The Bartlett School of Environment, Energy and Resources

MSc ESDA Title Page

UCL Candidate Code: PNXF3, **KVQS4, NQMJ9, KXBS7, MMVQ8**

Module Code: BENV0148

Module Title: Advanced Machine Learning for Energy Systems 24/25

Coursework Title: Forecasting the Impact of Climate Change on Great Britain's Offshore Wind Power Generation Under Different RCP Scenarios using Deep Learning

Module Leader: Dr. Amir Gharavi

Date: 29/04/2025

Word Count: 2996

By submitting this document, you are agreeing to the Statement of Authorship:

I/We certify that the attached coursework exercise has been completed by me/us and that each and every quotation, diagram or other piece of exposition which is copied from or based upon the work of other has its source clearly acknowledged in the text at the place where it appears.

I/We certify that all field work and/or laboratory work has been carried out by me/us with no more assistance from the members of the department than has been specified.

I/We certify that all additional assistance which I/we have received is indicated and referenced in the report.

Please note that penalties will be applied to coursework which is submitted late, or which exceeds the maximum word count. Information about penalties can be found in your Course Handbook which is available on Moodle:
https://moodle.ucl.ac.uk/course/view.php?id=29435&section=0

- ***Penalties for late submission***
- ***Penalties for going over the word count***

In the case of coursework that is submitted late and is also over length, then the greater of the two penalties shall apply. This includes research projects, dissertations and final reports.

I. Abstract

This study aims to evaluate the potential impacts of climate change on offshore wind generation in Great Britain, a key technology in its decarbonisation ambitions, under the Representative Concentration Pathway (RCP) 2.6 and 8.5 scenarios. Deep Learning (LSTM) forecasting models were developed for 21 windfarms connected to the GB grid using historical climate data as predictors and hourly generation data as the target variable, from 2019-2024. The models were used to predict hourly future wind generation using projected climate variables for the high (RCP 8.5) and low (RCP 2.6) emission scenarios. The models forecasted similar results for both scenarios. However, the RCP 2.6 forecasts had less fluctuations indicating less variability in climate patterns in this scenario.

II. Introduction

The UK government has ambitious goals for offshore wind, with a capacity target of 43-50GW by 2030, nearly double of its onshore wind capacity target (Government, UK, 2024). The UK has established itself as a global leader in offshore wind, second to China, with an installed capacity of 96GW across 123 projects (Norris, 2025), and a record generation of 83 TWh in 2024 (Dale, 2025).

This project presents a novel forecasting framework to investigate the potential impacts of climate change on offshore wind generation in Great Britain by leveraging machine learning (ML) and deep learning (DL) models. Traditionally, wind power curves, which are non-linear, are used to convert wind speeds at a windfarm into power output values. Our methodology uses deep learning models as a proxy for these wind power curves which incorporates extra climate variables such as temperature, dew point and pressure, to predict the power generated at windfarm level.

By training individual DL models on historical data for each windfarm and using projected climate variables given by climate simulations for different RCP forcing scenarios, the variation in future generation for offshore windfarms can be investigated for different climate scenarios. Our aim is to evaluate if DL techniques can be used to forecast future offshore windfarms and if so, investigate how future generation is affected under different RCP scenarios. This work is motivated by trying to improve on previous work done in Semester 1 which investigated the same issue but used machine learning only.

III. Literature Review

Traditional linear and simple nonlinear regression models fail to characterize the complexity of the non-linear features affecting offshore wind generation (Yun Wang, 2021). Deep Neural Networks (DNNs) have consequently gained further attention in this area, due to their ability to better characterize non-linear relationships. A prime example is Kaminski's paper (Mateusz KAMIŃSKI, 2023) which uses a neural network to incorporate extra features other than windspeed, such as air turbulence to predict power output with a 99.84% predictive accuracy. Likewise, Liu's study (Zongxu Liu, 2025) highlights how Long-short-term-memory (LSTM) networks or Convolutional Neural Networks (CNNs) models better characterise the sequential nature of wind power data to improve prediction accuracy.

While Deep Neural Networks (DNNs) have advantages over traditional ML models, they have their downsides. Hanifi's study (Shahram Hanifi, 2023) compared LSTM, CNN, and Gated Recurrent Unit (GRU) models with wavelet decomposition (WD) for offshore wind generation forecasting and concluded CNN models should be avoided due to their reliance on heavy preprocessing techniques such as WD, with LSTMs performing best. Considering this, our research will aim to use an LSTM model to improve on to machine learning based techniques used in our previous work.

IV. Methodology

A schematic diagram of our workflow is shown in **Fig. 1,** with our previous work informing the climate variables to consider.

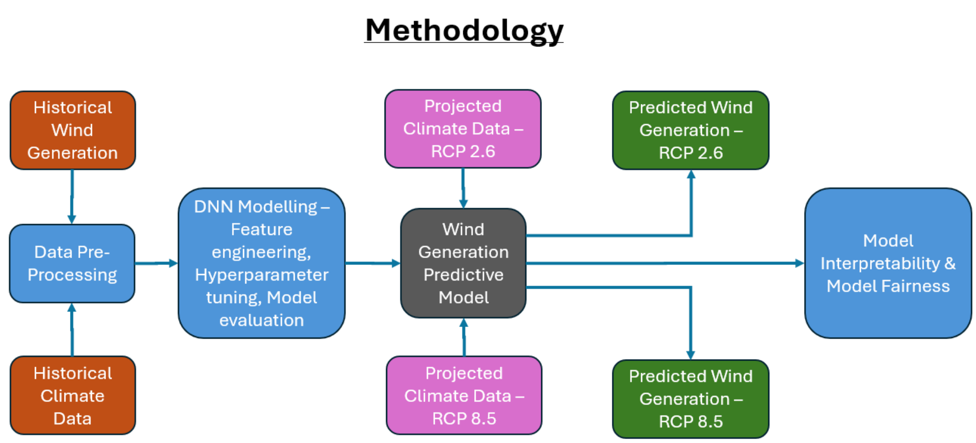

Fig. 1: *Methodology workflow for project*

i) Historical Wind Generation Data:

The first step was to collect a list of offshore windfarms in the UK, which was obtained from the Renewable Energy Planning Database (REPD), compiled by the UK's Department for Energy Security and Net Zero (DESNZ), which tracks the progress of UK renewable electricity projects over 150kW through the planning system (UK Government, 2025). For each offshore windfarm extracted, the Settlement Balancing Mechanism Units (BMU) associated with it were obtained by referencing to the Power Station Dictionary (Bourn, 2025). Offshore windfarms often have several BMUs associated with them, enabling more fine-tuned balancing by grid operators. The Elexon API was then used to obtain the historical generation data from 01/01/2019 to 31/12/2024 for each BMU unit collected.

The “Actual Generation Output Per Generation Unit (B1610) stream” endpoint was used, which provides the “actual metered volume output (MWh) per half-hourly Settlement Period for all BM units (Positive, Negative or zero MWh values)” (Elexon BSC, 2025). For windfarms operational after 01/01/2019, the generation data was collected from their "Operational" date. Although only 53/66 of BM units collected returned data when the Elexon API was queried, all offshore windfarms had at least one their associated BM units which returned generation data.

Generation data was summed together for any windfarm with generation data for several of its BMUs. The process left us with the generation data for 21 unique offshore windfarms, see **Fig. 2**. Not all offshore windfarms in the UK are signed up to the Balancing Mechanism; only those that are signed up have generation data available on Elexon.

Fig. 2: Map of the offshore windfarms connected to Great Britain’s power grid which returned generation data via the Elexon API.

ii) Historical Climate Data:

For each windfarm location obtained from the REPD database, the climate variables, described in **Table 2**, were extracted from 01/01/2019 to 31/12/2024, from the "ERA5 hourly data on single levels from 1940 to present" dataset, available on the Copernicus Climate Change Service’s (C3S) Climate Data Store (CDS) (Copernicus Climate Change Service, Climate Data Store, 2025).

Climate Variable:	Units:	Description:
10m u-component of wind	m s^{-1}	Eastward horizontal component of wind at 10m height. Can be combined with v-component to determine wind speed and direction.
10m v-component of wind	m s^{-1}	Northward horizontal component of wind at 10m height. Can be combined with u-component to determine wind speed and direction.
2m dewpoint temperature	°K	Temperature to which air at 2m height would need to be cooled for saturation. Can be combined with temperature to calculate relative humidity.
2m temperature	°K	Air temperature at 2m above surface of land, sea or inland waters.
Surface pressure	Pa	Atmospheric pressure at the Earth's surface. Represents weight of air column above a point.

Table 2: *ERA5 historic climate variables downloaded from the CDS.*

iii) Projected Climate Data:

The projections for our climate variables, (see **Table 3)**, were obtained from the "CORDEX regional climate model data on single levels" dataset from the CDS (Copernicus Climate Change Service, Climate Data Store, 2019). The CORDEX projections are obtained from Regional Climate Model (RCM) simulations driven by boundary conditions from Global Climate Models for different RCP forcing scenarios. RCMs have a higher resolution than GCMs. "Single level" variables are 2D-matrices computed at one vertical level. Simulations were done for different combinations of GCMs and RCMs, and multiple ensemble members for a GCM-RCM combination are created by running simulations with slightly different conditions. For more information, see (ECMWF, 2025).

The GCM selected was the Met Office Hadley Centre's (MOHC) **HadGEM2-ES** model (Met Office, 2025) and the RCM selected was the Royal Netherlands Meteorological Institute (KNMI) **RACMO22E** model (KNMI, 2025) This combination was chosen as it was the most local to the UK with all our variables of interest available to download for both RCP scenarios. Projections were collected for 2026 to 2045.

Climate Variable:	Units:	Description:
10m u-component of wind	m s^{-1}	Eastward horizontal component of wind at 10m height. 6h temporal resolution.
10m v-component of wind	m s^{-1}	Northward horizontal component of wind at 10m height. 6h temporal resolution.
10m wind speed	m s^{-1}	Magnitude of the two-dimensional horizontal air velocity at 10m height. The data is the mean value over the aggregation period. Daily temporal resolution.
2m relative humidity	%	Relative humidity is defined as the % ratio of water vapour mass to the water vapour mass at saturation point given the temperature at that location. The data is the mean value over the aggregation period at 2m height. Daily temporal resolution.

2m temperature	°K	Air temperature at 2m height. The data is the mean value over the aggregation period. Daily temporal resolution.
Surface pressure	Pa	Atmospheric pressure at the Earth's surface. Daily temporal resolution.

Table 3: *Climate variables from CORDEX projections downloaded from the CDS.*

The projections were saved for bounding boxes around each windfarm, as for the ERA5 data, with their width matching the horizontal resolution of the data. The average values of the variables in each bounding box at each time-step was computed and combined to form a final multivariate timeseries for the data.

iv) Data aggregation and pre-processing

The historical climate data for each wind farm from 2019 – 2024 is first converted from GRIB to CSV format and then concatenated to a single data frame for faster processing. The wind generation data is then aggregated from half-hourly to hourly data. The generation data for each BMU for a wind farm, if it had several, was summed to obtain its total generation. Finally, the historical climate and wind generation data, all at hourly resolution from 2019 – 2024, is merged to create a final dataset for each wind farm.

V. Exploratory Data Analysis (EDA)

To simplify the EDA, the data for all windfarms was combined into a single dataset. The assumption here is that the high-level distribution of features in the combined data frame is applicable to each wind farm.

i) Feature Distributions

The EDA revealed that most features followed a near-normal distribution – see **Fig. 3.** However, temperature was bi-modal, and generation was right skewed, with many zeros. These distributions are acceptable as DNNs are flexible and able to learn non-linear relationships.

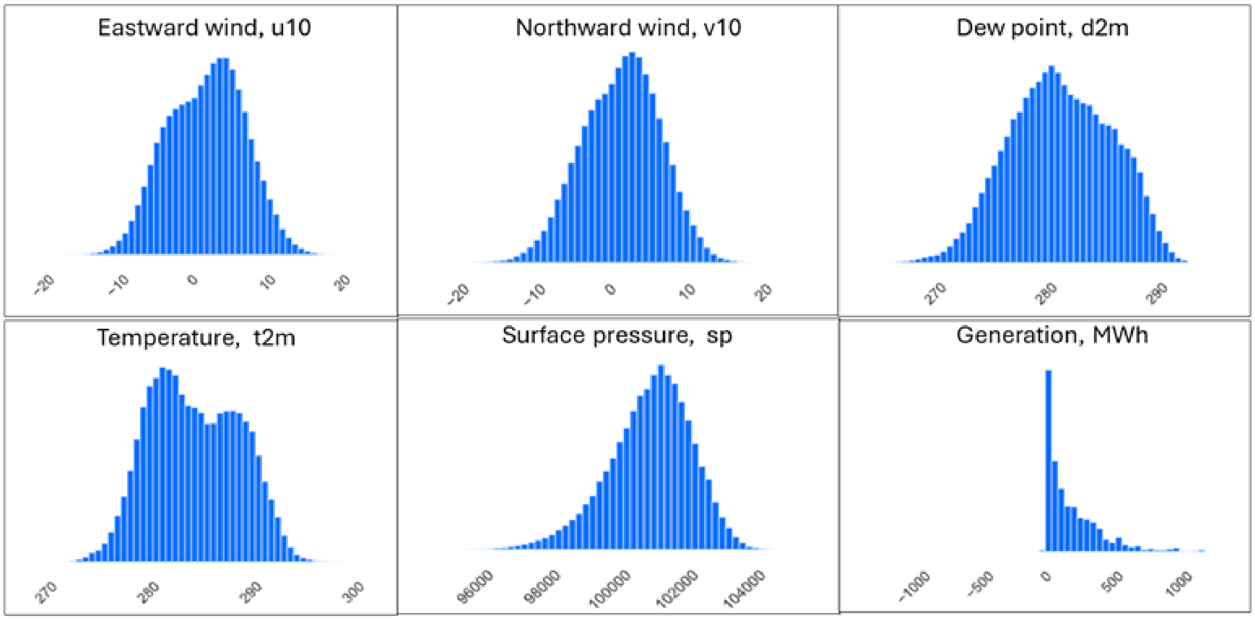

***Fig 3:** feature distributions*

The heat map in **Fig. 4** shows a high correlation between the temperature and dewpoint, as well as longitude, latitude and wind farm name, as expected.

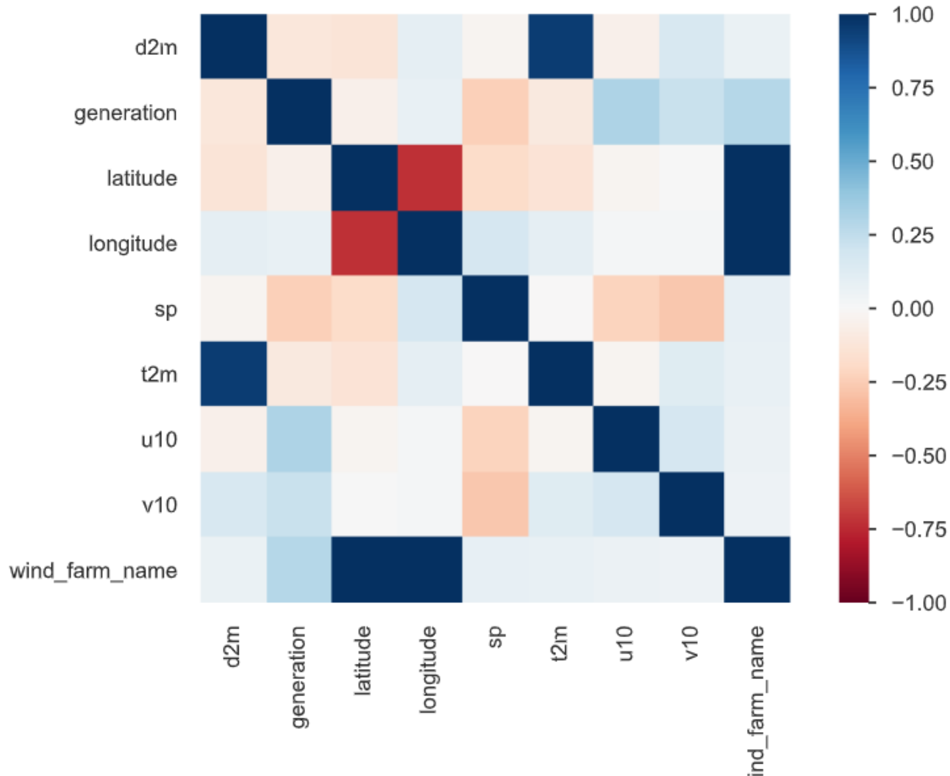

Fig 4: *Heatmap of features*

ii) Investigating Outliers

Fig. 5 and **Fig. 6** show the outliers in the features and target variables respectively.

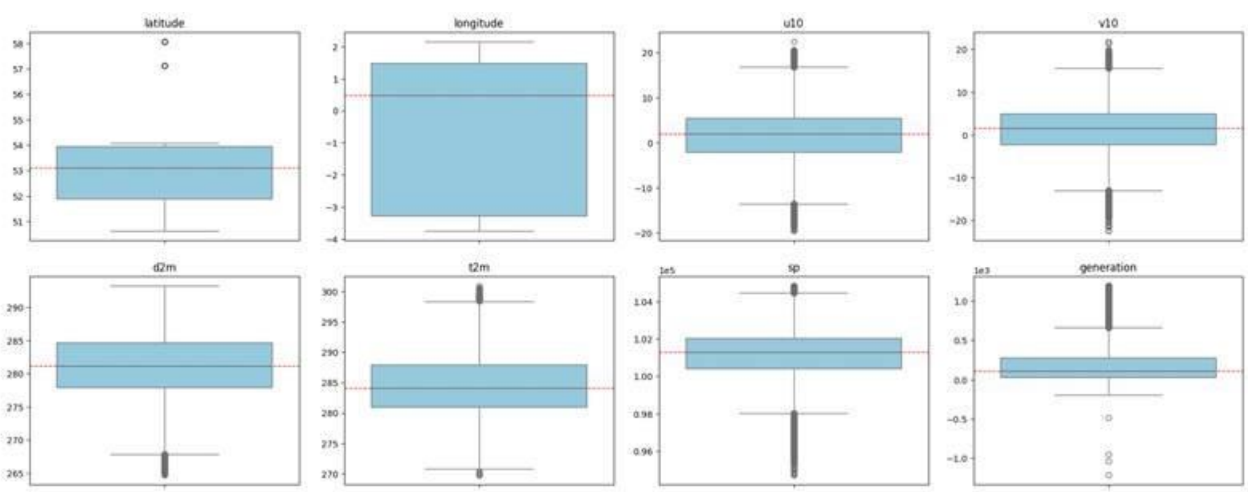

Fig 5: *Outliers in features*

To deal with outliers, zero and negative Generation values were removed, along with their corresponding observations for the other features. The minimum generation was capped at 0.1MWh. The total number of observations decreased from 1,022,571 to 924,325.

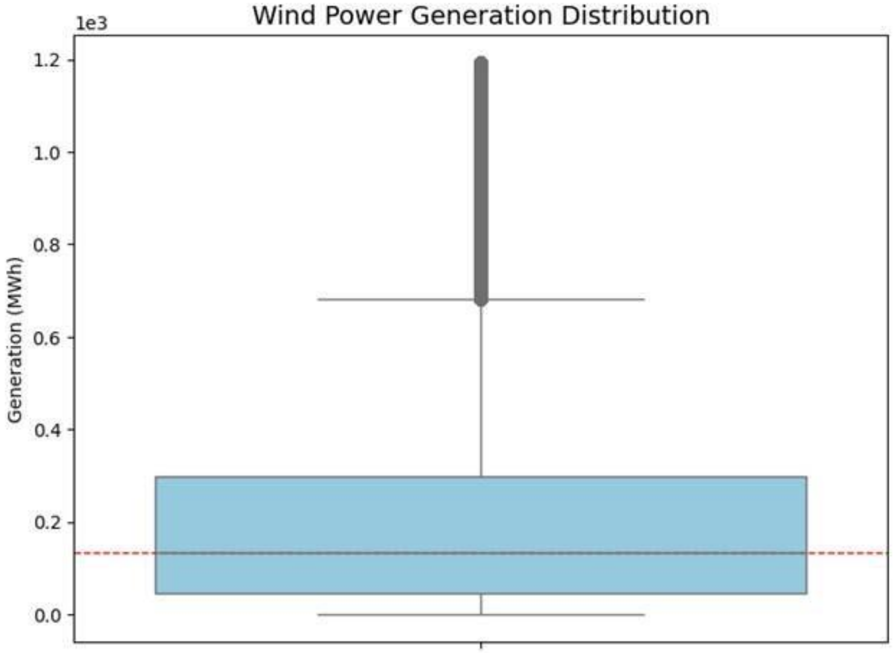

Fig. 6: *Example of outliers in target variable*

The upper outliers and the remaining near-zero generation values were kept, accounting for periods of high and low generation. Additionally, the Huber loss function was used during modelling process, as it is robust to outliers and helps train balanced models (Huber, 1964).

iii) Missing Values and Numerical Encoding

The EDA revealed there were no missing values, as shown in **Fig. 7**, which meets the first requirement for regression models.

Dataset statistics	
Number of variables	10
Number of observations	1,022,571
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	78.0 MiB
Average record size in memory	80.0 B

Variable types	
DateTime	1
Numeric	8
Categorical	1

Fig 7: *Dataset statistics*

However, the second requirement of numeric feature values was not satisfied. The dataset contained one Date-Time and one categorical variable, which were treated using encoding methods.

The categorical variable, the wind farm names in the combined dataset, was "Label Encoded" to produce a new feature called "wind_farm_label", which assigns a unique integer value to each wind farm. Cyclical Encoding was applied to the DateTime variable, i.e., the "hour" feature. This preserves the cyclical nature of features such as time, ensuring values are continuous, and is achieved by extracting the "Hour of day", "Month" and "Year" from the DateTime feature. These new features are then cyclically encoded which produces their respective sine and cosine components (hour_sin, hour_cos, month_sin, month_cos).

iv) Feature Engineering

Wind speed and wind direction features were also engineered as they had high feature importance in models generated from our Semester 1 work. The Wind Speed was engineered by applying Pythagoras' Theorem to the eastward and northward wind components as follows:

$$\text{Wind Speed} = \sqrt{u_{10}^2 + v_{10}^2}$$

Wind Direction was engineered applying the arctan function to the u- and v-wind components:

$$\text{Wind Direction} = \arctan\left(\frac{v_{10}}{u_{10}}\right)$$

The final list of processed features in the dataset and their types are seen in **Fig. 8.** All the initial features which were used to engineer new features were also retained in the dataset.

hour	datetime64[ns]
latitude	float64
longitude	float64
u10	float64
v10	float64
d2m	float64
t2m	float64
sp	float64
wind_farm_name	object
wind_farm_label	int32
hour_of_day	int32
month	int32
year	int32
hour_sin	float64
hour_cos	float64
month_sin	float64
month_cos	float64
wind_speed	float64
wind_dir	float64
generation	float64

Fig 8: *Final set of processed features and their data type.*

VI. Baseline Modelling

Using the combined dataset, which was used for the EDA, three baseline models were trained as outlined below:

1. Random Forest (RF) – to explore feature importance of predictor variables.
2. DNN – to learn non-linear relationships between the features and target variables.
3. LSTM – to potentially improve DNN performance as used in previous studies.

The combined dataset was used for baseline modelling to reduce computational processing times. All models were evaluated using the R-Squared (R^2), Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) metrics, as appropriate for our models which are regression models. For simpler comparison across different models, our discussion focuses on the R^2 metric.

- The DNN and LSTM models were compiled using the Huber loss, given our decision to retain outliers during our Data Cleaning to enable the model to predict high and low generation periods.
- Early stopping with a patience of 5 was used to prevent overfitting and save computational resources.
- ADAM optimizer with a learning rate of 0.001 was used to smoothen the learning process.
- ReLu activation function was used in the hidden layers to keep all values positive or zero, and a linear activation function was used in the output layer.
- Dropout regularization of 0.2 was used in all hidden layers to prevent overfitting.
- Layer normalization was applied to the final LSTM model to prevent exploding and vanishing gradients.

In all models, the features were scaled using a Standard Scaler to transform all features to the same range, ensuring that no feature outweighs any other during training. The features were split into train and test sets before applying the scaler to prevent data leakage. A test and validation size of 0.2 each was used for all models.

Before the baseline deep learning modelling was performed, a random forest regression was run; the feature importance is shown in **Fig 9.**

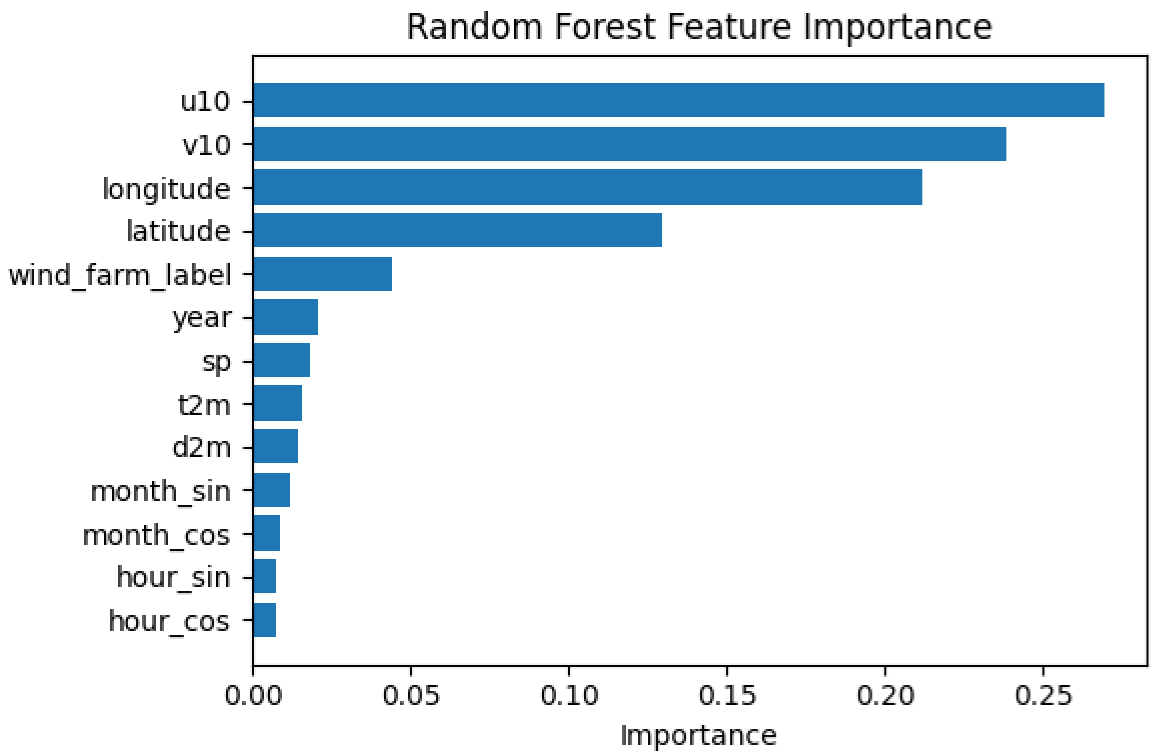

Fig. 9: *Random forest feature importance*

I) Baseline DNN

Several iterations of DNNs are trained using the same features as identified in the random forest regression. Using 3 hidden layers in the first iteration, this model achieved an R^2 of 0.84. Using 5 wider hidden layers in the second iteration improved the R^2 to 0.88.

To further improve the DNN performance, an advanced encoding method called Embedding is tested. Guo and Berkhahn (2016) introduced Entity Embedding of categorical variables which maps similar entities together and aids NNs in learning intrinsic properties of categories.

Despite the reduction of the hidden layers from 5 to 2 due to the long computational time of DNN2, DNN3 still achieved a slightly better R^2 of 0.89. This shows the potential of Neural Networks paired with more powerful computing resources.

II) Baseline LSTM

The LSTM performed worse than both RF and DNN architectures, achieving an R^2 of 0.31. The large gap between training and validation loss indicates underfitting.

This might be explained by the combined dataset interfering with the LSTM sequencing feature. The initial assumption is that since DNNs performed well, LSTMs would perform better at the wind farm level.

To conclude, the best performing features of our baseline modelling, results summarized in **Table 4**, were identified using RF1. DNN3 performed almost as well as RF1 model, showing room for improvement through hyperparameter tuning. The LSTM performed poorly; however, this could change at wind farm level. Further details on our iterations and model architectures can be found in our Jupyter notebook, see Appendix link.

Baseline Models' Results				
Model	**RMSE**	**MAE**	**R-Squared**	**Model Explanation**
RF1	55.981	33.552	0.929	raw features + cyclical encoding + label encoding wind farm names
RF2	89.691	58.739	0.817	testing wind speed and direction features, and target encoding of wind farm names
DNN1	87.71	55.88	0.82	3 hidden layers
DNN2	72.9	45.36	0.88	5 hidden layers
DNN3	68.78	41.81	0.89	Entity Embedding
LSTM	223.35	147.58	0.31	raw features + cyclical encoding

Table 4: *Baseline modelling results*

Fig 10. and **Fig 11.** show training and validation loss of the 3rd iteration of the DNN and the LSTM baseline models respectively.

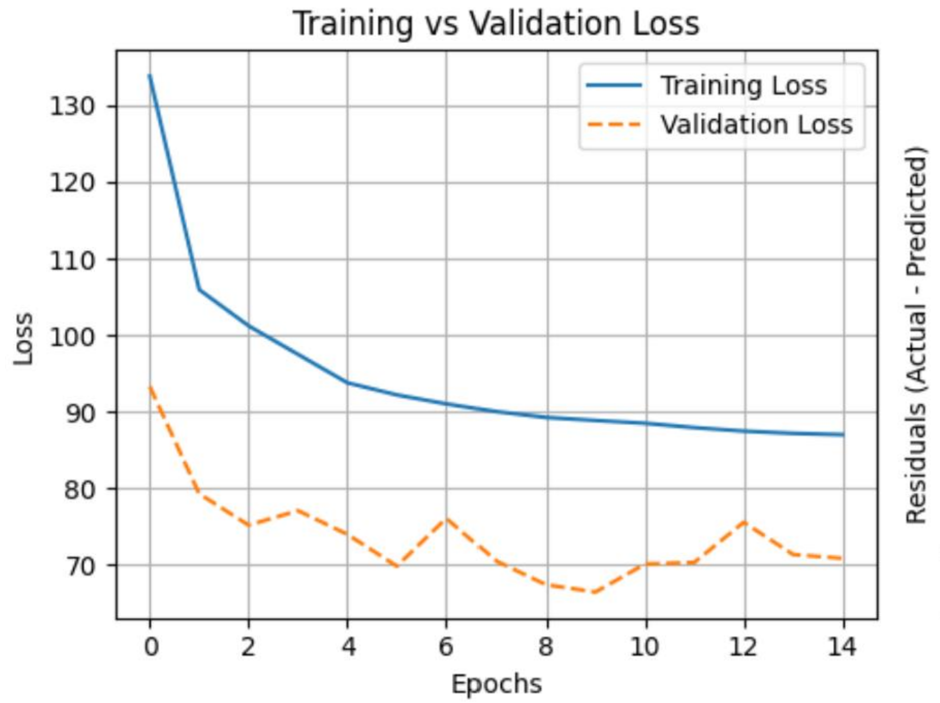

Fig 10: *DNN3 training and validation loss*

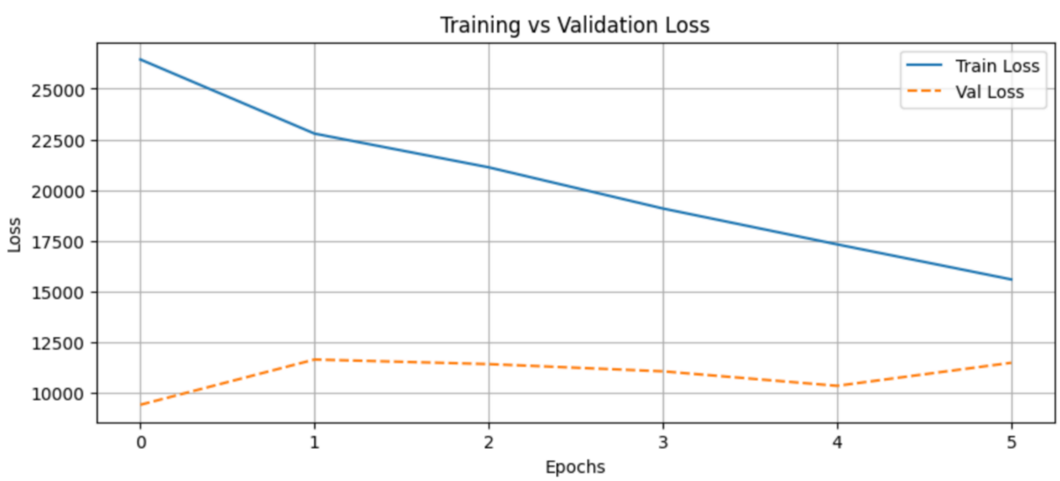

Fig 11: *LSTM training and validation loss*

VII. Windfarm-level modelling

Fig 12. and **Fig 13.** show the R^2 values for DNN and LSTM models trained for all windfarms respectively.

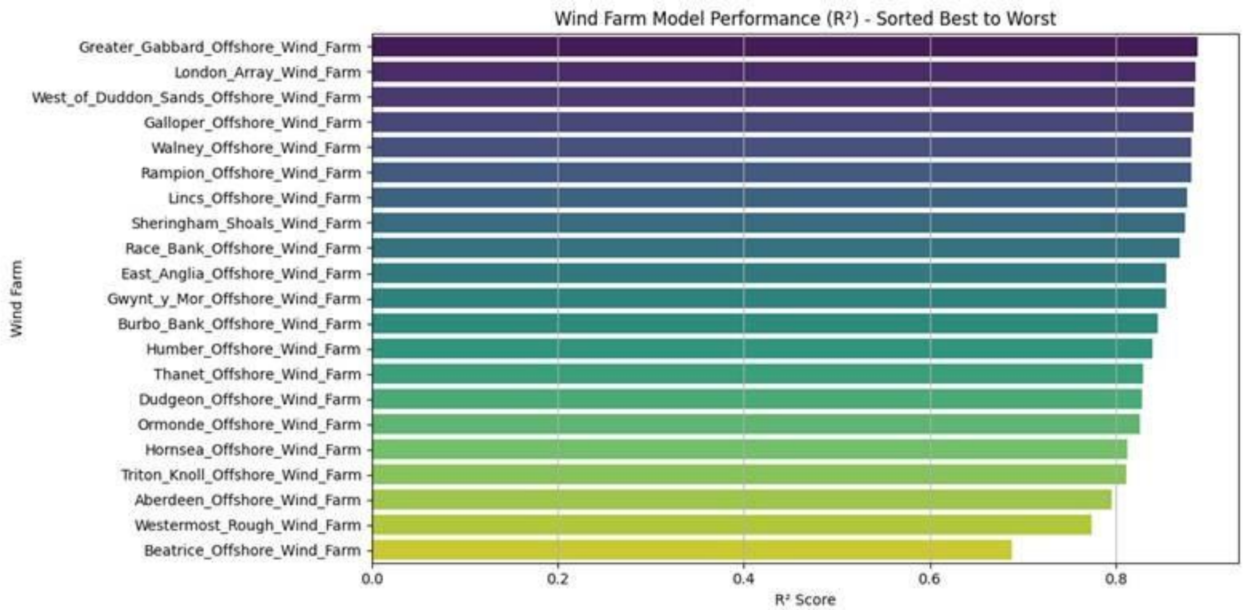

Fig 12: *DNN – Windfarm level*

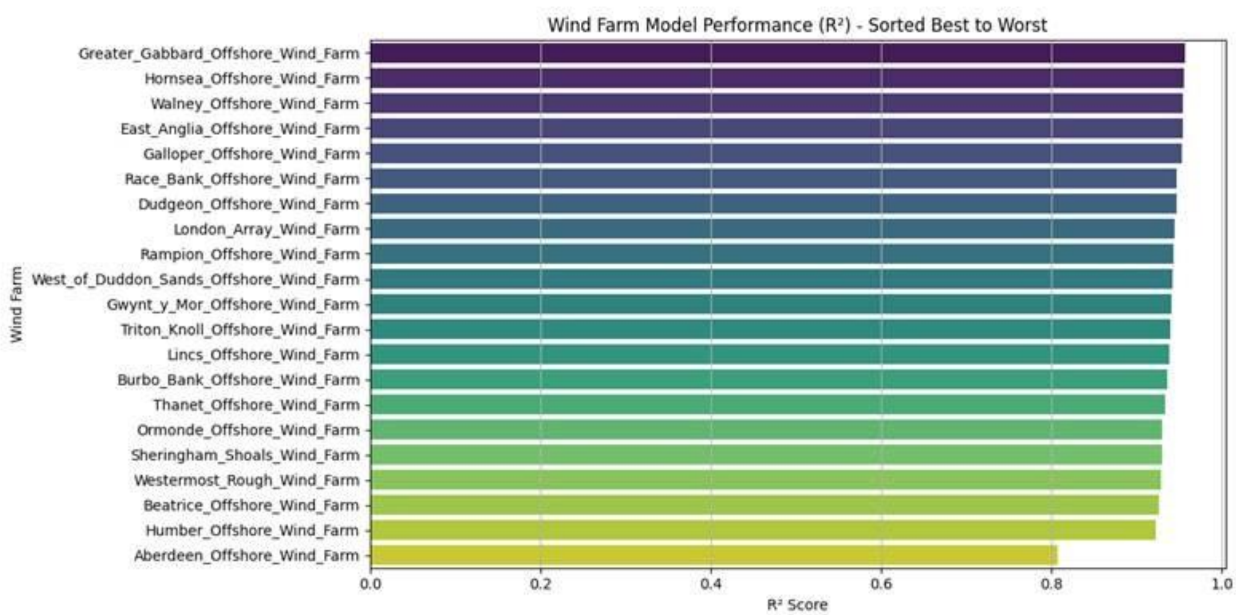

Fig 13: *LSTM – Windfarm level*

Because the LSTM consistently outperformed the DNN models across most windfarms, the LSTM models, and scalers, which were performed best were saved to use in our forecasting.

VIII. Results & Discussion - Wind power forecasting

Each wind farm's processed climate projection data was scaled using their respective scalers. Then, the saved LSTM models were used to forecast future wind generation under the RCP 8.5 and 2.6 scenarios. The models were trained on hourly generation data, hence the forecast results are expected hourly generation in each year.

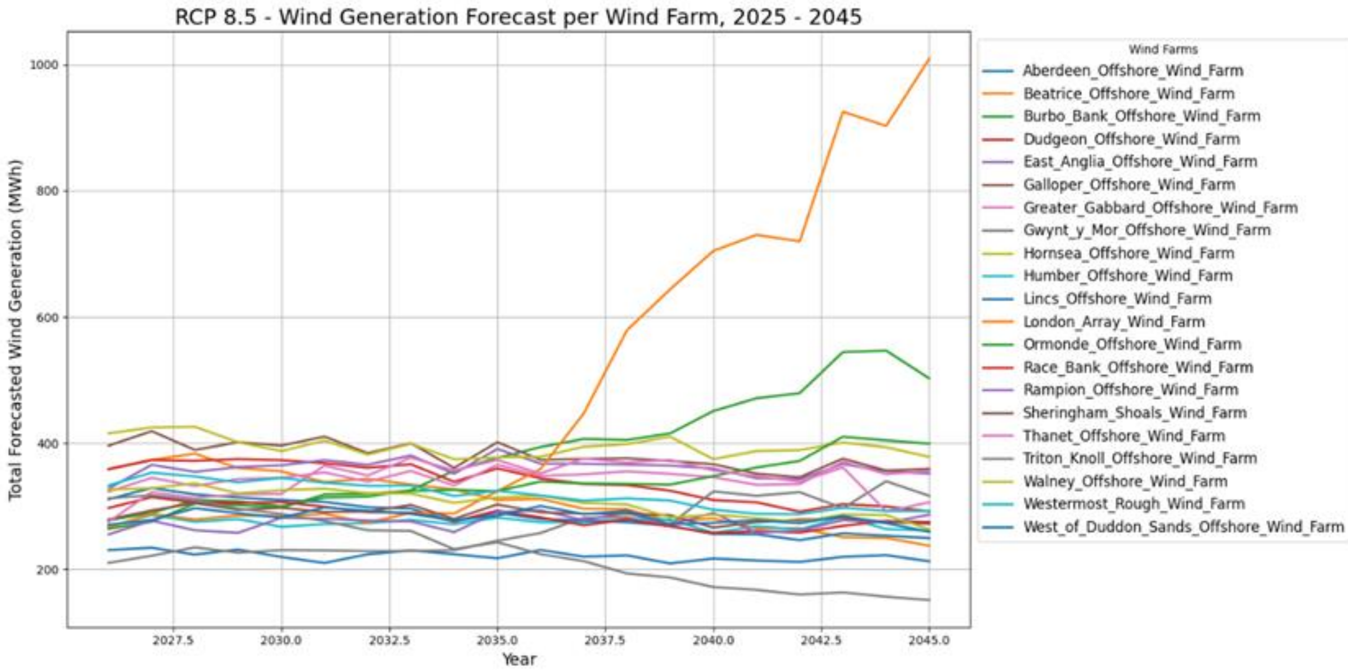

Fig 14: *Wind farm level forecasts under RCP 8.5 Scenario.*

RCP 8.5 Scenario:

Fig 14. presents the forecasted annual wind generation (MWh) from 2025 to 2045 under the RCP 8.5 scenario. Most wind farms exhibit relatively stable and consistent generation trends, clustering between 200- and 400-MWh per hour annually. The London Array model ($R^2 = 0.93$) projected a significant increase in generation starting around 2035, reaching a record high of over 1000 MW per hour by 2045. This suggests drastically increasing wind speeds due to unmitigated climate change. Similarly, the Burbo Bank model ($R^2 = 0.92$) projected the second highest increase in output, hitting a peak of about 550 MW in the mid 2040s. In contrast, the Triton Knoll model ($R^2 = 0.94$) projected a decrease from about 300 MW to 100MW between 2030 and 2045.

RCP 2.6 Scenario:

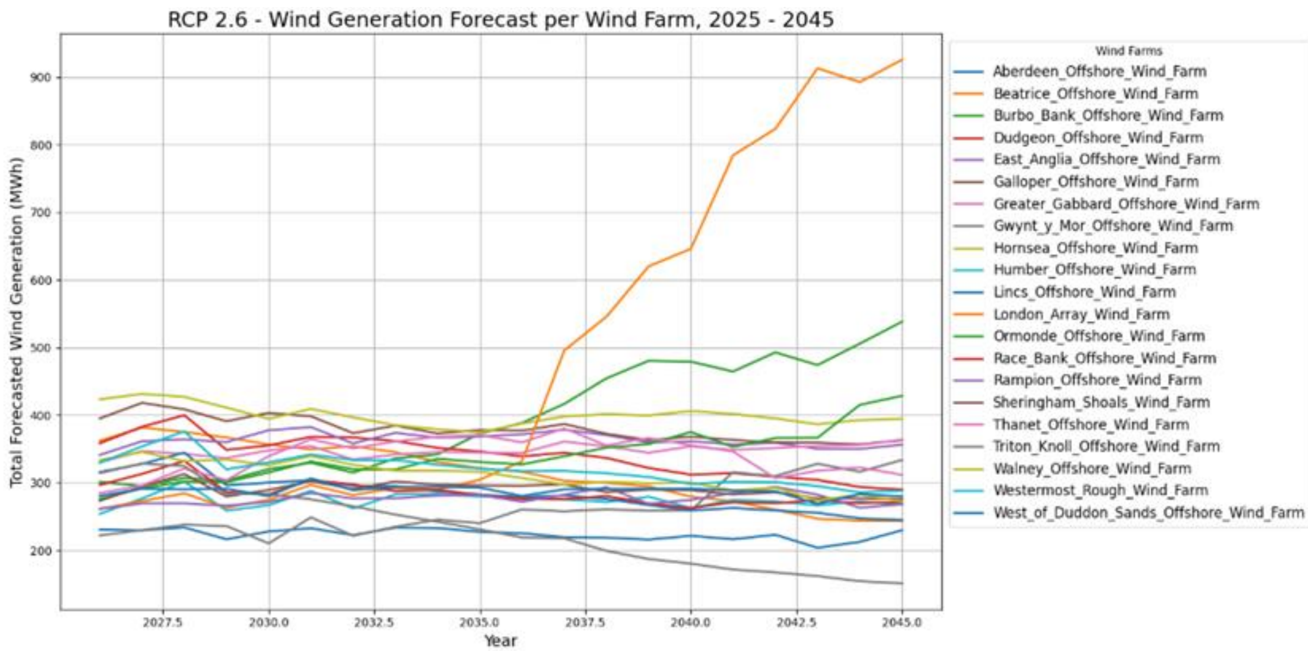

Fig 15: *RCP 2.6 Scenario.*

Fig 15. presents the forecasts for the RCP 2.6 scenario. The results are broadly similar to those under RCP 8.5. However, in the RCP 2.6 scenario, the projected increase in power output of London Array and Burbo Bank is about 100 MWh less per hour. Additionally, RCP 2.6 forecasts showed less fluctuations in generation patterns across the other wind farms, indicating stable weather patterns.

VI. Conclusion

This study found that LSTM models performed better than DNN and RF models in forecasting wind generation using climate data at the wind farm level.

The LSTM models forecasted similar results under both the RCP 8.5 and 2.6 scenarios. London Array and Burbo Bank wind farms had the greatest forecasted increase in generation in both scenarios, while Triton Knoll's generation was forecasted to decrease.

Overall, the more stable generation patterns under RCP 2.6 align with climate science expectations, where climate change mitigations lead to more stable weather conditions and reduced variability in wind patterns.

Future Work:

- Evaluate results using data from different combinations of climate models
- Evaluate other scenarios such as RCP 4.5 - medium stabilization scenario
- Train on more historical data
- Conduct similar studies for onshore wind generation
- Sensitivity analysis of how input features affect predictions

VII. Bibliography

- Arun Kumar M, R. J. K. S. R. C. D. E. J. K. D. M., 2024. Predicting wind power using LSTM, Transformer, and other techniques. Clean Technologies and Recycling, 4(2), pp. 125-145.
- Bourn, A., 2025. Power Station Dictionary. [Online] Available at: https://osuked.github.io/Power-Station-Dictionary/dictionary.html [Accessed 24 April 2025].
- Copernicus Climate Change Service, Climate Data Store, 2019. CORDEX regional climate model data on single levels. [Online] Available at: https://cds.climate.copernicus.eu/datasets/projections-cordex-domains-single-levels?tab=overview [Accessed 21 April 2025].
- Copernicus Climate Change Service, Climate Data Store, 2025. ERA5 hourly data on single levels from 1940 to present. [Online] Available at: https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels?tab=download [Accessed 22 April 2025].
- Dale, M. P. a. B., 2025. Record year for wind power in 2024. [Online] Available at: https://www.bbc.co.uk/news/articles/c3e3g9xv3ylo [Accessed 17 April 2025].
- ECMWF, 2025. CORDEX: Regional climate projections. [Online] Available at: https://confluence.ecmwf.int/display/CKB/CORDEX%3A+Regional+climate+projections [Accessed 22 April 2025].
- Elexon BSC, 2025. Actual Generation Output Per Generation Unit (B1610). [Online] Available at: https://bmrs.elexon.co.uk/api-documentation/endpoint/datasets/B1610 [Accessed 23 April 2025].
- Government, UK, 2024. Clean Power 2030 Action Plan: A new era of clean electricity, London.

- Huber, Peter J. “Robust Estimation of a Location Parameter.” *The Annals of Mathematical Statistics*, vol. 35, no. 1, 1964, pp. 73–101. *JSTOR*, http://www.jstor.org/stable/2238020. Accessed 29 Apr. 2025.
- KNMI, 2025. R & D Weather and Climate Models. [Online] Available at: https://www.knmi.nl/research/weather-climate-models [Accessed 25 April 2025].
- Mateusz KAMIŃSKI, M. K., 2023. Investigating the possibility of using a supervised neural network to predict the amount of electricity generated by wind farms. Przegląd Elektrotechniczny, 99(4), pp. 111-117.
- Met Office, 2025. HadGEM2 family: Met Office climate prediction model. [Online] Available at: https://www.metoffice.gov.uk/research/approach/modelling-systems/unified-model/climate-models/hadgem2 [Accessed 24 April 2025].
- Norris, R., 2025. Global operational offshore wind capacity grows by 15% in 12 months to 80 gigawatts. [Online] Available at: https://www.renewableuk.com/news-and-resources/press-releases/global-operational-offshore-wind-capacity-grows-by-15-in-12-months-to-80-gigawatts/ [Accessed 27 April 2025].
- Shahram Hanifi, H. Z.-B. A. C. S. L., 2023. Offshore wind power forecasting based on WPD and optimised deep learning methods. Renewable Energy, Volume 218, pp. 119241-119241.
- UK Government, 2025. Renewable Energy Planning Database: quarterly extract. [Online] Available at: https://www.gov.uk/government/publications/renewable-energy-planning-database-monthly-extract [Accessed 27 April 2025].
- Wanqing Zhang, Z. L. X. L., 2022. Short-term offshore wind power forecasting - A hybrid model based on Discrete Wavelet Transform (DWT), Seasonal Autoregressive Integrated Moving Average (SARIMA), and deep-learning-based Long Short-Term Memory (LSTM). Renewable Energy, Volume 185, pp. 611-628.
- Yun Wang, R. Z. F. L. L. Z. Q. L., 2021. A review of wind speed and wind power forecasting with deep neural networks. Applied Energy, Volume 304, pp. 117766-117766.
- Zongxu Liu, H. G. Y. Z. Z. Z., 2025. A Comprehensive Review of Wind Power Prediction Based on Machine Learning: Models, Applications, and Challenges. energies, 18(2), p. 350.

VIII. Appendix

1. Project GitHub Repository - https://github.com/OkeMoyo/BENV0148_Group_3
2. Contribution table

	Contribution Table					
Name:	UCL Candidate Codes	Role in Project	Major Contributions	Challenges faced & overcome	Hours contributed	Additional Notes
Ali	PNXF3	Literature Analysis	Introduction & Literature Review Data Extraction Final Write up: Introduction and Literature Review Bibliography	Data extraction workload- help reducing the time to extract historic data Examing exisiting literature comparing published papers and their findings with our aims	80	
Zhihan	KVQS4		Introduction & Literature Review Map Data Visualization	Identifying methods and algorithms to use. Processing data and visualizing data in a good way Interpreting the findings and outcomes	20	
Tony	NQMJ9	Data engineering	Conceptualization of data dictionary data structures to store windfarms level data. Aggregation and concatenation of raw climate and windfarm data into data dictionary of windfarm level data. Document review, editing and journal write up	working with data extraction API's Finding correct parameters for request API's in climate data store. Data format issues e.g. aggregating raw grib files was time consuming hence had to convert to csvs	80	
Moyo	KXBS7	Data Scientist	GitHub repository and workflow setup, including rules for compulsory code review before approving "push" requests. Historical climate data extraction for 2019-2023 Data preprocessing, Exploratory Data Analysis, Feature Engineering, Building Models including Random Forest, DNN, and LSTM models, and Model Evaluation. Climate projection data preprocessing for forecasting. Generating forecasts under the RCP 8.5 and 2.6 scenarios using the best performing models (LSTM). Drafted the EDA, Modelling, Results and Conclusion section of the report.	Long processing times for data extraction and model training. Identifying suitable methods to handle outliers in this context. Model evaluation and pruning to improve performance across iterations.	150	
Axel	MMVQ8	Data engineering	Finding datasets, extracting and processing data for: list of UK offshore windfarms from REPD database; windfarm BMU IDs from Power Station Dictionary; Elexon Generation data per BMU; CORDEX projections for the RCP scenarios. + Methodology & general report write up.	Finding appropriate data to use for this project, cross-referencing between open-source datasets, developing code to appropriately process data into final timseries for each windfarm, finding appropriate GCM + RCM climate model to obtain data projection. Learning ghow to use github.	127	