PN000229-RPT-001

DATA ANALYTICS IN THE OFFSHORE WIND INDUSTRY

PILOT CASE STUDY OUTCOMES

OPERATIONS & MAINTENANCE

AUTHOR // Lynsey Duguid
DATE // 27th March 2018

Document History

Revision	Date	Prepared by	Checked by	Approved by	Revision History
0	27/03/2018	L. Duguid	O. Murphy	J. Ugwu	First Draft

Disclaimer

Contents

1 Executive Summary

It is known that great volumes of data are collected in the wind energy industry which is often not utilised fully. This could be for a variety of reasons:

- The data volume is too large to handle effectively (processing, storing and managing),
- It's not clear how to use the data to answer questions of interest,
- The full potential of the value which could be extracted from the data is not known.

To address this, a case study was carried out by the Offshore Renewable Energy (ORE) Catapult using 6 months of historical turbine SCADA data from a typical Round 2 offshore wind farm. Working in conjunction with data analytics experts, a variety of analysis techniques were investigated for their suitability to extract value from the turbine data, a summary of which can be seen below.

Analysis Technique	Value Proposition
Temperature trending	Setting or adjusting alarm thresholds for more effective use of alarms.
Power curves	Quantitative assessment of the effects of a change (e.g. degradation or blade repair).
Nacelle calibration to North	Improve accuracy in measuring performance and diagnosing operational issues. Correct application of directionally dependent operational strategies (e.g. curtailment).
Energy loss calculations	Identification of underperforming turbines and quantification of energy loss, in overall terms or due to categories such as low winds or partial performance.
Dashboards	Allows disparate data sources to be integrated, interrogated and visualised in a single location. Flexible configuration allows adaptation to various use-cases and requirements.

Conclusions from the study include:

- Readily available data in the offshore wind industry allows for a great deal of valuable analysis to be applied without the need for additional sensors/equipment,
- Improving data management can enable easier analysis with more reliable results,
- Best practices should be explored and shared with the wider industry to raise awareness of the potential of the existing data.

Future analyses to be explored:

- Yaw misalignment identification and resulting lost production,
- Machine learning techniques and algorithms.

2 Abbreviations

BI	Business Intelligence
JIP	Joint Industry Project
KPI	Key Performance Indicator
OEM	Original Equipment Manufacturer
ORE	Offshore Renewable Energy
PCA	Principal Component Analysis
RDS-PP	Reference Designation System for Power Plants
SCADA	Supervisory Control and Data Acquisition
SVM	Support Vector Machine
SQL	Structured Query Language
WTG	Wind Turbine Generator

3 Introduction to the Case Study

3.1 Background

It is known that there are a range of levels of activity within the offshore wind industry with regards to data analytics, from some organisations showing very little activity to others having large teams of experts working on specific challenges. However, there is no visibility of what these activities are, how successful they have been or where the major challenges lie.

To gain a better understanding of the subject of offshore wind data analysis, a case study was carried out using a subset of historical operational data from an offshore wind farm operator. The operator was aware of the large volume of data that they generate and store, which they get limited value from as it is not currently used to inform decisions or strategies. An agreement was made to provide ORE Catapult with 6 months of data in return for suggestions on how to extract value from it.

ORE Catapult consulted with several data analytics experts whose expertise lay in a variety of areas from data mining to machine learning. The methodologies outlined in this report highlight the main outcomes of these collaborations.

These methodologies could be applied to any wind farm which has access to the required data.

3.2 Data Provided

Table 1 summarises the data which was made available to ORE Catapult for use in the case study.

Table 1 – Data Provided for Case Study

Data Received	Details
SCADA signals	These were provided in Microsoft Access files with each file containing one data table. Each data table consisted of related SCADA tags for one month; for example, one of the Access files included all temperature measurements from all wind farm components for the month of January. For each SCADA tag, the data table included 10-minute aggregate values which are maximum, minimum, mean and standard deviation over the 10-minute period. There were seven Access files per month and six months of data was provided, which resulted in a total of 42 individual tables. The number of signals in each category varied, with the total number across all at around 580. The tables were created from the WTG SCADA system.
Customised alarm log	An Excel file covering the full operational period of the wind farm. Each row describes a separate alarm event based on a clustering of overlapping alarms, with start and end timestamps, component classification and total downtime.
Service records	An Excel spreadsheet containing planned and actual annual service timelines for the full operational period of the wind farm.
Major system repairs log	An Excel file consisting of records of all major system repairs for the full operational period of the wind farm.
Wind turbine grid coordinates	Latitude and Longitude of each wind turbine.

4 Areas Explored

This section of the report details a variety of the areas which were explored over the course of the case study.

4.1 Data Preparation

For collaboration with data analytics specialists, it was necessary to convert the data to SQL format to enable quick and easy extraction of any required data which would seamlessly span all time periods as opposed to extracting the required data for each month separately. Once in this format, the required data could easily be imported into the analytics software.

All analysis methods must begin with some form of data preparation. If data is not quality checked and cleansed, inaccurate results are likely to emerge. By identifying and discarding erroneous data such as unrealistic outliers, repeating values or interpolated values, this can be avoided.

Indexing to define subcategories of data is a valuable task which aids filtering. Data points could be flagged as curtailment, derating or suboptimal for example, which allows the exclusion of whole subsets of data from analyses when they are not relevant to produce more targeted results. It can also be useful to view graphs with the indexed categories included but flagged as different colours so that the effect of the categorised turbine state can be clearly seen.

Figure 1 shows an example of two different graphs viewed side by side. The flagged categories give a clear indication of where the different operative states appear on each of the graphs. For example the graph on the right has a small cluster of blue dots which you otherwise may not know the meaning of. However these are clearly due to curtailment as can be seen in the graph on the left.

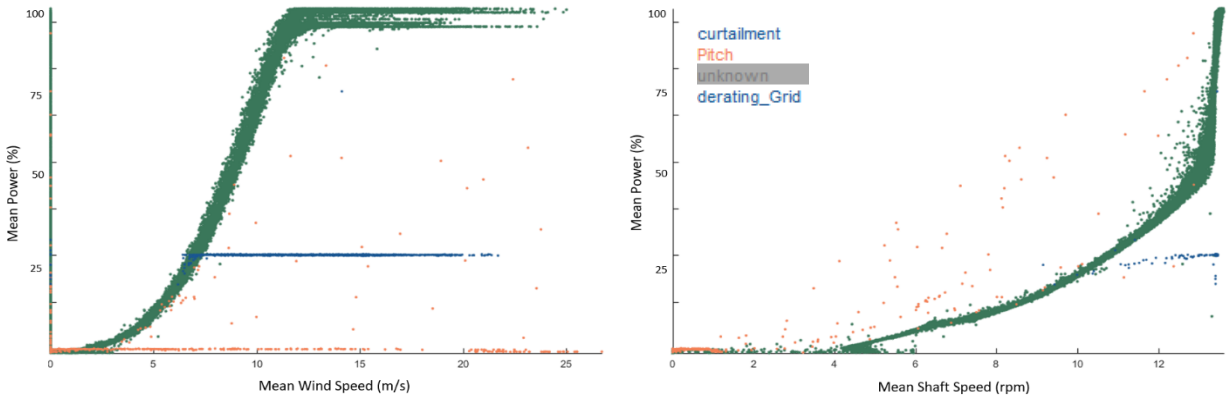

Figure 1 – Flagging Example

4.2 Temperature Trending Analysis

Temperature trending analysis is a useful method for identifying outliers or anomalies by creating a reference map of expected component temperature when a turbine is considered to be in a healthy state and comparing actual component temperatures against this reference. By inspecting the differential between the prediction and the actual state, it is possible to infer upon the component state.

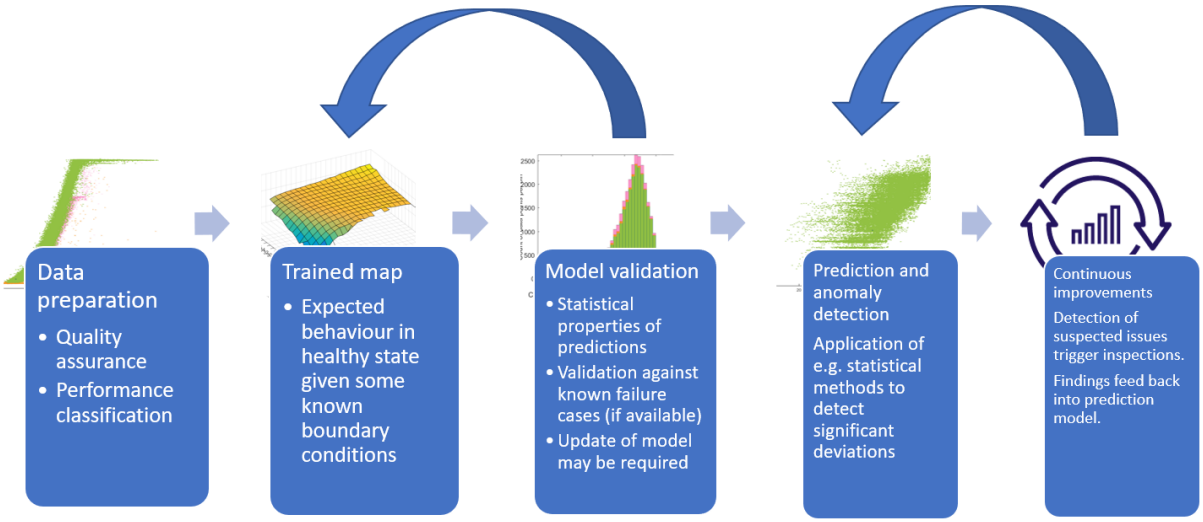

Figure 2 – Temperature Trending Analysis General Process

Figure 2 summarises the general process of temperature trending analysis. The process begins with data preparation as previously described.

Once the data is prepared, the trained map can be constructed. The objective of this is to build an empirical model that describes the expected characteristic (for example temperature) of the component of interest as a function of some other known data signal values. The selection of these parameters is based on both experience and knowledge of the system of interest, as well as statistical methods (for example covariance or principal component analysis).

Once the reference map is created with the filtered data, it should be inspected to assess whether it broadly looks as expected, i.e. does it match expectations given it is known how the system operates.

If the reference map is constructed from a small data set, then it may contain large gaps. If this is the case then it is possible to synthesise the data, which involves extrapolating values from nearby data to fill in these gaps. See Figure 3 for an example synthesised component temperature reference map. In this example, an offshore wind turbine main bearing temperature is modelled as a function of main shaft torque and ambient temperature.

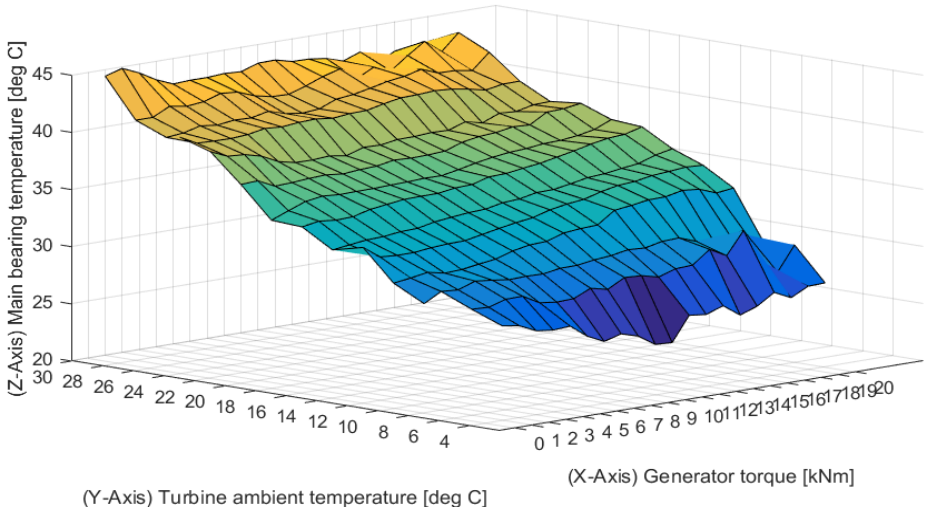

Figure 3 – Synthesised Reference Map

The next step in the process is model validation. In the example below, the model has been used to predict around 4 months immediately after the model training data period. Figure 4 displays a histogram of the number of correct predictions by calculating the 'residual' which is the deviation between the actual value and the predicted value – in this case of the main bearing temperature. The distribution shape can be assessed, with the expected result being that it should be normally distributed (thereby showing that the model is producing a high number of correct predictions with lower numbers of higher deviations). The overall aim should be to minimise this variance.

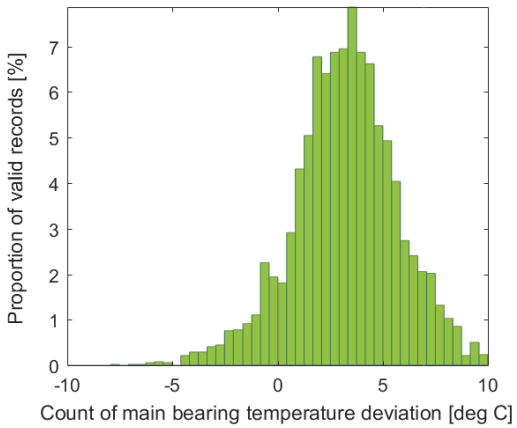

Figure 4 – Distribution of Valid Records

If the period of data includes known failures, the model's performance can be tested against those failures. This involves setting a threshold value and checking how well the model manages to predict these known failures. The results can be recorded in a confusion matrix. A confusion (or error) matrix is a type of table which allows the performance of an algorithm to be visualised. The table has two dimensions; actual results and predicted results. Each instance of a dimension is entered in the rows or columns, with the top left to bottom right values representing correct results (i.e. the actual result was

correctly predicted). Values to either side of this diagonal represent either false positive or false negative results. By utilising the confusion matrix at this stage, it is possible to visualise the accuracy of the model and make changes accordingly, i.e. if actual failures are often being predicted to be healthy state or vice versa, then it will be easily identified from this table.

By looking at the residuals on a time series plot, further assessment of the model can be carried out. If short but significant spikes are observed (Figure 5), it may indicate input data error, in which case observed and predicted data should be double checked. However, these spikes could also represent significant component or system damage.

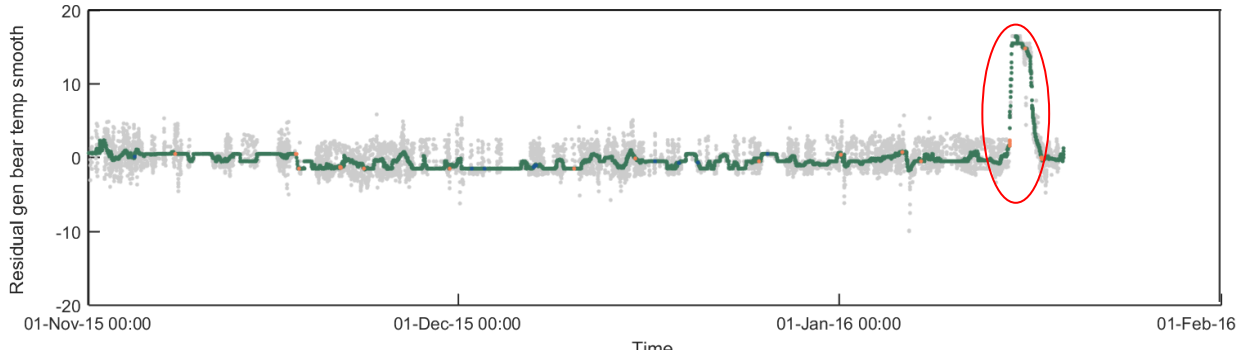

Figure 5 – Temperature Spike Observed

If a gradual increase in temperature over time is observed (Figure 6), it could indicate degradation of the component, gradual degradation of lubricity or reduction in cooling efficacy. If periodicity in the error is observed this is a sign that the model is failing to capture some cyclic phenomena.

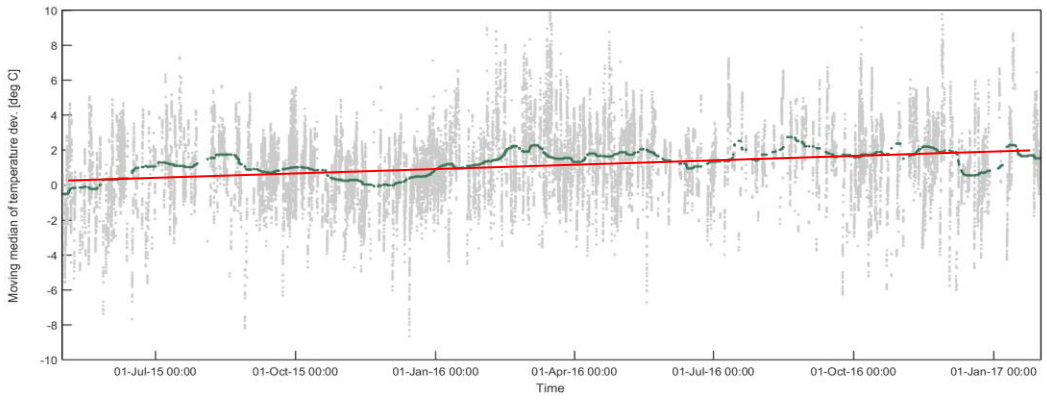

Figure 6 – Temperature Upward Trend Observed

After the model has been inspected and validated, the next step is to feed back to the trained map. Component temperature monitoring requires feedback from failure cases and inspection. Once the model is created, it can be used to identify fleet leaders (for example 3 high risk and 3 low risk turbines). Creating a heatmap of the median temperature differentials is a good way to identify these visually. See Figure 7 as an example.

Time peri	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15	T16	T17	T18	T19	T20	T21	T22	T23	T24	T25	T26	T27	T28	T29	T30
Jun-16	-1	-4	-2	-5	-3			-5	-3	-2	-4	-3	-5	-2	-5	-2	-2	-1	-3	-2	-1	-2	-3	-3	-2	-5	-4		-3	
Jul-16	0	-2	0	-4	-2	-2		-4	-2	-1	-3	-2	-3	-2	-2	-1	-1	0	-2	-1	-1	-1	-2	-1	0	-2	-3	-2	-2	-3
Aug-16	1	-1	1	-2	-1	0	0	-2	0	1	0	0	-2	0	0	1	1	2	0	0	1	1	0	1	0	-1	-1	0	0	-2
Sep-16	1	-1	0	-3	-1	0	0	-2	-1	0	-1	0	-2	-1	0	1	1	2	-1	0	1	0	0	0	0	-1	-1	-1	0	-1
Oct-16	1	-3	-1	-5	-2	-1	-1	-3	-2	0	-2	-1	-3	-2	-2	-1	0	1	-2	-1	0	-1	-1	-1	-1	-2	-1	-1	-1	-2
Nov-16	0	-3	-2	-5	-3	-1	-1	-4	-2	-1	-3	-2	-3	-2	-3	-1	0	0	-3	-1	-1	-2	-2	-2	-2	-2	-2	-1	-2	-2
Dec-16	1	-3	-1	-5	-3	-1	-2	-5	-3	-1	-3	-2	-3	-2	-3	0	-1	0	-3	-2	-1	-1	-1	-2	-2	-3	-2	-1	-1	-2
Jan-17	1	-3	-1	-6	-3	0	-1	-4	-2	-1	-3	-1	-3	-3	-3	0	-1	0	-3	-2	-1	-1	-1	-1	-2	-2	-2	-1	-1	-2
Feb-17	0	-3	-2	-5	-3	-1		-5	-2	-1	-3	-2	-3	-2	-2	0	-1		-4	-2	-1	-1	0	-2		-2	-2		-1	
Mar-17	0	-3	-1	-5	-3	-1	-1	-4	-2	-1	-3	-1	-3	-2	-2	0	-1	0	-3	-1	-1	-1	0	-1	-2	-2	-1	-1	-1	-2
Apr-17	1	-2	-2	-6	-2	-1	-1	-4	-2	-1	-2	-1	-3	-3	-3	0	-1	0	-2	-1	-1	-1	-1	-2	-2	-1	-2	-1	-1	-2
May-17	0	-2	-2	-6	-3	-1	-1	-5	-3	-1	-3	-2	-3	-3	-3	0	-1	0	-2	-1	-1	-1	-1	-2	-2	-2	-2	-1	-1	-2
Jun-17	0	-2	-1	-5	-2	0	-1	-4	-3	-1	-2	-1	-2	-2	-2	0	0	0	-2	-1	-1	-1	-1	-1	-2	-2	-2	-1	-1	-1
Jul-17	-1	-2	-2	-5	-2	-1	-2	-4	-3	-1	-3	-2	-2	-2	-2	-1	-1	0	-3	-2	0	-1	-1	-1	-2	-2	-2	-1	-1	-2
Aug-17	1	-2	-1	-4	-2	0	0	-3	-1	0	-2	-1	-1	-1	-2	0	0	1	-1	-1	0	0	0	-1	-1	-1	-1	-1	-1	-1
Sep-17	1	-1	0	-3	-1	0	0	-3	-1	-1	-1	-1	-1	-1	-1	0	0	1	-1	-1	0	0	-1	-1	-1	-1	-1	-1	-1	-1
Oct-17	0	-2	-1	-4	-2	0	-1	-4	-2	-2	-2	-2	-2	-1	-2	-1	-1	0	-2	-2	-1	-1	-1	-2	-2	-2	-2	-1	-2	-2
Nov-17	-1	-3	-2	-6	-3	-1	-2	-5	-3	-3	-3	-3	-3	-3	-3	-1	-2	0	-3	-2	-2	-2	-2	-2	-3	-2	-2	-2	-1	-3
Dec-17	0	-2	-1	-5	-2	0	-1	-4	-2	-2	-2	-2	-2	-2	-2	0	-1	1	-2	-1	-1	-1	-1	-2	-2	-2	-1	-1	-1	-2
Jan-18	0	-2	-2	-5	-2	-1	-1	-4	-2	-2	-3	-3	-3	-2	-3	-1	-2	0	-2	-2	-1	-1	-1	-2	-2	-2	-1	-1	-1	-3
Feb-18	1	-2	-1	-4	-2	0	-1	-4	-2	-1	-2	-1	-2	-1	-2	1	-1	1	-2	-1	0	0	0	-1	-1	-1	-1	0	0	-1

Figure 7 – Heatmap of Median Temperature Differentials

Once a model and thresholds which yield predictions have been created, the system is ready to be deployed. This involves storing the developed model in a location where the monitoring system has access to it. When new data is received from assets, a new prediction should be made and compared to the observed data. To validate the prediction, the updated set of data should be checked to ensure that it meets the detection thresholds.

The final stage in the process is continuous improvement. When alerts are triggered, inspections should be undertaken; the results of which should be fed back into the confusion matrix. At this point the detection thresholds should be re-tuned to balance the updated confusion matrix.

4.3 Power Curve Uses

The power curve is a simple graph of power as a function of wind speed. They are a standard and well-known method of analysis in offshore wind which have many uses, some examples of which are outlined below:

- Degradation identification – by periodically assessing the power curve of a wind turbine, it is possible to identify performance reduction over time. This will not classify a problem, but it can identify that there may be an issue which requires further investigation.
- Measurement of the impact of a change – if a change is planned such as a blade repair or the installation of Vortex generators for example, power curves can be compared before and after to measure the impact of the change, i.e. assess whether it has improved or degraded.
- Continuous monitoring – operational states can be identified by continuously monitoring power curves along with variations of the power curve. For example, by monitoring power against pitch it is possible to confirm that suspected points on a power curve are due to curtailment. This could be useful to identify times when turbines are being curtailed by the OEM which the owner may otherwise be unaware of. See example in Figure 8 and Figure 9.
- Outlier analysis – comparing individual wind turbines against the average of the rest of the farm is one method of benchmarking to identify poor performers.
- Forecasting production – By creating a reference power curve, it is possible to forecast energy production based on wind speed forecasts. This could be used to plan a new wind farm, inform

the electricity grid of near future production for energy balancing, or simply to compare the actual performance of a turbine against what was expected.

On the down side, power curves may give false alarms or confusing results due to the OEM's tendency to make changes without the operator's knowledge such as modifying transfer functions or swapping out wind instruments, which must be kept in mind when drawing conclusions.

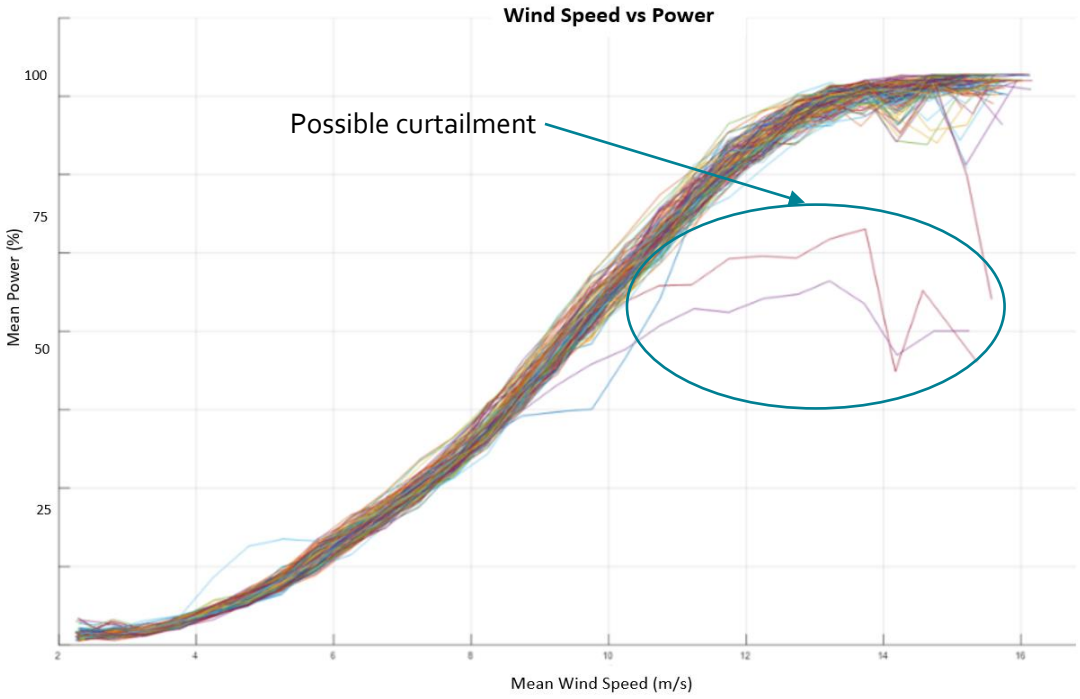

Figure 8 – Example Power Curve

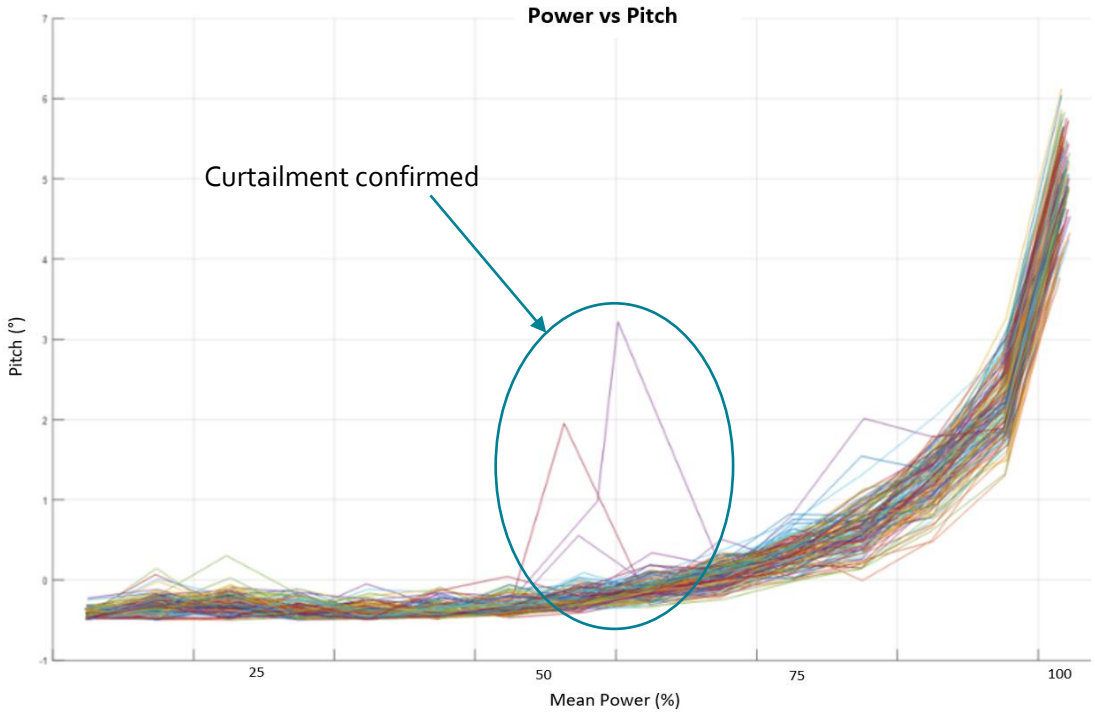

Figure 9 – Example Power V's Pitch

4.4 Nacelle Calibration to North

It is common that SCADA data signals which represent the geographic orientation of a wind turbine nacelle relative to some datum, typically North, is poorly calibrated. This may occur for several reasons, for example:

- Incorrect calibration applied during first configuration/commissioning,
- Inherent difficulty in aligning nacelle with datum to set calibration,
- Signal drift over time.

While the calibration error does not directly impact the operating efficiency of a wind turbine, poor calibration may have some secondary undesirable effects, such as:

- Elevated inaccuracy in measuring performance and diagnosing operational issues (such as directional power curves),
- Incorrect application of directionally dependent operational strategies (such as directional curtailment).

There are two known SCADA analysis methods which can be used to calibrate the nacelle direction signal relative to some datum. Each method has pros and cons, however they can both be used to complement one another:

- Method 1 uses polar plots to detect the 'signatures' of wind turbine wakes within SCADA data and matching these signatures to the bearing of nearby wind turbines.
- Method 2 is based on a simple cross correlation of the measured direction to some reference. This process is reliant on the absolute accuracy of the reference used.

Both of the following methods use data which is filtered to only include wind speeds in the range of around 5-10m/s as this is where wake impacts are most pronounced.

4.4.1 Method 1 – Calibration Based on Wake Signature

For two wind turbines located near one another, it is expected that in the absence of wake effects, the wind speeds measured at both wind turbine locations within the same 10-minute periods should be similar.

For certain wind directions where one wind turbine is in the wake of another, and the wind turbine producing the wake remains free from wakes itself, a significant difference in measured wind speeds should be observed.

By calculating the ratios of wind speeds measured at these two wind turbines and comparing this to the layout of a wind farm, a direction calibration error can be inferred.

As an example, two wind turbines have been selected at the South-Western side of a wind farm (see Figure 10). Wakes are represented by the orange and blue sections. Blue indicates where T02 wakes the other turbines and orange indicates where the other turbines wake T02. To simplify the visualisation, all the wake zones are centred around T02 instead of each individual turbine. Both turbines are exposed to

wakes from several different wind turbines in sectors 330°N to 180°N. Turbine T02 is located at a bearing of approximately 202°N from turbine T01. T02 is not waked by any other turbines in this sector.

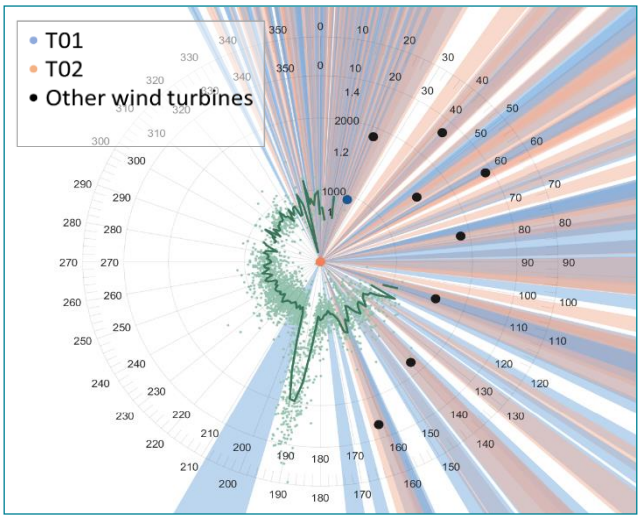

Figure 10 – Polar Plot of Wind Speed Ratios

Each light green dot represents a 10-minute sample of the ratio of the wind speed (taken from SCADA) at T02 divided by T01 (for a given nacelle orientation). The dark green line shows the median of the 10-minute sample ratios in each 2° sector. All this data is plotted with the graph centred at turbine T02, using the nacelle direction of T02 as the angle argument. Given the wake exposure of the two wind turbines, it would be expected that the wind speed at T02 should be significantly higher than that at T01 in a narrow direction sector centred at around 202°N. However, this is not what is observed; the spike in ratio occurs at a bearing of approximately 191°N. Therefore, the calibration error at turbine T02 is estimated to be -11° (see Figure 11).

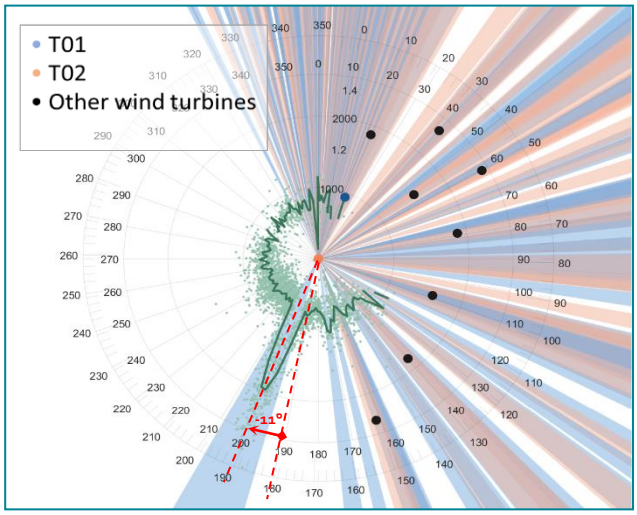

Figure 11 – Polar Plot Corrected by -11°

After applying an 11° offset, the spike in normalised wind speed aligns with the wake induced by T02.

4.4.2 Method 1 Alternative – Calibration Based on Turbulence Intensity

It is also often possible to use an estimate of the turbulence intensity to identify wakes from neighbouring wind turbines, particularly offshore where free-stream turbulence levels tend to be low. Figure 12 shows the turbulence by direction for turbine T01. The wake caused by T02 at 195°N to 208°N is clearly detectable. The measure of turbulence intensity used here was calculated as the standard deviation of wind speed divided by the mean wind speed. Again, the angle of each point on the polar plot is the nacelle orientation taken from SCADA.

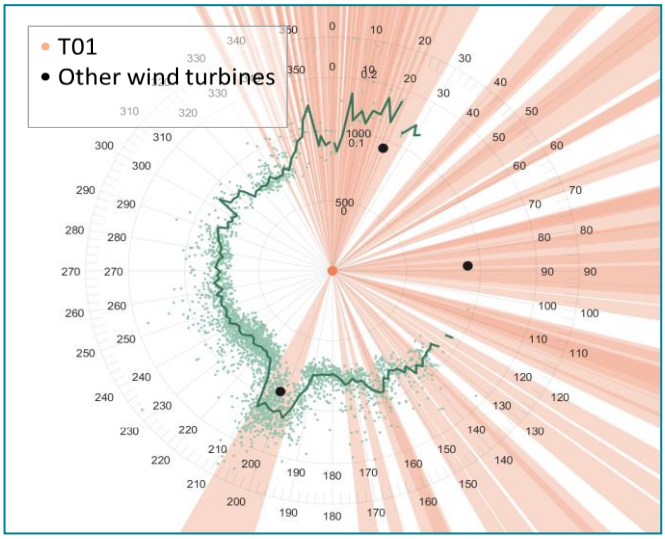

Figure 12 – Nacelle Calibration Using Turbulence Intensity

The wind speed ratio and turbulence intensity methods both have their merits and drawbacks:

- The wind speed ratio method tends to provide a high signal-to-noise ratio and is therefore more robustly detected. It does however rely on having good data from a neighbouring wind turbine and that the neighbouring wind turbine is free from wakes in a sector in which the test turbine is waked.
- The turbulence intensity method uses data from one individual wind turbine. This can be of benefit in some data management/analysis systems where linking data from different turbines may be prohibitively challenging. The signal-to-noise ratio does however tend to be lower than that of the wind speed ratio version. For offshore sites this may not be an issue but at onshore sites it can be, where terrain and ground cover geometry are non-simple.

4.4.3 Method 2 – Calibration Relative to a Reference

Calibration method 2 is to infer the calibration error relative to some other trusted reference. The reference may be data from a trusted mast, turbine absolute wind direction or nacelle direction data calibrated using the wake signature approach. It may be beneficial to derive the reference direction as the median value of the calibrated direction signal from several turbines. Then the difference between the observed direction and the reference direction is calculated.

See example in Figure 13 below. If direction data from T1 is accurate, it can be inferred that that the direction at T2 is offset by approximately 25°.

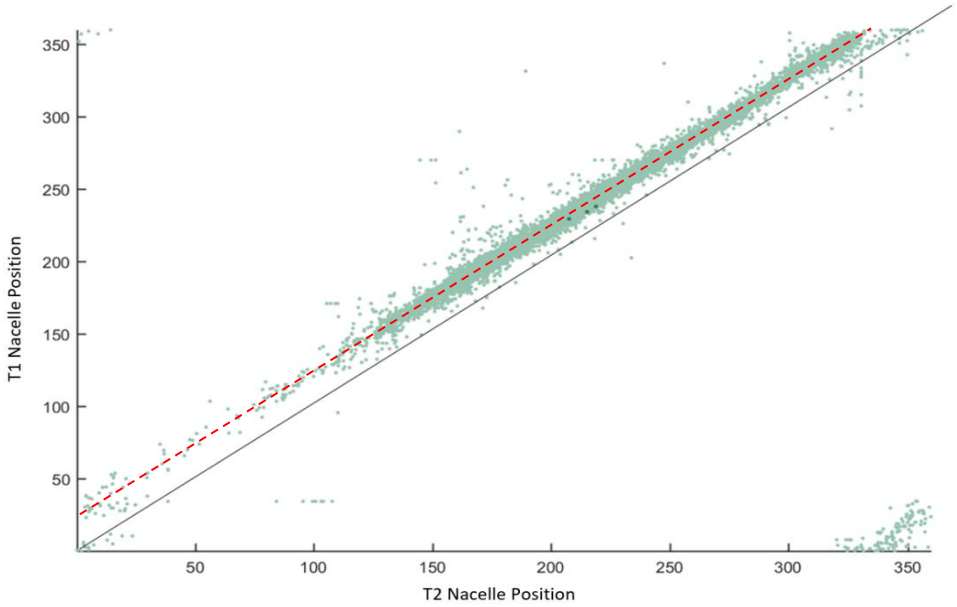

Figure 13 – Nacelle Positions T1 Vs T2

Figure 14 shows another example. In this scatter graph, there are two distinctive clusters of data points. The points falling on the diagonal from the origin indicate accurate data and the other points indicate an offset from the accurate reference.

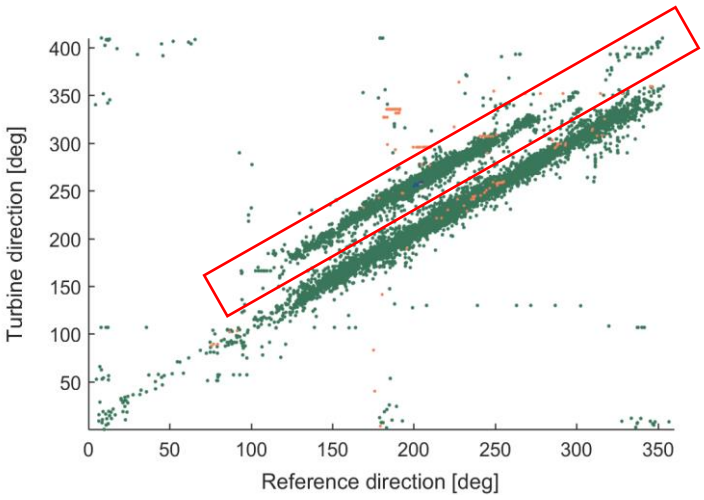

Figure 14 – Observed Offset

When these points are plotted on a timeseries chart (see Figure 15), it shows that the turbine has a period of approximately 1 month with a nacelle direction offset of approximately +50°

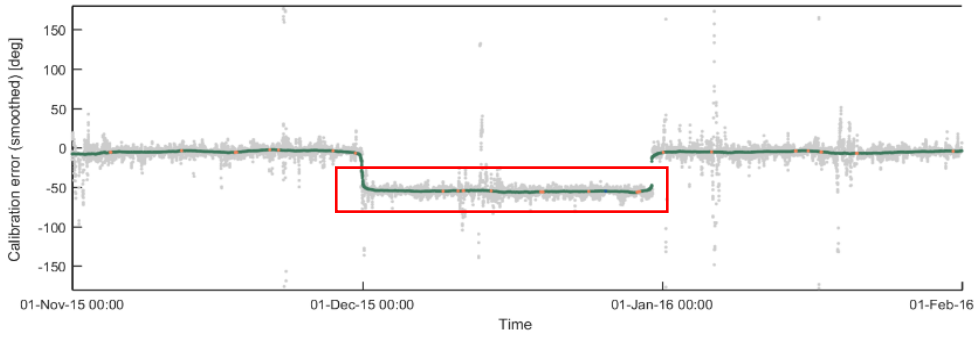

Figure 15 – Observed Offset – Timeseries

4.5 Energy Loss Calculations

Calculating energy lost allows identification of underperforming turbines. This method is based on the development of reference power curves derived using data representing normal operation. Energy losses can then be calculated as the difference between the observed power and the power expected from the reference power curve, and the calculated losses can be aggregated over time, per turbine, and by flag.

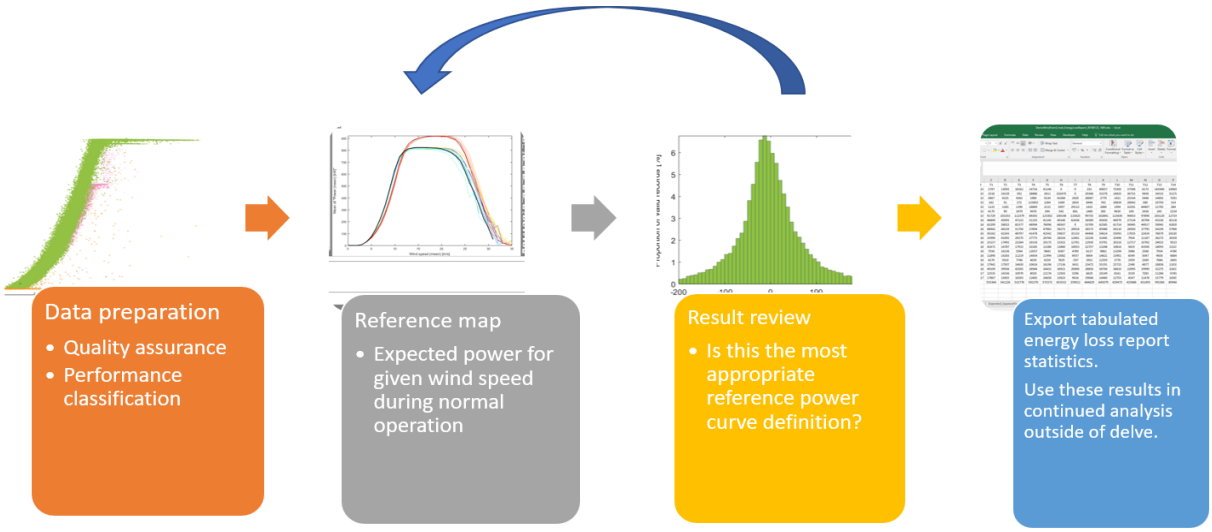

Figure 16 – Energy Loss Analysis General Process

As with all analyses, the process begins with data preparation – classifying, flagging and filtering to identify the values which represent normal operation.

The next step is to create reference power curves for each wind turbine. This is essentially an average power curve for normal operation over the period the data covers. Where gaps exist, the data can be synthesized by taking an average of the points around it (see Figure 17 and Figure 18).

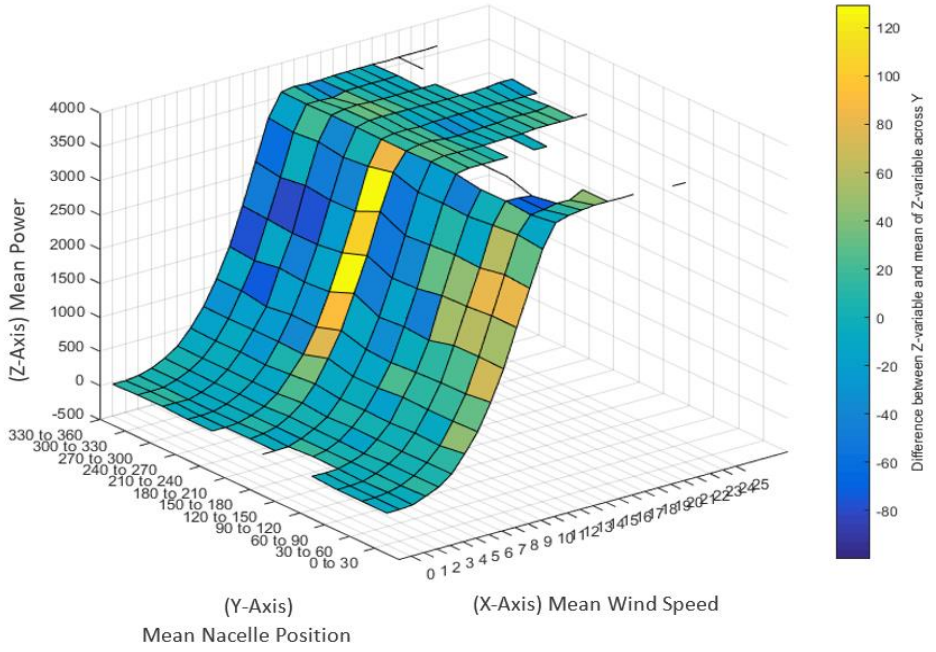

Figure 17 – Un-synthesised Power Curve

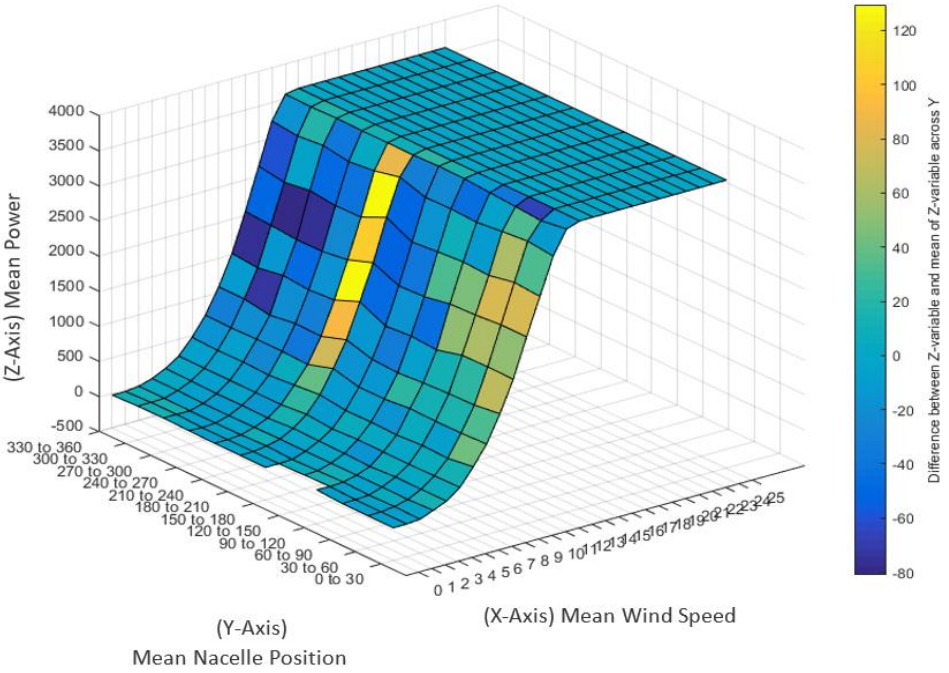

Figure 18 – Synthesised Power Curve

The reference power curve suitability should be checked and updated if required. One method for this is to plot the expected power against the observed power on a scatter chart (see Figure 19). The aim is to minimise the scatter of the points which represent normal operation; if there is a high level of spread, then this suggests some changes should be made to the reference map.

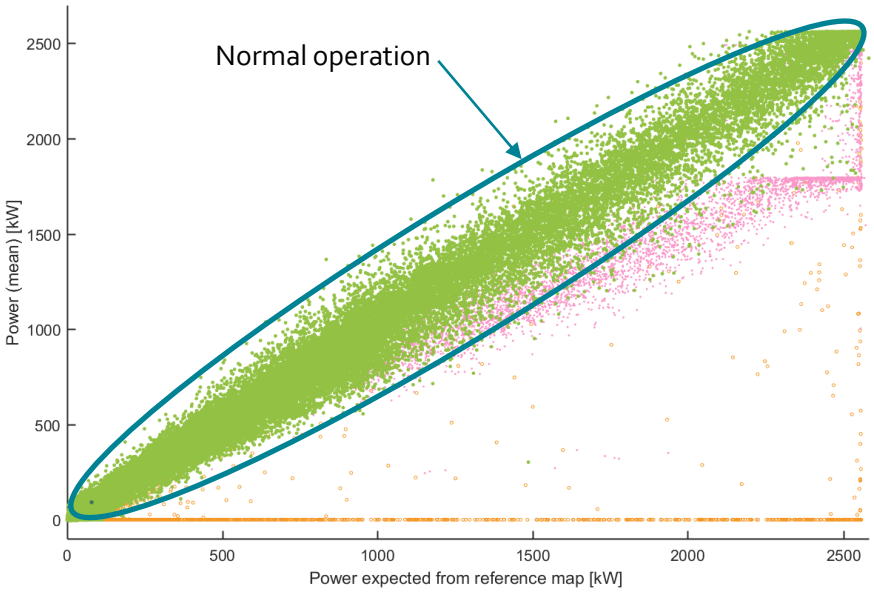

Figure 19 – Verification Method for Reference Map

A more quantifiable method is to calculate the deviation (residual) between the predicted and observed values. Visualising this in a histogram should show a normal distribution centred at zero if the reference map is of good quality. Again, this will highlight if changes are required to the reference power curve.

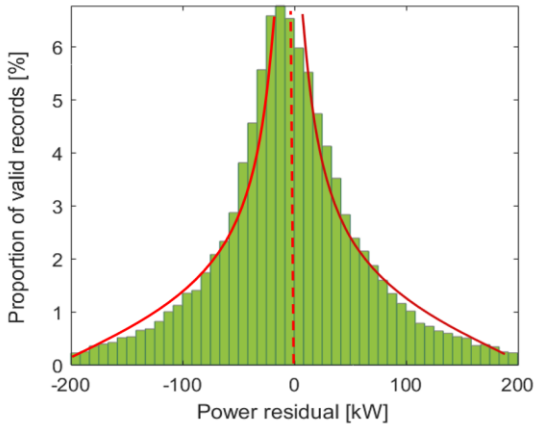

Figure 20 – Verification Method 2 for Reference Map

With a finalised reference power curve, it is possible to calculate the expected energy production across the period in question. Each timestamp has a known wind speed, and using the reference power curve, the energy production associated with this wind speed can be looked up and recorded as the expected energy production for that timestamp. This is repeated for each timestamp and each wind turbine until a full picture of expected energy production is created.

The final step is simply to compare the expected energy with the actual energy from SCADA data; the difference between the two being the energy loss.

Depending on how the data was originally flagged, the results can be seen in categories. i.e. if curtailment is flagged at the start, then it will be possible to see the energy loss associated specifically with say, curtailment for example.

4.6 Dashboards

Using a business intelligence (BI) tool, dynamic and interactive dashboards can be created, which allows disparate data sources to be tied together. With a wider variety of data sources in a single file, more context can be given to the data viewed. For example, a graph which shows a wind turbine with a long period of downtime has more value if it can also provide the reason for the downtime. This information may be available in alarm logs, so to include this in the dashboard allows for deeper interrogation of the information and 'root cause' investigation.

As an example, the following data sources were combined in a BI tool: SCADA, alarm logs, service records, site targets, wind speeds (met mast and modelled) and wind turbine grid coordinates. This involved joining the tables into one relational database to allow all the data to be queried together.

Visualisations are completely configurable and dynamic. This allows dashboards to be created in any way the user requires and it allows for drilling down to the lower level details. When an element of a chart is clicked, or a slider is moved, all other charts, tables, maps etc. automatically update to show the relevant data.

The value in dashboards is the bringing together, configuration and visualisation of various data sources. These individual pieces of data would otherwise have to be reviewed in isolation, which makes it difficult to see the full picture. They can be used on a static monthly basis or linked to live streaming data which would allow continuous monitoring.

An example Production Summary dashboard can be seen in Figure 21 (note that some information has been obfuscated to preserve anonymity for the wind farm). In this dashboard, a few key high level KPI's are displayed. On the left is a heat map of the wind farm, showing the performance of individual turbines. Simply by adding a data table containing the turbine latitudes and longitudes and linking this to the existing datasets, it has allowed this powerful visual to be incorporated. By cycling through the three options (energy produced, capacity factor and normalised capacity factor), the heat map can provide a very quick and efficient way of identifying outliers by looking for the colour which represents

poor performance. Elsewhere in the dashboard, the monthly energy production compared to the target and capacity factor for the period selected can be seen.

Figure 21 – Production Summary Dashboard

Figure 22 shows how the dashboard dynamically updates when a user clicks on specific parameters for further analysis. The month of December has been selected and as a result, the capacity factor and heat map update automatically to reflect this subset of data.

Figure 22 – Production Summary Dashboard using Dynamic Links

A Downtime Summary dashboard can be seen in Figure 23. This is in a similar style to the Production Dashboard; however, the focus is this time on categorised downtime. This dashboard allows the user to interrogate individual turbines or view aggregated group data. It also allows the user to focus on a categorisation of downtime to see how much of the total downtime is associated with that category.

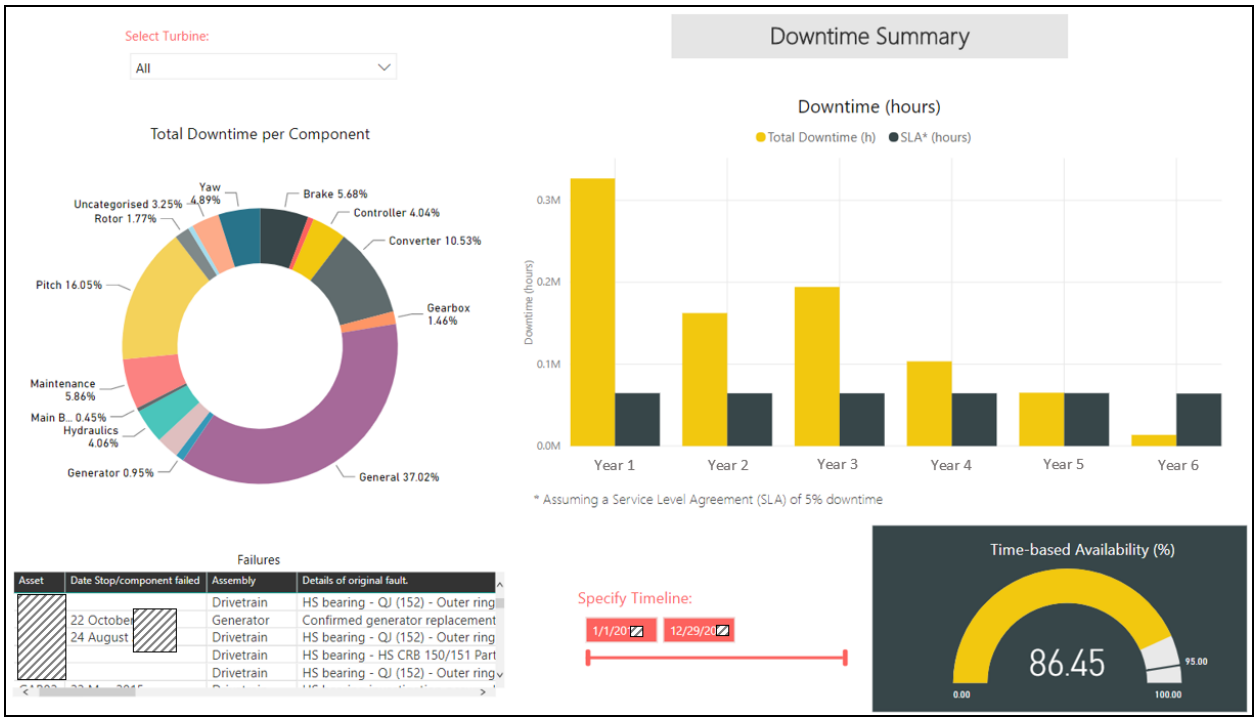

Figure 23 – Downtime Summary Dashboard

Figure 24 shows the effect of pitch systems on total downtime for the wind farm. By clicking on the 'Pitch' segment of the pie chart, it has filtered all the related data to show pitch specific data as a subset. This is useful for visualising the effect of an individual element against the full dataset.

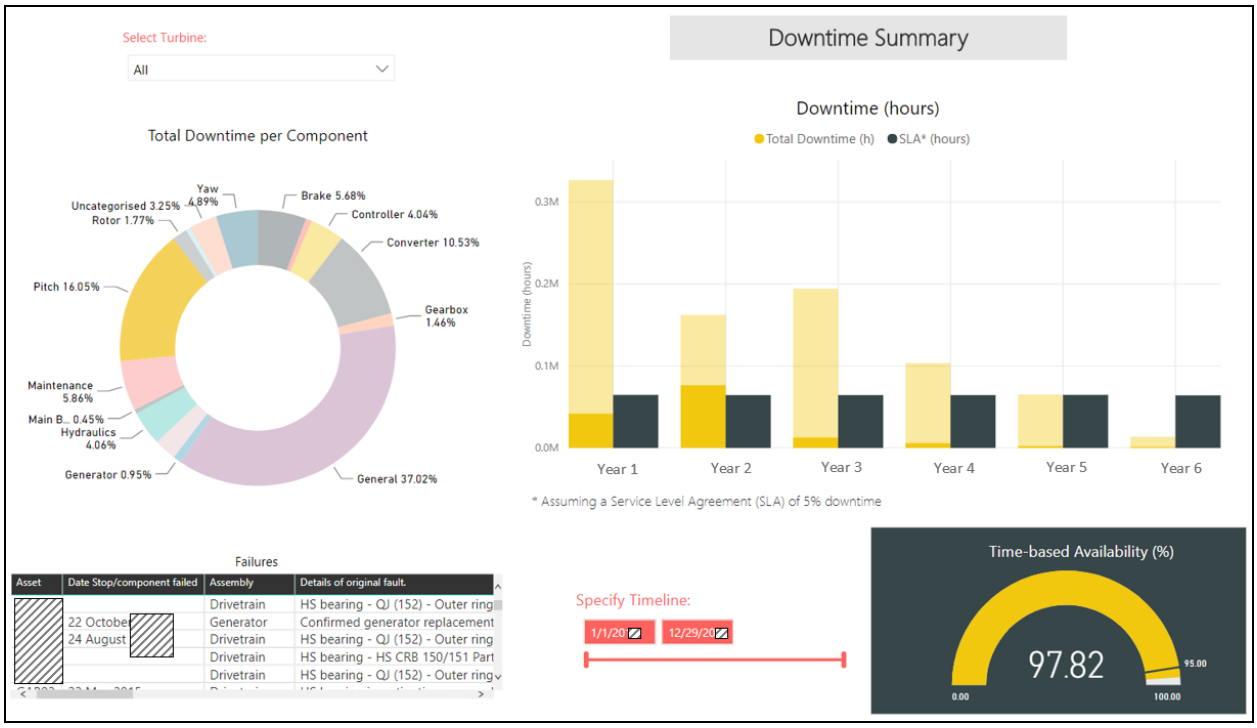

Figure 24 – Downtime Summary Dashboard Interrogation of Pitch Failures

Figure 25 is a final example of a Service History Dashboard which has the purpose of allowing management to monitor service progress over the course of the year.

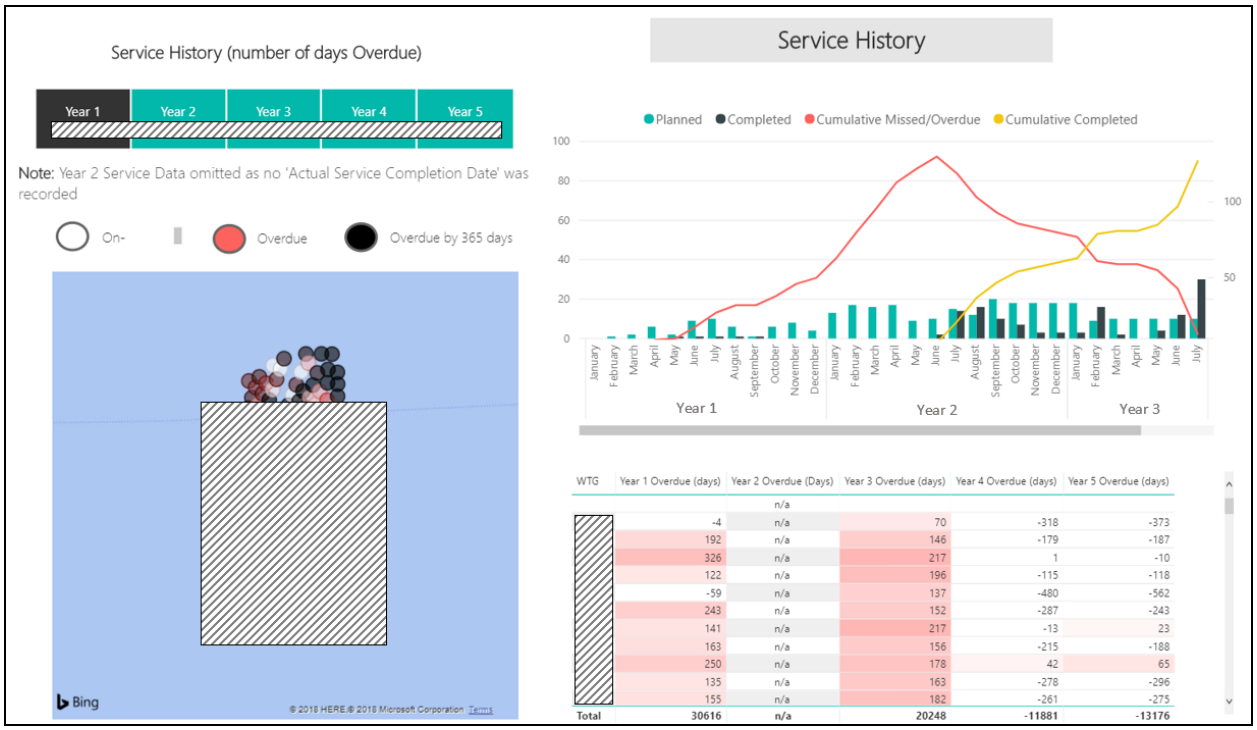

Figure 25 – Service History Dashboard

These example dashboards would be useful for management level, where very high-level insight can be obtained. More granular data can also be used for deeper analytical dashboards which would be aimed at data analyst roles. These could allow analysis of alarm trends over time at the component level and include other parameters such as active power signals for correlation.

5 Barriers

5.1 SCADA Data Signals

Analysis was attempted on yaw alignment to wind direction as it is believed that this could be an industry wide issue. If this is the case, then it could be having a direct and significant effect on energy production. If a turbine's control system believes it to be facing into the wind, but is in fact misaligned, then it will not be extracting all the available energy.

However, to use the methodology suggested by an analytics expert, SCADA signal 'yaw error' was required which was not provided to ORE Catapult. If similar studies are to be carried out in the future, yaw misalignment will be investigated further. The aim would be to find a methodology to identify when misalignment is occurring, quantify the lost energy per degree of misalignment and propose both control and instrument calibration-based solutions.

5.2 Data Quantity and Quality

Although it was possible to replicate these analytics methodologies to test how the processes work in practice, with more data it would be possible to perform deeper analysis of the wind farm data itself to, for example, identify any unknown faults which require attention.

With higher quality data (i.e. better categorised), it would also be possible to give deeper insight to any findings. With the current dataset, a lot of data is categorised as 'general' instead of a specific systems or components which means analysis results lose the root cause aspect which is very valuable.

6 Recommendations

Data analytics is a vast subject of which only the surface could be scratched in this pilot study. However, several findings have emerged which serve as key recommendations.

These recommendations have been divided into categories as follows:

1. Quick to implement actions aimed at general data management. The ORE Catapult has capability to provide advice and support these internal improvements.
2. Actions which could take more time to implement aimed at business processes.
3. Higher level, industry wide suggestions.

Category 1

- **Assess and redefine alarm categories for better root cause analysis of wind farm operations.** During this case study, it became apparent that alarm categorisation was not always accurate or consistent, leading to less reliable results. By improving the categorisations, it will allow full trust in the results and therefore increase the value to be gained from the data. For example, many alarms are categorised as 'general' which gives very little insight when analysed. If these alarms were assessed in more detail and assigned a specific category (such as pitch system), then a true picture of the state of wind farm operations could be viewed.
- **Automate data source creation for consistency and accuracy of use.** It has been seen that many potentially valuable data sources are managed manually across a variety of spreadsheets (for example service records and major system repairs logs). This increases the likelihood of error or omissions which reduces the reliability of results. If these reports were automatically generated from system data, then the quality of data analysis would be improved.
- **Integrate data sources.** Following on from the above point, a further step which could be taken is to integrate or link these sources in some way which would aid interrogation of the data. For example, storing it in a relational database such as SQL Server. This can reveal hidden patterns in the data and unlocks value.

Category 2

- **Store adequate volume of historical data.** To perform predictive maintenance analyses, it is advised that a minimum of 3 years of data is used for training algorithms. The less data available, the lower the chances of achieving valuable results.
- **Consider changing stored data format.** Storing data in restrictive formats such as Microsoft Access files can increase file size and introduce formatting complications when converting data into a more suitable format for handling. It could improve the efficiency of processes if data was converted directly to a different format (such as SQL Server) straight from source in a single step. The ORE Catapult has capability to advise on and provide support for this transition.
- **Start dialogue with OEM's.** Wind turbines record a great deal of data which isn't always available as standard from the SCADA system. Some of this data would be valuable for analysis

and therefore would be worthwhile acquiring from the OEM (for example yaw error, which can be used to assess the yaw misalignment).

- **Research and invest in appropriate software packages.** There are a great deal of commercial tools available which have been created specifically for data analytics, such as Breeze Wind Farm Management System, Bazefield, Power BI, Tableau, Delve Wind and Wolfram Mathematica. By understanding business requirements, off the shelf software could be selected which may reduce the time and cost required in building in-house tools.

Category 3

- **Understand best practice in data analytics.** There is a need to understand best practices in these data analytics areas and showcase how to get value from the large volumes of data generated by offshore wind farms. Sharing findings from studies like this would be beneficial to the industry.
- **Standardise component classifications.** With clear and standard component classifications, interrogation of analysis results could be enriched due to clearer root causes. The recommendation would be to utilise the RDS-PP standard for classifying components.
- **Start a digital transformation joint industry project (JIP).** Sharing of best practice is critical for the offshore wind industry to benefit from digital opportunities. By creating a JIP, progress for everyone involved could be greatly escalated. There is lots to be learnt from each other, within the wind industry as well as from other industries.

7 Future Work

Machine Learning Techniques and Algorithms Applied to Wind Energy

The ORE Catapult plan to work closely with machine learning experts to investigate the effectiveness of applying various machine learning techniques to wind energy industry data. The following section describes a variety of machine learning techniques and their potential application in the wind energy industry. By applying these techniques, it could be possible to identify problems with a wind turbine prior to the event occurring which would allow early intervention and therefore reduced downtime and cost savings.

A key first step to applying machine learning techniques to wind energy data is to determine the exact desired outcome. Four options are listed below. The option chosen will determine the machine learning technique that is most appropriate to use for a specific case. Common desired outcomes are:

1. Finding unusual data points
2. Predicting categories/classes
3. Predicting values
4. Finding structure

1. **Finding unusual data points:** If the desired outcome is to find unusual data points in a set of data or to determine if a sub-set of data is unusual or not, anomaly detection algorithms should be used. Two examples of anomaly detection algorithms are one-class support vector machines (SVMs) and PCA-based anomaly detection.

   If an operator has a large amount of normal behaviour data, (such as gearbox bearing temperature data from normal operation across the full range of power production and ambient temperature possibilities for tens or hundreds of turbines) that data can be labelled as 'normal' and can be used to train an anomaly detection algorithm. When new data from a turbine of unknown health is used as an input to the trained algorithm, unusual data points will be recognised by the anomaly detection algorithm and flagged to the operator. The anomaly detection algorithms may recognise relationships and dependencies between input variables (such as relationships between bearing temperature, power production, wind speed, generator rpm, nacelle temperature, ambient temperature) that current industry standard trending cannot.

2. **Predicting Categories:** If the desired outcome is to predict the category/class that a data set most resembles, classification algorithms should be used. If the data of interest should be assigned to one of two categories a two-class classification algorithm is used. However, if the data of interest should be assigned to one of more than two categories a multi-class classification algorithm should be used. Examples of two-class classification algorithms are: Two class neural networks, two class averaged perceptron, two class support vector machine and two-class boosted decision tree.

Examples of multi-class classification algorithms are: Multi-Class logistic regression, multi-class neural network and multi class decision forest.

If as well as normal behaviour data, an operator also has a large amount of failure data, both data sets can then be labelled as 'healthy' or 'unhealthy' respectively and used to train a two-class classification algorithm. The trained two class classification algorithm can then be used to predict if a new set of data (such as bearing temperature, power production, wind speed, generator rpm, nacelle temperature or ambient temperature etc.) from a turbine of unknown health belongs to the healthy or unhealthy category.

For multi-class classification instead of labelling the data from the two months before failure as 'unhealthy', the data could be split and labelled as 'two months to failure' and 'one month to failure'. Once combined with the data labelled as healthy, the data sets can be used to train a multi-class classification algorithm. The trained algorithm will then be able to predict if a new data set resembles data that is 'healthy', 'two months to failure' or 'one month to failure' (see example in Figure 26)

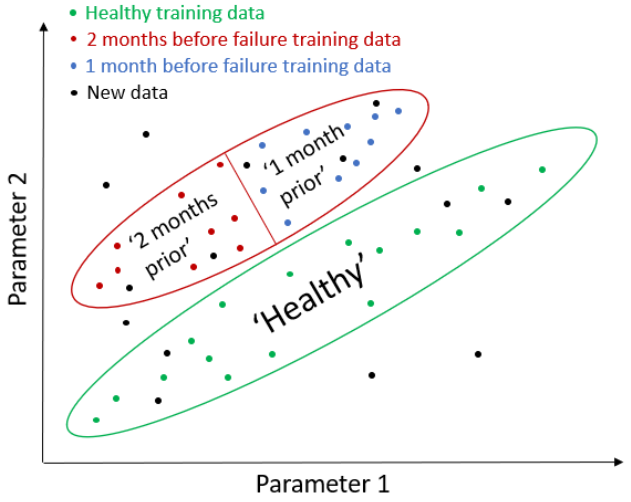

Figure 26 – Example Multi-Class Classification

3. **Predicting Values:** If the desired outcome is to predict a value based on several input variables, regression algorithms should be chosen. Examples of regression algorithms are: Linear regression and decision forest regression.

   If an operator has one year of data (such as bearing temperature, power production, wind speed, generator rpm, nacelle temperature, ambient temperature or time before failure in days) in the period before a number of gearbox bearing failures, the operator can use that data to train a regression model. In general, the more examples of the failure that the operator has the more accurate the model will be. When an operator is using the trained model with new data that has an unknown remaining useful life, they will require 6 of the 7 data points listed above (minus the remaining life data) as an input. The output of the trained regression model

will then be an estimated remaining useful life in days based on the other 6 inputs with a measure of confidence in the prediction.

4. **Finding structure:** If the desired outcome is to find structure in a data set, clustering algorithms should be chosen. An example of a clustering algorithms is K-means clustering.

   If an operator is trying to predict how long before a certain failure type will occur based on a dataset, the accuracy of their algorithms may be improved if trained and used on turbines that operate in a similar manner or in a similar environment. A clustering algorithm could be used to group similar wind turbines based on their capacity factor, environmental conditions or other variables that may influence the remaining useful life of components.

Table 2 – Summary of the aims, techniques and algorithms previously discussed

Aim	Technique	Example Algorithm
Finding unusual data points	Anomaly detection	One-class support vector machine, PCA based anomaly detection
Predicting categories/classes	Classification	Two-class neural network, two-class support vector machine, multi-class logistic regression
Predicting values	Regression	Linear regression, decision forest regression
Finding structure	Clustering	K-means clustering

Contact

ORE Catapult

Inovo

121 George Street
Glasgow
G1 1RD, UK
T: +44 (0)333 004 1400

National Renewable Energy Centre

Albert Street, Blyth
Northumberland
NE24 1LZ, UK
T: +44 (0)1670 359 555

Fife Renewables Innovation Centre

Ajax Way
Leven
KY8 3RS
T: +44 (0)1670 357 649

O&M Centre of Excellence

Room 241, 2nd Floor
Wilberforce Building
University of Hull
HU6 7RX

ore.catapult.org.uk
Tweet us: @ORECatapult // @CatapultBlyth
info@ore.catapult.org.uk