaken for a different wind farm than that test in this paper, however, this is considered a useful comparative metric for predicting wind speed.

Table 7. Wind Speed Predictions [48]

Test	Wind Speed RMSE (1 Week)
Q1 2018	1.76
Q2 2018	1.25
Q3 2018	0.88
Q4 2018	0.90

Hub-height wind speed predictions utilising the kNN hub-height wind speed model produced a lower RMSE in both Q1 and Q2 in both 2019 and 2020, though RMSE was higher in Q3 and Q4 2019 and 2020. Overall, it is considered that the ability to upscale wind speed from an easily accessible source of wind speed data, make predictions over long timescales and the simplicity of the kNN model may potentially be worth the trade-off in the correct circumstances, such as in situations with limited data availability and where ease of model training is required.

Despite this, there are potential methods that could be utilised to improve performance. The use of alternative weather forecasting services [56] may give a more accurate representation of wind speed at the site of the wind turbine. The inclusion of other factors beyond wind speed and power generation may prove beneficial if available in alternative datasets. This includes temperature, pressure, wind direction and humidity [32]. Alternative approaches could be tested, such as deep neural networks [32] and hybrid approaches combining physics, statistics and machine learning techniques [28].

It is envisaged that the SimTwin framework demonstrated in this paper could act as an alternative system to SCADA systems, reducing reliance on wind speed sensors by providing reasonable wind speed and power generation forecasts for use by wind turbine operators. The relative simplicity of the system should also help with ease of deployment. Additionally, it is anticipated that the outputs of the SimTwin would provide a useful tool in research and development environments, particularly when access to live wind speed and power generation forecasts are needed but inaccessible. This includes research into the fatigue effect of wind loading, the management of power generation and the storage of additional power generation, which can be undertaken in real-time.

Future work should look to include the potential improvements highlighted and test the model for different wind turbine models in differing locations, including both onshore and offshore installations.

5 Conclusion

Wind power is a key pillar for the decarbonisation of electricity generation. Digital Twin (DT) technology allows increased integration between physical assets and virtual models including live monitoring, updates and predictions, allowing suitable and informed actions to be undertaken. It has been demonstrated that a power generation forecasting Simulated Digital Twin (SimTwin) is achievable, providing hub-height wind speed and power generation predictions beyond the timeframe of data availability via the use of a weather forecast API and machine learning in the form of k-nearest neighbour (kNN) regression-based models.

The use of kNN regression for hub-height wind speed prediction was seen to have comparable or better results when compared to the Wind Speed Power Law (WSPL), Decision Tree Regression (DTR) and Extreme Gradient Boosting (XGBoost) regression. Additionally, it has been demonstrated that this approach can provide reasonable predictions for short and long timescales. This reduces reliance on Supervisory Control and Data Acquisition (SCADA) systems and associated physical wind turbine sensors and provides a useful tool in DT research where dataset availability may be limited. Furthermore, by using live openly available weather data, power generation predictions can be expanded to entire wind farms, as well as for regional, national or global ranges of wind turbine power production.

Future work should look to provide improvements to hub-height wind speed predictions and power generation forecasts. This may be achievable by using different machine learning techniques, alternative weather prediction sources, hybrid approaches and more detailed historical power generation and wind speed datasets.

We would like to thank the Aura Centre for Doctoral Training in Offshore Wind Energy and the Environment and the University of Hull for their support. Funding is provided by EPSRC and NERC. Grant number EP/S023763/1, project reference 2609887.

References

1. United Nations Framework Convention on Climate Change. “The Paris Agreement. What is The Paris Agreement?”. United Nations Framework Convention on Climate Change (2023). https://unfccc.int/process-and-meetings/the-paris-agreement.
2. Z. Ren, A. S. Verma, Y. Li, J. J. E. Teuwen, Z. Jiang. "Offshore wind turbine operations and maintenance: A state-of-the-art review”. Renewable and Sustainable Energy Reviews, **144** (2021).
3. M. Lydia, S. S. Kumar, A. I. Selvakumar, G. E. Prem Kumar. "A comprehensive review on wind turbine power curve modeling techniques”. Renewable and Sustainable Energy Reviews, **30**, 452-460 (2014).
4. S. A. Vargas, G. R. T. Esteves, P. M. Maçaira, B. Q. Bastos, F. L. Cyrino Oliveira, R. C. Souza. "Wind power generation: A review and a research agenda”. Journal of Cleaner Production, **218**, 850-870 (2019).
5. S. D. Ahmed, F. S. M. Al-Ismail, M. Shafiullah, F. A. Al-Sulaiman, I. M. El-Amin. "Grid Integration Challenges of Wind Energy: A Review". IEEE Access, **8**, 10857-10878 (2020).
6. E. Pean, M. Pirouti, M. Qadrdan. "Role of the GB-France electricity interconnectors in integration of variable renewable generation". Renewable Energy, **99**, 307-314 (2016).
7. P. H. A. Barra, W. C. de Carvalho, T. S. Menezes, R. A. S. Fernandes, D. V. Coury. "A review on wind power smoothing using high-power energy storage systems". Renewable and Sustainable Energy Reviews, **137** (2021).
8. A. K. Bashir, S. Khan, B. Prabadevi, N. Deepa, W.S. Alnumay, T. R. Gadekallu, P. K. R. Maddikunta. "Comparative analysis of machine learning algorithms for prediction of smart grid stability". Electrical Energy Systems, International Transactions on Electrical Energy Systems, **31**, 9 (2021).
9. M. Fahim, V. Sharma, T. V. Cao, B. Canberk, T. Q. Duong. Machine Learning-Based Digital Twin for Predictive Modeling in Wind Turbines”. IEEE Access, **10**, 14184-14194 (2022).
10. A. Javaid, U. Javaid, M. Sajid, M. Rashid, E. Uddin, Y. Ayaz, A. Waqas. “Forecasting hydrogen production from wind energy in a suburban environment using machine learning”. Energies, **15**, 23 (2022).
11. S. Hanifi, X. Liu, Z. Lin, S. Lotfian. "A critical review of wind power forecasting methods—past, present and future". Energies, **13**, 15, (2020).
12. D. Karamichailidou, V. Kaloutsa, A. Alexandridis. "A critical review of wind power forecasting methods—past, present and future". Renewable Energy, **163**, 2137-2152 (2021),

13. W. Udo, Y. Muhammad. "Data-driven predictive maintenance of wind turbine based on SCADA data". IEEE Access, **9**, 162370-162388 (2021)/
14. BVG Associates "Guide to an offshore wind farm". BVG Associates (2019) https://www.thecrownestate.co.uk/media/2861/guide-to-offshore-wind-farm-2019.pdf.
15. H. Rashid, E. Khalaji, J. Rasheed, C. Batunlu. "Fault prediction of wind turbine gearbox based on SCADA data and machine learning". IEEE, *2020 10th International Conference on Advanced Computer Information Technologies,* 391-395 (2020).
16. L. Xiang, P. Wang, X. Yang, A. Hu, H. Su. "Fault detection of wind turbine based on SCADA data analysis using CNN and LSTM with attention mechanism". Measurement, **175** (2021).
17. Q. Lu, W. Ye, L. Yin. "ResDenIncepNet-CBAM with principal component analysis for wind turbine blade cracking fault prediction with only short time scale SCADA data". Measurement, **212** (2023).
18. A. Zaher, S. D. J., McArthur, D. G. Infield. "Online wind turbine fault detection through automated SCADA data analysis". Wind Energy, 12, 574-593 (2009).
19. S. O. Erikstad. "Merging physics, big data analytics and simulation for the next-generation digital twins". *HIPER 2017* (2017).
20. A. Coraddu, L. Oneto, F. Baldi, F. Cipollini, M. Atlar, S. Savio. "Data-driven ship digital twin for estimating the speed loss caused by the marine fouling". Ocean Engineering, **186** (2019).
21. Z. Lin, X. Liu, M. Collu. "Wind power prediction based on high-frequency SCADA data along with isolation forest and deep learning neural networks". International Journal of Electrical Power & Energy Systems, **118** (2020).
22. I. Delgado, M. Fahim. "Wind turbine data analysis and LSTM-based prediction in SCADA system". Energies, **14**, 1 (2021).
23. S-X. Lv, L. Wang. "Deep learning combined wind speed forecasting with hybrid time series decomposition and multi-objective parameter optimization". Applied Energy, **31** (2022).
24. S-h. Hur. "Short-term wind speed prediction using Extended Kalman filter and machine learning". Energy Reports, **7**, 1046-1054 (2021).
25. M. Yesilbudak, S. Sagiroglu, S. I. Colak. "A novel implementation of kNN classifier based on multi-tupled meterological input data for wind power prediction". Energy Conversion and Management, **135**, 434-444 (2017).
26. A. D. Mehr, R. Farhangi, A. R. Ghiasi. "The validity of deep learning computational model for wind speed simulation". IEE, 2021 7th International Confrence on Control, Instrumentation and Automation (2021).
27. M. Singh, E. Fuenmayor, E. P. Hinchy, Y. Qiao, N. Murray, D. Devine. "Digital Twin: origin to future". Applied System Innovation, **4**, 2 (2021).
28. I. Errandonea, S. Beltrán, S. Arrizabalaga. "Digital Twin for maintenance: A literature review". Computers in Industry, **123** (2020).
29. C. Semeraro, M. Lezoche, H. Panetto, M. Dassist. "Digital twin paradigm: A systematic literature review". Computers in Industry, **130** (2021).
30. J. Xia, G. Zou. "Operation and maintenance optimization of offshore wind farms based on digital twin: A review". Ocean Engineering, **268** (2023).
31. AMRC. "Untangling the requirements of a Digital Twin" (2020) https://www.amrc.co.uk/files/document/404/1604658922_AMRC_Digital_Twin_AW.pdf.
32. Y. Wang, R. Zou, F. Liu, L. Zhang, Q. Liu. "A review of wind speed and wind power forecasting with deep neural networks". Applied Energy, **304** (2021).
33. F. C. Mehlan, E. Pedersen, A. R. Nejad. "Modelling of wind turbine gear stages for Digital Twin and real-time virtual sensing using bond graphs". J. Phys.: Conf. Ser., **2265** (2022).
34. M. Wang, C. Wang, A. Hnydiuk-Stefan, S. Feng, I. Atilla, Z. Li. "Recent progress on reliability analysis of offshore wind turbine support structures considering digital twin solutions". Ocean Engineering, **232** (2021).
35. K. Sivalingam, M. Sepulveda, M. Spring, P. Davies. "A Review and Methodology Development for Remaining Useful Life Prediction of Offshore Fixed and Floating Wind turbine Power Converter with Digital Twin Technology Perspective". IEE Explore. *2018 2nd International Conference on Green Energy and Applications* (2018).
36. C. Kim, M-C, Dinh, H-J. Sung, K-H. Kim, J-H, Choi, L, G. I-K, Yu. M. Park. "Design, implementation, and evaluation of an output prediction model of the 10 MW floating offshore wind turbine for a Digital Twin. Energies, **15**, 17.
37. C. Plumley. "Kelmarsh wind farm data". Zenodo (2022) https://zenodo.org/record/5841834.
38. OpenWeather. OpenWeather – Weather forecasts, nowcasts and history in a fast and elegant way". OpenWeather (2023) https://openweathermap.org/.
39. O. Eyecioglu, B. Hangun, K. Kayisli, M. Yesilbudak. Performance comparison of different machine learning algorithms on the prediction of wind Turbine power generation". IEE. *2019 8th International Conference on Renewable Energy Research and Applications* (2019).
40. S. Tajmouati, B. El Wahbi, A. Bedoui, A. Abarda, M. Dakkoun. "Applying k-nearest neighbors to time series forecasting : two new approaches". arXiv.2103.14200.
41. M. Yesilbudak, S. Sagiroglu, I. Colak. "A new approach to very short term wind speed prediction using k-nearest neighbor classification". Energy Conversion and Management, **69**, 77-86 (2013).
42. U. Singh, M. Rizwan, M. Alaraj, I. Alsaidan. "A machine learning-based gradient boosting regression approach for wind power production forecasting: A step towards smart grid environments". Energies, **14**, 16 (2021).
43. R. Becker and D. Thrän, "Completion of wind turbine data sets for wind integration studies applying random forests and k-nearest neighbors," Applied Energy, **208**, 252-262 (2017).
44. K. Chomboon, P. Chujai, P. Teerarassamee, K. Kerdprasop, N. Kerdprasop. "Completion of wind turbine data sets for wind integration studies applying random forests and k-nearest neighbors". *Proceedings of the 3rd international conference on industrial application engineering* (2015).
45. SciPy. "Distance computations" (2023) https://docs.scipy.org/doc/scipy/reference/spatial.distance.html.
46. C. Jung, D. Schindler. International Journal of Energy Research, **45**, 6..
47. Nvidia. "XGBoost". Nvidia (2023) https://www.nvidia.com/en-us/glossary/data-science/xgboost/.
48. Open Weather. "One Call API 3.0". OpenWeather. (2023) https://openweathermap.org/api/one-call-3.
49. OpenWeather. "Create New History Bulk". OpenWeather. (2023) https://home.openweathermap.org/history_bulks/new.
50. Microsoft. "Learn about twin models and how to define them in Azure Digital Twins". Microsoft

(2023) https://learn.microsoft.com/en-us/azure/digital-twins/concepts-models.

51. Microsoft. “Tutorial: Coding with the Azure Digital Twins SDK”. Microsoft (2023) https://learn.microsoft.com/en-us/azure/digital-twins/tutorial-code.
52. LazaUK “IndustrySolutions-WindFarm/3D_Models”. Github (2023) https://github.com/LazaUK/IndustrySolutions-WindFarm/tree/main/3D_Models.
53. Microsoft. “Ingest IoT Hub telemetry into Azure Digital Twins”. Microsoft (2023) https://learn.microsoft.com/en-us/azure/digital-twins/how-to-ingest-iot-hub-data.
54. M. Santhosh, C. Venkaiah, D. M. Vinod Kumar. "Current advances and approaches in wind speed and wind power forecasting for improved renewable energy integration: A review”. Engineering Reports, **2**, 6 (2020).
55. T. Chai R. R. Draxler. “Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature". Geosci. Model Dev., **7**, 3, 1247-1250 (2014).
56. Meteoblue. Weather API+”. Meteoblue (2023) https://www.meteoblue.com/en/weather-api.

Authors' background

Your Name	Title*	Research Field	Personal website
Eamonn Tuton	Phd candidate	Digital Twin use for wind energy operations and maintenance	
Xinhui Ma	Lecturer	Data science, machine learning, Digital Twins	https://www.hull.ac.uk/staff-directory/Xinhui-Ma
Nina Dethlefs	Senior lecturer	Artificial intelligence, interactive systems, environmental modelling, offshore wind	https://www.hull.ac.uk/staff-directory/nina-dethlefs

*This form helps us to understand your paper better, the form itself will not be published.

*Title can be chosen from: master student, Phd candidate, assistant professor, lecture, senior lecture, associate professor, full professor